FROM PROTOTYPE TO CLINICAL WORKFLOW: MOVING MACHINE LEARNING FOR LESION QUANTIFICATION INTO NEURORADIOLOGICAL PRACTICE

EDITED BY: Raphael Meier, Jayashree Kalpathy-Cramer, Susanne Wegener, Benedikt Wiestler, Richard McKinley and Spyridon Bakas

PUBLISHED IN: Frontiers in Medicine, Frontiers in Neuroscience and Frontiers in Neuroimaging

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-679-6
DOI 10.3389/978-2-88976-679-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

FROM PROTOTYPE TO CLINICAL WORKFLOW: MOVING MACHINE LEARNING FOR LESION QUANTIFICATION INTO NEURORADIOLOGICAL PRACTICE

Topic Editors:
Raphael Meier, Bern University Hospital, Switzerland
Jayashree Kalpathy-Cramer, Harvard Medical School, United States
Susanne Wegener, University of Zurich, Switzerland
Benedikt Wiestler, Technical University of Munich, Germany
Richard McKinley, Bern University Hospital, Switzerland
Spyridon Bakas, University of Pennsylvania, United States

Citation: Meier, R., Kalpathy-Cramer, J., Wegener, S., Wiestler, B., McKinley, R., Bakas, S., eds. (2022). From Prototype to Clinical Workflow: Moving Machine Learning for Lesion Quantification into Neuroradiological Practice. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-679-6

Table of Contents

frontiers in Medicine

ORIGINAL RESEARCH
published: 19 August 2021
doi: 10.3389/fmed.2021.705071

Joint Modeling of RNAseq and Radiomics Data for Glioma Molecular Characterization and Prediction

Zeina A. Shboul[1], *Norou Diawara*[2], *Arastoo Vossough*[3], *James Y. Chen*[4] and *Khan M. Iftekharuddin*[1]*

[1] Vision Lab, Department of Electrical & Computer Engineering, Old Dominion University, Norfolk, VA, United States, [2] Department of Mathematics & Statistics, Old Dominion University, Norfolk, VA, United States, [3] Department of Radiology, Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, United States, [4] University of California, San Diego Health System, San Diego, CA, United States

RNA sequencing (RNAseq) is a recent technology that profiles gene expression by measuring the relative frequency of the RNAseq reads. RNAseq read counts data is increasingly used in oncologic care and while radiology features (radiomics) have also been gaining utility in radiology practice such as disease diagnosis, monitoring, and treatment planning. However, contemporary literature lacks appropriate *RNA-radiomics* (henceforth, ***radiogenomics***) joint modeling where RNAseq distribution is adaptive and also preserves the nature of RNAseq read counts data for glioma grading and prediction. The Negative Binomial (NB) distribution may be useful to model RNAseq read counts data that addresses potential shortcomings. In this study, we propose a novel radiogenomics-NB model for glioma grading and prediction. Our radiogenomics-NB model is developed based on differentially expressed RNAseq and selected radiomics/volumetric features which characterize tumor volume and sub-regions. The NB distribution is fitted to RNAseq counts data, and a log-linear regression model is assumed to link between the estimated NB mean and radiomics. Three radiogenomics-NB molecular mutation models (e.g., *IDH* mutation, *1p/19q codeletion*, and *ATRX* mutation) are investigated. Additionally, we explore gender-specific effects on the radiogenomics-NB models. Finally, we compare the performance of the proposed three mutation prediction radiogenomics-NB models with different well-known methods in the literature: Negative Binomial Linear Discriminant Analysis (NBLDA), differentially expressed RNAseq with Random Forest (RF-genomics), radiomics and differentially expressed RNAseq with Random Forest (RF-radiogenomics), and Voom-based count transformation combined with the nearest shrinkage classifier (VoomNSC). Our analysis shows that the proposed radiogenomics-NB model significantly outperforms (ANOVA test, $p < 0.05$) for prediction of *IDH* and *ATRX* mutations and offers similar performance for prediction of *1p/19q codeletion*, when compared to the competing models in the literature, respectively.

Keywords: RNA sequencing, radiomics, radiogenomics, negative binomial, molecular mutation

OPEN ACCESS

Edited by:
*Benedikt Wiestler,
Technical University of
Munich, Germany*

Reviewed by:
*Yu Qian,
University of Texas MD Anderson
Cancer Center, United States
Chunlin Ou,
Central South University, China*

***Correspondence:**
*Khan M. Iftekharuddin
kiftekha@odu.edu*

Specialty section:
*This article was submitted to
Translational Medicine,
a section of the journal
Frontiers in Medicine*

Received: *13 May 2021*
Accepted: *20 July 2021*
Published: *19 August 2021*

Citation:
*Shboul ZA, Diawara N, Vossough A,
Chen JY and Iftekharuddin KM (2021)
Joint Modeling of RNAseq and
Radiomics Data for Glioma Molecular
Characterization and Prediction.
Front. Med. 8:705071.
doi: 10.3389/fmed.2021.705071*

INTRODUCTION

Radiomics is increasingly being applied to radiology practice in disease diagnosis, grading, monitoring, and treatment planning (1, 2). Radiomics is extracted from various radiological images of a targeted area of the disease. Fusing the important radiomics and genomics information in the proper computational machine learning (ML) model may helpto achieve a more comprehensive disease diagnosis, prognosis, and treatment planning scheme (3–5). Different studies have evaluated the association between glioma molecular subtypes and radiomics (e.g., tumor shape and size) (6–8), or between different form of genomics (e.g., RNA sequencing (RNAseq) gene expression, protein expression, copy number, molecular mutations, or DNA methylation) and glioma subtypes (9–11).

Conventional ML models do not adequately model the count-based nature of the RNA-sequence data as these models are usually designed to work with data that has a normal distribution. In order to alleviate the lack of appropriate ML models, researchers propose to transform the RNAseq read-count data to approximate a normal distribution. The transformation to normal distribution allows the use of existing methods such as the nearest shrinkage method (12, 13) or Random Forest for classification. However, such transformation removes the count-based nature of the RNAseq read counts data, and hence, lacks the ability to fully preserve the strong mean-variance relationship that is otherwise useful for glioma classification and prediction (14, 15). In order to appropriately model RNAseq read-count data, Negative Binomial (NB) and Poisson distributions are commonly used (16). The Poisson distribution is a single parameter distribution with its mean equals to its variance, which makes it rather restrictive. On the other hand, NB is similar to a Poisson distribution with an additional parameter called "dispersion" that allows the NB distribution to modify its variance without affecting the mean.

RNAseq uses high-throughput or next-generation sequencing technology (NGS) and has emerged as a novel alternative to microarray-based techniques for quantifying gene expression. The microarray technique is known to suffer from background noise. Gene expression level is measured as the relative frequency of the RNAseq reads that are mapped to one gene (17). RNAseq is a very sensitive technique that provides high resolution and a thorough understanding of the transcriptome and has revealed many novel gene structures.

RNAseq distribution requires an appropriate model that adapts and preserves the nature of RNAseq read counts data, and such classification models that preserve the nature of RNAseq are lacking in the traditional ML literature. The NB distribution is an appropriate choice to model such discrete reads counts data (16). Even though traditional ML tools that are developed based on NB are lacking, the choice of using NB distribution in differential gene expression and RNAseq analysis has been adapted by different studies in the literature such as in EdgeR (18–20), DESeq (21), and NBPSeq (22).

An example of a count-based classifier that fits a NB distribution is the Negative Binomial Linear Discriminant Analysis (NBLDA). NBLDA is a well-known classifier that is developed by fitting NB to RNAseq and the mean and dispersion parameter are estimated from the RNAseq data (23). A different type of classifier, known as VoomNSC, is developed based on the transformed count data. VoomNSC is a combination of Voom (an acronym for mean-variance modeling at the observational level) transformation (12) and the nearest shrunken centroids classifier (NSC) (24).

Consequently, the aim of this work is to implement a joint radiogenomics-NB model that predicts and classifies glioma molecular mutations following the 2016 World Health Organization's (WHO) updated guidelines for classification of tumors of the Central Nervous System (CNS) (including high grade and diffuse low-grade gliomas) (25). This work is critical especially when the RNAseq of some cases are unknown and a careful assessment is needed to avoid mischaracterization of lower grade gliomas. In this work, we utilize both volumetric features (radiomics) and RNAseq to implement and learn a radiogenomics-NB model. Then, the trained radiogenomics model is used to predict and classify the unknown RNAseq data. In the proposed model, a log-linear regression modeling is fitted to the estimated mean of the NB distribution and is linked with radiomics. We introduce this step to fuse the continuous radiomics data with the RNAseq count-based data without the need to transform RNAseq data into a normal distribution. Finally, we compare our radiogenomics-NB model performance with that of different genomics and radiogenomics state-of-the-art methods in the literature.

The rest of the paper is organized as follows. A complete step-by-step mathematical derivation of the radiogenomics-NB model and parameters' estimations are presented in section Methodology. Section Experimental Results addresses the dataset used in this study, the data preparation, and the effect of using different numbers of differentially expressed genes in the radiogenomics-NB model. Furthermore, in section Experimental Results, a comparative analysis is discussed in which we compare the proposed radiogenomics-NB model's performance with different well-known methods in the literature. Moreover, in section Experimental Results, we investigate the effect of gender by developing a gender-specific radiogenomics-NB model for glioma molecular grading. Finally, the study's discussion is addressed in section Discussion.

METHODOLOGY

In this study, we propose a radiogenomics-NB method for glioma molecular grading and prediction. **Figure 1** illustrates an overall flow diagram of the proposed radiogenomics-NB model. In **Figure 1A**, we fit the NB distribution to RNAseq read counts of the training dataset and estimate the model mean and dispersion parameter. Then, we use the estimated mean along with the predictor radiomics vector in a log-linear regression model to estimate the model regression coefficients. The dispersion parameter is estimated using the weighted likelihood empirical Bayes method (19). In **Figure 1B**, the estimated parameters of regression coefficients and the dispersion parameters along with

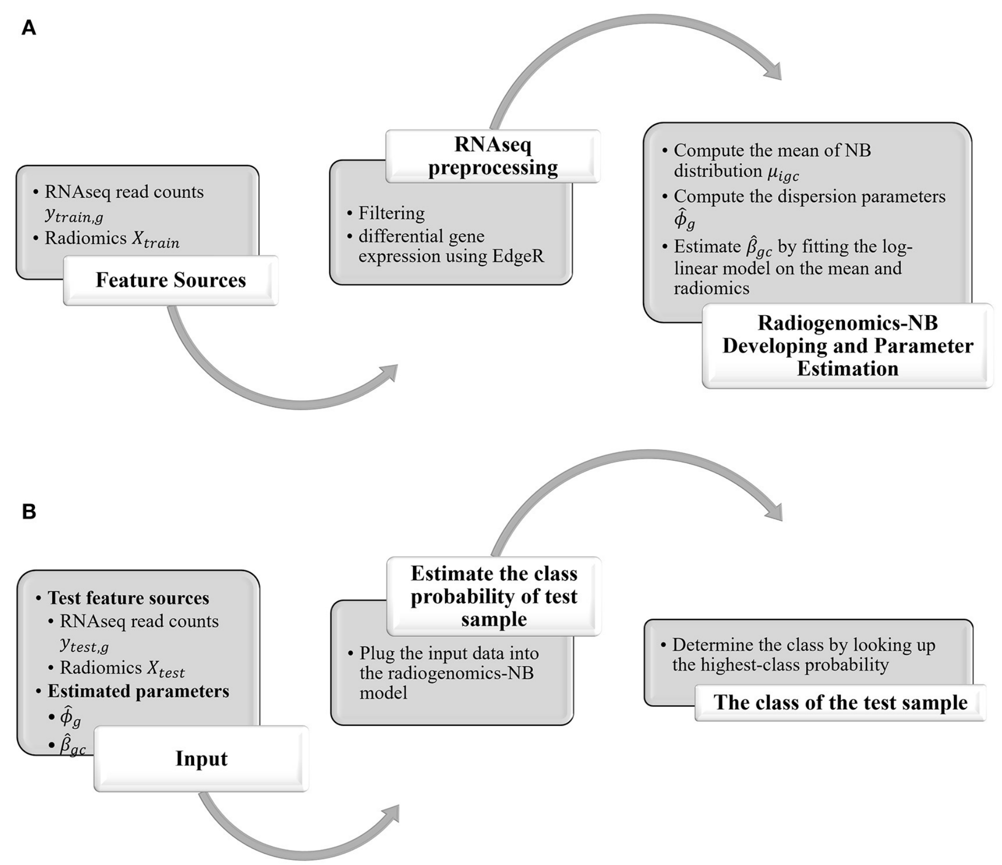

A
• RNAseq read counts $y_{train,g}$
• Radiomics X_{train}
Feature Sources
RNAseq preprocessing
• Filtering
• differential gene expression using EdgeR
• Compute the mean of NB distribution μ_{igc}
• Compute the dispersion parameters $\hat{\phi}_g$
• Estimate $\hat{\beta}_{gc}$ by fitting the log-linear model on the mean and radiomics
Radiogenomics-NB Developing and Parameter Estimation
B
• Test feature sources
• RNAseq read counts $y_{test,g}$
• Radiomics X_{test}
• Estimated parameters
• $\hat{\phi}_g$
• $\hat{\beta}_{gc}$
Input
Estimate the class probability of test sample
• Plug the input data into the radiogenomics-NB model
• Determine the class by looking up the highest-class probability
The class of the test sample

FIGURE 1 | Overall Flow diagram of the proposed radiogenomics-NB prediction model. **(A)** radiogenomics-NB model utilizing the training data. **(B)** class prediction of a test sample using the developed radiogenomics-NB model.

the sample radiomics and its RNAseq read counts are utilized to predict the class label of a future test sample. A complete mathematical derivation of the radiogenomics-NB model is presented in the following subsection.

Prediction Using Negative Binomial Regression Model

To fuse radiomics with RNAseq read counts data in an NB model, the following parametrization is defined:

Let C be the total number of classes, and $I_c \in (1, \ldots, n_c)$ be the indices of samples in class c for $c = 1, \ldots, C$. The examples of different classes include:

IDH mutated vs. wildtype *IDH* ($C = 2$),

1p/19q codeletion: *codeletion* vs. *non-codeletion* ($C = 2$),

Mutated *ATRX* vs. wildtype ($C = 2$).

Let $Y_i = \left(y_{i1}, y_{i2}, \ldots, y_{iG}\right)$ be the RNAseq read counts training sample in the class label c and G is the total number of RNAseq. The purpose of this study is to predict the class label c of a future observation Y_t using training samples associated with known class labels: $p\left(c|Y_t\right) \propto p\left(Y_t|c\right)p_c$, where p_c is the probability of class c.

Using Bayes' rule, we have,

$$p\left(c|Y_i\right) \propto p\left(Y_i|c\right)p_c; \tag{1}$$

where, $p(Y_i|\ c)$ is the pdf of the sample Y_i in class c, and p_c is the prior probability that one sample comes from class c. The pdf of class-specific c of RNAseq read counts of sample Y_i and of RNAseq g is,

$$P\left(Y_{ig} = y_{ig}|c\right) = \frac{\Gamma\left(\phi_g^{-1} + y_{ig}\right)}{\Gamma\left(\phi_g^{-1}\right)y_{ig}!}\left(\frac{\phi_g\mu_{igc}}{1+\phi_g\mu_{igc}}\right)^{y_{ig}} \left(\frac{1}{1+\phi_g\mu_{igc}}\right)^{\phi_g^{-1}}. \tag{2}$$

In this parameterization, Y_{ig} represents a count response of RNAseq, where μ_{igc} represents the mean, ϕ_g represents the dispersion parameter, $E\left(Y_{ig}\right) = \mu_{igc}$, and $Var\left(Y_{ig}\right) = \mu_{igc} + \mu_{igc}\phi_g^2$. Note we assume that all RNAseq are independent of each other, so we have,

$$p\left(Y_i|c\right) = \prod_{g=1}^{G} P\left(Y_{ig} = y_{ig}\right). \tag{3}$$

Evaluating Equation (1) requires an estimation of $p\left(Y_i|c\right)$ and p_c. The model in Equation (2) states that $Y_{ig} \sim NB\left(\mu_{igc}, \phi_g\right)$. We first estimate $\phi_1,\ \phi_2,\ \ldots, \phi_G$, and $\mu_{i1c},\ \mu_{i2c},\ \ldots,\ \mu_{iGc}$ of all the training samples n_c, and all RNAseq G. The mean is estimated as $\mu_{igc} = s_{ic}\lambda_{gc}$, where s_{ic} is the size factor (26, 27) which is used to scale RNAseq counts for the *ith* sample (in class c), λ_{gc} is the total number of reads of RNAseq g across all samples in class c. For prior p_c, we assume all classes are equally likely, $p_c = 1/C$. Note that $\mu_{igc},\ s_{ic},\ and\ \lambda_{gc}$ are estimated for each class c.

Next, plugging these estimates into Equation (2) and using the assumption of independent RNAseq, Equation (1) yields,

$$\log(p(c|Y_i)) \ = \ \log(p(Y_i|c) \ + \ \log(p_c). \tag{4}$$

The log-likelihood $\log\left(p\left(Y_i|c\right)\right)$ is written as,

$$\log\left(p\left(Y_i|c\right)\right) = \log\left(\prod\nolimits_{g=1}^{G} P\left(Y_{ig} = y_{ig}|c\right)\right)$$

$$= \log\left(\prod\nolimits_{g=1}^{G} \frac{\Gamma\left(\phi_g^{-1} + y_{ig}\right)}{\Gamma\left(\phi_g^{-1}\right) y_{ig}!} \times \left(\frac{\phi_g\mu_{igc}}{1+\phi_g\mu_{igc}}\right)^{y_{ig}} \times \left(\frac{1}{1+\phi_g\mu_{igc}}\right)^{\phi_g^{-1}}\right). \tag{5}$$

Equation (5) can be written as,

$$\begin{aligned} log\left(p\left(Y_i|c\right)\right) \ = \ & \sum\nolimits_{g=1}^{G} log\left(\frac{\phi_g\mu_{igc}}{1+\phi_g\mu_{igc}}\right)^{y_{ig}} \\ & + \sum\nolimits_{g=1}^{G} log\left(\frac{1}{1+\phi_g\mu_{igc}}\right)^{\phi_g^{-1}} \\ & + \sum\nolimits_{g=1}^{G} log\left(\frac{\Gamma\left(\phi_g^{-1} + y_{ig}\right)}{\Gamma\left(\phi_g^{-1}\right) y_{ig}!}\right). \end{aligned} \tag{6}$$

Rewriting Equation (6) yields,

$$\begin{aligned} \log\left(p\left(Y_i|c\right)\right) \ = \ & \sum_{g=1}^{G} y_{ig}\log\left(\phi_g\mu_{igc}\right) \ - \ \sum_{g=1}^{G} y_{ig}\log\left(1+\phi_g\mu_{igc}\right) \\ & - \sum_{g=1}^{G} \frac{1}{\phi_g}\log\left(1+\phi_g\mu_{igc}\right) \\ & + \ \sum_{g=1}^{G} \log\left(\frac{\Gamma\left(\phi_g^{-1} + y_{ig}\right)}{\Gamma\left(\phi_g^{-1}\right) y_{ig}!}\right). \end{aligned} \tag{7}$$

The proposed NB model of genomics relates to the radiomics (imaging features) **X** through the mean parameters μ_{igc} (estimated mean of an *ith* sample and RNAseq g in class c). We assume a log-linear regression model for estimating the mean μ_{igc} in terms of the radiomics (imaging features) is given as follows:

$$\begin{aligned} \log\left(\mu_{igc}\right) \ &= \ X_i\beta_{gc}; \qquad (8.a) \\ \log\left(s_{ic}\lambda_{gc}\right) \ &= \ X_i\beta_{gc}; \qquad (8.b) \end{aligned}$$

where X_i is a p-dimensional of radiomics, β_{gc} is a p-dimensional vector of unknown regression coefficients (translate the relationship between X and Y through μ_{igc}). The estimation of β_{gc} depends on class c and gene g of the *ith* sample. Hence, if there are two classes, we will need to estimate $\beta_{g1}$and β_{g2} (one from each class).

Plugging Equations (8.a) into Equation (7), yields,

$$\begin{aligned} \log\left(p\left(Y_i|c\right)\right) \ = \ & \sum\nolimits_{g=1}^{G} y_{ig}\log\left(\phi_g\exp\left(X_i\beta_{gc}\right)\right) \\ & - \ \sum\nolimits_{g=1}^{G} y_{ig}\log\left(1+\phi_g\exp\left(X_i\beta_{gc}\right)\right) \\ & - \ \sum\nolimits_{g=1}^{G} \frac{1}{\phi_g}\log\left(1+\phi_g\exp\left(X_i\beta_{gc}\right)\right) \\ & + \ \sum\nolimits_{g=1}^{G} log\left(\frac{\Gamma\left(\phi_g^{-1} + y_{ig}\right)}{\Gamma\left(\phi_g^{-1}\right) y_{ig}!}\right). \end{aligned} \tag{9}$$

Using the estimated $\hat{\beta}_{gc}$, and $\hat{\phi}_g$ from the *training* data, we classify a *test* observation Y_t as follows,

$$\log(p(c|Y_t)) \ = \ \log(p\left(Y_t|c\right) + \ log\left(p_c\right); \tag{10}$$

and,

$$\begin{aligned} log\left(p\left(c|Y_t\right)\right) \ = \ & \sum\nolimits_{g=1}^{G} y_{tg}\ log\left(\hat{\phi}_g\exp\left(X_t\hat{\beta}_{gc}\right)\right) \\ & - \ \sum\nolimits_{g=1}^{G} y_{tg}\ log\left(1+\hat{\phi}_g\exp\left(X_t\hat{\beta}_{gc}\right)\right) \\ & - \ \sum\nolimits_{g=1}^{G} \frac{1}{\phi_g}\ log\left(1+\hat{\phi}_g\exp\left(X_t\hat{\beta}_{gc}\right)\right) \\ & + \ \sum\nolimits_{g=1}^{G} log\left(\frac{\Gamma\left(\hat{\phi}_g^{-1} + y_{tg}\right)}{\Gamma\left(\hat{\phi}_g^{-1}\right) y_{tg}!}\right) \ + \ log\left(p_c\right). \end{aligned} \tag{11}$$

Radiogenomics-NB Model Parameter Estimation

Estimating Dispersion ϕ_g Using Weighted Likelihood Empirical Bayes

Various methods for estimating the dispersion parameter are proposed in the literature. The EdgeR method applies a weighted conditional log-likelihood method to estimate the dispersion parameter (19). The weighted conditional log-likelihood (WL) for ϕ_g is defined as a weighted combination of the individual (per-gene) likelihood $l_g\left(\phi_g\right)$ and common $l_C\left(\phi_g\right)$ likelihood:

$$WL\left(\hat{\phi}_g\right) = \ l_g\left(\phi_g\right) \ + \ \alpha l_C\left(\phi_g\right); \tag{12}$$

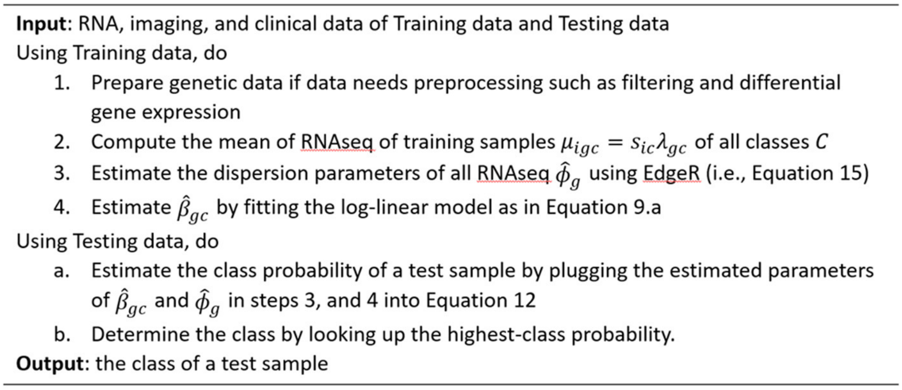
where α is the weight of $l_C\left(\phi_g\ \right)$.

In EdgeR, $\hat{\phi}_g$ is assumed to be normally distributed with means ϕ_g and known variance τ^2, and has the following hierarchical model:

$$\hat{\phi}_g|\phi_g \sim N(\phi_g, \tau^2),\ and\ \phi_g \sim N(\phi_0, \tau_0^2). \tag{13}$$

Under this hierarchical normal model, the maximum weighted conditional log-likelihood estimator is given as:

$$\hat{\phi}_g^{WL} = \frac{\hat{\phi}_g/\tau^2 \ + \alpha\sum\nolimits_{i=1}^{G}\hat{\phi}_i/\tau_i^2}{1/\tau^2 \ + \alpha\sum\nolimits_{i=1}^{G} 1/\tau_i^2}; \tag{14}$$



```
Input: RNA, imaging, and clinical data of Training data and Testing data
Using Training data, do
    1. Prepare genetic data if data needs preprocessing such as filtering and differential
       gene expression
    2. Compute the mean of RNAseq of training samples $\mu_{igc} = s_{ic}\lambda_{gc}$ of all classes $C$
    3. Estimate the dispersion parameters of all RNAseq $\hat{\phi}_g$ using EdgeR (i.e., Equation 15)
    4. Estimate $\hat{\beta}_{gc}$ by fitting the log-linear model as in Equation 9.a
Using Testing data, do
    a. Estimate the class probability of a test sample by plugging the estimated parameters
       of $\hat{\beta}_{gc}$ and $\hat{\phi}_g$ in steps 3, and 4 into Equation 12
    b. Determine the class by looking up the highest-class probability.
Output: the class of a test sample
```

FIGURE 2 | Algorithm of prediction using radiogenomics Negative Binomial classification model.

where,

$$1/\alpha = \sum_{i=1}^{G} \tau_0^2/\tau_i^2 \tag{15}$$

and,

$$\phi_0 = \hat{\phi}_0 = \frac{\sum_{i=1}^{G} \hat{\phi}_i/\tau_i^2}{\sum_{i=1}^{G} 1/\tau_i^2}. \tag{16}$$

Computation of the Mean of RNAseq μ_{igc}

The size factor s_{ic} of sample i and class c is the total number of RNAseq read counts of that sample divided by the total number of all RNAseq read counts across all training samples (in class c). The size factor estimation is vital to account for the different sequencing depth (library size) that may be used to sequence different samples and is computed as follows:

$$s_{ic} = \frac{\sum_{g=1}^{G} y_{igc}}{\sum_{i=1}^{n_c}\sum_{g=1}^{G} y_{igc}}; \tag{17}$$

where, y_{igc} is the RNAseq read count of sample i and RNAseq g in class c, and n_c is the total number of samples in class c.

The mean μ_{igc} of sample i and RNAseq g in class c is then estimated as $\mu_{igc} = s_{ic}\lambda_{gc}$, where λ_{gc} is the total number of reads per RNAseq in class c, and is computed as follows:

$$\lambda_{gc} = \sum_{i=1}^{N_c} y_{igc}. \tag{18}$$

Using the estimated value of μ_{igc}, the values of β_{gc} are computed using equation 8.a as follows:

$$\beta_{gc} = X_i \log\left(\frac{\sum_{g=1}^{G} y_{igc}}{\sum_{i=1}^{n_c}\sum_{g=1}^{G} y_{igc}} \sum{}_{i=1}^{N_c} y_{igc}\right). \tag{19}$$

The algorithm in **Figure 2** illustrates the steps of estimating the different parameters in the radiogenomics-NB classification model.

EXPERIMENTAL RESULTS

Dataset

The dataset in this study consists of 108 pre-operative lower grade glioma (LGG) patients that are described in Menze et al. (28), Bakas et al. (29), and Bakas et al. (30). Four sequences of the MRI are provided with the dataset: pre-contrast T1-weighted (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR). These scans are skull-stripped, re-sampled to 1 mm^3 resolution, and co-registered to the T1 template. The dataset provides the segmented sub-regions of the LGG: Gadolinium enhancing tumor (ET), the peritumoral edema (ED), and necrosis along with non-enhancing tumor (NCR/NET).

RNAseq read counts data (with a total number of 56830 RNAseq), molecular alterations (*IDH* mutation, *1p/19q codeletion*, and *ATRX*), grade (II and III), and the clinical dataset can be found and downloaded from The Cancer Genome Atlas (TCGA) dataset in the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/). RNAseq are primarily obtained from solid portions of tumor. The clinical dataset is de-identified in compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA). The distribution of the data is as follows: (i) *IDH* mutation: 85 Mutant and 23 wildtype (WT), (ii) *1p/19q codeletion*: 27 *codeletion* and 81 *non-codeletion*, and (iii) *ATRX* status: 43 Mutant and 65 WT. The range of the patients' age at diagnosis is 20–75 years, and the median age is 46.5 years.

Data Preparation

In this study we first filter RNAseq read counts to remove RNAseq with very low value of read counts before performing any statistical analysis. RNAseq with very low read counts hold very little information because an RNAseq of biological importance needs to be expressed at some minimal level. We utilize a quantile filter (31) with a quantile threshold of 0.25. This step returns each RNAseq that has a mean across all samples higher than the defined quantile threshold of 0.25. Then, we reduce the number of RNAseq that are used in the radiogenomics-NB models, by

TABLE 1 | Radiomics features description and their ANOVA *p-value* association with *IDH* mutations, *1p/19q codeletion*, and *ATRX* mutations.

Feature number	Feature description	*p-value* of *IDH* mutation	*p-value* of *1p/19q codeletion*	*p-value* of *ATRX* mutation
1	the size of the enhancing tumor to the necrosis size	<0.005	0. 393	0.178
2	the size of the enhancing tumor to the size of enhancing tumor and necrosis	0.8630	0.070	0.239
3	the size of the enhancing tumor to the edema size	<0.005	0.600	<0.005
4	the size of the enhancing tumor to the whole tumor size	<0.005	0.707	0.027
5	the size of the edema to the necrosis size	0.188	0.996	0.114
6	the size of the edema to the size of enhancing tumor and necrosis	0.138	0.789	0.0237
7	the size of the edema to the whole tumor size	<0.005	0.131	<0.005
8	and the size of the necrosis to the whole tumor size	<0.005	0.221	<0.005

utilizing EdgeR (18–20) to extract the differentially expressed RNAseq (DERs). DERs reflect the significance of a gene in a certain biological condition. In this study, we select the top 10, 20, 30, 50, 100, and 150 DERs (see **Supplementary Table 1**).

Furthermore, we use eight volumetric radiomics features as illustrated in **Table 1**. ANOVA analysis for radiomics in **Table 1** shows that feature numbers 1, 3, 4, 7, and 8 are significantly associated (ANOVA test, $p < 0.05$) with *IDH* mutations as illustrated in **Figure 3A**. Our analysis also indicates that feature number 2 is marginally associated (ANOVA test, $p = 0.07$) with *1p/19q codeletion*. Furthermore, our analysis indicates that feature numbers 3, 4, 6, 7, and 8 are significantly associated (ANOVA test, $p < 0.05$) with *ATRX* mutations as illustrated in **Figure 3B**. Additionally, our analysis reveals that thresholding feature number 6 around the mean creates an ordinal feature that is significantly associated (ANOVA test, $p < 0.05$) with *IDH* mutations, *1p/19q codeletion*, and *ATRX* mutations. Likewise, thresholding feature numbers 1, 3, 5, 7, and 8 around their means converts these features into ordinal features that are significantly associated (ANOVA test, $p < 0.05$) with *IDH* and *ATRX* mutations. Moreover, thresholding feature numbers 5, 6, 7, and 8 around their median converts these features into ordinal features that are significantly associated (ANOVA test, $p < 0.05$) with *IDH* and *ATRX* mutations.

Few other studies suggest that these volumetric imaging features and their ratios are associated with and predictive of several mutations in gliomas (32–35).

The 108 LGG cases are randomly split into 80% training and 20% testing sets, and a balanced distribution of the target molecular alteration is ensured in the training and testing sets in each molecular classifier. The trained model classifier is developed using the training set. Model performance prediction is estimated and reported using the testing sets in terms of accuracy, balanced accuracy, F1 score, sensitivity, specificity, negative predictive value, and positive predictive value. The training set is utilized to build our radiogenomics-NB classifier as shown in steps **1-4** in **Figure 2**. The testing set is used to estimate the performance of the classifier as shown in steps **a** and **b** in **Figure 2**. Authors In Dong et al. (23), Maufroy et al. (36), Pan et al. (37), and Vabalas et al. (38) repeat training and testing analysis for a specific number of times to ensure the robustness of the model performance. Consequently, in this work, we repeat the whole procedure 100 times independently for the 3 molecular alterations and then report the mean and standard deviation of the classifiers' performance using the testing sets.

Model performance parameters are computed based on the confusion matrix in **Figure 4** as follows:

$$Accuracy = TP + \frac{TN}{TP} + TN + FP + TN; \tag{20}$$

$$Sensitivity = \frac{TP}{TP} + FN; \tag{21}$$

$$Specificity = \frac{TN}{FP} + TN; \tag{22}$$

$$Positive\ predictive\ value = \frac{TP}{TP} + FP; \tag{23}$$

$$Negative\ predictive\ value = \frac{TN}{FN} + TN; \tag{24}$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}, \text{and} \tag{25}$$

$$F1\ score = TP\left(TP + \frac{FP + FN}{2}\right); \tag{26}$$

where TP is the true positive, TN is the true negative, FP is the false positive, and TN is the true negative.

Radiogenomics-NB Models Using Different Number of Differentially Expressed RNAs

In this section, we investigate the importance of using different numbers of DERs on the performance of the radiogenomics-NB model. LGG radiogenomics-NB mutation prediction models are developed based on the top 10, 20, 30, 50, 100, and 150 DERs. The performance of the radiogenomics-NB *IDH* model using the top 10 DERs achieves slightly higher performance. However, such improvement is not statistically significant (ANOVA test, $p > 0.05$) when compared to the performance of the *IDH* models with the other number of DERs (**Figure 5A**) except for negative predictive value (NPV) performance when using the top 20 DERs. Using the top 20 DERs in the *IDH* model achieves significantly worse NPV when compared to the NPV achieved using the top 10 DERs (ANOVA test, $p < 0.05$). Radiogenomics-NB *IDH* model with the top 10 DERs (red line in **Figure 5A**) achieves an overall accuracy (Acc) of 0.92 ± 0.06, sensitivity (Sens) of 0.94 ± 0.07, specificity (Spec)

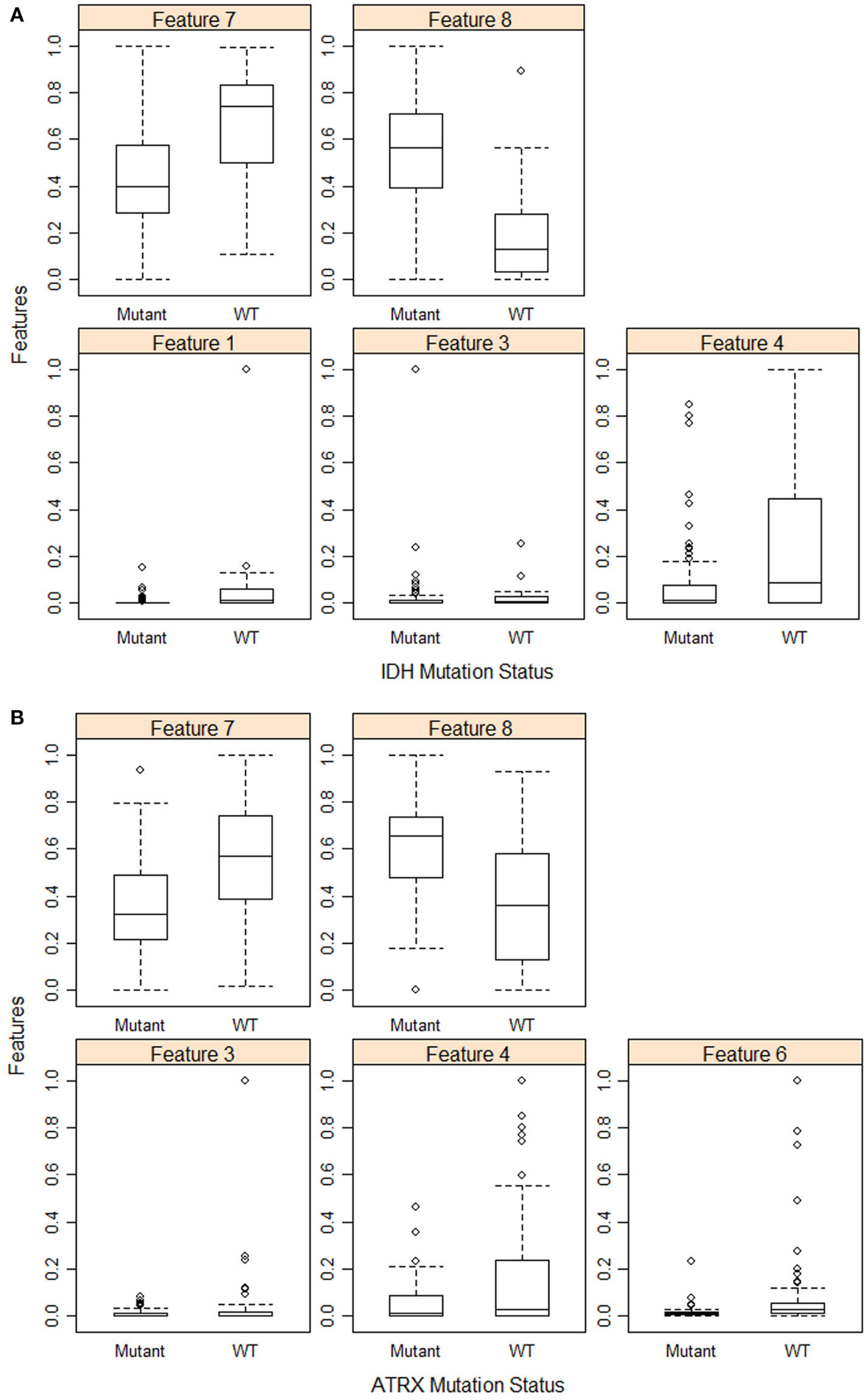
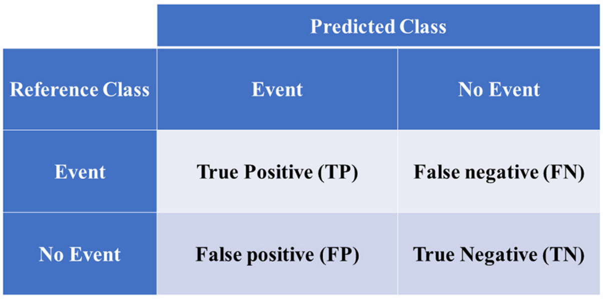


FIGURE 3 | Feature distribution plot of the significant volumetric radiomic associated with **(A)** *IDH* mutations, and **(B)** *ATRX* mutations.

of 0.83 ± 0.18, positive predictive value (PPV) of 0.96 ± 0.04, negative predictive value (NPV) of 0.82 ± 0.17, F1 score of 0.95 ± 0.04, and balanced accuracy (B. Acc) of 0.88 ± 0.09, respectively.

Radiogenomics-NB *codeletion* models achieve similar performance (ANOVA test, $p > 0.05$) using the top 10, 20, 30, and 50 DERs as shown in **Figure 5B**. Furthermore, using the top 100 and 150 DERs in the *codeletion* model achieves significantly



	Predicted Class	
Reference Class	Event	No Event
Event	True Positive (TP)	False negative (FN)
No Event	False positive (FP)	True Negative (TN)

FIGURE 4 | Confusion matrix of binary classification.

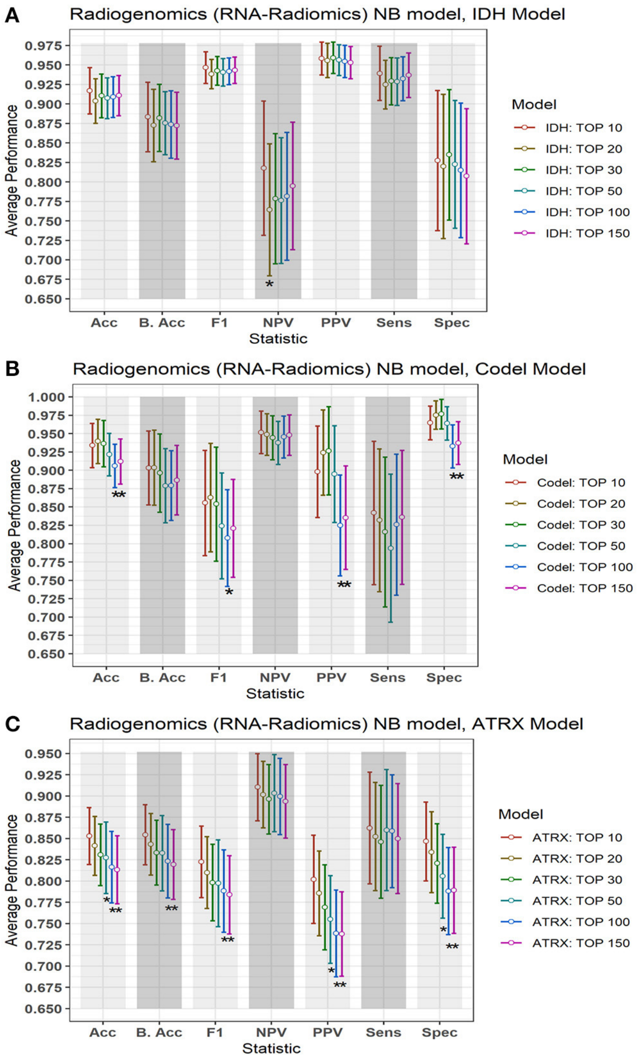


FIGURE 5 | Performance of the proposed radiogenomics-NB model using a different number of DERs. **(A)** Radiogenomics-NB *IDH*, **(B)** Radiogenomics-NB *Codeletion*, and **(C)** Radiogenomics-NB *ATRX* models. The average performance (of the Acc, B. Acc, F1, NPV, PPV, Sens, and Spec) is computed across 100 testing sets/splits. Y-axis represents the average performance of the different statistics on the X-axis. Different colors represent the radiogenomics-NB model with different numbers of DERs. The error bar represents one standard deviation. Asterisk "*" represents a statically significant difference between the performance achieved when using the top 10 DERs (in red) and using the number of DER where the star is located.

worse performance when compared to the performance of using the top 10 DERs (ANOVA test, $p < 0.05$). Using the top 10 DERs, the radiogenomics-NB *codeletion* model achieves an accuracy of 0.93 ± 0.06, a balanced accuracy of 0.90 ± 0.10, F1 score of 0.86 ± 0.14, a sensitivity of 0.84 ± 0.19, a specificity of 0.96 ± 0.04, an NPV of 0.95 ± 0.06, and a PPV of 0.90 ± 0.12, respectively.

Radiogenomics-NB *ATRX* model also achieves similar performance (ANOVA test, $p > 0.05$) using the top 10, 20, and 30 DERs, even though the performance when using the top 10 DERs is slightly better as illustrated in **Figure 5C**. Using the top 10 DERs, the *ATRX* model achieves an accuracy of 0.85 ± 0.07, a balanced accuracy of 0.85 ± 0.07, an F1 score of 0.82 ± 0.08, a sensitivity of 0.86 ± 0.13, a specificity of 0.85 ± 0.09, an NPV of 0.91 ± 0.08, and a PPV of 0.80 ± 0.10, respectively.

Comparative Analysis

Figure 6 illustrates a graphical performance comparison between our radiogenomics-NB model with that of four different classifiers in the literature: NBLDA (23), VoomNSC (12, 13), RF-genomics where we first log-transformed (20) the RNAseq into a normal distribution, and RF-radiogenomics. Note that the number of DERs that we apply to develop these classifiers is 10 DERs. Moreover, when developing these classifiers, the 108 LGG cases are randomly split into 80% training and 20% testing sets, and balanced distribution is ensured when developing the different classifiers. The trained model classifier is developed using the training set, and 10-fold cross-validation is performed to identify the tuning parameters in the different classifiers. Model performance prediction is estimated and reported using the testing sets. Additionally, to ensure the robustness of the different classifiers' performance, we repeat the whole procedure 100 times independently and every training/testing set is utilized to develop and estimate the performance of each classifier.

The NBLDA (23) classifier is developed by fitting NB to the top 10 DERs; then the mean and dispersion parameter are estimated from these DERs. In RF-genomics, the top 10 DERs of the training sets are first log-transformed into normal distribution and then fed into RF to build the RF-genomics classifier. In RF-radiogenomics, radiomics (eight volumetric features described previously in section Data Preparation) are utilized with the log-transformed DERs and then fed into RF to build the RF-radiogenomics classifier. VoomNSC (12, 24) is developed by first applying the Voom-based transformation on the 10 DERs and then applying the NSC classifier as illustrated in Zararsiz et al. (12) and Tibshirani et al. (24).

Comparing the performance of our radiogenomics-NB *IDH* model with that of NBLDA, RF-genomics, and VoomNSC, the radiogenomics-NB *IDH* significantly outperforms (ANOVA test, $p < 0.05$) these methods as shown in **Figure 6A** and **Table 2**. Additionally, our radiogenomics-NB *IDH* model significantly

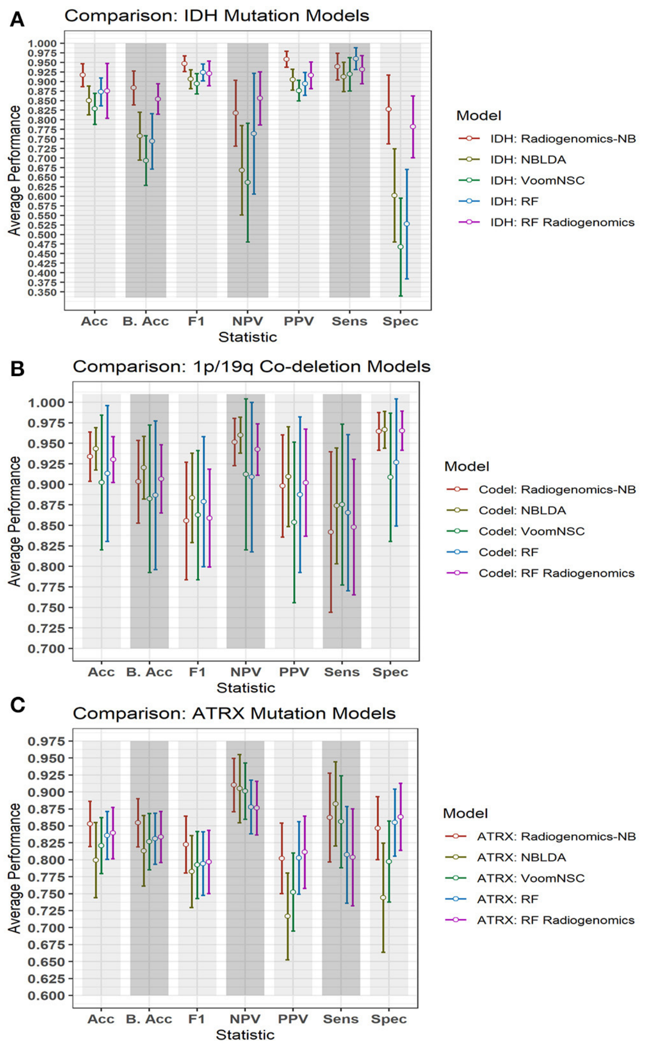


FIGURE 6 | Comparison of performance between our radiogenomics-NB model and different classifiers. The comparison is performed using the **(A)** *IDH* mutations, **(B)** *1p/19q codeletion*, and **(C)** *ATRX* mutations dataset. The average performance (of the Acc, B. Acc, F1, NPV, PPV, Sens, and Spec) is computed across 100 test sets. The error bar represents one standard deviation. RNAseq that are used in developing all classifiers represent the top 10 DERs in the training sets between mutated and WT *IDH* group, codeleted and non-codeleted groups, and mutated and WT *ATRX* mutation, respectively. *Y*-axis represents the average performance of the different statistics on the *X*-axis. Different colors represent different classifiers.

outperforms (ANOVA test, $p < 0.05$) the F1 score, balanced accuracy, and PPV performance of the RF-radiogenomics method whereas it achieves a similar (ANOVA test, $p > 0.05$) accuracy, sensitivity, and specificity. Our radiogenomics-NB *IDH* model archives an accuracy of 0.92 ± 0.06, a sensitivity of 0.94 ± 0.07, a specificity of 0.93 ± 0.18, an F1 score of 0.95 ± 0.04, and a balanced accuracy of 0.88 ± 0.09, respectively. The RF-radiogenomics-*IDH* model achieves an accuracy of 0.88 ± 0.17, a sensitivity of 0.93 ± 0.07, a specificity of 0.78 ± 0.16, an F1 score of 0.92 ± 0.06, and a balanced accuracy of 0.85 ± 0.08, respectively.

Our radiogenomics-NB *codeletion* model (**Figure 6B** and **Table 3**) performance is similar to NBLDA, RF-genomics, VoomNSC, and RF-radiogenomics models, except for the specificity and NPV performance when using RF-genomics and VoomNSC. The specificity and NPV of our model are significantly higher than those achieved by RF-genomics and VoomNSC. Our radiogenomics-NB *codeletion* model achieves an accuracy of 0.93 ± 0.06, a sensitivity of 0.84 ± 0.20, a specificity of 0.96 ± 0.5, an F1 score of 0.86 ± 0.14, and a balanced accuracy of 0.90 ± 0.10, respectively.

The performance of our radiogenomics-NB *ATRX* model as shown in **Figure 6C** and **Table 4** outperforms both NBLDA and VoomNSC significantly (ANOVA test, $p < 0.05$). However, comparing our *ATRX* model to RF-genomics, our model achieves significantly better balanced-accuracy, F1 score, NPV, and sensitivity. Additionally, comparing our *ATRX* model to RF-radiogenomics, our model achieves significantly (ANOVA test, $p < 0.05$) better sensitivity but achieves similar accuracy, balanced-accuracy, F1 score, and sensitivity. Our radiogenomics-NB *ATRX* model achieves an accuracy of 0.85 ± 0.07, a sensitivity of 0.86 ± 0.13, a specificity of 0.85 ± 0.09, an F1 score of 0.82 ± 0.08, and a balanced accuracy of 0.85 ± 0.07, respectively. The RF-radiogenomics *ATRX* model achieves an accuracy of 0.84 ± 0.08, a sensitivity of 0.80 ± 0.14, a specificity of 0.86 ± 0.10, an F1 score of 0.80 ± 0.09, and a balanced accuracy of 0.83 ± 0.08, respectively.

Gender–Specific Effect Analysis of Radiogenomics-NB

In our LGG dataset, *IDH* mutated patients, unlike *IDH* WT patients, have significantly longer survival (65.7 vs. 19.9 months, log-rank test $p = 0.004$). The association between *IDH* status and overall survival remains significant after stratifying for gender (likelihood ratio test $p = 0.015$). However, the association between *1p/19q codeletion* and *ATRX* status and overall survival is not significant. Additionally, the chi-square test shows no significant association ($p > 0.05$) between gender and *IDH* status, *1p/19q codeletion*, and *ATRX* status. **Table 5** shows patient *IDH* status, *1p/19q codeletion*, and *ATRX* status distribution based on gender.

To explore the gender-specific effect in the performance of the radiogenomics-NB, we build two radiogenomics-NB models based on gender; male-specific radiogenomics-NB and female-specific radiogenomics-NB. Our analysis indicates that female-specific models significantly outperform (ANOVA test, $p < 0.05$) male-specific models as illustrated in **Figure 7**. In the radiogenomics-NB *IDH*, female-specific model achieves an accuracy of 0.93 ± 0.08, a sensitivity of 0.93 ± 0.09, a specificity of 0.91 ± 0.10, a PPV of 0.97 ± 0.05, an NPV of 0.83 ± 0.21, and a balanced accuracy of 0.92 ± 0.11, respectively. The male specific *IDH* model achieves an accuracy of 0.85 ± 0.08, a sensitivity of 0.97 ± 0.06, a specificity of 0.35 ± 0.33, a PPV of 0.86 ± 0.07, an NPV of 0.55 ± 0.48, and a balanced accuracy of 0.66 ± 0.17, respectively.

In the radiogenomics-NB *codeletion*, female-specific model achieves an accuracy of 0.91 ± 0.09, a sensitivity of 0.77 ±

TABLE 2 | Probability of significant difference using ANOVA test between the differentially expressed radiogenomics-NB model and different classifiers using the *IDH* dataset.

IDH	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	Balanced accuracy
radiogenomics-NB vs. NBLDA	**0.000**	**0.010**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
radiogenomics-NB vs. VoomNSC	**0.000**	0.075	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
radiogenomics-NB vs. RF	**0.001**	0.023	**0.000**	**0.000**	0.138	**0.000**	**0.000**
radiogenomics-NB vs. RF-radiogenomics	0.069	0.432	0.061	**0.000**	0.084	**0.001**	**0.01**

A statistically significant difference exists if p < 0.05. Values in bold show a significant improvement of our radiogenomics-NB IDH over the compared one.

TABLE 3 | Probability of significant difference using ANOVA test between the differentially expressed radiogenomics-NB model and different models using the *1p/19q codeletion* dataset.

CODEL	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	Balanced accuracy
radiogenomics-NB vs. NBLDA	0.232	0.186	0.756	0.514	0.253	0.123	0.181
radiogenomics-NB vs. VoomNSC	0.072	0.228	**0.001**	0.057	**0.042**	0.742	0.317
radiogenomics-NB vs. RF	0.242	0.390	**0.020**	0.636	**0.027**	0.271	0.42
radiogenomics-NB vs. RF-radiogenomics	0.671	0.815	0.893	0.825	0.282	0.855	0.792

A statistically significant difference exists if p < 0.05. Values in bold show a significant improvement of our radiogenomics-NB codeletion over the compared one.

TABLE 4 | Probability of significant difference using ANOVA test between the differentially expressed radiogenomics-NB model and different models using the *ATRX* dataset.

ATRX	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	Balanced accuracy
Radiogenomics-NB vs. NBLDA	**0.000**	0.269	**0.000**	**0.000**	0.677	**0.004**	**0.001**
Radiogenomics-NB vs. VoomNSC	**0.003**	0.741	**0.001**	**0.002**	0.432	**0.021**	**0.012**
Radiogenomics-NB vs. RF	0.083	**0.005**	0.540	0.960	**0.004**	**0.026**	**0.025**
Radiogenomics-NB vs. RF-radiogenomics	0.183	**0.003**	0.215	0.561	**0.003**	0.052	0.053

A statistically significant difference exists if p < 0.05. Values in bold show a significant improvement of our radiogenomics-NB ATRX over the compared one.

TABLE 5 | Gender-based distribution of *IDH* status, *1p/19q codeletion*, and *ATRX* status in the LGG dataset.

	IDH status		1p/19q codeletion		*ATRX* status	
	Mutant	WT	Codeletion	Non-codeletion	Mutant	WT
Female	43	14	14	43	24	33
Male	42	9	13	38	19	32

0.31, a specificity of 0.96 ± 0.07, a PPV of 0.80 ± 0.32, an NPV of 0.93 ± 0.07, and a balanced accuracy of 0.84 ± 0.17, respectively. The male specific *codeletion* model achieves an accuracy of 0.84 ± 0.10, a sensitivity of 0.56 ± 0.32, a specificity of 0.95 ± 0.08, a PPV of 0.79 ± 0.32, an NPV of 0.86 ± 0.10, and a balanced accuracy of 0.77 ± 0.17, respectively.

In the radiogenomics-NB *ATRX*, female-specific model achieves an accuracy of 0.80 ± 0.11, a sensitivity of 0.79 ± 0.20, a specificity of 0.81 ± 0.15, a PPV of 0.76 ± 0.16, an NPV of 0.87 ± 0.12, and a balanced accuracy of 0.80 ± 0.12, respectively. The male specific *ATRX* model achieves an accuracy of 0.76 ± 0.12, a sensitivity of 0.69 ± 0.23, a specificity of 0.81 ± 0.14, a PPV of 0.73 ± 0.18, an NPV of 0.81 ± 0.12, and a balanced accuracy of 0.75 ± 0.13, respectively.

DISCUSSION

In this study, we propose a novel radiogenomics-NB model to fuse radiomics (imaging features) with RNAseq (genes) for glioma grading and prediction. NB distribution is appropriate for modeling RNAseq discrete read counts data and for preserving the count-based nature of this data. In the proposed radiogenomics-NB model, log-linear regression modeling is fitted to the estimated mean of the NB distribution and is linked with radiomics. We introduce this step to fuse the continuous radiomics data with the RNAseq count-based data without the need to transform the RNAseq data into a normal distribution.

The NB, unlike a Poisson distribution, has two parameters; the mean (e.g., the expected value of the RNAseq read counts data) and dispersion (e.g., a parameter that helps in capturing

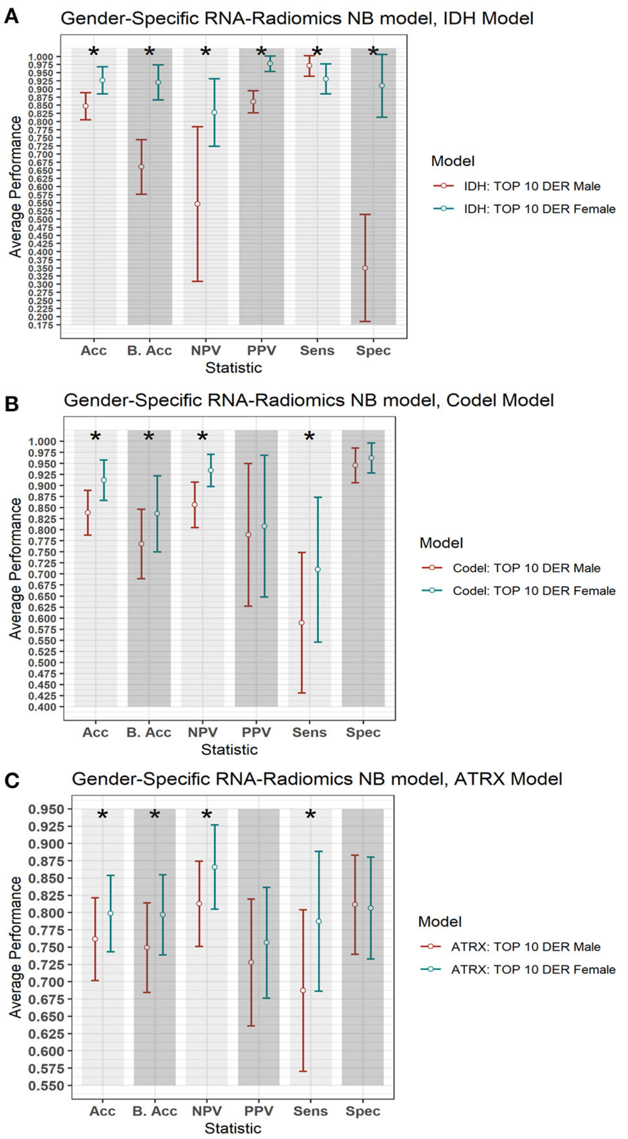


FIGURE 7 | Gender-based radiogenomics-NB models performance. **(A)** *IDH* mutations, **(B)** *1p/19q codeletion*, and **(C)** *ATRX* mutations which are computed across 100 testing sets. The error bar represents one standard deviation. The asterisk * illustrates a significant difference between the two measurements. *Y*-axis represents the average performance of the different statistics on the *X*-axis. Different colors represent the female- and male-specific radiogenomics-NB models.

the variability of the RNAseq read counts). If the dispersion of NB is zero, the model reduces to Poisson distribution. In Poisson distribution, the mean is equal to the variance, which makes it rather restrictive. However, variation is usually observed in the real data of RNAseq counts data that the Poisson distribution cannot handle properly. On the other hand, NB has an additional parameter called the "dispersion" that allows the NB distribution of RNAseq counts data to modify its variance without affecting the mean. Thus, NB serves as a practical approximation to model RNAseq count data with variability different from its mean.

The mean of the proposed radiogenomics-NB model is estimated as the size factor multiplied by the total number of reads per RNAseq. Moreover, we utilize EdgeR to estimate the dispersion of the proposed radiogenomics-NB assuming RNAseq variability is assessed using the weighted conditional log-likelihood model. In the weighted conditional model, RNAseq counts data is assumed to have a distinct and individual dispersion for each RNAseq in addition to a common dispersion. Such an assumption can be more reliable when estimating the dispersion of real data of RNAseq counts data.

The performance evaluation of the proposed work indicates that linking simple, clinically feasible radiomics (i.e., tumor volumetric features) to RNAseq improves the performance of *IDH* and *ATRX* mutations prediction. The radiomics features utilized in the proposed radiogenomics-NB model that are described in **Table 1** mainly depend on volumetric features. Our analysis shows that these features are associated with particular glioma mutations. This outcome supports previous studies that show the association between volumetric features and glioma mutations (32–35). The efficacy of the proposed radiogenomics-NB model is further investigated using the top 10, 20, 30, 50, 100, and 150 DERs, respectively. Our analysis shows that the smaller the number of DERs (fewer than 30 DERs) utilized in radiogenomics-NB, the better is the radiogenomics-NB model performance. Our analyses indicate that using fewer than 30 DERs in our analysis offers the best performance (statically significant) in the radiogenomics-NB codeletion and ATRX prediction model. This suggests that using large numbers of DERs (more than 30) in the proposed radiogenomics-NB would over parametrize the dataset and create model fitting problems and thus degrade the performance.

Comparing our radiogenomics-NB model to NBLDA, RF-genomics, FR-radiogenomics, and VoomNSC, our model significantly outperforms NBLDA, RF-genomics, and VoomNSC for prediction of *IDH* and *ATRX* mutations. Our radiogenomics-NB model offers similar performance as NBLDA, RF-genomics, RF-radiogenomics, and VoomNSC models for prediction of *1p/19q codeletion*. Specifically, for prediction of *IDH* mutations, while the proposed radiogenomics-NB model achieves significantly better balanced-accuracy, F1 score, and PPV than RF-radiogenomics, our model achieves similar accuracy, sensitivity, and specificity. Such results indicate the power of fusing radiomics and genomics data to develop radiogenomics models for classification and prediction models. The findings in this work indicate that the radiomics volumetric features may be vital for the prediction of *IDH* and *ATRX* mutations along with the genomics.

Different studies have revealed that gender is a significant factor in identifying cancer survival, prognosis, and treatment response (39–41). Hence, improved glioma molecular mutation prediction may require the development of gender-specific models. In this study, we explore the gender-specific effect on the radiogenomics-NB models. Our analysis reveals that *IDH* mutated patients remain significant after stratifying for gender, unlike *1p/19q codeletion* and *ATRX* status. Moreover, our analysis indicates that no association is found between gender and the three specific mutations (*IDH* mutations, *1p/19q codeletion,* and *ATRX* status) using the Chi-square test. This result is in agreement with the findings in Brat et al. (42), Li et al. (43), and Ebrahimi et al. (44). However, our gender-specific modeling shows that female-specific radiogenomics-NB models significantly outperform the male-specific radiogenomics-NB models for prediction of *IDH* status, *1p/19q codeletion*, and *ATRX* status, respectively.

In conclusion, we present a glioma mutations radiogenomics-NB prediction model that preserves the count nature of RNAseq counts data in the NB model and utilizes radiomics to develop a complete and a better characterization prediction model of patient data. Our analysis shows the superiority of utilizing both genomics and clinically feasible radiomics data when compared to only genomics models. Use of tumor volumetrics can be more easily and reproducibly implemented in clinical practice compared to more complex radiomics metrics, such as higher order texture analysis features. Finally, this study shows the efficacy of volumetric radiomics features in the radiogenomics-NB model for glioma molecular characterization and prediction. This study is a first step toward implementing joint modeling of RNAseq and MRI patient data for glioma grading. However, further investigation is needed with a larger dataset with both RNAseq and full multimodality MRI dataset for each patient in a cohort. In the future, larger prospective studies may be needed to investigate specific radiomics features and their association with the different mutations and RNAseq read counts data for implementation into clinical workflow. Furthermore, it will be interesting to investigate the cause of superior performance of female-specific radiogenomics-NB models when compared to that of the male-specific radiogenomics-NB models for prediction of *IDH* status, *1p/19q codeletion*, and *ATRX* status. Also, these models may be further investigated in treatment response and survival prediction in the future.

DATA AVAILABILITY STATEMENT

Radiomics data are available at doi: 10.7937/K9/TCIA.2017.GJQ7R0EF. Genomics and mutations data are available at: https://portal.gdc.cancer.gov/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Old Dominion University IRB. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

ZS and KI: conception and design and development of methodology. ZS, ND, and KI: analysis and interpretation of data. ZS, ND, AV, JC, and KI: drafting the article and/or revising. KI: funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This work is supported by National Institutes of Health (R01 EB020683).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.705071/full#supplementary-material

REFERENCES

1. Weissleder R, Schwaiger MC, Gambhir SS, Hricak H. Imaging approaches to optimize molecular therapies. *Sci Transl Med.* (2016) 8:355ps316-355ps316. doi: 10.1126/scitranslmed.aaf3936
2. O'Connor JP, Aboagye EO, Adams JE, Aerts HJ, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol.* (2017) 14:169. doi: 10.1038/nrclinonc.2016.162
3. Reza SM, Samad MD, Shboul ZA, Jones KA, Iftekharuddin KM. Glioma grading using structural magnetic resonance imaging and molecular data. *J Med Imaging.* (2019) 6:024501. doi: 10.1117/1.JMI.6.2.024501
4. Shboul ZA, Iftekharuddin KM. Prediction of low-grade glioma progression using MR imaging. In: *Medical Imaging 2019: Computer-Aided Diagnosis*: *International Society for Optics and Photonics.* San Diego, CA (2019).
5. Shboul ZA, Iftekharuddin KM. Efficacy of radiomics and genomics in predicting TP53 mutations in diffuse lower grade glioma. In: *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*: *International Society for Optics and Photonics.* Houston, TX (2020).
6. Kickingereder P, Bonekamp D, Nowosielski M, Kratz A, Sill M, Burth S, et al. Radiogenomics of glioblastoma: machine learning–based classification of molecular characteristics by using multiparametric and multiregional MR imaging features. *Radiology.* (2016) 281:907–18. doi: 10.1148/radiol.2016161382
7. Mazurowski MA, Clark K, Czarnek NM, Shamsesfandabadi P, Peters KB, Saha A. Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with The Cancer Genome Atlas data. *J Neurooncol.* (2017) 133:27–35. doi: 10.1007/s11060-017-2420-1
8. Rathore S, Akbari H, Rozycki M, Abdullah KG, Nasrallah MP, Binder ZA, et al. Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. *Sci Rep.* (2018) 8:5087. doi: 10.1038/s41598-018-22739-2
9. Iwadate Y, Sakaida T, Hiwasa T, Nagai Y, Ishikura H, Takiguchi M, et al. Molecular classification and survival prediction in human gliomas based on proteome analysis. *Cancer Res.* (2004) 64:2496–501. doi: 10.1158/0008-5472.CAN-03-1254
10. Zhang X, Sun S, Pu JKS, Tsang ACO, Lee D, Man VOY, et al. Long non-coding RNA expression profiles predict clinical phenotypes in glioma. *Neurobiol Dis.* (2012) 48:1–8. doi: 10.1016/j.nbd.2012.06.004
11. Zeng W-J, Yang Y-L, Liu Z-Z, Wen Z-P, Chen Y-H, Hu X-L, et al. Integrative analysis of DNA methylation and gene expression identify a three-gene signature for predicting prognosis in lower-grade gliomas. *Cellular Physiology and Biochemistry.* (2018) 47:428–39. doi: 10.1159/000489954
12. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* (2014) 15:R29. doi: 10.1186/gb-2014-15-2-r29
13. Zararsiz G, Goksuluk D, Klaus B, Korkmaz S, Eldem V, Karabulut E, et al. voomDDA: discovery of diagnostic biomarkers and classification of RNA-seq data. *PeerJ.* (2017) 5:e3890. doi: 10.7717/peerj.3890

14. Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. *Genome Biol.* (2010) 11:1–10. doi: 10.1186/gb-2010-11-12-220
15. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* (2012) 40:4288–97. doi: 10.1093/nar/gks042
16. Gardner W, Mulvey EP, Shaw EC. Regression analyses of counts and rates: poisson, overdispersed Poisson, and negative binomial models. *Psychol Bull.* (1995) 118:392. doi: 10.1037/0033-2909.118.3.392
17. Kukurba KR, Montgomery SB. RNA sequencing and analysis. *Cold Spring Harbor Prot.* (2015) 2015 :pdb. top084970. doi: 10.1101/pdb.top084970
18. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics.* (2007) 9:321–32. doi: 10.1093/biostatistics/kxm030
19. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics.* (2007) 23:2881–7. doi: 10.1093/bioinformatics/btm453
20. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616
21. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* (2010) 11:R106. doi: 10.1186/gb-2010-11-10-r106
22. Di Y, Schafer DW, Cumbie JS, Chang JH. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat Appl Genet Mol Biol.* (2011) 10:1–28. doi: 10.2202/1544-6115.1637
23. Dong K, Zhao H, Tong T, Wan X. NBLDA: negative binomial linear discriminant analysis for RNA-Seq data. *BMC Bioinformatics.* (2016) 17:369. doi: 10.1186/s12859-016-1208-1
24. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science.* (2003) 18:104–17. doi: 10.1214/ss/1056397488
25. Louis DN, Perry A, Reifenberger G, Von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* (2016) 131:803–20. doi: 10.1007/s00401-016-1545-1
26. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* (2008) 18:1509–17. doi: 10.1101/gr.079558.108
27. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* (2008) 5:621. doi: 10.1038/nmeth.1226
28. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging.* (2015) 34:1993–2024. doi: 10.1109/TMI.2014.2377694
29. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Nat Sci Data*. (2017) 4:170117. doi: 10.1038/sdata.2017.117
30. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby J, et al. Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection. *Cancer Imaging Arch.* (2017) 286. doi: 10.7937/K9/TCIA.2017.GJQ7R0EF
31. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* (2015) 44:e71. doi: 10.1093/nar/gkv1507
32. Metellus P, Coulibaly B, Colin C, De Paula AM, Vasiljevic A, Taieb D, et al. Absence of IDH mutation identifies a novel radiologic and molecular subtype of WHO grade II gliomas with dismal prognosis. *Acta Neuropathol.* (2010) 120:719–29. doi: 10.1007/s00401-010-0777-8
33. Gutman DA, Dunn WD, Grossmann P, Cooper LA, Holder CA, Ligon KL, et al. Somatic mutations associated with MRI-derived volumetric features in glioblastoma. *Neuroradiology.* (2015) 57:1227–37. doi: 10.1007/s00234-015-1576-7
34. Park Y, Han K, Ahn S, Bae S, Choi Y, Chang J, et al. Prediction of IDH1-mutation and 1p/19q-codeletion status using preoperative MR imaging phenotypes in lower grade gliomas. *Am J Neuroradiol.* (2018) 39:37–42. doi: 10.3174/ajnr.A5421
35. Thust S, Hassanein S, Bisdas S, Rees J, Hyare H, Maynard J, et al. Apparent diffusion coefficient for molecular subtyping of non-gadolinium-enhancing WHO grade II/III glioma: volumetric segmentation versus two-dimensional region of interest analysis. *Eur Radiol.* (2018) 28:3779–88. doi: 10.1007/s00330-018-5351-0
36. Maufroy A, Chassot E, Joo R, Kaplan DM. Large-scale examination of spatio-temporal patterns of drifting fish aggregating devices (dFADs) from tropical tuna fisheries of the Indian and Atlantic Oceans. *PLoS ONE.* (2015) 10:e0128023. doi: 10.1371/journal.pone.0128023
37. Pan L, Liu G, Lin F, Zhong S, Xia H, Sun X, et al. Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. *Sci Rep.* (2017) 7:1–9. doi: 10.1038/s41598-017-07408-0
38. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLoS ONE.* (2019) 14:e0224365. doi: 10.1371/journal.pone.0224365
39. Yuan Y, Liu L, Chen H, Wang Y, Xu Y, Mao H, et al. Comprehensive characterization of molecular differences in cancer between male and female patients. *Cancer Cell.* (2016) 29:711–22. doi: 10.1016/j.ccell.2016.04.001
40. Ippolito JE, Yim AK-Y, Luo J, Chinnaiyan P, Rubin JB. Sexual dimorphism in glioma glycolysis underlies sex differences in survival. *JCI Insight.* (2017) 2: 92142. doi: 10.1172/jci.insight.92142
41. Yang W, Warrington NM, Taylor SJ, Whitmire P, Carrasco E, Singleton KW, et al. Sex differences in GBM revealed by analysis of patient imaging, transcriptome, and survival data. *Sci Transl Med.* (2019) 11:eaao.5253. doi: 10.1126/scitranslmed.aao5253
42. Brat DJ, Verhaak RG, Aldape KD, Yung WA, Salama SR, Cooper LA, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New Engl J Med.* (2015) 372:2481–98. doi: 10.1056/NEJMoa1402121
43. Li M-Y, Wang Y-Y, Cai J-Q, Zhang C-B, Wang K-Y, Cheng W, et al. Isocitrate dehydrogenase 1 gene mutation is associated with prognosis in clinical low-grade gliomas. *PLoS ONE.* (2015) 10:e0130872. doi: 10.1371/journal.pone.0130872
44. Ebrahimi A, Skardelly M, Bonzheim I, Ott I, Mühleisen H, Eckert F, et al. ATRX immunostaining predicts IDH and H3F3A status in gliomas. *Acta Neuropathol Commun.* (2016) 4:60. doi: 10.1186/s40478-016-0331-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Medicine

METHODS
published: 05 October 2021
doi: 10.3389/fmed.2021.738425

EASE: Clinical Implementation of Automated Tumor Segmentation and Volume Quantification for Adult Low-Grade Glioma

Karin A. van Garderen [1,2,3], Sebastian R. van der Voort [1], Adriaan Versteeg [1], Marcel Koek [1], Andrea Gutierrez [1], Marcel van Straten [1], Mart Rentmeester [1], Stefan Klein [1] and Marion Smits [1,2,3]*

[1] Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, Netherlands, [2] Brain Tumor Center, Erasmus MC Cancer Institute, Rotterdam, Netherlands, [3] Medical Delta, Delft, Netherlands

The growth rate of non-enhancing low-grade glioma has prognostic value for both malignant progression and survival, but quantification of growth is difficult due to the irregular shape of the tumor. Volumetric assessment could provide a reliable quantification of tumor growth, but is only feasible if fully automated. Recent advances in automated tumor segmentation have made such a volume quantification possible, and this work describes the clinical implementation of automated volume quantification in an application named EASE: Erasmus Automated SEgmentation. The visual quality control of segmentations by the radiologist is an important step in this process, as errors in the segmentation are still possible. Additionally, to ensure patient safety and quality of care, protocols were established for the usage of volume measurements in clinical diagnosis and for future updates to the algorithm. Upon the introduction of EASE into clinical practice, we evaluated the individual segmentation success rate and impact on diagnosis. In its first 3 months of usage, it was applied to a total of 55 patients, and in 36 of those the radiologist was able to make a volume-based diagnosis using three successful consecutive measurements from EASE. In all cases the volume-based diagnosis was in line with the conventional visual diagnosis. This first cautious introduction of EASE in our clinic is a valuable step in the translation of automatic segmentation methods to clinical practice.

Keywords: brain tumor, low-grade glioma (LGG), segmentation (image processing), magnetic resonance imaging (MRI), clinical translation, lesion quantification

OPEN ACCESS

Edited by:
Benedikt Wiestler,
Technical University of Munich, Germany

Reviewed by:
Barbara Tomasino,
Eugenio Medea (IRCCS), Italy
Xiyue Wang,
University of Southern California, United States

****Correspondence:***
Marion Smits
marion.smits@erasmusmc.nl

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *08 July 2021*
Accepted: *13 September 2021*
Published: *05 October 2021*

Citation:
van Garderen KA, van der Voort SR, Versteeg A, Koek M, Gutierrez A, van Straten M, Rentmeester M, Klein S and Smits M (2021) EASE: Clinical Implementation of Automated Tumor Segmentation and Volume Quantification for Adult Low-Grade Glioma. Front. Med. 8:738425. doi: 10.3389/fmed.2021.738425

INTRODUCTION

Magnetic resonance (MR) imaging plays a key role in the management of low-grade glioma (LGG) as a method for measuring treatment response and for regular surveillance during periods of watchful waiting. LGG are known to show constant slow growth (1), until—in adults—they inevitably transform to a more malignant type. The early growth rate of the T2-weighted hyperintense region is a known prognostic factor for malignant progression (2) and overall survival (3), so the reliable quantification of growth may be a valuable tool for clinical decision making (4). However, due to the anisotropic growth and irregular size it can be difficult to evaluate slow growth on consecutive imaging using a visual assessment or 2D measurement (5). Volumetric measurements are preferred for the assessment of early growth due to their reproducibility and sensitivity to subtle changes (6), but a manual segmentation would require an effort that is unrealistic in clinical practice.

Automatic segmentation of glioma has shown great advances in recent years due to the release of public datasets and the development of artificial intelligence (7). A recent method described in Kickingereder et al. (8) has been shown to be a reliable alternative for the prognostication of glioma, comparable to the current clinical standard of 2D measurement according to the RANO criteria. Although these criteria apply specifically to high-grade glioma and the measurement of enhancing tumor (6), the performance evaluation in Kickingereder et al. also shows an almost perfect quantification of non-enhancing abnormalities on T2-weighted FLAIR imaging. This makes it potentially suitable for the assessment of volume changes in non-enhancing low-grade glioma.

Due to the clear clinical need of volume quantification in LGG, we decided to implement a segmentation pipeline and integrate it in the existing clinical workflow of the Brain Tumor Center, Erasmus MC Cancer Institute, Rotterdam. This introduced a new measurement tool in the radiologists' toolbox, which we named EASE: Erasmus Automated SEgmentation. With a new tool come potential risks to patient safety and quality of care, which need to be considered in the design of the software and protocols for its use.

For the clinical implementation of this segmentation pipeline, we identified potential risks and practical challenges. The main concern was that of incorrect tumor segmentations resulting in incorrect volume measurements. Further risks were found in software updates over time, potentially leading to unreliable or inconsistent volume measurements, and finally in the incorrect interpretation of volume measurements at time of diagnosis. These risks and the design choices to address these are described in more detail in sections Materials and Equipment and Methods, and an overview is shown in **Table 1**.

This work describes the design of both the technical implementation of EASE and its integration into the clinical workflow, to ensure quality of results and prevent incorrect interpretation of the resulting volume measurements. Furthermore, an initial evaluation of the software was performed in which both the success rate and clinical impact of the volumetric assessment were measured.

TABLE 1 | Overview of identified risks and measures to address those risks.

Risk	Measure
Segmentation errors	Quality check in annotation interface (section Quality Assessment)
Inconsistencies due to updates	Reference dataset and version control (section Validation and Version Control)
Incorrect interpretation of volumes	Design guidelines for usage (section Diagnosis)
	Storage of segmentations in PACS (section Reporting)

MATERIALS AND EQUIPMENT

This section describes the software implementation of EASE. Each scan assessed with EASE goes through a number of processing steps: (1) The images (pre- and post-contrast T1-weighted, T2-weighted, and T2-weighted FLAIR) are received and stored (section Data Management); (2) The segmentation is generated (section Segmentation); (3) The segmentation is checked by a radiologist (section Quality Assessment); (4) A report is generated and sent back to the PACS (section Reporting). A data and state management tool is used to manage the state of each scan and launch processing tasks, in order to balance the workload on the server and enable monitoring of errors in the process. The global software design and data flow are shown in **Figure 1**. The software components for data management, processing and annotation are all open-source, both as separate components and as an adaptable containerized framework[1] using Docker (11).

Data Management

The scan is sent from the PACS (Vue PACS, Carestream Health, v12.2.2.1025) to a dedicated workstation where the scan protocol is automatically checked and the required MR sequences (see section Segmentation) are automatically selected. The images are then stored on a local XNAT database (v1.7) (9), which forms the common database for all further processing steps. The images are stored for a maximum of 6 months to allow for monitoring of the algorithm performance over time, while avoiding unnecessary risk to patient privacy.

Segmentation

The input for the segmentation consists of four MR sequences: pre- and post-contrast T1-weighted, T2-weighted and T2-weighted FLAIR imaging. The pipeline consists of the following steps: first, the images are converted from DICOM to Nifty images using dcm2niix (v1.0.20171215) (12) and co-registered to the postcontrast T1-weighted scan using Elastix (v4.8) (13). Then, they are skull-stripped using HD-BET (git commit 98339a2) (14) and MR bias fields are corrected using N4ITK (using SimpleITK v2.0.2 for Python) (15). The resulting images are used as input for HD-GLIO (v1.5) (14, 16), producing the final delineation of both the enhancing tumor and non-enhancing hyperintensities on T2-weighted FLAIR. Although bias correction is not included in the recommended preprocessing for HD-GLIO, initial tests showed that this improves the performance of the segmentations for scans from our clinic. This pipeline was found, in initial experiments, to perform well on representative images in our center. The Fastr workflow engine (v3.2) (10) was used to integrate these different tools in a robust pipeline.

Quality Assessment

Although the underlying segmentation algorithm, HD-GLIO, was evaluated in a large number of scans and found to be reliable (8), an initial evaluation in our center found that our pipeline

[1] https://gitlab.com/radiology/infrastructure/medical-imaging-demo.

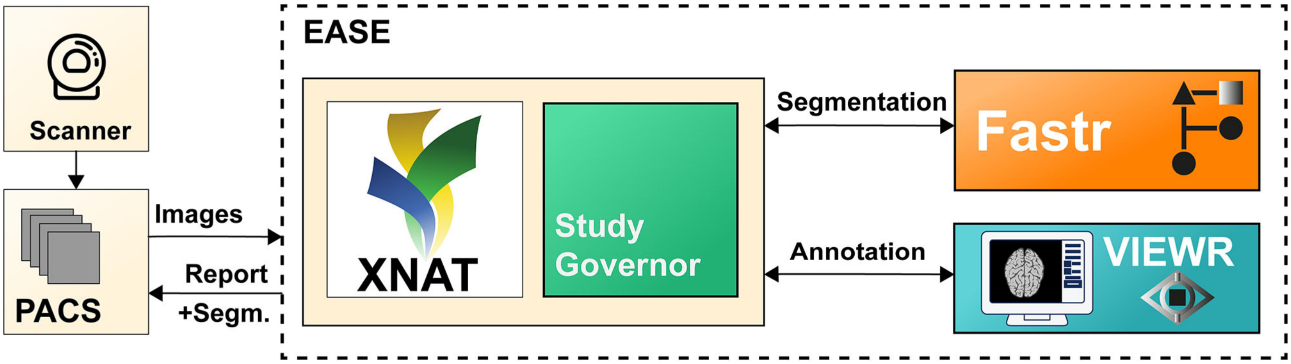


FIGURE 1 | Illustration of the different components of EASE. Images are sent from the PACS and added to the XNAT (9) database. The data and state manager (Study Governor) triggers the processing using Fastr (10). After successful processing, the results can be checked in the VIEWR. A report, including the delineations, is sent back to the PACS.

does not provide perfect segmentations in all scans of low-grade glioma (see section Validation and Version Control). The manual quality assessment of segmentations is therefore essential for the use of EASE in clinical practice. To enable this assessment within a clinical workflow, a dedicated interface was developed for the radiologist to easily assess the segmentation.

The main purpose of the quality assessment is to prevent failed segmentations from being used for a volume-based diagnosis. Additionally, the same quality assessment can be used for the initial validation of the algorithm, prospective evaluation, and continuous monitoring of the segmentation quality. Therefore, besides a binary check on the usability of the segmentation, a more refined quality assessment scoring system was included. Important factors in the design were usability and prevention of human errors.

The interface shows the segmentation as an overlay over all four co-registered scans, and allows for basic interaction through scrolling, manipulation of the contrast, and selecting sequences and imaging planes. The radiologist is asked to evaluate the segmentation both in a binary way (ACCEPTABLE/UNACCEPTABLE) and on an ordinal scale (rating of 1–5, where 5 is the best score). As an additional sanity check, specifically to prevent unnoticed false positives, the interface also lists the number of connected components in the segmentation together with their volumes. Segmentations deemed UNACCEPTABLE cannot be used for diagnosis. A screenshot of the interface is shown in **Figure 2**.

Reporting

Results of the EASE assessment are sent back to the PACS in the form of a report (see **Figure 3**) exported as DICOM file. This report contains the quality assessment, current software version and details of the scan session. Volume measurements are included only if the segmentation is deemed acceptable, to make sure rejected segmentations are not used for diagnosis. In addition to the report, the segmentations are shown as delineations on the T2-weighted FLAIR and post-contrast T1-weighted scan. It would have been possible to store results as a DICOM Structured Report and DICOM SEG respectively, but conventional DICOM images were preferred as not all viewers used in the clinic supported these formats.

METHODS

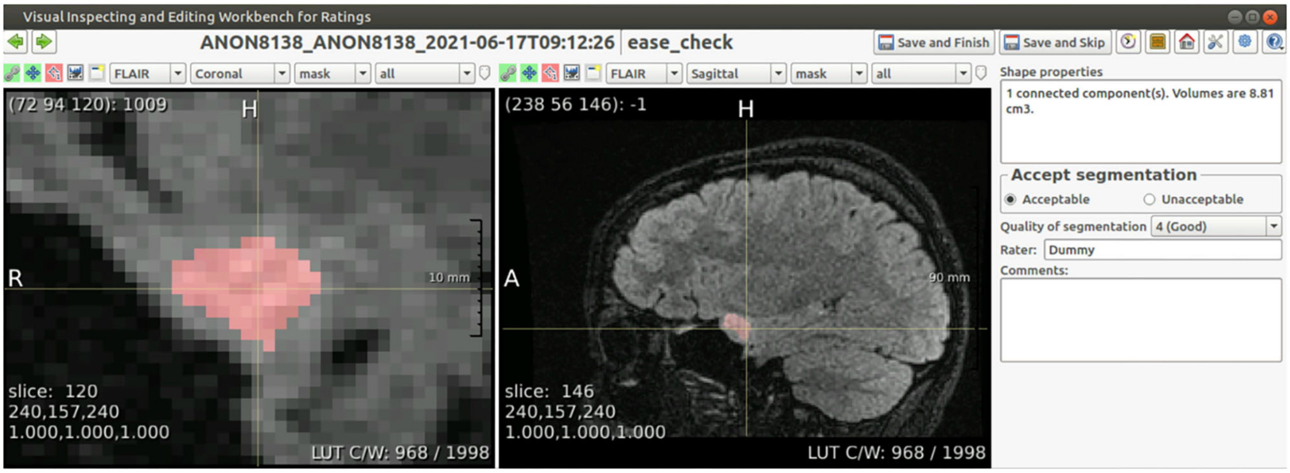
This section describes the protocols for usage of EASE in diagnosis (section Diagnosis), the measures for software validation and version control (section Validation and Version Control), and the method for initial evaluation in clinical practice (section Evaluation in Clinical Practice).

Diagnosis

The purpose of volume measurements produced by EASE is to assess therapy response or progression by estimating tumor growth. The standard clinical procedure for estimating growth is to compare the current measurement to two previous measurements and measure the difference in size, with a manual quantitative measurement of two perpendicular diameters if possible, as described in the RANO guidelines (6). The EASE software provides an automated 3D alternative to the existing measurement. However, as the EASE software has not been tested extensively in this setting, we decided that the existing 2D method should still be performed before using EASE. The volume measurements provided by EASE can lead to further insight and even a different diagnosis, but if there is a discrepancy between the two assessment methods leading to a different conclusion, the diagnosis should be made in consensus with a second radiologist.

The following protocol is in place for the interpretation of automatic volume measurements in clinical practice. The complete workflow is illustrated in **Figure 4**.

1. Two prior reference scans are selected for the assessment (in addition to the current scan).
2. The radiologist assesses the scan using the routine 2D RANO measurement.
3. EASE is applied to all three scans and the segmentations are checked for quality and acceptance. If any scan was already processed and checked previously, this does not have to be repeated.



Visual Inspecting and Editing Workbench for Ratings

ANON8138_ANON8138_2021-06-17T09:12:26 | ease_check

FLAIR | Coronal | mask | all

FLAIR | Sagittal | mask | all

Shape properties: 1 connected component(s). Volumes are 8.81 cm3.

Accept segmentation: ● Acceptable ○ Unacceptable

Quality of segmentation: 4 (Good)

Rater: Dummy

Comments:

FIGURE 2 | Screenshot of the annotation interface. Both image panels can be controlled to show different scan sequences, imaging planes, to change the contrast, to zoom in or out, or to set the overlay transparency. Besides the required annotation of quality, the panel on the right shows the volume of each connected component in the segmentation, and allows for free text comments that are included in the report.

4. If any of the segmentations are rejected, a volumetric assessment is not possible.
5. If all segmentations are accepted, the volumes can be compared.
6. If the volume measurements lead to a change in interpretation compared to the initial assessment after step 2, a second radiologist must be consulted. This second rater first forms an independent opinion of the diagnosis. If this is in line with the first radiologist's opinion, this finalizes the conclusion. If not, both radiologists discuss together how their findings are best described in the report, clearly indicating the uncertainty regarding the findings.

The radiological report clearly describes how each assessment is done (2D RANO, 3D EASE) and how the conclusion is reached. If there was a discrepancy between the two methods, leading to a consensus diagnosis, this should be reflected in the report.

Validation and Version Control

Before deploying the EASE workflow/pipeline, and after any subsequent update, the segmentation quality should be tested in a reference dataset that is representative of the target domain. For this purpose, 20 scans were selected of patients with non-enhancing LGG. All sessions were surveillance scans of patients who had undergone surgical resection, but no further treatment, of LGG. For these scans, the same quality assessment as described in section Quality Assessment was performed by an experienced neuroradiologist.

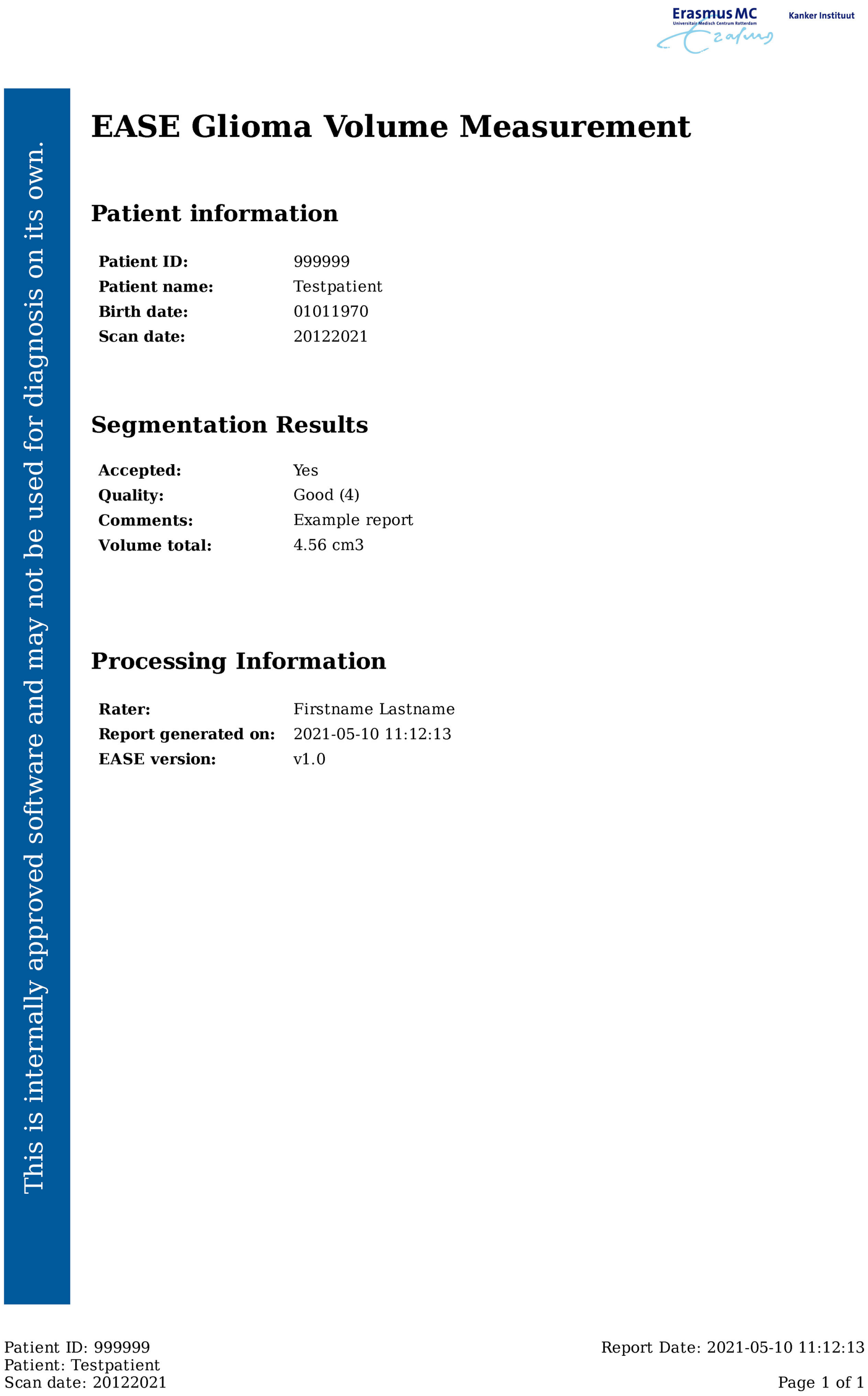
It is essential that updates to the software do not cause a bias in volume that might skew the diagnosis. Therefore, a protocol for software updates was established that allows updates of the processing pipeline while ensuring the continued quality and consistency of the volume measurements. The protocol is as follows:

1. In case of an update, the reference dataset of 20 segmentations is processed again with EASE.
2. The segmentation results are compared to earlier versions of the software. If there is no change in the segmentation, the update can be deployed.
3. If there is a change in results, the manual validation is repeated with the new results.
4. If the qualitative scores are equal or improved with respect to the previous version, the update can be deployed.
5. If the update causes substantial differences in volume (defined as a difference >25%) in any of the accepted segmentations in the reference dataset, the new version is considered incompatible with previous versions and volume results cannot be compared between versions. A warning is included in subsequent EASE reports, so that radiologists know when they have to re-assess previously segmented reference scans with the updated version of EASE.

Evaluation in Clinical Practice

To evaluate the impact of automated segmentation and volume quantification, an observational study was performed for 3 months from first introduction of the software in the clinic. The study protocol was reviewed and approved by the internal review board (MEC-2021-0530). Users were asked to complete a survey after each patient in whom EASE was applied, to measure the success rate of EASE in practice and the rate at which volume quantification leads to a change in diagnosis.

To assess the treatment response or tumor progression in non-enhancing LGG three consecutive volume measurements are required, as the standard clinical procedure is to compare the current scan to two former scans. Therefore, patients were excluded if EASE was applied to the first scan after surgery. Furthermore, patients were excluded if any contrast enhancement was found, which would automatically



Erasmus MC Kanker Instituut

This is internally approved software and may not be used for diagnosis on its own.

EASE Glioma Volume Measurement

Patient information

Patient ID:	999999
Patient name:	Testpatient
Birth date:	01011970
Scan date:	20122021

Segmentation Results

Accepted:	Yes
Quality:	Good (4)
Comments:	Example report
Volume total:	4.56 cm3

Processing Information

Rater:	Firstname Lastname
Report generated on:	2021-05-10 11:12:13
EASE version:	v1.0

Patient ID: 999999
Patient: Testpatient
Scan date: 20122021

Report Date: 2021-05-10 11:12:13

Page 1 of 1

FIGURE 3 | Example of the EASE report.

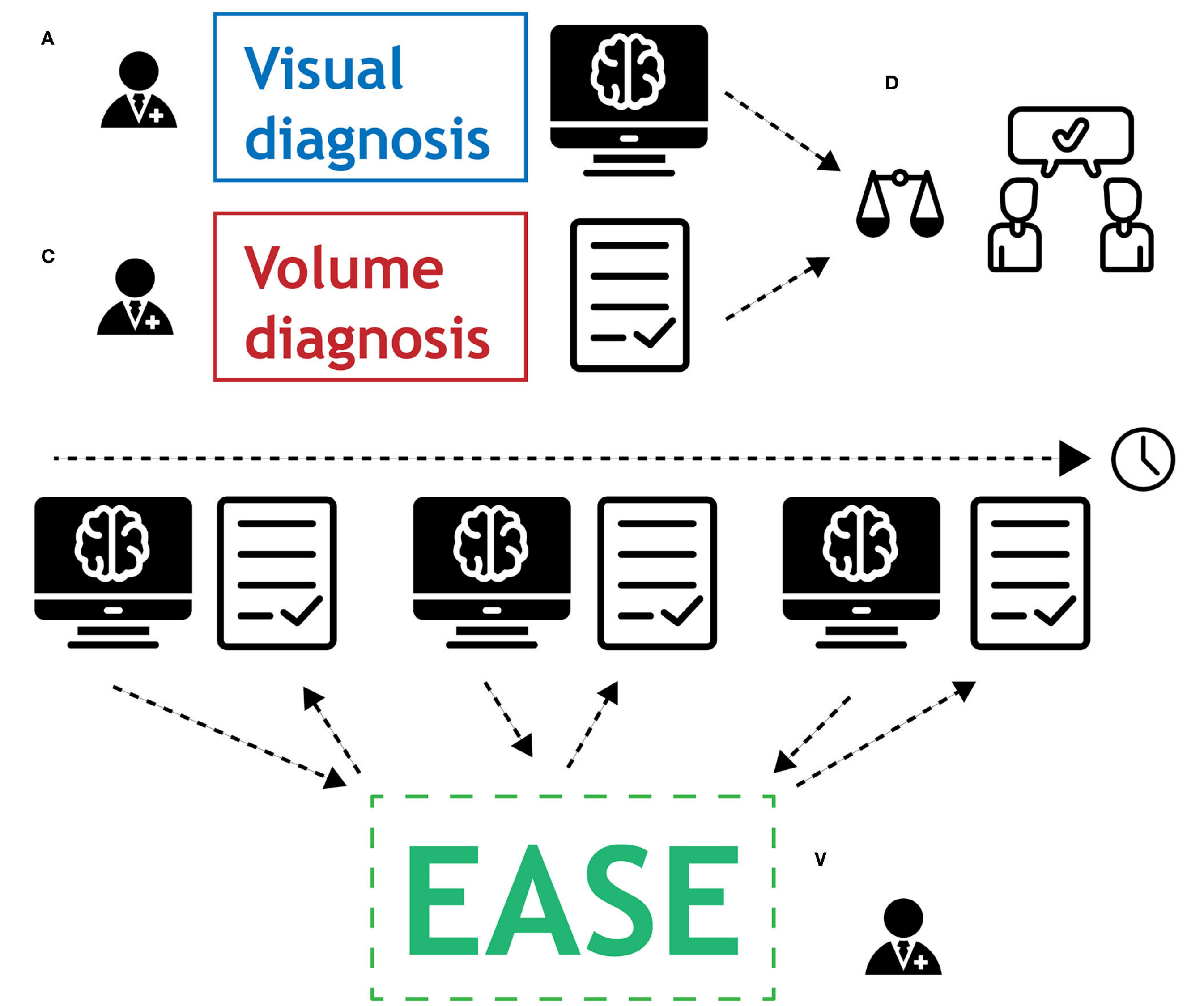


FIGURE 4 | Graphical representation of the protocol for the use of EASE in clinical practice. **(A)** Radiologist applies conventional method for visual diagnosis. **(B)** Segmentations are produced by EASE and assessed separately. **(C)** Radiologist interprets volume measurements. **(D)** If volume measurements lead to change in diagnosis, a second radiologist is consulted for a consensus conclusion.

lead to a diagnosis of tumor progression irrespective of volume measurements.

For each of the included patients, the radiologist was first asked whether EASE had led to a successful diagnosis. Although the success rate of a single segmentation can be extracted from the quality assessments made in the user interface, the success of a full diagnosis requires three accepted segmentations from the same patient. If the diagnosis was unsuccessful, the user was asked to submit the reason for failure.

When the volumetric diagnosis was successful, the radiologist was asked to categorize both the visual (2D) diagnosis and the volume-based diagnosis (through EASE) as progression (PD), stable disease (SD) or treatment response. These results, combined with the quality assessments made in EASE for the individual scans, were used to measure the success rate of EASE and the impact on the clinical diagnosis. The full user survey is shown in **Figure 5** in the form of a flowchart.

Additionally, for the purpose of a quantitative comparison, measurements were made according to the 2D RANO-LGG guidelines (6) if at all possible, measuring two perpendicular diameters of the lesion. As these lesions are often irregular in shape, the diameters were measured in the portion of the lesion that could be measured most reliably.

RESULTS

Of the 20 scans in the reference set, which were processed and evaluated before deployment of EASE, 13 (65%) were considered

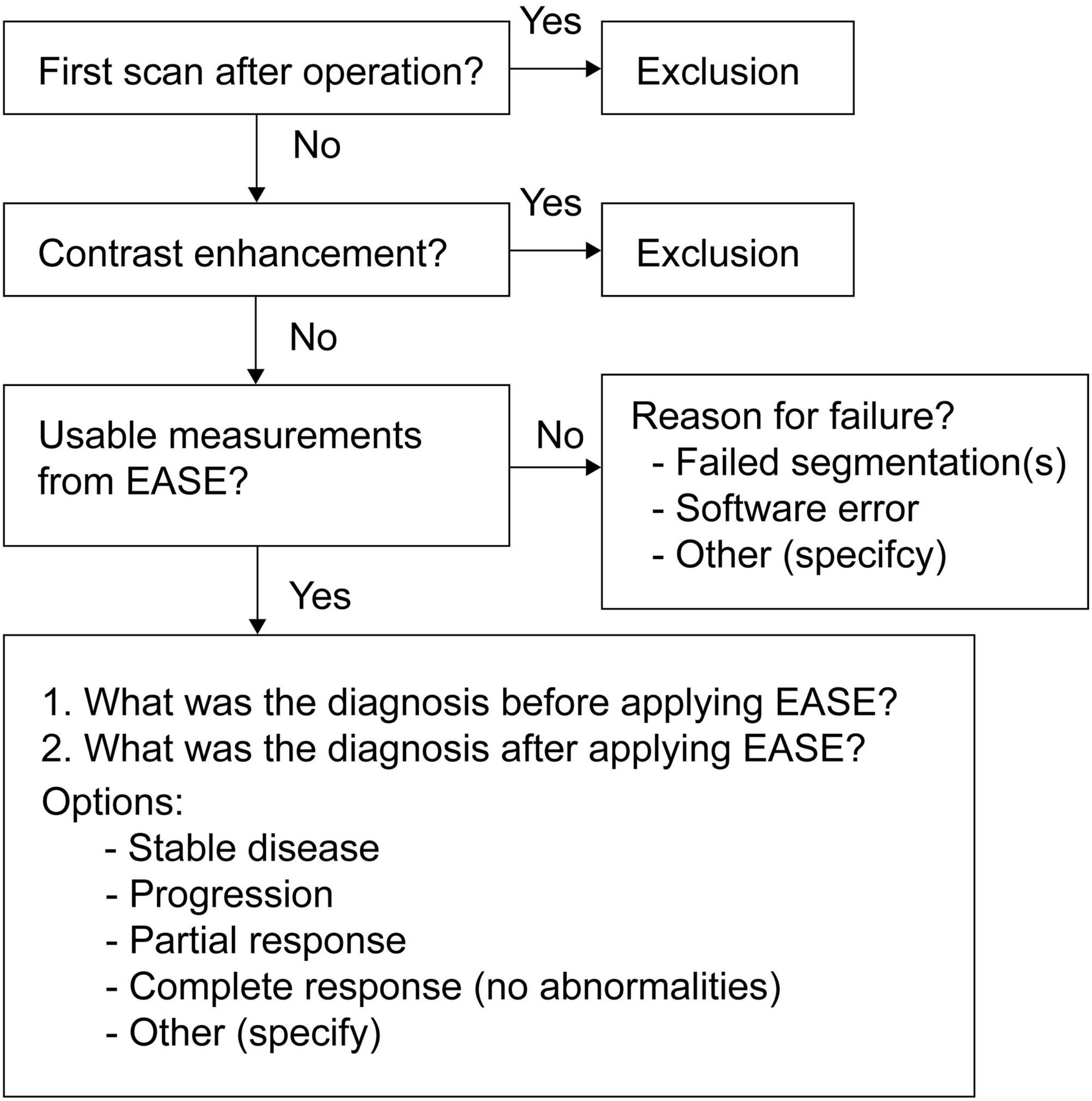


FIGURE 5 | Flowchart of the survey on the usage of EASE in clinical practice.

acceptable for clinical volume measurement. The quality scores are summarized in **Table 2**.

EASE was released for local use in Erasmus MC on 25 May 2021, and the evaluation in clinical practice was performed from 1 June 2021 until 19 August 2021.

During the evaluation period, 55 patients were included in the clinical evaluation, meaning that their visual diagnosis was performed and a volume-based diagnosis was attempted. The patient characteristics are summarized in **Table 3**. A successful diagnosis requires three consecutive scans per patient, and in total 162 scans were segmented by EASE and checked by a radiologist. In one of the patients, the two reference scans were not submitted to EASE after the first segmentation was already rejected and in another scan the segmentation failed due to a software error.

Of the 162 segmentations generated by EASE, 124 (77%) were accepted by the radiologist. The distribution of quality scores can be found in **Table 4**. A successful volume-based diagnosis was

TABLE 2 | Results of reference dataset of 20 MR scans at initial release of EASE.

Acceptance		
	ACCEPTABLE	13 (65%)
	UNACCEPTABLE	7 (35%)
Quality		
	Perfect (5)	0 (0%)
	Good (4)	10 (50%)
	Fair (3)	6 (30%)
	Poor (2)	2 (10%)
	Terrible (1)	2 (10%)

Scans were annotated for acceptability and quality by an experienced neuroradiologist.

TABLE 3 | Characteristics of 55 patients included in the evaluation of EASE in clinical practice.

Patient characteristics: (total)	55
Age (years)	
Median (minimum–maximum)	54 (26–76)
Sex	
Female	24
Male	31
Tumor type	
Oligodendroglioma	19
Astrocytoma	25
Oligo-astrocytoma	2
Presumed low-grade glioma (no tissue diagnosis)	9
Time after surgery (months)	
Median (minimum–maximum)	80 (5–307)
Time after last treatment (months)	
Median (minimum–maximum)	67 (5–307)
Treatment	
Radiotherapy	33
Chemotherapy	33
Surgical resection	41
Time between scans from current scan (months)	
vs. first reference scan, median (minimum–maximum)	14 (7–32)
vs. second reference scan, median (minimum–maximum)	7 (3–20)
Tumor volume found in successful volume-based diagnosis (mL)	
Median (minimum–maximum)	13.2 (1.3–77.1)
Oligodendroglioma, median (minimum–maximum)	18.3 (2.1–77.1)
Astrocytoma, median (minimum–maximum)	16.1 (2.1–60.0)
Oligo-astrocytoma	27.8 (9.5–46.1)
Presumed low-grade glioma, median (minimum–maximum)	2.0 (1.3–14.9)

TABLE 4 | Results of annotations entered in EASE in clinical practice.

Acceptance		
	ACCEPTABLE	124 (77%)
	UNACCEPTABLE	38 (23%)
Quality		
	Perfect (5)	15 (9%)
	Good (4)	87 (54%)
	Fair (3)	33 (20%)
	Poor (2)	17 (10%)
	Terrible (1)	10 (6%)

During the first 3 months of usage, 162 scans were annotated for acceptability, and quality by five different radiologists.

TABLE 5 | Results of evaluation of EASE in clinical practice.

Total number of patients assessed	55
Volume-based diagnosis same as visual diagnosis	36
- Stable disease	32
- Progression	4
Volume-based diagnosis different from visual diagnosis	0
No usable results	19
- Segmentation unacceptable	17
- Inconsistent segmentations	1
- Software error	1

Radiologists were asked to fill in a questionnaire after assessing a patient with EASE, which requires the successful segmentation of three consecutive scans.

reached in 36 out of 55 patients. Results of the questionnaire are summarized in **Table 5**. In all patients where volume-based diagnosis was successful, the volume-based diagnosis made by the radiologist was the same as the conventional visual diagnosis, even though in some cases there was a discrepancy between 2D and 3D measurements as shown in **Figure 7**. **Figure 6** shows an overview of the volume differences detected by EASE, separated by diagnosis (stable disease vs. progression). **Figure 7** shows a comparison to the 2D RANO measurements for those patients in whom both measurements were possible. Three patients are not included in this figure because the lesion was too small to measure according to RANO guidelines. In four patients, EASE measurements indicated a volume increase of more than 40% while the final diagnosis was SD. These differences in volume could be explained by inconsistencies between the segmentations, possibly caused by differences in intensities on T2-FLAIR, and therefore the radiologist maintained the original visual diagnosis of SD. There were no other reported reasons for considering volumetric measurements longitudinally unreliable.

Of the failed cases, 19 could be attributed to the rejection of one of the segmentations and two failed diagnoses were attributed to a different reason. Specifically, in one case a segmentation was missing due to a software error, and in another case all segmentations were accepted by the radiologist but the final volume results were considered unusable due to inconsistencies between the segmentations across the three timepoints. **Figure 8** shows examples of segmentations made by EASE: two consecutive delineations that were considered inconsistent and two consecutive delineations from a successful volume-based diagnosis.

DISCUSSION

A clinical segmentation pipeline 'EASE' was implemented to perform automated 3D volume measurements in LGG. As the effect of such a measurement on clinical decision making is still

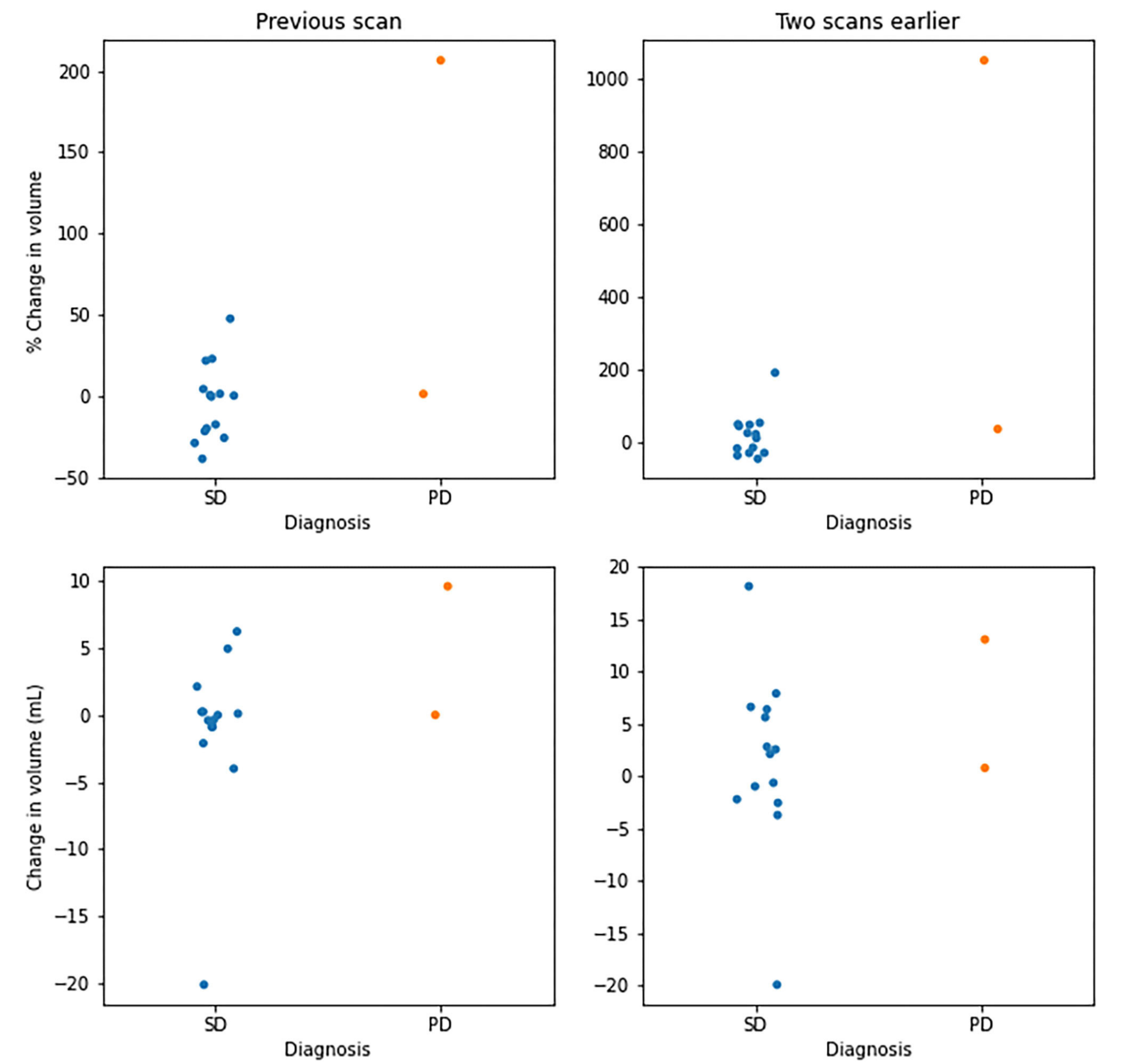


FIGURE 6 | Overview of volume changes in successful volume-based diagnosis. Changes per patient with respect to the previous scan (left) and two scans earlier (right). Values are given in percentage change (top) and change in volume (bottom), separated by diagnoses categorized as stable disease (SD) and progressive disease (PD).

unknown, and perfect performance of the algorithm cannot be expected, several steps were taken to ensure patient safety and monitor results.

The main purpose of this work is to establish the protocols and tools to allow the first introduction of a new, potentially valuable diagnostic tool into clinical practice. From the initial reference dataset, with 7 out of 20 segmentations rejected, it is clear that the quality assessment remains an essential step in the usage of EASE. First results from clinical practice indicate a similar success rate of 74% for individual scans, and approximately half of the patients could be successfully diagnosed with three consecutive volume measurements. However, since the sample size is limited, with almost exclusively diagnoses of stable disease, so further validation of the performance is required to draw firm conclusions on the expected success rate. It must be noted that the segmentation of non-enhancing LGG is particularly difficult due to their diffuse border and varying signal intensity on particularly T2-FLAIR imaging. Furthermore,

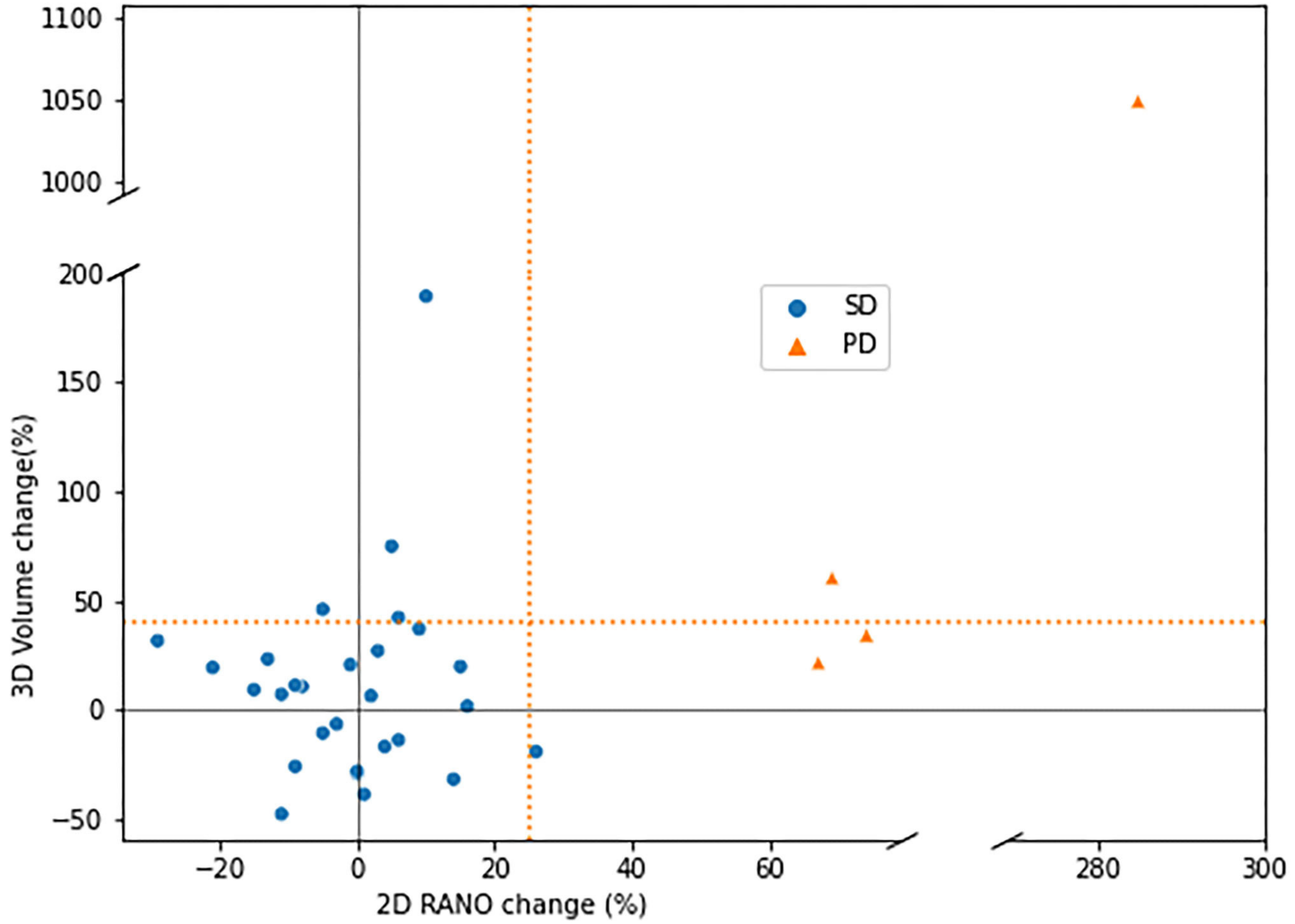


FIGURE 7 | Comparison of measurements using EASE (volume) and 2D RANO (product of two diameters), in percentage change with respect to the first (*t*−2) scan, for patients where both measurements were successful (33 patients). Dotted lines indicate the recommended thresholds for diagnosis of PD.

the underlying deep learning solution, HD-GLIO, was evaluated mostly on high-grade glioma. The current application is therefore aimed at a different, and possibly more challenging patient group and while our results show that a clinical application is feasible, but a more reliable segmentation is needed to facilitate efficient diagnosis.

The results confirm that automatic segmentation of low-grade glioma during follow-up is not a solved problem, and therefore highlight the importance of the quality assurance protocols and manual checks that are presented in this work, and which are ideally part of any introduction of new assessment tools into clinical practice. EASE facilitates a quantitative measurement of lesions that are often impossible to measure accurately even in 2D, due to their irregular shape, and therefore serves a long-standing wish from the neuro-oncological community to move to a potentially more accurate 3D measurement. In this light, a successful diagnosis in over half of the patients is already a valuable step forward.

The initial evaluation in clinical practice provides valuable feedback on the use of automatic segmentation in low-grade glioma. Notably, it shows that an automatic segmentation method is no guarantee for consistent results. Even though the inter-rater variation is removed through automation, the diffuse border of low-grade glioma can still cause ambiguity in the segmentation. Ideally, an automatic segmentation method would be consistent in its choice of where to set the border, but results from EASE show that slight variations in image intensities between consecutive scans can lead to longitudinal inconsistencies. This means that a critical assessment by the radiologist is still needed even if all segmentations are checked and accepted on an individual basis. In EASE, this is ensured by a workflow that can be easily applied in the clinical routine and the protocol for clinical decision-making described in section Diagnosis. Future technical improvements in the automatic segmentation of LGG should focus not only on improving the quality of individual segmentations, but also on longitudinal stability. For this, assessing the reproducibility of the entire process from scan to measurement would be of value, although this would require repeated measurements within a close enough timeframe to assume no change in tumor volume. Such a set-up is not consistent with clinical practice and would require a dedicated study with funding for additional scanning procedures

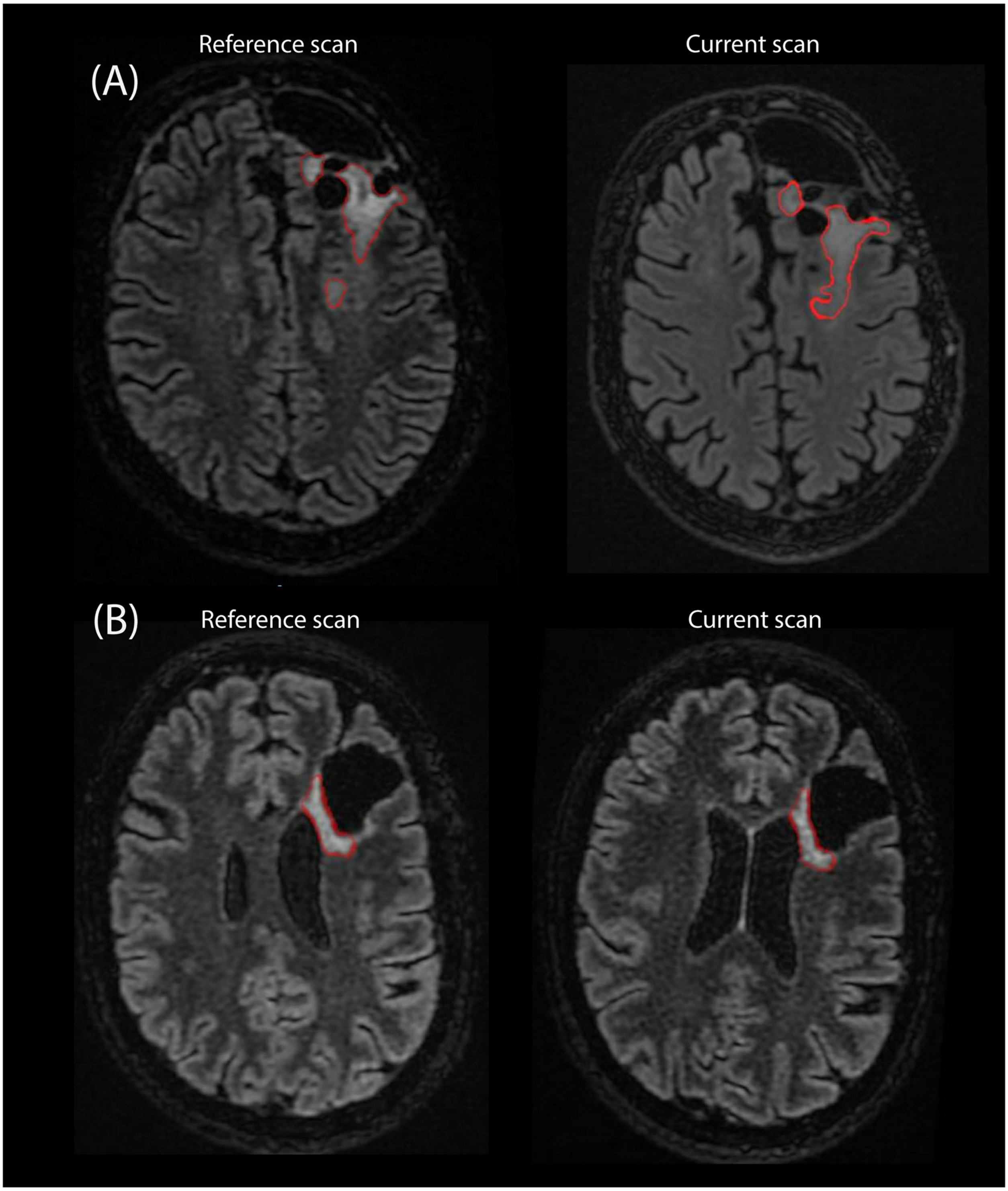


FIGURE 8 | Example of segmentations as they are stored in PACS as an overlay on the T2-FLAIR scan from two consecutive timepoints. **(A)** Two consecutive scans of patient where EASE segmentations were considered inconsistent by the radiologist. **(B)** Two consecutive scans where a volume-based diagnosis of stable disease could be made.

and full consideration of whether the burden this incurs on patients is justified reproducibility of the entire process from scan to measurement.

This work describes a first and careful implementation of automatic segmentation of LGG in clinical practice. Although the results leave room for improvement for the segmentation method, it is already being applied successfully in approximately half of the patients. In all patients diagnosed thus far, the volume measurements confirm the conventional visual diagnosis, as would be expected, but the volume quantification increases confidence in the diagnosis. Essentially, results show that radiologists are cautious in their use of the measurements. The fact that the segmentations are verified and stored for future reference not only decreases the risk of a false diagnosis, but also increases the confidence of the radiologist when using such deep learning solutions in their clinical practice.

Only four patients were included with a diagnosis of progressive disease (PD), which can be attributed to the fact that the most common sign of PD is the presence of contrast enhancement. This is often accompanied by concurrent volume increase, but these cases were excluded from the study in order to address the diagnostic uncertainty regarding non-enhancing lesions. When comparing the volume change between patients with SD and PD, there is no clear threshold to separate the two categories. Although the RANO guidelines recommend a threshold of 25% change for 2D measurements, which would correspond to a 40% change in volume, the final interpretation is left to the discretion of the radiologist and may depend on other factors, such as baseline volume, the presence or absence of treatment-related white matter abnormalities and the consistency of segmentations longitudinally.

When looking at the 2D RANO measurements there is a clear distinction between SD and PD, even though these measurements do not capture the full extent of the irregular shape and diffuse infiltration of these lesions. From these results it seems that the existing visual diagnosis is still being used as the primary tool to determine tumor growth, but are too few patients showing progression in either method to draw a firm conclusion. Also, it must be noted that these results were gathered in the first months after EASE was released for clinical use.

EASE was put into service prior to the date of application of EU regulation 2017/745 on medical devices (MDR). We are aware that in case of substantial changes in the design or intended purpose of EASE, the requirements of this regulation are applicable. Our approach to ensure quality of results and prevent incorrect interpretation is already in line with the general aim of the MDR.

We think this implementation provides a potential benefit to both the clinicians and researchers, as radiologist receive a valuable tool for the quantification of glioma volume, even if not fully perfected, while researchers receive valuable feedback from clinical practice. In its current form, EASE does not allow for correction of failed segmentations through manual intervention of the radiologist, as this is not feasible in clinical practice. However, the feedback from clinical practice could enable further improvement in the segmentation, whether that is in the preprocessing or by improving the HD-GLIO model in a transfer learning approach, while the clearly defined protocol for software updates ensures patient safety during such future improvements.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by MEC-2021-0530 (Erasmus MC Rotterdam). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

The manuscript was written by KG and reviewed by SvdV, MSm, MSt, and SK. Software development was done by KvG, AV, AG, and SvdV. The protocol for use of the software in the clinic was designed by MSm, MSt, and MR. The evaluation in clinical practice was conceptualized by MSm and KvG and implemented by KvG. All authors contributed to the article and approved the submitted version.

FUNDING

KvG was funded by the Dutch Cancer society (project number 11026, GLASS-NL) and the Dutch Medical Delta.

ACKNOWLEDGMENTS

We would like to thank radiologists M. Vernooij, H. Ahmad, R. Gahrmann and A. van der Eerden for contributing to the evaluation in clinical practice.

REFERENCES

1. Mandonnet E, Delattre J-Y, Tanguy M-L, Swanson KR, Carpentier AF, Duffau H, et al. Continuous growth of mean tumor diameter in a subset of grade II gliomas. *Ann Neurol.* (2003) 53:524–8. doi: 10.1002/ana.10528
2. Rees J, Watt H, Jäger HR, Benton C, Tozer D, Tofts P, et al. Volumes and growth rates of untreated adult low-grade gliomas indicate risk of early malignant transformation. *Eur J Radiol.* (2009) 72:54–64. doi: 10.1016/j.ejrad.2008.06.013
3. Brasil Caseiras G, Ciccarelli O, Altmann DR, Benton CE, Tozer DJ, Tofts PS, et al. Low-grade gliomas: six-month tumor growth predicts patient outcome better than admission tumor volume, relative cerebral blood volume, and apparent diffusion coefficient. *Radiology.* (2009) 253:505–12. doi: 10.1148/radiol.2532081623

4. Duffau H, Taillandier L. New concepts in the management of diffuse low-grade glioma: Proposal of a multistage and individualized therapeutic approach. *Neuro Oncol.* (2015) 17:332–42. doi: 10.1093/neuonc/nou153
5. Jakola AS, Moen KG, Solheim O, Kvistad KA. "No growth" on serial MRI scans of a low grade glioma? *Acta Neurochir (Wien).* (2013) 155:2243–4. doi: 10.1007/s00701-013-1914-7
6. Van den Bent MJ, Wefel JS, Schiff D, Taphoorn MJB, Jaeckle K, Junck L, et al. Response assessment in neuro-oncology (a report of the RANO group): Assessment of outcome in trials of diffuse low-grade gliomas. *Lancet Oncol.* (2011) 12:583–93. doi: 10.1016/S1470-2045(11)70057-2
7. Kofler F, Berger C, Waldmannstetter D, Lipkova J, Ezhov I, Tetteh G, et al. BraTS Toolkit: translating BraTS brain tumor segmentation algorithms into clinical and scientific practice. *Front Neurosci.* (2020) 14:125. doi: 10.3389/fnins.2020.00125
8. Kickingereder P, Isensee F, Tursunova I, Petersen J, Neuberger U, Bonekamp D, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* (2019) 20:728–40. doi: 10.1016/S1470-2045(19)30098-1
9. Marcus DS, Olsen TR, Ramaratnam M, Buckner RL. The extensible neuroimaging archive toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics.* (2007) 5:11–33. doi: 10.1385/NI:5:1:11
10. Achterberg HC, Koek M, Niessen WJ. Fastr: A workflow engine for advanced data flows in medical image analysis. *Front ICT*. (2016) 3:24. doi: 10.3389/fict.2016.00015
11. Merkel D. *Docker: Lightweight Linux Containers for Consistent Development and Deployment.* (2014). Available online at: http://www.docker.io (accessed July 5, 2021).
12. Li X, Morgan PS, Ashburner J, Smith J, Rorden C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J Neurosci Methods.* (2016) 264:47–56. doi: 10.1016/j.jneumeth.2016.03.001
13. Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. Elastix: a toolbox for intensity-based medical image registration. *IEEE Trans Med Imaging.* (2010) 29:196–205. doi: 10.1109/TMI.2009.2035616
14. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Hum Brain Mapp.* (2019) 40:4952–64. doi: 10.1002/hbm.24750
15. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Trans Med Imaging.* (1998) 17:87–97. doi: 10.1109/42.668698
16. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Neuroscience

ORIGINAL RESEARCH
published: 06 October 2021
doi: 10.3389/fnins.2021.740353

Stratification by Tumor Grade Groups in a Holistic Evaluation of Machine Learning for Brain Tumor Segmentation

Snehal Prabhudesai[1*†], *Nicholas Chandler Wang*[2*†], *Vinayak Ahluwalia*[3], *Xun Huan*[4], *Jayapalli Rajiv Bapuraj*[5], *Nikola Banovic*[1] *and Arvind Rao*[2,6,7,8*]

[1] Computer Science and Engineering, University of Michigan, Ann Arbor, MI, United States, [2] Computational Medicine and Bioinformatics, Michigan Medicine, Ann Arbor, MI, United States, [3] Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, United States, [4] Mechanical Engineering, University of Michigan, Ann Arbor, MI, United States, [5] Department of Radiology, University of Michigan, Ann Arbor, MI, United States, [6] Department of Biostatistics, University of Michigan, Ann Arbor, MI, United States, [7] Department of Radiation Oncology, University of Michigan, Ann Arbor, MI, United States, [8] Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, United States

OPEN ACCESS

Edited by:
Spyridon Bakas,
University of Pennsylvania,
United States

Reviewed by:
David Haynor,
University of Washington,
United States
Francesco Grussu,
Vall d'Hebron Institute of Oncology
(VHIO), Spain
Zeina A. Shboul,
Old Dominion University, United States

***Correspondence:**
Snehal Prabhudesai
snehalbp@umich.edu
Nicholas Chandler Wang
ncwang@med.umich.edu
Arvind Rao
ukarvind@umich.edu

†*These authors share first authorship*

Specialty section:
This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

Received: *12 July 2021*
Accepted: *01 September 2021*
Published: *06 October 2021*

Citation:
Prabhudesai S, Wang NC, Ahluwalia V, Huan X, Bapuraj JR, Banovic N and Rao A (2021) Stratification by Tumor Grade Groups in a Holistic Evaluation of Machine Learning for Brain Tumor Segmentation.
Front. Neurosci. 15:740353.
doi: 10.3389/fnins.2021.740353

Accurate and consistent segmentation plays an important role in the diagnosis, treatment planning, and monitoring of both High Grade Glioma (HGG), including Glioblastoma Multiforme (GBM), and Low Grade Glioma (LGG). Accuracy of segmentation can be affected by the imaging presentation of glioma, which greatly varies between the two tumor grade groups. In recent years, researchers have used Machine Learning (ML) to segment tumor rapidly and consistently, as compared to manual segmentation. However, existing ML validation relies heavily on computing summary statistics and rarely tests the generalizability of an algorithm on clinically heterogeneous data. In this work, our goal is to investigate how to holistically evaluate the performance of ML algorithms on a brain tumor segmentation task. We address the need for rigorous evaluation of ML algorithms and present four axes of model evaluation—diagnostic performance, model confidence, robustness, and data quality. We perform a comprehensive evaluation of a glioma segmentation ML algorithm by stratifying data by specific tumor grade groups (GBM and LGG) and evaluate these algorithms on each of the four axes. The main takeaways of our work are—(1) ML algorithms need to be evaluated on out-of-distribution data to assess generalizability, reflective of tumor heterogeneity. (2) Segmentation metrics alone are limited to evaluate the errors made by ML algorithms and their describe their consequences. (3) Adoption of tools in other domains such as robustness (adversarial attacks) and model uncertainty (prediction intervals) lead to a more comprehensive performance evaluation. Such a holistic evaluation framework could shed light on an algorithm's clinical utility and help it evolve into a more clinically valuable tool.

Keywords: medical AI, evaluation, brain imaging, segmentation, GBM, LGG

1. INTRODUCTION

Accurate and consistent segmentation of gliomas (Chen et al., 2017), is important for diagnosis, treatment planning, and post treatment evaluation. Glioblastoma Multiforme (GBM), the most aggressive of high grade gliomas, has the worst prognosis with a 5-year survival rate of <5% and a median survival of approximately a year even with treatment (Tamimi and Juweid, 2017; Witthayanuwat et al., 2018). Low grade gliomas (LGG), though less aggressive than GBM, reportedly undergo anaplastic progression into higher grade tumors around 70% of the time within 5–10 years of diagnosis. The median survival from initial diagnosis is ∼7 years (Claus et al., 2015).

Current standard of care for High Grade Glioma (HGG), for example GBM, is surgical resection of the tumor followed by radiotherapy combined with the chemotherapeutic agent temozolomide (Tan et al., 2020). Segmentation for the surgical resection for gliomas should be effective for total gross resection or reduction in tumor bulk, without affecting the surrounding normal functional brain tissue. Radiation therapy requires accurate delineation of tumor margins to ensure effective dosage to tumor region. Due to the relative low aggressiveness of LGG, a more conservative management ("wait-and-watch") approach (Whittle, 2004) is sometimes adopted. Segmentation is important in this scenario also to monitor temporal morphological and volumetric alterations of the tumors during observation, prior to elective tumor resection (Larsen et al., 2017).

However, the imaging presentation of gliomas varies between LGG and HGG, which could affect the accuracy of their segmentation. Most HGGs, such as GBMs, have a heterogeneous appearance on T1-weighted pre-contrast imaging and typically show a heterogeneous thick-walled rim-enhancing appearance on the T1 post-contrast (T1-Gd) sequence, with a surrounding low attenuation of perifocal edema. The overall appearance of HGGs on T2-weighted fluid-attenuated inversion recovery (FLAIR) sequence is heterogeneously hyperintense, with areas corresponding to enhancing and non-enhancing components as seen on T1-weighted post contrast sequence. The advancing non contrast-enhancing FLAIR hyperintense portions of the tumor are of concern to clinicians because it is believed to contain active tumor remote from the apparent enhanced portions of the aggressive core. On the other hand, low grade tumors appear hyperintense on a FLAIR sequence with or without clear margins. On the pre-contrast T1-weighted sequences, the lesions tend to be hypointense and typically do not enhance following administration of gadolinium based agents (Forst et al., 2014; Bulakbaşı and Paksoy, 2019).

Manually defining the margins of the tumor and surrounding non-enhancing perifocal region remains challenging due to tumor heterogeneity, ill-defined margins, and the varying degrees of perifocal edema. This makes segmentation an arduous task with questionable consistency. In recent years, Machine Learning (ML) techniques have shown potential to assist in tumor segmentation for correct diagnosis and efficient treatment planning (Wadhwa et al., 2019; Bajaj and Chouhan, 2020; Kocher et al., 2020; Nazar et al., 2020). While both HGG, including GBM, and LGG, benefit from accurate segmentation, existing ML validation rarely tests if an algorithm generalizes well to out-of-distribution data that reflects this tumor heterogeneity. Rebsamen et al. (2019) have shown that implicitly incorporating high-vs.-low tumor grade information in model training could improve model performance. While recent work has evaluated for tumor heterogeneity across geographic populations (McKinney et al., 2020), hospital systems (Zech et al., 2018), and federated learning settings (Sheller et al., 2020), this has yet to be done considering differences between HGG, for example GBM and LGG imaging presentations.

In this work, we address the need for rigorous evaluation of ML algorithms for brain tumor segmentation. We propose a holistic evaluation framework (**Figure 1**) that takes into account tumor heterogeneity, robustness, and confidence of the ML algorithm, and batch effects that may arise from the data. We demonstrate this framework with a cross-sectional study design similar to Zech et al. (2018) and analyze how well an ML algorithm trained on one glioma type (either HGG, exemplified by GBM or LGG) generalizes to another, out-of-distribution glioma type. We conduct four experiments and holistically evaluate an ML algorithm for the problem of tumor segmentation:

Diagnostic Performance: We compute standard segmentation metrics to objectively compare the ML algorithm's segmentation performance against radiologist-annotated ground truth. Results indicate that metrics such as Dice and AUROC do not sufficiently capture differences in generalizability, although the classification matrix reveals clear differences.

Model Confidence: We measure model confidence in segmentation performance by computing prediction intervals for the brain as well as tumor region. Results indicate that ML algorithms trained on LGG data is more confident than the rest on all homogeneous as well as mixed data.

Robustness: We measure the ML algorithm's ability to maintain performance despite adversarial perturbations to test their reliability comparably. Results indicate that the ML algorithm trained only on GBM data was least robust when segmenting tumor corrupted with high levels of noise. Testing performance of the model across out of distribution data, was performed in all the experiments, but can be considered an extension of robustness testing.

Data Quality (Batch Effects): We measure the degree to which MRI scan quality influences segmentation metrics. Results found that scan quality features are not significantly correlated with performance, but that there were some batch effect differences, primarily between LGG and GBM sites.

Our results demonstrate the limitations of segmentation metrics, and caution that metrics alone do not capture all aspects of an ML algorithm's performance. We discuss how our findings relate to recent literature in segmentation metrics. We further discuss how such a holistic evaluation framework could shed light on the algorithm's clinical utility in post-deployment scenarios and help it evolve into a more clinically valuable tool (Recht et al., 2020).

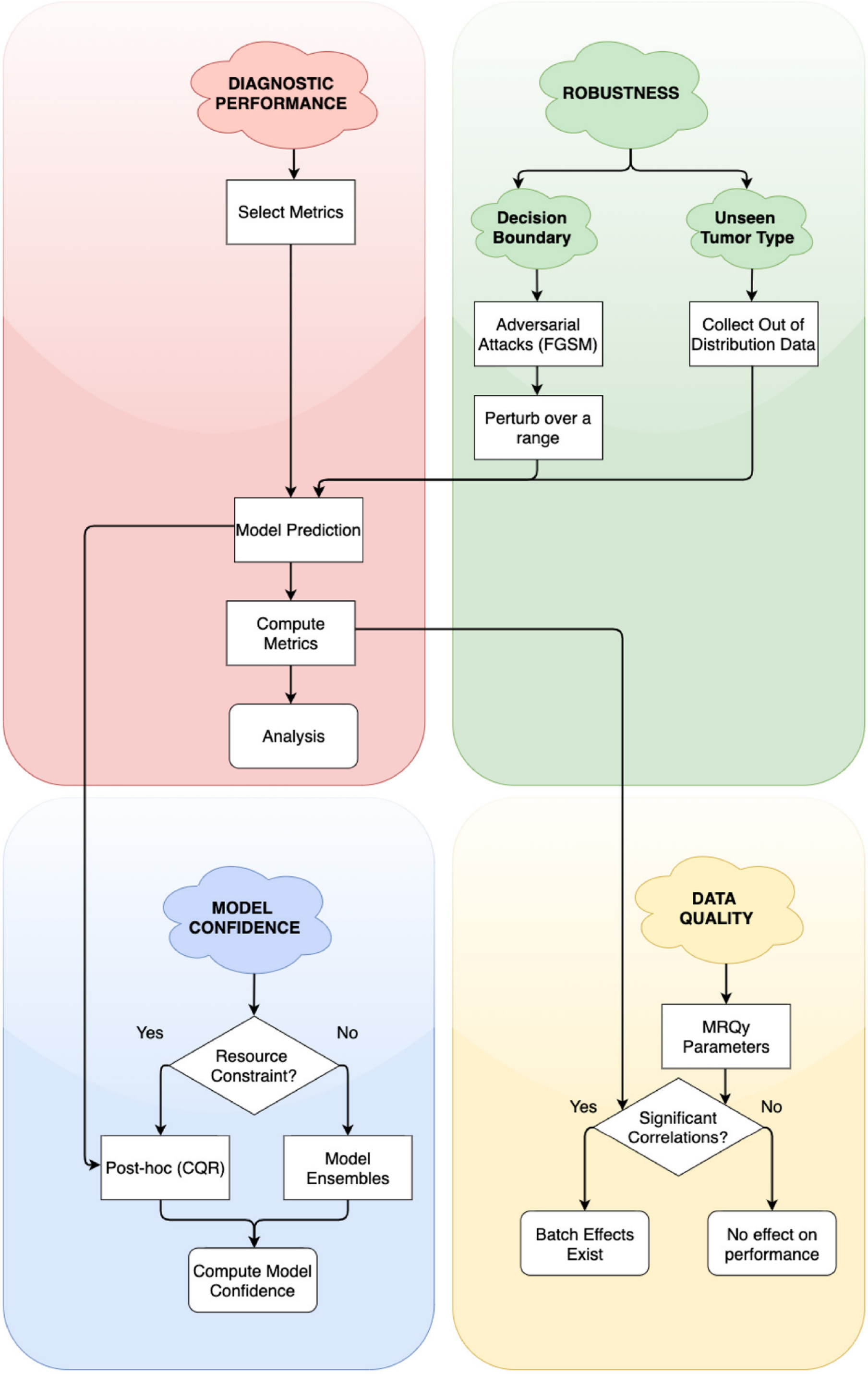


FIGURE 1 | Simplified flowchart of different axes of holistic evaluation—diagnostic performance, robustness, model confidence, and data quality. Axes are ordered by dependency and relation with each other. We recommend models to be evaluated with atleast one experiment on each of these axes. We evaluate two aspects of robustness, namely, closeness to decision boundary and generalizability on unseen glioma type. Decision points in the framework lead to alternate paths for researchers to follow.

TABLE 1 | Split of patients in each of the three datasets.

Dataset	GBM patients	LGG patients	ALL patients
Train	102 (14,688)	65 (9,360)	167 (24,048)
Validation	16 (2,304)	21 (3,024)	37 (5,328)
Test	17 (2,448)	22 (3,168)	39 (5,616)

Values in brackets (.) indicate the total number of images available in the dataset for 2D segmentation. Note that henceforth, we refer to the test dataset as D_{GBM} (GBM patients only), D_{LGG} (LGG patients only), and D_{ALL} (All patients—GBM and LGG patients).

2. MATERIALS AND METHODS

The aim of this work is to propose a framework to evaluate model performance across four axes—diagnostic performance, model confidence, robustness, and data quality. To demonstrate this framework, we first train ML algorithms by considering tumor heterogeneity. We use publicly accessible code for algorithm development and perform *post-hoc* calibration.

2.1. Dataset

We used publicly available Magnetic Resonance Imaging (MRI) from The Cancer Genome Atlas (TCGA) (Clark et al., 2013). Glioblastoma Multiforme (GBM) and Low Grade Glioma (LGG) collection (Bakas et al., 2017a,b). This included the skull-stripped and co-registered MICCAI-BraTS 2018 Test Dataset (Menze et al., 2015; Bakas et al., 2017c). The data consisted of pre-operative multimodal MR imaging sequences (i.e., T1, T1-Gd, T2, T2-FLAIR) along with their whole-tumor segmentation labels composed of edema, enhancing tumor, and non-enhancing tumor. We combined these labels into a single whole tumor for this study. Number of patients in GBM BraTS Test Dataset and LGG BraTS Test Dataset were split approximately in half and allotted to validation and test datasets. The GBM and LGG data were merged across the three categories to form an ALL dataset. Each patient was associated with 144 pre-operative MRI scans, which were treated as independent data points for 2D segmentation. These MRI scans were cropped to 144 × 144 pixels and further pre-processed the data by pixel-intensity normalization. **Table 1** describes the total number of patients and total number of MRI scans available in each dataset. The training datasets were used for model development (section 2.2), validation datasets were used to determine hyperparameters and calibrate the models (section 2.3), and test datasets (D_{GBM}, D_{LGG}, D_{ALL}) were used to perform subsequent experiments (section 3).

2.2. Network Architecture and Training

We used the state-of-the-art U-Net architecture (Ronneberger et al., 2015) to develop three tumor segmentation models using the GBM, LGG, and ALL train datasets. The U-Net architecture consists of an encoder, decoder, and skip connections. Each module of the encoder consists of 2D Convolution layers, followed by Batch Normalization and MaxPooling layers. Four such modules make up the encoder. The decoder consists of four modules of Conv2DTranspose layers followed by Concatenate layers. The network performs slice-wise (2D) segmentation with multi-modal MRI scans provided as the input. Models were trained with Dice Loss function for 100 epochs on 8 GPUs. Adam optimizer (Kingma, 2015) was used with a learning rate of 1×10^{-4} and a batch size of 128. Data augmentation was used while training each of the models to improve generalization. This consisted of random rotations (0–25° degrees range), random zooming (value = 0.2, zooms image by 80–120% range), width shift (value = 0.2, horizontal translation of images by 0.2 percent), height shift (value = 0.2, vertical translation of images by 0.2 percent), shear (value = 0.2, clips the image in counter-clockwise direction) and random horizontal flips. We referred to publicly available code for model development, model training, and data augmentation (Dong et al., 2017; Ojika et al., 2020).

TABLE 2 | We first compute calibration metrics on a patient-level, then aggregated by mean.

Metrics	M_{GBM}		M_{LGG}		M_{ALL}	
	Before	After	Before	After	Before	After
NLL	0.038212	0.013506	0.070146	**0.022842**	0.056573	0.018483
BS	0.003519	**0.002970**	0.006020	0.005263	0.004533	0.003862
ECE%	0.3413	0.1439	0.5877	0.3141	0.4454	**0.1876**
MCE%	36.4552	14.0762	31.9731	14.3702	37.0614	**13.8812**

We consider only pixels in the skull-stripped brain to compute these metrics. ECE and MCE are presented in %. Metrics should ideally reduce upon calibration. Columns under each model indicate metric values before and after calibration. Bold values indicate best % decrease or increase as compared to the "before" column. All models improved after calibration.

2.3. Model Calibration

The goal of model calibration is to align the algorithm's predicted probabilities align with the observed (ground truth) outcomes (Guo et al., 2017). Calibration process ensure that algorithms do not overstate or understate their confidence in prediction of tumor (Jungo and Reyes, 2019; Mehrtash et al., 2020). Models that have been already trained can be calibrated with *post-hoc* methods (Rousseau et al., 2021). Guo et al. (2017) recommend performing post-hoc calibration with the same validation dataset (**Table 1**) used for model development. We use Platt Scaling technique (Platt, 1999) for post-hoc calibration due to its simplicity and ease of implementation. To ensure models are properly calibrated, we compute and report common calibration metrics. Negative Log Likelihood (NLL) measures a probabilistic model's quality and is also known as cross-entropy loss. Brier Score (BS) measures the accuracy of probabilistic predictors. Percentage Expected Calibration Error (ECE%) partitions the model's predictions into equally spaced bins and takes a weighted average of the difference between accuracy and model confidence across bins. Percentage maximum calibration error (MCE%) estimates the worst-case deviation between confidence and accuracy. For metric definitions and more information, we refer readers to Mehrtash et al. (2020) and Guo et al. (2017). **Table 2** indicates that all models are properly calibrated.

3. EXPERIMENTS

Here, we perform an experiment on each of the four axes of our evaluation framework. We compute metrics to summarize diagnostic performance, measure model confidence by computing prediction intervals, simulate adversarial attacks to assess robustness and use MRQy package to analyze batch effects in data. For each experiment, we point to related work, and provide details on the experiment procedure. Then, in section 4, we provide the outcome of these experiments. We evaluate each of the calibrated ML algorithms (M_{GBM}, M_{LGG}, and M_{ALL}) on each of the three test datasets (D_{GBM}, D_{LGG}, and D_{ALL}). Thus, we evaluate 3 (models) $\times$ 3 (datasets) $= 9$ conditions.

3.1. Metrics for Segmentation Performance

There exist a plethora of metrics to evaluate the performance of a medical image segmentation algorithm (Udupa et al., 2006; Taha and Hanbury, 2015). Each metric focuses on a specific aspect of the algorithm's performance, and is thus limited in capability to describe the algorithm's performance by itself. Several metrics are necessary to describe comprehensive characteristics of segmentation performance (Renard et al., 2020).

We perform this experiment as a baseline, reflective of the current standard practice for evaluation. We follow the guidelines described by Taha and Hanbury (2015) and select eight metrics to evaluate segmentation performance. Sensitivity (Sens) measures the proportion of tumor pixels that are correctly identified as tumor (foreground). Specificity (Spec) measures the proportion of benign pixels that are correctly identified as benign (background). Positive Predictive Value (PPV) measures the probability that pixels classified as benign truly belong to parts of the patients' brain without a tumor. Negative Predictive Value (NPV) measures the probability that pixels classified as tumor truly belong to parts of the patients' brain with a tumor. While accuracy can be skewed due to the paucity of tumor pixels in the tumor class, Balanced Accuracy (BAcc) takes into account class imbalance. Dice Coefficient (Dice) and Jaccard Coefficient (Jac.C) both measure the overlap between tumor annotated by the different sources (ML algorithm and the radiologists' manual annotations). Area under Receiver Operating Characteristics curve (AUROC) describes the probability that a randomly selected tumor pixel will have a higher predicted probability of being a tumor than a randomly selected benign pixel. We eliminate any extra-cranial regions and only consider the skull-stripped brain for computing the metrics. We compute metrics on a per-patient level, as it offers more granularity than at a population-level.

3.2. Prediction Intervals for Model Confidence

Prediction Intervals (PIs) are often reported and considered for medical decision-making (Kümmel et al., 2018). In radiation oncology, Chan et al. (2008) used prediction intervals to capture uncertainty in tumor and organ movement. While a confidence interval measures the precision of a predicted value, PIs measure the expected range where a future observation would fall, given what has already been observed. The width of the PI is directly proportional to the model uncertainty at that region (Kabir et al., 2018). We use prediction intervals to quantify uncertainty in tumor segmentation.

We use Conformal Quantile Regression (CQR) (Romano et al., 2019) to compute PIs. Construction of PIs is difficult, as PIs can be too small that they don't capture the true magnitude (Type 1 error) or too large that they are uninformative (Type 2 error) (Elder et al., 2021). The CQR method guarantees construction of PI such that the target value is contained within the PI by error probability α (valid coverage) and that the PIs are informative.

We used the CQR method to compute PIs in a *post-hoc* manner. The method uses a dataset for training the CQR models and a separate test dataset to compute the PIs. To reduce computational cost, we selected summary images (image with the largest tumor) for each patient in the validation and test datasets (**Table 1**). We designed a setup to generate prediction intervals around the calibrated model values. We first obtained logits (model output before the calibration) for the selected summary images for patients in both datasets. The CQR models were trained on validation dataset logits and the corresponding calibrated model predictions as target values. The trained CQR models were then used to compute prediction intervals for test dataset logits. We followed the method described by Romano et al. (2019) to compute average prediction intervals (API) per-patient in the test set. We then generated API box plots for all nine conditions.

3.3. Adversarial Attacks for Robustness

This experiment was designed to test the impact of data quality and potential batch effects on the predictions of the model. There has been a lot of work in other domains on evaluating the adversarial robustness of ML algorithms. The application of imperceptible noise can change the prediction of image classification system from correctly identifying a panda to confidently miscalling the image a gibbon (Goodfellow et al., 2015). There are now a variety of adversarial attack techniques, from white-box techniques that can look inside the algorithm to those that can build attacks simply by testing inputs and outputs. These techniques can provide a useful framework for evaluating the robustness of a medical imaging machine learning system. In tumor imaging in general, Zwanenburg et al. (2019) showed how radiomics features can be evaluated for robustness by perturbing the tumor mask. Understanding how vulnerable ML algorithms are to noise, and how easily they change their decisions in response, gives a sense of how these ML algorithms might fail.

The adversarial attack used in this experiment was fast gradient signed method (FGSM), described by Goodfellow et al. (2015). This technique is a white-box method which takes the calculated gradient of the neural network to find the direction of the smallest change that will affect the label of the output. This gradient adversarial noise is multiplied by a factor of epsilon, to vary the strength of the attack. In these experiments the epsilon factor was varied over a range of 0–1 (0, 0.005, 0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0), with more examples on the lower end of the range to evaluate small perturbations.

We performed the FGSM attack on each of the test datasets (D_{GBM}, D_{LGG}, and D_{ALL}), for all three ML algorithms (M_{GBM}, M_{LGG}, and M_{ALL}). The full panel of metrics was computed for each of these experiments. The performance of the ML algorithms was expected to decay as epsilon decreased, but the relative robustness of each of the ML algorithms and the way that they decayed was studied as well. The chosen epsilon values were (0, 0.005, 0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, and 1). An epsilon of 0 indicates that no change was made to the image.

3.4. MRQy for Analyzing Batch Effects

Magnetic resonance imaging has many strengths in studying and monitoring cancer status, including a variety of sequences to investigate different aspects of tumors. However, the flexibility it provides to radiologists can lead to inconsistencies in protocol and scan quality. MRQy is MRI quality package that provides a variety of features that assess the quality of a scan, and other effects that might be considered batch effects (Sadri et al., 2020). The complexity of machine learning algorithms makes it possible for them to pick up on batch effects between sites rather than the underlying biology of a problem.

These MRQy factors were used to audit the susceptibility of the different ML algorithms to scan quality factors. For each of the MRI sequences, MRQy features were calculated independently on the original NIFTI files. The features used per modality were: MEAN, RNG, VAR, CV, PSNR, SNR1, SNR2, SNR3, SNR4, CNR, CVP, CJV, EFC, TSNEX, TSNEY, UMAPX, UMAPY (For metric definitions, Sadri et al., 2020). The metadata and size features were excluded as they were not available, and the sizing was consistent across all the images. The average true positive probability of a tumor pixel having a tumor label was calculated, as well as for true negative, false positive and false negative pixels. These were calculated on a per patient level and then averaged across all the patients in the test set. These values along with Dice score and AUROC were then assessed for their correlation with the MRQy features using Spearman correlation coefficient. MRQy features that are correlated with model performance are potential quality control metrics that might be used to flag problematic cases. False discovery rate (FDR) correction was then performed using Benjamini-Hochberg correction at an alpha of 0.25 (Benjamini and Hochberg, 1995). We used this correction as it is less stringent than a more aggressive Bonferroni correction and was still found to eliminate the uncorrected p-values.

Additionally, independent of the metrics, batch effects were investigated using the MRQy parameters to compare TCGA site codes in the combined testing data set (D_{ALL}). The MRQy features were normalized then decomposed using principal component analysis (Tipping and Bishop, 1999). The first two MRQy principal components and their relationship to institution were investigated using ANOVA and paired T-tests in the statsmodels python package (Seabold and Perktold, 2010). We hypothesized that some site differences within the data sets might be captured by this dimensionality reduction.

4. RESULTS

In this section, we present and analyze the results of the four experiments in section 3. We discuss their implications in section 6. Note that we perform these experiments for the pixels within the skull-stripped brain.

4.1. Metrics Alone Do Not Sufficiently Describe the Nature and Severity of Segmentation Mistakes

True Negative (TN) panel in **Figure 2** indicates all models perform equally well in identifying benign pixels. M_{ALL} has the highest percentage TP, indicating the best performance at correctly identifying tumor pixels. On average, due to a higher percentage of False Negatives than False Positives, all algorithms (M_{LGG},M_{GBM},M_{ALL}) under-segment tumor more often than they over-segment. The FP value is highest for M_{LGG}. Thus, out of all models, M_{LGG} classifies benign regions as tumor the most (over-segments). The FN value is highest for M_{GBM}, on average. M_{GBM} thus, under-estimates tumor pixels and classifies them as benign (under-segments).

The training of the algorithms further explains these findings. M_{LGG} learns to pick up subtle patterns in the training phase, and when evaluated on D_{GBM}, classifies normal-appearing tissue as part of a tumor. In contrast, M_{GBM} is used to seeing dominant contrast patterns, which explains why it misses a lot of tumor pixels in LGG.

In **Figure 3**, all models have similarly worse performance on some patients, indicated by red rows. This is visible across all test datasets. This could be due to multiple confounding variables such as different vendors, field strengths, parameters of imaging, strength of the imaging magnet, type of machine, and it is difficult to pinpoint the contributing factor. Metrics show similar trends in all conditions. Models have a high specificity, low sensitivity, and a high AUROC. There is an overall trend of NPV being higher than PPV. These findings reflect the effect of class imbalance in the dataset, and the models' ability to recognize benign areas much more easily than tumor regions.

4.2. Example Illustrations

Here, we present example patients (**Figures 4–7**) with the Ground Truth (GT) tumor and tumor segmentation contours of M_{GBM}, M_{LGG}, and M_{ALL}. We selected good and bad segmentation examples from D_{GBM} and D_{LGG} each for qualitative analysis. One of the authors, who is a board-certified neuroradiologist of more than a decade of experience in brain tumor diagnosis, interpreted these images.

4.3. M_{LGG} Has the Greatest Confidence for Segmentation Across All Datasets

Violin plots were constructed to analyze average model confidence across all patients. **Figure 8** depicts the average prediction intervals for the skull-stripped brain region. Models have approximately the same median average prediction intervals (API) on each test dataset. **Figure 9** represents model confidence

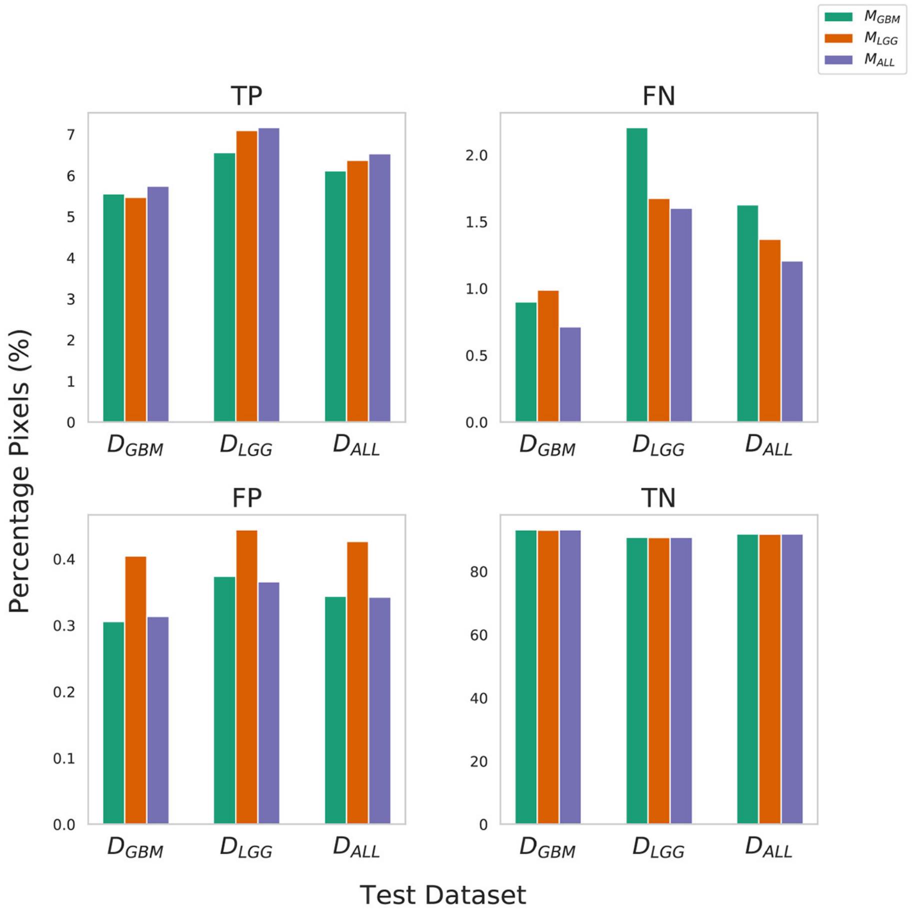


FIGURE 2 | Confusion Matrix to assess the performance of M_{GBM}, M_{LGG}, and M_{ALL} across stratified and composite datasets. The y-axis denotes percentage of total pixels in a test dataset classified as TP, FN, FP, TN. M_{LGG} has the tendency to over-segment (high %FP), while M_{GBM} has the tendency to under-segment tumor(high %FN), relative to each other. Note that metrics such as Dice coefficient routinely ignore the background (TN) in a segmentation context, so a 0.1% difference in false positives should be understood relative to the 6–9% of the volume that is tumor.

while identifying tumor regions. Models have wider interquartile range and greater variability compared to **Figure 8**. This indicates models have low confidence in identifying tumors as compared to non-tumor. M_{GBM} and M_{ALL} have similar distributions of API across patients, indicating both models are similarly confident while segmenting both GBM and LGG tumor. M_{LGG} has the lowest median prediction interval widths, and their distribution has the lowest variability and highest concordance. This indicates M_{LGG} is the most confident model while segmenting both LGG and GBM patients. Out of all models, M_{LGG} is consistently confident while making predictions.

M_{LGG} has the highest confidence, even though it makes mistakes (over-segments) in segmentation, suggestive of an aggressive approach. M_{GBM} also makes mistakes (under-segments) but has lower confidence, which suggests a cautious approach. LGG may be monitored for a longer period of time, so a high rate of false positives can overburden clinicians, going against the goal of reducing their burden. If mistakes are very obvious, it can cause a high degree of frustration and eventual abandonment of the algorithm (Beede et al., 2020). Previous works have proposed monitoring cases with low confidence (Kompa et al., 2021). However, in a case where a model makes mistakes with high confidence, a confidence-based screening approach might cause the reviewer to miss important areas of model failure.

4.4. Models Trained on D_{GBM} Deteriorated the Most Under Adversarial Attacks

The three models (M_{GBM}, M_{LGG}, M_{ALL}) were each evaluated on the three test datasets under FGSM attack across a range of epsilons from 0 to 1. The 95% confidence intervals are also included for each of the metrics that were evaluated on a per patient level. M_{GBM} was the least robust to this type of FGSM attack, across all three test datasets for AUROC, Dice score, and Sensitivity. This might be due to the somewhat consistent imaging presentation of glioblastomas. It was marginally more robust to attack on its own datatype (D_{GBM}). All three models failed by losing sensitivity instead of specificity, indicating that

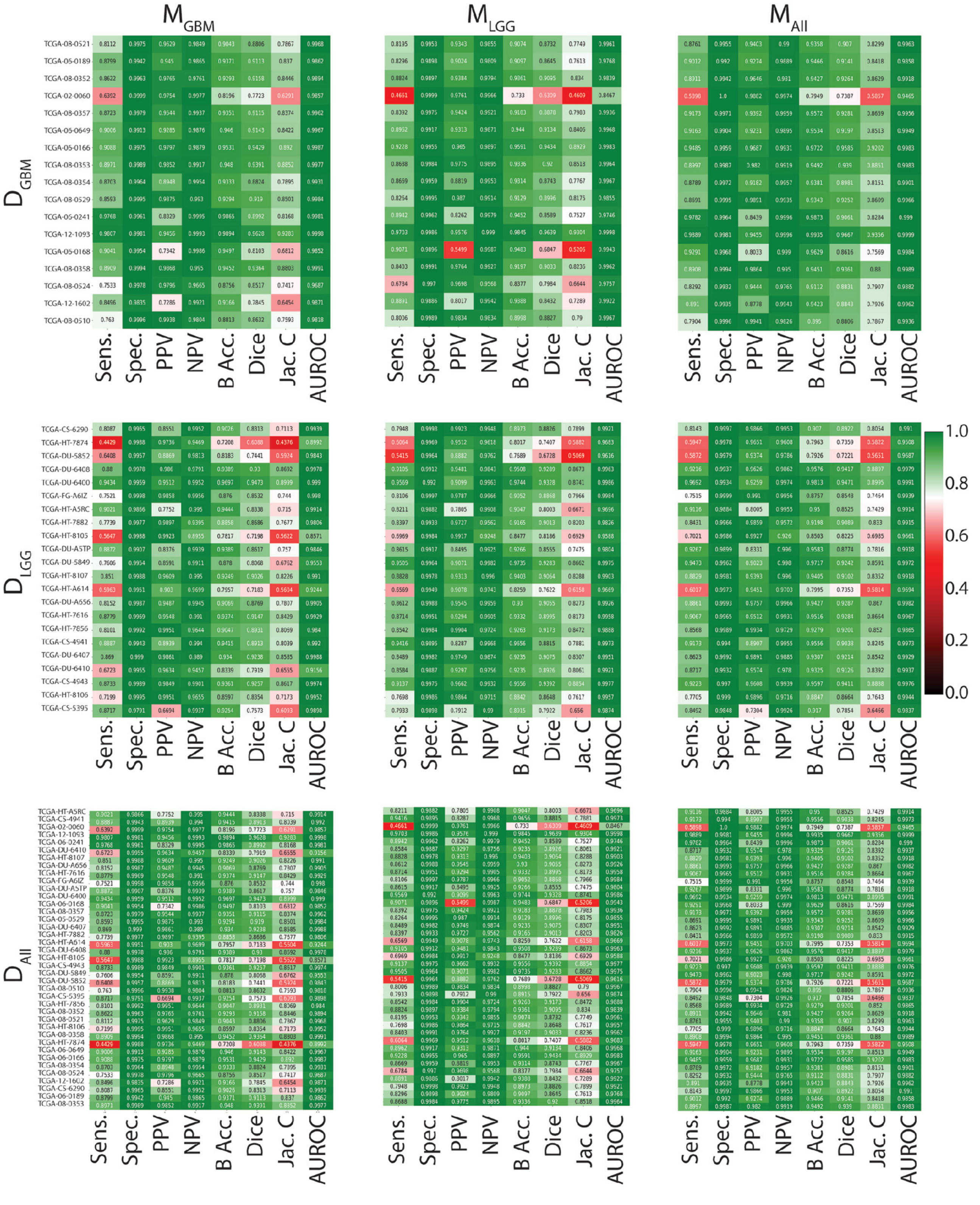


FIGURE 3 | Heat maps indicating patient-level performance metrics. Rows represent test datasets (D_{GBM}, D_{LGG}, D_{ALL}) and columns represent ML algorithms (M_{GBM}, M_{LGG}, M_{ALL}). D_{ALL} is formed by concatenating the first two rows. In each individual heat map, rows represent model performance on a particular test dataset and columns represent segmentation metrics. Patients for whom all models perform similarly worse are indicated in red.

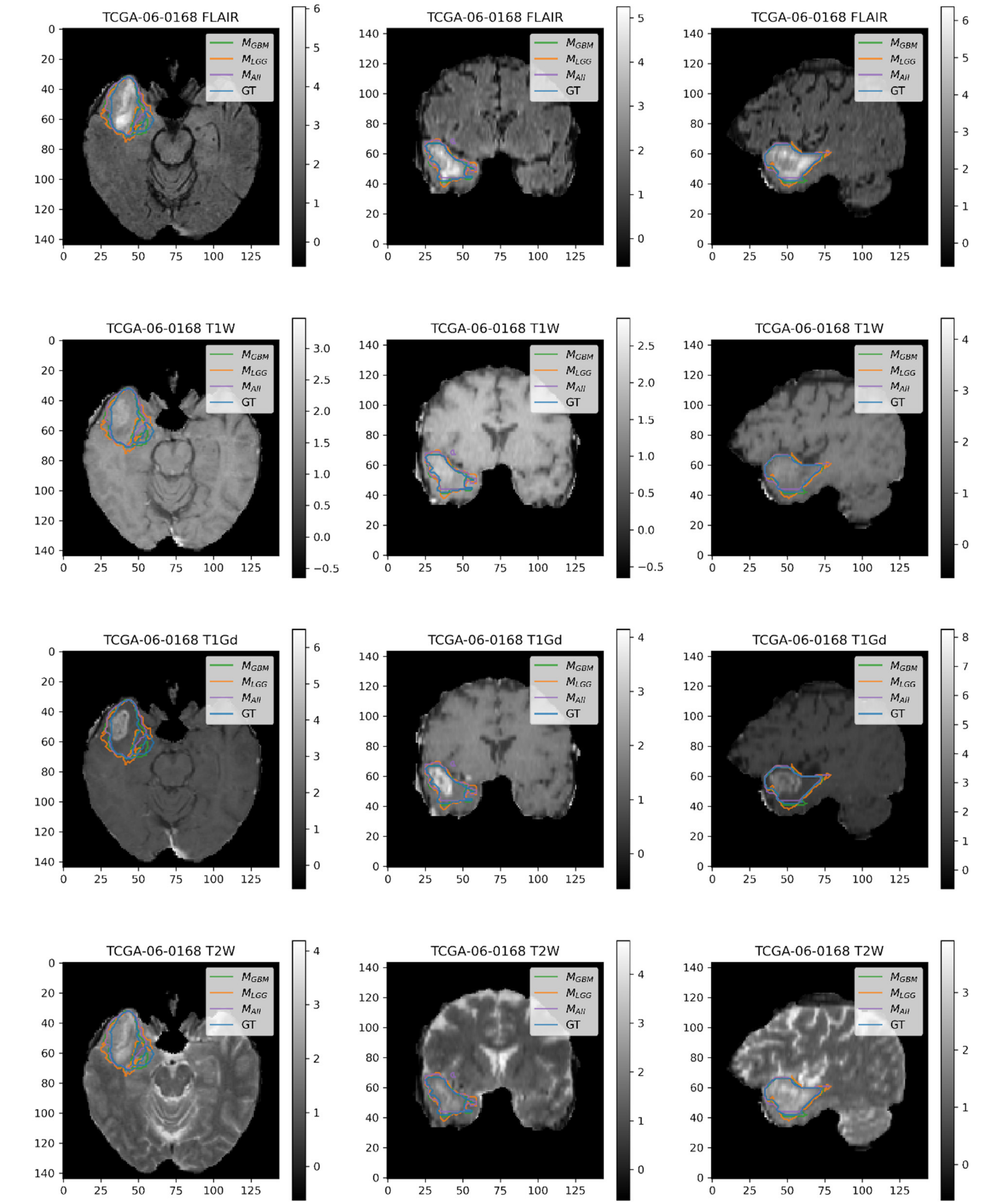


FIGURE 4 | Patient TCGA-06-0168 is diagnosed with GBM in the right temporal operculum. M_{LGG} has low performance on Dice Coefficient (Dice = 0.6847) than M_{GBM} (Dice = 0.8103) and M_{ALL} (Dice = 0.8616). AUROC for all models is high despite unequal performance. The boundary of the edema on FLAIR sequence shows where M_{LGG} over-segments and M_{GBM} under-segments tumor.

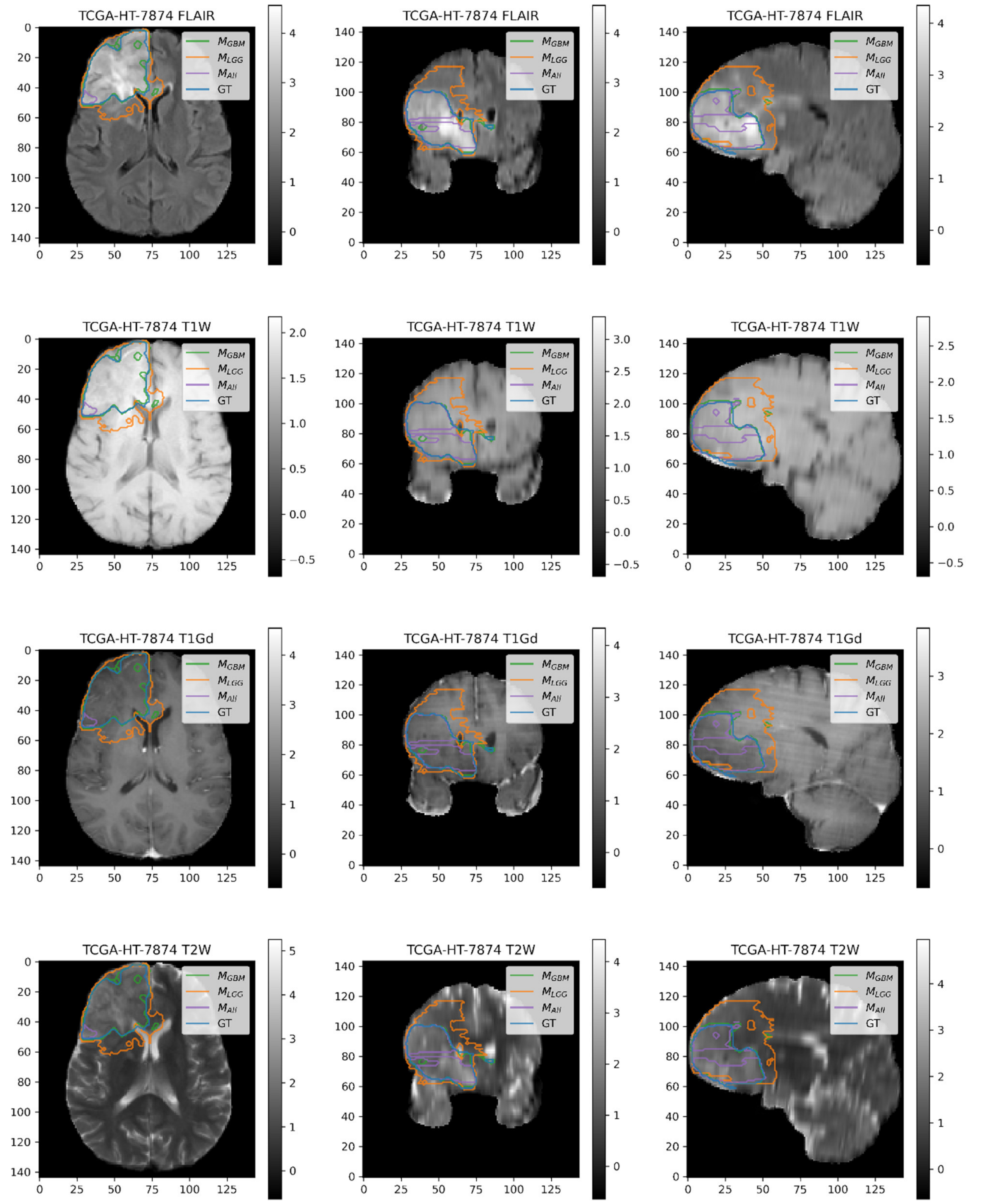


FIGURE 5 | Patient TCGA-HT-7874 belongs to D_{LGG} and has a tumor in the right frontal lobe. We selected this patient as it has consistently worse performance for metrics (Sens, B.Acc, Dice, Jac.C) across all models. Segmentation plot indicates M_{All} and M_{GBM} under-segment in this case, whereas M_{LGG} over-segments. M_{ALL} appears to be missing a central part of the tumor, as seen in the coronal and sagittal image planes. M_{LGG} appears to extend well beyond the region of FLAIR enhancement to over-segment the tumor. This LGG was significantly larger than most LGGs, and that may contribute to the difficulty of segmentation.

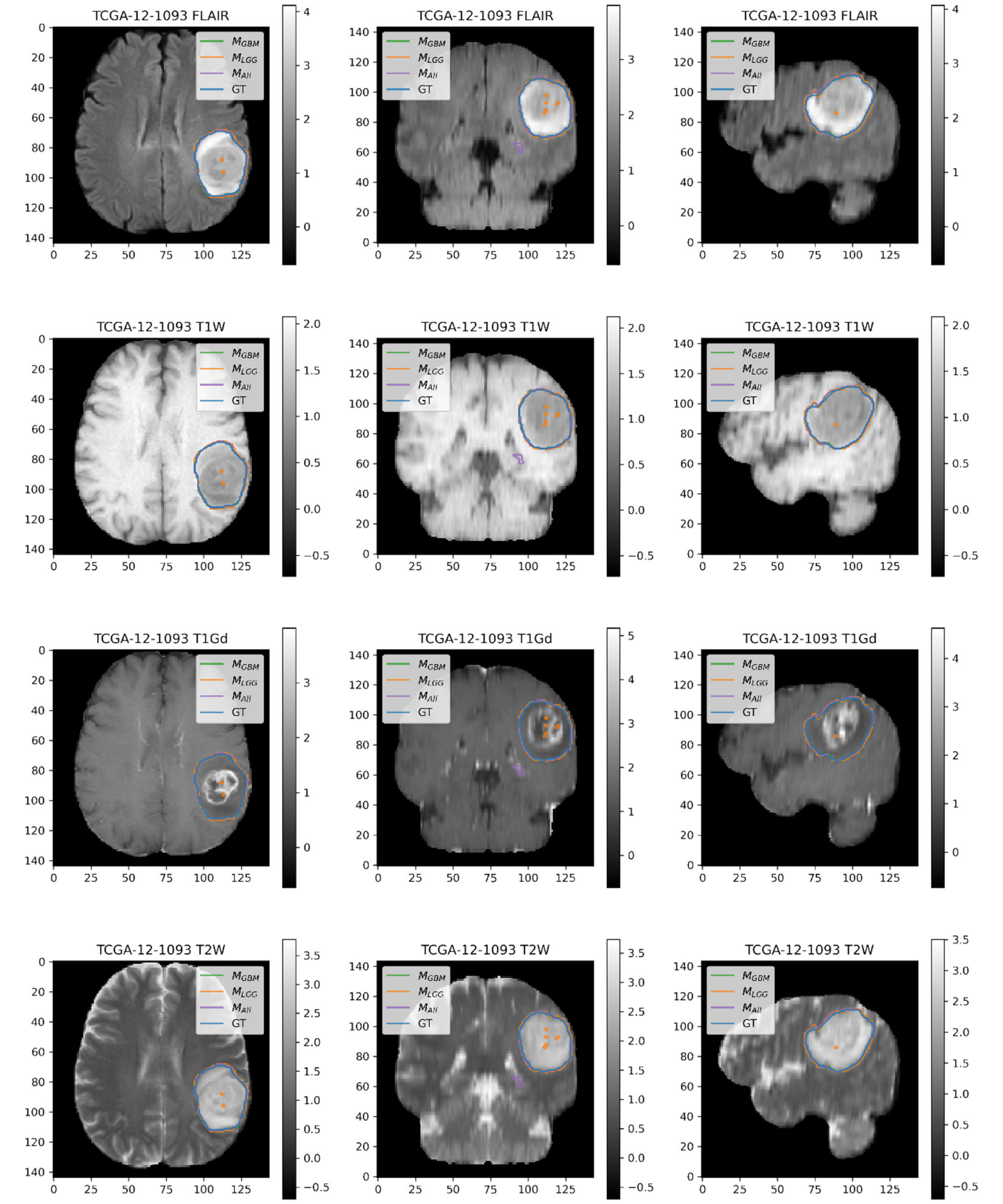


FIGURE 6 | Patient TCGA-12-1093 belongs to D_{GBM} and has a tumor in the left parietal lobe. We selected this patient as an example because it has consistently good performance for metrics (Sens, B.Acc, Dice, Jac.C) across all models. This GBM has clear margins, and a sharp boundary on FLAIR enhancing regions. The enhancing tumor core is central and distinct, and the models all perform relatively consistently in segmentation.

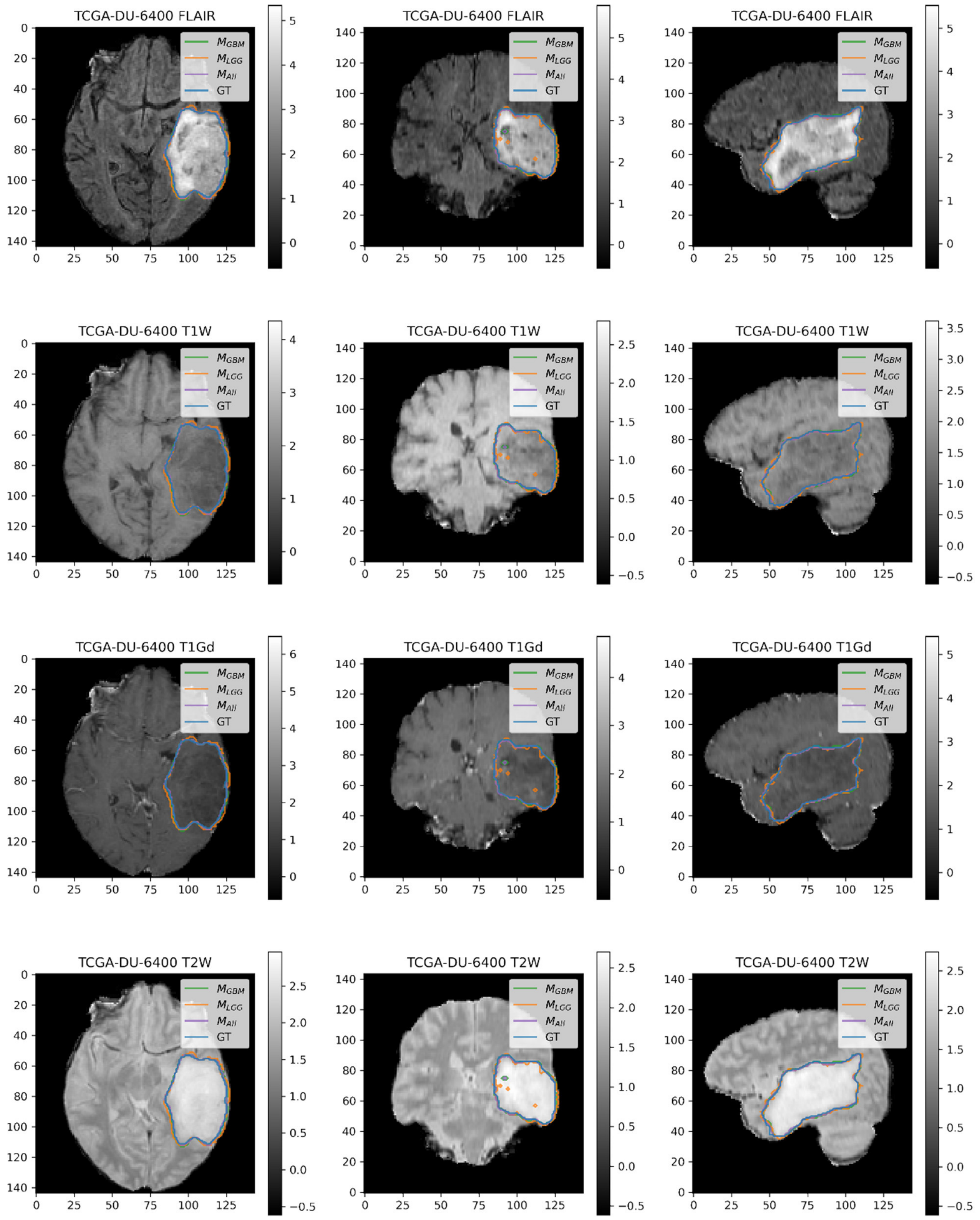


FIGURE 7 | Patient TCGA-DU-6400 belongs to D_{LGG} and has a tumor in the left temporal parietal region. We selected this patient as an example because it has consistently good performance for metrics (Sens, B.Acc, Dice, Jac.C) across all models. This LGG has clear margins, and the classic signature of FLAIR enhancement and no T1-Gd enhancement.

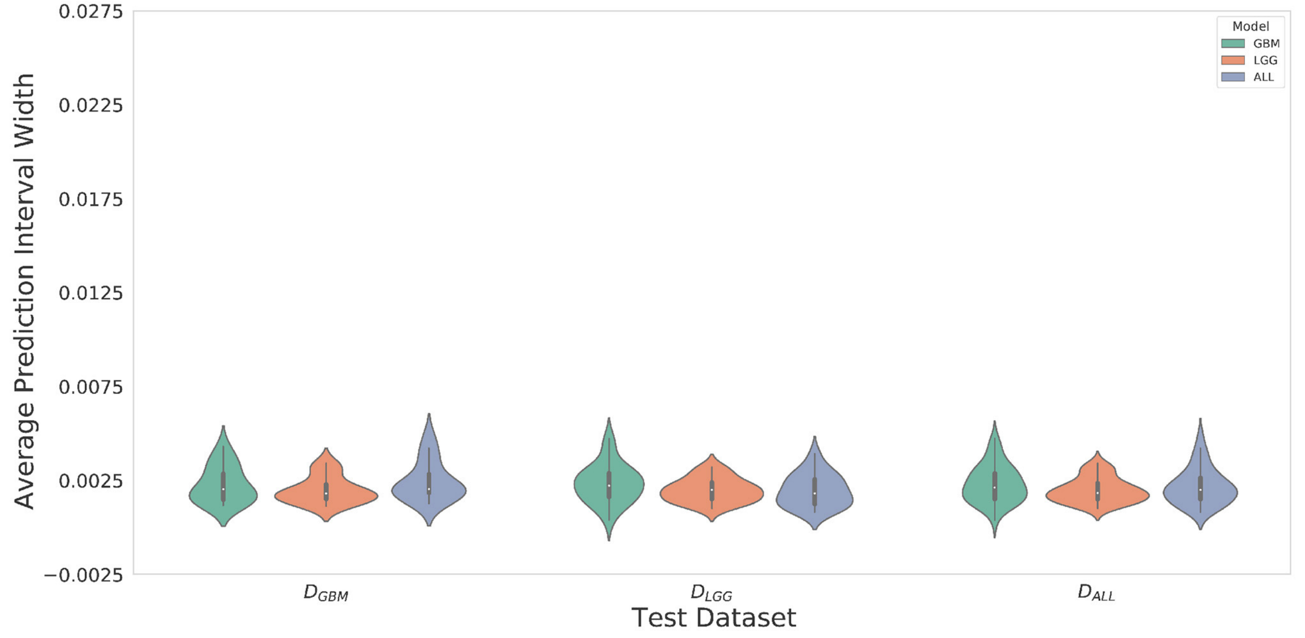


FIGURE 8 | The violin plot indicates models have equal median confidence while segmenting GBM and LGG patients due to greater number of non-tumor pixels in the datasets. The x-axis represents the datasets. The y-axis represents average prediction intervals. Models are sorted by hues and grouped together by test dataset.

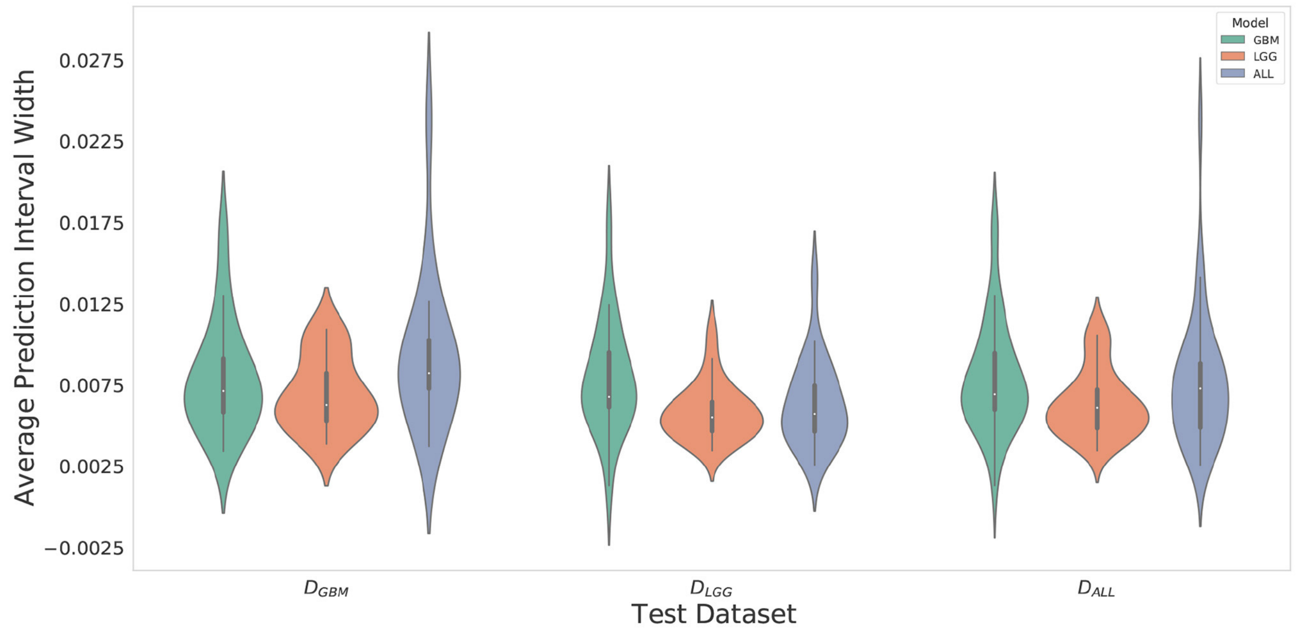


FIGURE 9 | Violin plots constructed to correct for effects of class imbalance and analyze model confidence while identifying tumor pixels only. Plots indicate models confidence is less consistent in identifying tumors due to wider inter-quartile range and greater spread of prediction interval distribution. Plot indicates M_{LGG} is the most confident model while segmenting both LGG and GBM patients.

the models began drastically under-segmenting the tumor under high levels of noise. **Figure 10** highlights the model behavior under different levels of noise. Under smaller amounts of noise (**Figure 11**), the all model had the best performance generally, though not significantly. M_{LGG} and M_{GBM} had the highest AUROC values of the three models for D_{LGG} and, D_{GBM} respectively, though the differences did not reach the significance threshold of ($p < 0.05$).

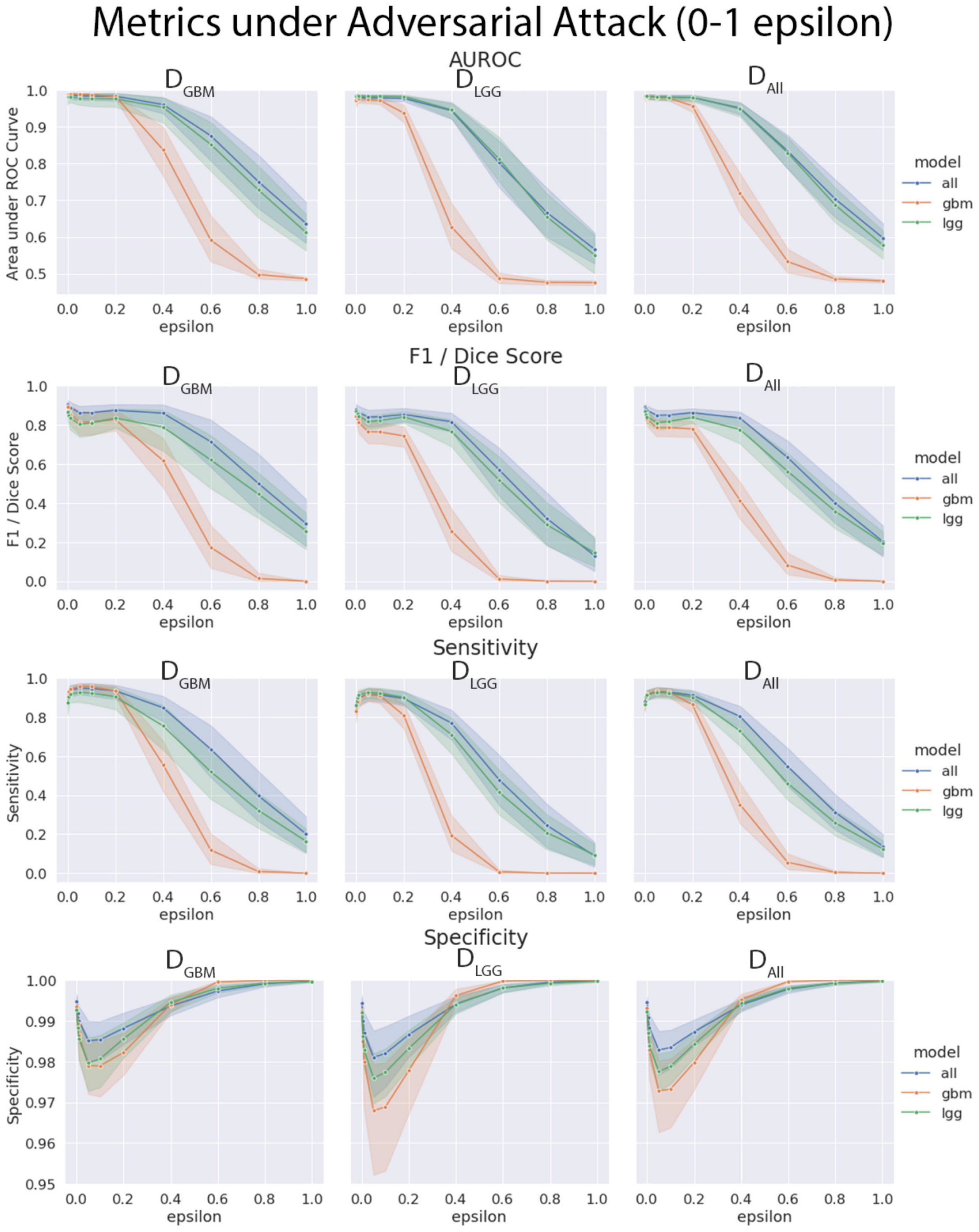


FIGURE 10 | Robustness of each model under FGSM attack, across the full range of epsilons (0–1.0) for four selected metrics. Ninety-five percent confidence intervals are provided to each model, and each of the three data sets were evaluated. M_{GBM} was least robust to FGSM attack at higher epsilon values with regard to AUROC, Dice score, and sensitivity.

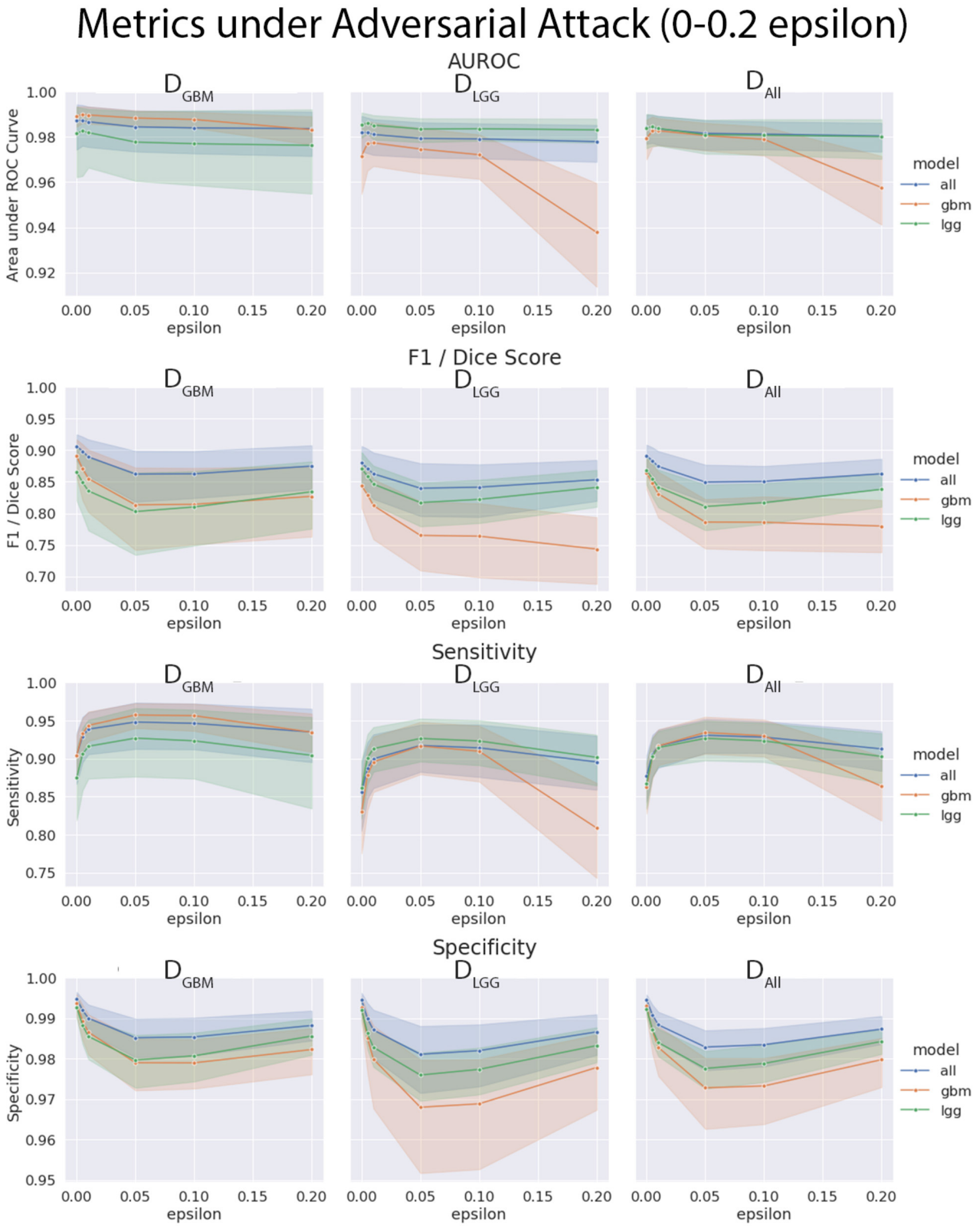


FIGURE 11 | Robustness of each model under FGSM attack, zoomed in on the early range of epsilons (0–0.2) for four selected metrics. Ninety-five percent confidence intervals are provided to each model, and each of the three data sets were evaluated. Models had more similar performance in the less aggressive levels of attack, with all model having marginally better performance, except with M_{LGG} and M_{GBM} models performing better with AUROC on their own test data sets.

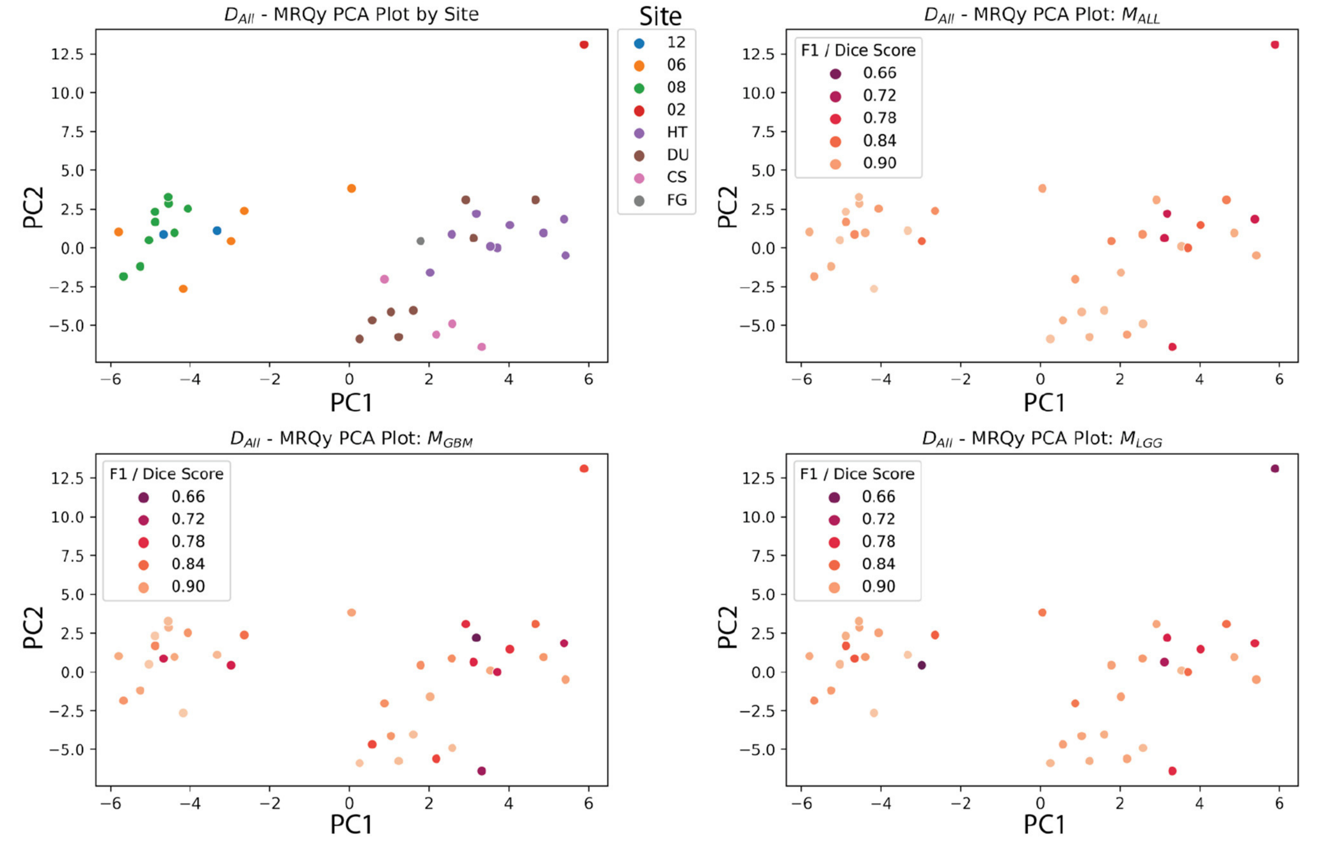


FIGURE 12 | MRQy features after principal component analysis, plotted against site code, and Dice scores of three models. The D_{ALL} test dataset has the first two principal components of the MRQy features plotted. Pairwise *t*-tests found that three clusters had significant differences in PCA space (06, 08, 12), (HT, DU, CS, and FG), and (02). Notably, these numeric codes happen to correspond to GBM studies, and the letter codes happen to correspond to LGG studies.

We found that the models trained only on D_{GBM} were less robust to adversarial noise, particularly at high levels of adversarial noise. These levels of noise may be extreme, but do give some sense of the performance of the models under duress. Other types of attacks that might be worthwhile to investigate include: adversarial patch attacks, Carlini and Wagner attacks, projected gradient descent, as well as GAN based attacks (Carlini and Wagner, 2017; Brown et al., 2018; Ren et al., 2020). This is not the only way of assessing robustness of models, as it assumes a motivated attacker to guide attacks, as opposed to natural sources of error, but it addresses how the margins of the tumor are affected on a consistent scale across the models. Natural sources of error are less coherent, comparable, and not as well computationally modeled in MRI as the body of work on adversarial attacks.

4.5. MRQy Features Vary Between Data Sets and Institutions, but Are Not Significantly Correlated With Metrics

The calibrated models' metrics and probabilities were evaluated for correlations with MRQy parameters, across the different test datasets. While there were some limited parameters that had significant correlations with model metrics, this was before FDR correction. One Thousand two hundred and twenty-four parameter to metric comparisons (17 MRQy parameters, 4 sequences, 6 metrics, 3 models) were performed, and none of the parameter-metric pairs were significantly correlated after FDR correction ($p < 0.05$). The MRQy features were collected before preprocessing, and were shown to be different across different institutions. However, the model used preprocessed data, and the MRQy features were not significantly correlated with the models' predictions and performance. This negative result adds more confidence to the predictions of the machine learning pipeline.

The PCA analysis showed that there were significant differences between three groups of site codes. The first cluster of institutions was 12, 06, and 08, the second was HT, DU, CS and FG, and the last was 02. Paired *t*-tests showed that the first principal component created splits with significant differences ($p < 0.05$). Notably, the numerical codes (02, 06, 08, 12) correspond to GBM studies, and alpha codes corresponded to LGG studies (HT, DU, CS, FG). However, within these clusters, the differences didn't reach significance. **Figure 12** shows the site codes plotted in PCA space, and then the three models with Dice coefficient. The fact that Henry Ford Hospital (06 for GBM and DU for LGG) had more in common with other GBM and LGG sites than between those two groups is notable, though hard to explain with such a limited sample size. Site 02 was also an outlier from both other clusters in this PCA space, and had relatively poor performance, though with one case it's hard to draw a firm conclusion.

The BraTS 2018 test datasets (Menze et al., 2015) did not have significant correlations after FDR correction between scan quality and metrics. This could be due to the high-fidelity curation and good consistency of the dataset. Another potential explanation could be the limited size of the dataset. Still, these data quality metrics show significant correlations with TCGA sites after PCA analysis, indicating batch effect differences, at least between the GBM and LGG datasets. Other data quality issues that models should be tested for include bias based on race, sex, and socioeconomic status. The rise of federated learning models makes this more urgent, because they allow for training models across collaborators without sharing data (Kairouz et al., 2021). Since sensitive data is not shared between sites, tracking batch effects and sources of bias requires more work and planning than if all the data were shared and managed centrally.

5. DISCUSSION

In this work, we used publicly available data and compared three U-Net-based algorithms in a stratified manner. Our main finding is that traditional segmentation performance metrics do not capture all aspects of an algorithm's performance, and can be potentially misleading. In this section, we first discuss the limitations of segmentation metrics, and how our proposed evaluation framework leads to a better understanding of model performance. We discuss the four axes of evaluation—diagnostic performance, model confidence, robustness, and analysis of batch effects in detail. Finally, we address the practical utility of our framework and list recommendations for model evaluation.

5.1. Limitations of Segmentation Metrics

Despite the technological advancements of Machine Learning (ML), the adoption of Ml in clinical workflows remains limited (Caruana et al., 2015; Strickland, 2019; Beede et al., 2020). This divide between the development and adoption of ML algorithms has been termed the "translation gap" (Steiner et al., 2021). This limitation is in part due to lack of holistic evaluation of the performance of those ML systems.

Majority of existing algorithms are statistically validated only using segmentation metrics (van Kempen et al., 2021), such as Dice Coefficient (Dice, 1945). In our experiments, we followed guidelines (Taha and Hanbury, 2015) to compute several segmentation metrics and test the differences between segmentation of GBM and LGG patients. We hypothesized that segmentation of LGG patients would be more difficult than GBM patients. LGG is diffuse and has low proliferation, which makes accurate segmentation of submicroscopic tumor tissues and tendrils, a difficult task. In contrast, GBM has greater signal intensity and characteristic presence of necrotic cavities, which makes segmentation comparatively more obvious. Our results found that metrics alone were insufficient to highlight the severity of mistakes that models make in segmentation. Only when segmentation contours were interpreted by a board-certified neuroradiologist, the degree, and types of errors of these models were evident. Similarly, in a recent systematic review of glioma segmentation algorithms, van Kempen et al. (2021) expected to find performance differences in segmentation of HGGs and LGGs but found that reported metrics could not capture such differences.

This points to a bigger concern raised by Reinke et al. (2021) that metrics alone are insufficient to evaluate all aspects of segmentation performance. While metrics are important for objective performance evaluation, they have several limitations for clinical utility (Maier-Hein et al., 2018). Difference in consequences of an algorithm's errors cannot be uncovered by metrics alone, and requires a clinical expert to elucidate them. For example, the consequences of under-segmenting in D_{GBM} might be more severe than under-segmenting in D_{LGG} due to the prognosis and management of the two diseases. As LGGs may merit a more conservatory, "wait-and-watch" approach, tumor that might be previously missed can be caught with additional tests. However, segmentation in case of GBM has more immediate consequences for resection and radiotherapy. Under-segmentation in this case would result in non-total resection, and perhaps if tumor tissue remains, would increase the likelihood of recurrence. Over-segmentation on the other hand would cause removal of non-tumor regions of the brain, or subject them to higher levels of radiotherapy, potentially causing functional impairments for patients. In case of glioma, the Dice Coefficient has a limited utility for evaluation of multifocal lesions (Giannopoulos and Kyritsis, 2010) because it cannot represent over-segmentation and under-segmentation (Yeghiazaryan and Voiculescu, 2018), does not support segmentation of multiple structures (Yeghiazaryan and Voiculescu, 2018), and is not immune to imaging artifacts and shape differences (Reinke et al., 2021). This serves as a cautionary tale that metrics alone are insufficient for reporting model performance, and there is clearly a need for better evaluation and reporting standards (Nagendran et al., 2020).

Since medical data is tightly controlled to protect patient privacy, federated learning has risen as a methodology to train models without exposing data. However, while the cross-site training structure has it's advantages, it requires thoughtful planning of model evaluation since model designers will not have access to the underlying data from other sites. Any metrics, quality control features, and batch effect monitoring will have to be carefully pre-planned to judge any resulting models. Thorough and holistic evaluation is especially important as site variability in protocol and patient populations is a known confounding factor. Our framework also helps illuminate the axes on which a federated learning network should judge their models beyond simple metrics like accuracy or AUROC.

5.2. Dimensions of the Evaluation Framework

The goal of our work is to inform how researchers can holistically evaluate their segmentation algorithms, and consider other axes of model performance than metrics alone. A problem faced by model developers in this domain is the lack of large datasets to effectively train and evaluate their algorithms. To realistically recreate this, we worked with smaller test datasets from TCGA-GBM and TCGA-LGG. Our work explores the effects of working with limited data, and informs how to interpret results meaningfully in such scenarios. Our experiments and methodology stand independently of whether the model evaluator has pre-built models, or is yet to train them. Our framework considers tumor heterogeneity, limitations of metrics and evaluates other axes such as model confidence, robustness, and batch effects. We don't suggest completely abandoning metrics—they would be important as a start, to get some level of insight. However, we caution against solely relying on metrics, and propose a more holistic evaluation of algorithms. In **Figure 1**, we map the axes of evaluation onto the standard ML pipeline. We provide other potential experiments that researchers can choose for model evaluation along specific axes. For example, techniques such as model ensembles and k-fold cross validation can be used to evaluate model confidence.

In our experiments, we evaluate model robustness with adversarial attacks. Recent work has shown the importance to evaluate the models' abilities to withstand adversarial attacks, especially in high-stakes scenarios such as radiology (Wetstein et al., 2020). These attacks can arise due to strong financial interests or technical infrastructure. We designed this experiment to test how and in what way could models fail in deployment under such an attack. This could lead to appropriate safeguards being put in place. Adversarial attacks also help shed light on the decision boundary of a neural network (Woods et al., 2019), which is otherwise something of a black box. Other sources of noise could be added, but have their own complications. Adding Gaussian noise to the inputs can be difficult to calibrate and variable due to randomness. Addition of artifacts, such as motion artifacts, is complex to model, and tools for doing so are not publicly available. Further research should investigate models using these failure modes, but is outside the scope of this paper. Another axes we investigate is analyzing the dataset for batch effects. In the context of tumor segmentation, batch effects could occur when image acquisition parameters or technical variations correlate with measurement quantity (Sadri et al., 2020). This may become a major problem when it leads to incorrect conclusions (Leek et al., 2010), especially when ML algorithms learn to pick up on these patterns. Analyzing for batch effects thus becomes important, as model predictions can be correlated with confounding factors. Our experiments found that pre-processing might help in making MRI scans more homogeneous and reduce these correlations.

We demonstrated our evaluation framework on ML algorithms trained with reliable, high-fidelity. expert-annotated BraTS Datasets. To further simplify the process of model development, we used straightforward implementations such as fixed dataset split (testing/validation) and 2D segmentation to work with limited data. Model developers can certainly use more sophisticated techniques that result in higher accuracy. Despite these limitations, our experiments are aligned to the overall goal of this work. Another limitation is we consider LGG for evaluation of generalizability. While there are significant imaging differences as compared to GBM, LGG is a broad category consisting of a range of tumor types. A more clinically useful investigation would be to evaluate performance on WHO recognized genetic subtypes such as IDH-mutant vs IDH-wt or 1p/19q codeleted tumors, as the literature on tumor subtypes evolves (Louis et al., 2016). However, we defer this as future work.

5.3. Recommendations for Evaluation of Tumor Segmentation Algorithms

Here, we summarize our work and presented the following recommendations for holistic evaluation of ML algorithms:

Accounting for tumor heterogeneity in evaluation: We focus on a specific problem of glioma, and evaluate for differences in models trained by stratification of GBM and LGG Data. The first stage in standard of care for glioma is the identification of the type, which further dictates the prognosis and treatment planning. However, there is high variability in this stage, and experts often don't reach immediate consensus. It is thus important for ML algorithms to generalize well across all tumor grades. We set out to investigate this question, by performing holistic evaluation on LGG, GBM, and mixed data. Researchers should consider unique imaging presentations of each patient and evaluate on a patient-level, as important differences might be diminished upon aggregation of data. Researchers should avoid evaluation on a dataset-level.

Adoption of tools in other domains to investigate glioma segmentation: Domains such as adversarial robustness and statistics have highly specialized tools (e.g., FGSM, conformal prediction intervals) to interrogate different aspects of model performance. In this work, we demonstrate the value of adopting such tools for the problem of performance evaluation of glioma segmentation. Our results indicate clear differences in these experiments. We found model trained on LGG Data to be more confident, and model trained on GBM to suffer the most under adversarial attacks. Researchers should evaluate their algorithms on each of the evaluation axes, by performing at least one experiment on each of the axes (**Figure 1**).

Exploring limitations of metrics in clinical utility: In recent years, the community has started to acknowledge the clinical limitations of standard segmentation metrics. Our work demonstrates why evaluation by metrics alone is limiting in investigating heterogeneity in clinical populations (i.e., GBM vs. LGG patients), and our findings further support recent literature. Researchers should avoid relying solely on metrics to evaluate their models.

The framework can further shed light on the practical utility of an algorithm, and serve as a decision-support tool. It is not meant to replace the triaging mechanisms already in place. Since the action that accompanies a decision is different, researchers should know the situations and the patient case before use of these algorithms. If the algorithm's prediction would be followed by a high-stakes action component such as surgery, tumor resection, or radiation therapy, accuracy of segmentation is critical. Our results indicate that algorithms trained on a specific glioma grade group do not generalize well out of distribution, so it is best to use specifically-trained models. For example, if a patient with GBM is to undergo surgery, use of M_{GBM} as a decision-support tool would be best. In low-stakes scenarios such as accessing the extent of tumor infiltration, generalizability is more important at the cost of accuracy. The use of M_{ALL}, which has knowledge of all glioma grade groups, would be best in this scenario.

Establishing a close collaboration with a clinical expert is crucial to ensure that results of the framework are appropriately interpreted. In this work, the authors collaborated with experts in neuroradiology and radiation oncology to deep-dive into the problem of brain tumor segmentation and present the limitations of metrics in a clinically meaningful way. Researchers should similarly consult a clinical expert to understand how tumor heterogeneity manifests in imaging presentations between the subgroups of the tumor they are interested to investigate. The use of this framework in other domains would thus require a close collaboration between ML researchers and clinicians for effective investigation.

6. CONCLUSION

In this work, we proposed a framework to evaluate the performance of tumor segmentation algorithms. To illustrate the framework, we investigated the generalizability of algorithms in different glioma grade groups. Institutions such as the American College of Radiology, Data Science Institute (ACR DSI) often lay out guidelines to researchers for best practices before model deployment. However, it is often not clear to researchers on how to evaluate models. We take a more granular view and present a tutorial of sorts, in addition to proposing a holistic framework for better model evaluation. In addition, we provide the following recommendations to researchers: (1) Perform at least one experiment on model confidence, diagnostic performance, data quality and robustness. (2) Perform analysis on a per-patient basis. (3) Gather representative images informed by the results of such analysis. (4) Collaborate with a clinical expert to perform qualitative evaluation of these images to get deeper insight on model performance.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.med.upenn.edu/sbia/brats2018/data.html.

AUTHOR CONTRIBUTIONS

NCW: data acquisition and pre-processing. VA, NCW, and SP: design of the experiments. SP and NCW: performing experiments and data analysis. AR, NCW, and SP: results interpretation. SP, NCW, NB, and XH: writing of the manuscript. AR: conception and design of study project and supervision. NB and XH: co-advising. JRB: clinical interpretation and guidance. All authors contributed to the article and approved the submitted version.

FUNDING

SP was supported by UM-MICDE Catalyst Grant (to XH and AR). AR was supported by CCSG Bioinformatics Shared Resource 5 P30 CA046592, a Research Scholar Grant from the American Cancer Society (RSG-16-005-01), and a Precision health Investigator award from U-M Precision Health to AR (along with L. Rozek and M. Sartor). SP and AR were also partially supported by the NCI Grant R37-CA214955. AR, NCW, XH, and NB were also supported by the University of Michigan (U-M) startup institutional research funds.

ACKNOWLEDGMENTS

Experiments were performed on Armis2 HPC Clusters provided by UM's Precision Health Initiative. Some advice on clinical perspective on gliomas was provided by Dr. Ashok Srinivasan. We acknowledge the support from the Center for Ethics, Society, and Computing (ESC) at the University of Michigan, Ann Arbor.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.740353/full#supplementary-material

REFERENCES

Bajaj, A. S., and Chouhan, U. (2020). A review of various machine learning techniques for brain tumor detection from MRI images. *Curr. Med. Imag.* 16, 937–945. doi: 10.2174/1573405615666190903144419

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). Segmentation labels for the pre-operative scans of the TCGA-GBM collection. [Data Set]. The Cancer Imaging Archive. doi: 10.7937/K9/TCIA.2017.KLXWJJ1Q

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). Segmentation labels for the pre-operative scans of the TCGA-LGG collection. [Data Set]. The Cancer Imaging Archive. doi: 10.7937/K9/TCIA.2017.GJQ7R0EF

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Advancing the Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., et al. (2020). "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, (New York, NY: ACM), 1–12. doi: 10.1145/3313831.3376718

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Brown, T. B., Mané, D., Roy, A., Abadi, M., and Gilmer, J. (2018). Adversarial patch. *arXiv [preprint]*. Available online at: http://arxiv.org/abs/1712.09665 (accessed June 6, 2021).

Bulakbaşı, N., and Paksoy, Y. (2019). Advanced imaging in adult diffusely infiltrating low-grade gliomas. *Insights Imaging* 10:122. doi: 10.1186/s13244-019-0793-8

Carlini, N., and Wagner, D. (2017). "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)* (San Jose, CA: IEEE Computer Society), 39–57. doi: 10.1109/SP.2017.49

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). "Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission," in *Association for Computing Machinery* (New York, NY), 1721–1730. doi: 10.1145/2783258.2788613

Chan, P., Dinniwell, R., Haider, M. A., Cho, Y. B., Jaffray, D., Lockwood, G., et al. (2008). Inter- and intrafractional tumor and organ movement in patients with cervical cancer undergoing radiotherapy: a cinematic-MRI point-of-interest study. *Int. J. Radiat. Oncol. Biol. Phys.* 70, 1507–1515. doi: 10.1016/j.ijrobp.2007.08.055

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

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., et al. (2013). The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* 26, 1045–1057. doi: 10.1007/s10278-013-9622-7

Claus, E. B., Walsh, K. M., Wiencke, J., Molinaro, A. M., Wiemels, J. L., Schildkraut, J. M., et al. (2015). Survival and low grade glioma: the emergence of genetic information. *Neurosurg. Focus* 38:E6. doi: 10.3171/2014.10.FOCUS12367

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302. doi: 10.2307/1932409

Dong, H., Supratak, A., Mai, L., Liu, F., Oehmichen, A., Yu, S., et al. (2017). "TensorLayer: a versatile library for efficient deep learning development," in *Proceedings of the 25th ACM International Conference on Multimedia* (New York, NY: ACM), 1201–1204. doi: 10.1145/3123266.3129391

Elder, B., Arnold, M., Murthi, A., and Navrátil, J. (2021). Learning prediction intervals for model performance. *arXiv [Preprint]*. Available online at: http://arxiv.org/abs/2012.08625

Forst, D. A., Nahed, B. V., Loeffler, J. S., and Batchelor, T. T. (2014). Low-grade gliomas. *Oncologist* 19, 403–413. doi: 10.1634/theoncologist.2013-0345

Giannopoulos, S., and Kyritsis, A. (2010). Diagnosis and management of multifocal gliomas. *Oncology* 79, 306–312. doi: 10.1159/000323492

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations* (San Diego, CA).

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). "On calibration of modern neural networks," in *International Conference on Machine Learning* (Sydney, NSW: PMLR) 70, 1321–1330.

Jungo, A., and Reyes, M. (2019). "Assessing reliability and challenges of uncertainty estimations for medical image segmentation," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Shenzhen: Springer), 48–56. doi: 10.1007/978-3-030-32245-8_6

Kabir, H. M., Khosravi, A., Hosen, M. A., and Nahavandi, S. (2018). Neural network-based uncertainty quantification: a survey of methodologies and applications. *IEEE Access* 6, 36218–36234. doi: 10.1109/ACCESS.2018.2836917

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., et al. (2021). Advances and Open Problems in Federated Learning. *Found. Trends Mach. Learn.* 14, 1–210. doi: 10.1561/9781680837896

Kingma, D. P. (2015). Adam: a method for stochastic optimization. *arXiv [Preprint]*. Available online at: http://arxiv.org/abs/1412.6980.

Kocher, M., Ruge, M. I., Galldiks, N., and Lohmann, P. (2020). Applications of radiomics and machine learning for radiotherapy of malignant brain tumors. *Strahlenther. Onkol.* 196, 856–867. doi: 10.1007/s00066-020-01626-8

Kompa, B., Snoek, J., and Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit. Med.* 4, 1–6. doi: 10.1038/s41746-020-00367-3

Kümmel, A., Bonate, P. L., Dingemanse, J., and Krause, A. (2018). Confidence and prediction intervals for pharmacometric models. *Pharmacometr. Syst. Pharmacol.* 7, 360–373. doi: 10.1002/psp4.12286

Larsen, J., Wharton, S. B., McKevitt, F., Romanowski, C., Bridgewater, C., Zaki, H., et al. (2017). 'Low grade glioma': an update for radiologists. *Br. J. Radiol.* 90:1070. doi: 10.1259/bjr.20160600

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739. doi: 10.1038/nrg2825

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., et al. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* 9, 1–13. doi: 10.1038/s41467-018-07619-7

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94. doi: 10.1038/s41586-019-1799-6

Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., and Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Med. Imaging* 39, 3868–3878. doi: 10.1109/TMI.2020.3006437

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., et al. (2020). Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ* 368:m689. doi: 10.1136/bmj.m689

Nazar, U., Khan, M. A., Lali, I. U., Lin, H., Ali, H., Ashraf, I., et al. (2020). Review of automated computerized methods for brain tumor segmentation and classification. *Curr. Med. Imaging* 16, 823–834. doi: 10.2174/1573405615666191120110855

Ojika, D., Patel, B., Reina, G. A., Boyer, T., Martin, C., and Shah, P. (2020). Addressing the memory bottleneck in AI model training. *arXiv [Preprint].* Available online at: http://arxiv.org/abs/2003.08732.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* 10, 61–74.

Rebsamen, M., Knecht, U., Reyes, M., Wiest, R., Meier, R., and McKinley, R. (2019). Divide and conquer: stratifying training data by tumor grade improves deep learning-based brain tumor segmentation. *Front. Neurosci.* 13:1182. doi: 10.3389/fnins.2019.01182

Recht, M. P., Dewey, M., Dreyer, K., Langlotz, C., Niessen, W., Prainsack, B., et al. (2020). Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *Eur. Radiol.* 30, 3576–3584. doi: 10.1007/s00330-020-06672-5

Reinke, A., Eisenmann, M., Tizabi, M. D., Sudre, C. H., Rädsch, T., Antonelli, M., et al. (2021). Common limitations of image processing metrics: a picture story. *arXiv [Preprint].* Available online at: http://arxiv.org/abs/2104.05642.

Ren, K., Zheng, T., Qin, Z., and Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Engineering* 6, 346–360. doi: 10.1016/j.eng.2019.12.012

Renard, F., Guedria, S., Palma, N. D., and Vuillerme, N. (2020). Variability and reproducibility in deep learning for medical image segmentation. *Sci. Rep.* 10, 1–16. doi: 10.1038/s41598-020-69920-0

Romano, Y., Patterson, E., and Candés, E. J. (2019). Conformalized quantile regression. *arXiv [Preprint]*. Available online at: http://arxiv.org/abs/1905.03222.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 9351* (Munich: Springer Verlag), 234–241. doi: 10.1007/978-3-319-24574-4_28

Rousseau, A.-J., Becker, T., Bertels, J., Blaschko, M. B., and Valkenborg, D. (2021). "Post training uncertainty calibration of deep networks for medical image segmentation," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* (Nice: IEEE), 1052–1056. doi: 10.1109/ISBI48211.2021.9434131

Sadri, A. R., Janowczyk, A., Zou, R., Verma, R., Antunes, J., Madabhushi, A., et al. (2020). MRQy: an open-source tool for quality control of MR imaging data. *arXiv [Preprint].* Available online at: http://arxiv.org/abs/2004.04871.

Seabold, S., and Perktold, J. (2010). "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference* (Austin, TX), 92–96. doi: 10.25080/Majora-92bf1922-011

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., et al. (2020). Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep*. 10:12598. doi: 10.1038/s41598-020-69250-1

Steiner, D. F., Chen, P.-H. C., and Mermel, C. H. (2021). Closing the translation gap: AI applications in digital pathology. *Biochim. Biophys. Acta* 1875:188452. doi: 10.1016/j.bbcan.2020.188452

Strickland, E. (2019). IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. *IEEE Spectr*. 56, 24–31. doi: 10.1109/MSPEC.2019.8678513

Taha, A. A., and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* 15:29. doi: 10.1186/s12880-015-0068-x

Tamimi, A. F., and Juweid, M. (2017). "Epidemiology and outcome of glioblastoma," in *Glioblastoma*, ed S. De Vleeschouwer (Brisbane, AU: Codon Publications), 143–153. doi: 10.15586/codon.glioblastoma.2017.ch8

Tan, A. C., Ashley, D. M., López, G. Y., Malinzak, M., Friedman, H. S., and Khasraw, M. (2020). Management of glioblastoma: state of the art and future directions. *Cancer J. Clin*. 70, 299–312. doi: 10.3322/caac.21613

Tipping, M. E., and Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. *Neural Comput*. 11, 443–482. doi: 10.1162/089976699300016728

Udupa, J. K., LeBlanc, V. R., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L. M., et al. (2006). A framework for evaluating image segmentation algorithms. *Comput. Med. Imaging Graph*. 30, 75–87. doi: 10.1016/j.compmedimag.2005.12.001

van Kempen, E. J., Post, M., Mannil, M., Witkam, R. L., ter Laan, M., Patel, A., et al. (2021). Performance of machine learning algorithms for glioma segmentation of brain MRI: a systematic literature review and meta-analysis. *Eur. Radiol.* doi: 10.1007/s00330-021-08035-0

Wadhwa, A., Bhardwaj, A., and Singh Verma, V. (2019). A review on brain tumor segmentation of MRI images. *Magn. Reson. Imaging* 61, 247–259. doi: 10.1016/j.mri.2019.05.043

Wetstein, S., González-Gonzalo, C., Bortsova, G., Liefers, B., Dubost, F., Katramados, I., et al. (2020). Adversarial attack vulnerability of medical image analysis systems: unexplored factors.

Whittle, I. R. (2004). The dilemma of low grade glioma. *J. Neurol. Neurosurg. Psychiatry* 75, 31–36. doi: 10.1136/jnnp.2004.040501

Witthayanuwat, S., Pesee, M., Supaadirek, C., Supakalin, N., Thamronganantasakul, K., and Krusun, S. (2018). Survival analysis of glioblastoma multiforme. *Asian Pac. J. Cancer Prevent*. 19, 2613–2617.

Woods, W., Chen, J., and Teuscher, C. (2019). Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nat. Mach. Intell*. 1, 508–516. doi: 10.1038/s42256-019-0104-6

Yeghiazaryan, V., and Voiculescu, I. (2018). Family of boundary overlap metrics for the evaluation of medical image segmentation. *J. Med. Imaging* 5:1. doi: 10.1117/1.JMI.5.1.015006

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 15:e1002683. doi: 10.1371/journal.pmed.1002683

Zwanenburg, A., Leger, S., Agolli, L., Pilz, K., Troost, E. G. C., Richter, C., et al. (2019). Assessing robustness of radiomic features by image perturbation. *Sci. Rep*. 9, 1–10. doi: 10.1038/s41598-018-36938-4

Conflict of Interest: NCW was a founder and shareholder of Prenovo, EIQ, and AMI healthcare technology startups. AR has a consulting agreement with Voxel analytics LLC. and consults for Genophyll, LLC.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Medicine

ORIGINAL RESEARCH
published: 03 November 2021
doi: 10.3389/fmed.2021.740248

A Clinically-Compatible Workflow for Computer-Aided Assessment of Brain Disease Activity in Multiple Sclerosis Patients

Benoit Combès[1*†], *Anne Kerbrat*[1,2†], *Guillaume Pasquier*[3], *Olivier Commowick*[1], *Brandon Le Bon*[1], *Francesca Galassi*[1], *Philippe L'Hostis*[4], *Nora El Graoui*[1,5], *Raphael Chouteau*[2], *Emmanuel Cordonnier*[3], *Gilles Edan*[1,2] and *Jean-Christophe Ferré*[1,5]

[1] Univ Rennes, Inria, CNRS, Inserm IRISA UMR 6074, Empenn ERL U 1228, Rennes, France, [2] Neurology Department, Rennes University Hospital, Rennes, France, [3] IRT bcom, Rennes, France, [4] Biotrial, Rennes, France, [5] CHU Rennes, Department of Neuroradiology, Rennes, France

Over the last 10 years, the number of approved disease modifying drugs acting on the focal inflammatory process in Multiple Sclerosis (MS) has increased from 3 to 10. This wide choice offers the opportunity of a personalized medicine with the objective of no clinical and radiological activity for each patient. This new paradigm requires the optimization of the detection of new FLAIR lesions on longitudinal MRI. In this paper, we describe a complete workflow—that we developed, implemented, deployed, and evaluated—to facilitate the monitoring of new FLAIR lesions on longitudinal MRI of MS patients. This workflow has been designed to be usable by both hospital and private neurologists and radiologists in France. It consists of three main components: (i) a software component that allows for automated and secured anonymization and transfer of MRI data from the clinical Picture Archive and Communication System (PACS) to a processing server (and vice-versa); (ii) a fully automated segmentation core that enables detection of focal longitudinal changes in patients from T1-weighted, T2-weighted and FLAIR brain MRI scans, and (iii) a dedicated web viewer that provides an intuitive visualization of new lesions to radiologists and neurologists. We first present these different components. Then, we evaluate the workflow on 54 pairs of longitudinal MRI scans that were analyzed by 3 experts (1 neuroradiologist, 1 radiologist, and 1 neurologist) with and without the proposed workflow. We show that our workflow provided a valuable aid to clinicians in detecting new MS lesions both in terms of accuracy (mean number of detected lesions per patient and per expert 1.8 without the workflow vs. 2.3 with the workflow, $p = 5.10^{-4}$) and of time dedicated by the experts (mean time difference $2'45''$, $p = 10^{-4}$). This increase in the number of detected lesions has implications in the classification of MS patients as stable or active, even for the most experienced neuroradiologist (mean sensitivity was 0.74 without the workflow and 0.90 with the workflow, p-value for no difference $= 0.003$). It therefore has potential consequences on the therapeutic management of MS patients.

Keywords: computer aided diagnosis, radiology, lesion activity, MRI, Multiple Sclerosis

OPEN ACCESS

Edited by:
Richard McKinley,
Bern University Hospital, Switzerland

Reviewed by:
Ilya Nasrallah,
University of Pennsylvania,
United States
Stefano Cerri,
Technical University of
Denmark, Denmark

Correspondence:
Benoit Combès
benoit.combes@inria.fr

In Memoriam:
Dedicated to the memory of Christian Barillot

[†] These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *12 July 2021*
Accepted: *30 September 2021*
Published: *03 November 2021*

Citation:
Combès B, Kerbrat A, Pasquier G, Commowick O, Le Bon B, Galassi F, L'Hostis P, El Graoui N, Chouteau R, Cordonnier E, Edan G and Ferré J-C (2021) A Clinically-Compatible Workflow for Computer-Aided Assessment of Brain Disease Activity in Multiple Sclerosis Patients. Front. Med. 8:740248. doi: 10.3389/fmed.2021.740248

INTRODUCTION

Magnetic Resonance Imaging (MRI) currently plays a central role in the diagnosis, prognosis and follow-up of patients with Multiple Sclerosis (MS) (1). In particular, the identification of new FLAIR hyperintense lesions between two longitudinal MRI scans allows: (i) to confirm the diagnosis of relapsing-remitting MS if the criterion of dissemination in time is not met on the first MRI scan (2); (ii) to provide information on the prognosis of the disease (3); (iii) to evaluate for each patient the current efficacy of its disease modifying treatment. Indeed, in recent years, the number of disease-modifying treatments for MS has increased significantly (1). In particular, highly effective second-line immunosuppressive treatments have become available and the number of first-line treatments has increased. However, these treatments are not without potential side-effects. The challenge is therefore to prescribe the right treatment to the right patient and to monitor its effectiveness closely. In this context, the concept of No Evidence of Disease Activity (NEDA) has emerged (4) and implies that MS patients have neither clinical relapse nor new FLAIR lesions on their follow-up MRI under treatment. An annual follow-up by brain MRI is therefore currently recommended, at least during the first year of treatment (5, 6), and the comparison of annual MRI scans is frequently performed by the radiologists and neurologists in charge of the follow-up of MS patients. However, this comparison is a complex and mentally demanding task that often leads to an underestimation of lesion accumulation, even for most experienced radiologists (7). Consequently, there is a need for dedicated systems that can provide clinicians, regardless of their level of expertise, an aid for accurate and robust detection of new FLAIR MS lesions. The ultimate goal of these systems will be to reduce the underestimation of patients wrongly reported as having no or few new lesions as well as the associated expert dependencies, resulting in better therapeutic decisions. For many years, different methods have been proposed to address this issue (8).

More recently, standardization of MR imaging acquisitions and data-transfer protocols as well as advances in computer vision methods have offered the premises for an end-to-end workflow for computer-aided comparative analysis of longitudinal MRI data. In particular, thanks to the development of deep learning techniques, powerful tools for the automatic segmentation of new MS lesions have been proposed in the context of academic research on the one hand [e.g., (7, 9, 10)], and integrated into commercial products on the other hand (11). However, the added-value of these tools in clinical practice is not well-documented, especially regarding therapeutic strategy and disability progression. In addition, the question of their integration into clinical practice is generally not addressed. Finally, commercially available solutions based on Artificial Intelligence often lack available scientific evidence in peer-reviewed Journals (11) and their high cost limits their deployment for patient care. Consequently, such tools have not yet been adopted in routine clinical practice by the majority of radiologists and neurologists.

Within this context, we launched the MUSIC project (an acronym for MUltiple Sclerosis Image Checkout) in 2017 in Brittany, a region in the north-west of France. The objective of this project was to develop, deploy and evaluate a fully-integrated clinical workflow allowing to improve detection of new brain lesions in MS patients. The system has been designed to be usable by both hospital and private radiologists and neurologists in Brittany. The MUSIC project also included centralized storage of MS patients' MRI data so that their data could be accessed and compared even if they moved from one center to another for their MRI or neurological follow-up. The first "proof of concept" phase of the project reported in this article was deployed in 5 centers (2 university hospitals, 2 local hospitals, 1 private radiology center).

The MUSIC workflow consists of three main components: (i) a software component that allows for automated and secured anonymization and transfer of MRI data from the clinical Picture Archive and Communication System (PACS) to a processing server (and vice versa); (ii) a fully automated MR image segmentation core that enables detection of new lesions from patients T1 weighted, T2 weighted and FLAIR brain acquisitions, and (iii) a dedicated web viewer that provides an intuitive visualization of new lesions to the clinical staff, easy to show to the patients. These elements allow clinicians to access and visualize enhanced patient data scanned in any connected clinical center, even without being linked to a clinical PACS. In the present paper, we first illustrate the MUSIC project workflow. Second, we assess the performance of three clinicians, with different levels of expertise, in identifying new MS lesions on follow-up MRI scans, with and without the proposed workflow. For this evaluation, we used longitudinal pairs of scans from 54 MS patients.

MATERIALS AND METHODS

Figure 1 summarizes the overall MUSIC project workflow. Briefly, after being stored in the clinical local PACS, MR images are pseudonymized and securely transferred into a processing hosting, where images are processed and new lesions are automatically segmented using a deep neural network. Then, the processed images and corresponding segmentation maps are transferred back to the clinical hosting from which they can be efficiently visualized in a dedicated web MRI viewer. In the following, we describe the three main elements of the workflow: the transfer and storage modules (section The transfer and storage modules: Servers interoperability and data access), the segmentation module (section The segmentation module: Detection of new lesions from longitudinal brain MR images) and the visualization module (section The visualization module: Efficient and adapted reporting). Then, in section Evaluation of the MUSIC workflow, we present a set of experiments that we designed and carried out to evaluate the radiologist and the neurologist performances in identifying new FLAIR lesions between two sets of MRIs of MS patients with and without the workflow.

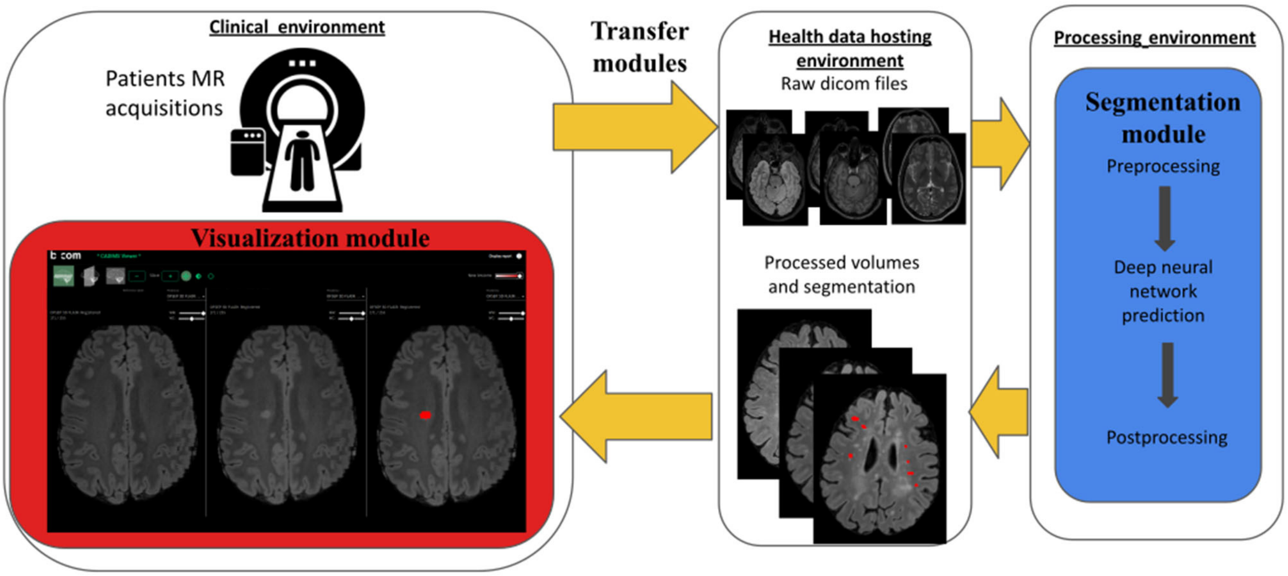


FIGURE 1 | Workflow overview. Colored elements were specifically designed and developed in the context of our MUSIC project and consists of (i) a set of Transfer and Storage Modules (yellow), (ii) a Segmentation Module (blue) and (iii) a Visualization Module (red). Briefly, after being stored in the clinical PACS, MR images are pseudonymised and securely transferred into a processing hosting, where images are processed so that new lesions are automatically segmented. Then the resulting processed images and associated new lesions segmentation maps are returned to the clinical data hosting platform where they can be visualized in a dedicated web MRI viewer.

Workflow Description

The Transfer and Storage Modules: Servers Interoperability and Data Access

In order to process the images outside the hospitals, a set of tools to pseudonymize, securely transfer and reidentify data has been set up. Overall, this module is composed of five main components: the hospital PACS, the centralized PACS, a telemedicine platform, the NodeJS transfer server and the research PACS. These elements and their interconnections are presented in **Figure 2**. The telemedicine platform is responsible for transferring the medical images from the hospital PACS to the centralized PACS. The latter gathers images coming from different hospitals and is hosted in a certified health data hosting provider. At this stage, patient data are still identified. From here, a clinical research assistant initiates the pseudonymization procedure. An HTTP request is sent to a NodeJS server, also deployed in the same location, with the Universally Unique IDentifier (UUID) of the study to be transferred and the new patient identifier. The server, developed using NodeJS, a JavaScript runtime to develop modular network applications, is a simple server which listens to incoming HTTP requests. It can answer two specific requests: "transfer data" and "import results." Once it receives a HTTP "transfer data" request, it retrieves the images from the PACS using a DICOMweb^{TM} WADO-RS (Web Access to DICOM Objects Retrieved Study) request, de-identifies the images according to DICOM recommendations (DICOM Supplement 142), and sends the de-identified images to the Research PACS over an HTTPS connection to prevent any attack, using DICOMweb^{TM} STOW-RS (STore Over the Web Retrieved Study). Thanks to this procedure, the pseudonymization is performed inside the health data hosting provider and no identifying data goes out.

To process a set of patient acquisitions, a web application has been developed. This application lists the patients available and is able to run a segmentation over one or multiple selected patients. It also communicates with the research PACS using DICOMweb^{TM}, locally downloads the data temporarily and runs the segmentation algorithms from a Python script. Once the images have been processed and segmented, the NodeJS server is notified that results are available by an "import result" HTTP request. It retrieves the images from the research PACS using a DICOMweb^{TM} WADO-RS request, reidentifies the new images using a patented method (ID EP3756123), and stores them in the centralized PACS. The images are finally exported to the radiologist and neurologist to be analyzed via the telemedicine platform. In practice, follow-up data from each patient is thus accessible from any connected clinical environment. The overall transfer time to perform these various tasks is about 15 min per subject (excluding segmentation).

Overall, this workflow has been designed to use only standardized requests for interoperability purposes and can be connected to any telemedicine platform.

The Segmentation Module: Detection of New Lesions From Longitudinal Brain MR Images

The visual identification of new lesions in MRI requires the mental processing of a large amount of 3D information and it is common for radiologists to miss notable lesions emerging from one acquisition to another, even for highly-experienced radiologists (7). The segmentation module thus aims

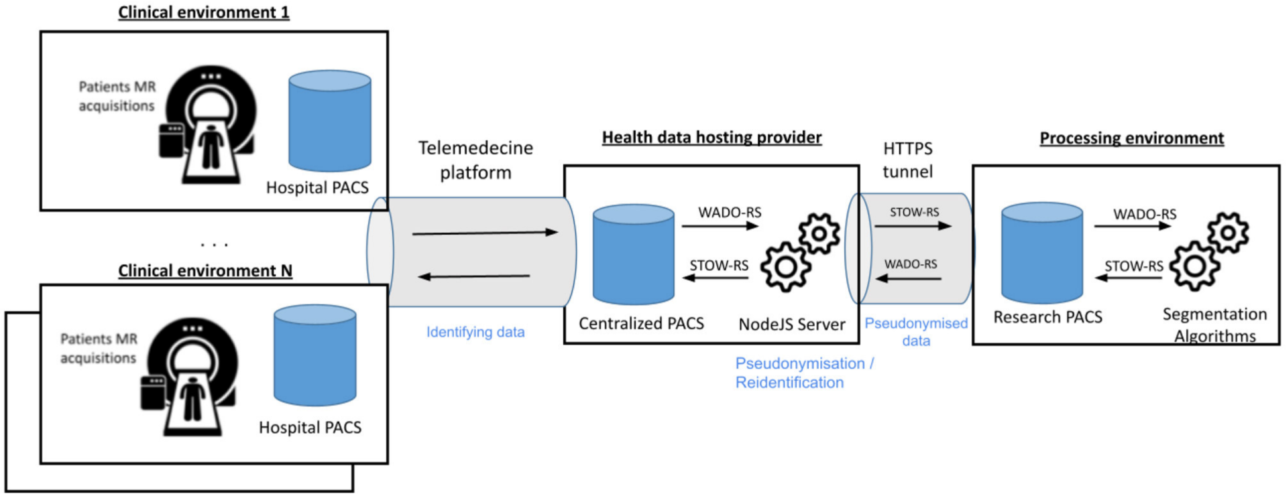


FIGURE 2 | Detailed architecture and interconnections in the MUSIC network. The system is composed of 2 main interconnected PACS (the centralized and the research PACS). The centralized PACS is interconnected to each connected hospital PACS to download the acquisitions to process and to upload the results. The research PACS is dedicated to data processing. The different connections are based on standardized HTTP requests (DICOMweb^{TM}) and are secured by an HTTPS tunnel.

at automatically extracting candidate new lesions that will then be highlighted in a dedicated viewer accessible to experts. This design comes with two consequences:

- First, we accept a reasonable amount of false positive candidates that will be naturally considered as irrelevant by image readers, with the counterpart that it increases our chances to detect relevant changes (in other terms, we favor sensitivity over specificity).
- Second, we accept to segment both growing and new lesions without distinction and let the image readers assess the relevance of including each of them into their radiological analysis.

A first natural solution to detection and segmentation of new lesions consists in first, independently segmenting the lesions for each of the two time-points of interest using a dedicated algorithm and second, comparing the resulting segmentation masks (or their associated probability maps) to infer a mask associated to the presence of new lesions. The main advantages of such an approach consist in its ability to stand on the numerous methods developed in the last decades to segment lesions from brain MRI (12) and on the availability of the associated annotated databases. However, by splitting the original problem into two subtasks, this approach disregards the temporal correlation in the images, which may lead to inaccurate segmentations for small lesions or subtle changes. A second fruitful approach thus consists in inferring notable signal changes due to lesions from one acquisition to another directly from the MRI volumes of interest at the two time points, instead of from the two lesion maps (9, 13–18). Such a solution has the advantage of benefiting from all the information available at once and thus maximizing its ability to detect relevant signals of interest. Intuitively, by comparing scans from one session to another, we alleviate the problem from a part of confounding factors due to interindividual anatomical differences. Nonetheless, such a method needs databases of serial MR scans acquired at different time steps and with manually segmented new lesions, which are relatively uncommon as of now.

In this project, we chose to develop a method following this second approach. This method is briefly detailed in the four next subsections. First, we designed a training/testing dataset consisting of a set of pairs of FLAIR, T2-weighted (T2w) and T1-weighted (T1w) acquisitions from 41 MS patients. Second, we set out a pre-processing pipeline so that data of a given patient are appropriately aligned and signal intensity is comparable from one acquisition to another. Third, we trained a deep neural network whose inputs consist, for a given patient, of the two sets of T1w, T2w, and FLAIR images and output consists of the softmax output map associated with the presence of new lesions at each voxel. Fourth, we implemented a few post-processing steps to produce a binary segmentation mask from the network softmax layer. The resulting trained model achieved a true positive rate of 0.83 and an overall rate of false positive of 0.09 on our testing dataset (17 patients, 41 new lesions). Moreover, comparable results were obtained on an additional set of 10 data consisting of acquisitions on Philips and General Electric 3T MRI scanners.

Building the Training and Testing Dataset

We designed our segmentation module using a dataset from a previous clinical project (ClinicalTrials ID: NCT02117375) consisting of a set of MR scans from 41 patients acquired on two Siemens 3T MRI scanners (Magnetom Verio, VB17). For each of these patients, data consists of 3D T1w, 2D axial T2w, and 3D FLAIR imaging at two times temporally distant by 1 year. Acquisition parameters were: for 3D T1w: 4 min 30, 1 × 1 × 1 mm, TR = 1,900, TE = 2.26, TI = 900, FA = 9, matrix = 256

× 256, GRAPPA2; for Axial DPw T2w: 2 min 12, 0.7 × 0.7 × 3 mm, TR = 6,530, TE = 9.4/84, FA = 150, matrix = 320 × 320, GRAPPA2; for 3D FLAIR: 5 min, 1 × 1 × 1.1 mm, TR = 5,000, TE = 399, TI = 1,800, matrix = 256 × 256, GRAPPA2.

New lesions from the first acquisition to the second one were manually segmented by an expert and reviewed by another expert using the ITK-SNAP software (http://www.itksnap.org/). This procedure provides the delineation of 152 new lesions. The dataset was split into a learning dataset (24 patients, 111 lesions) and a testing dataset (17 patients, 41 lesions). Data splitting was achieved by stratifying lesions according to their locations (deep white matter, periventricular, juxtacortical, brainstem, cerebellum) and optimizing patients repartition to achieve balanced (60%/40%) training and testing groups with respect to these characteristics.

Data Pre-processing

Our preprocessing pipeline, close to the subtraction pipeline proposed in Ref. (16), aims at preparing the T1w, T2w, and FLAIR data so that voxel-wise differences between consecutive scans were as meaningful as possible. Briefly, firstly MR volumes are reoriented in RAS coordinates. Secondly, skulls and skin tissues are removed from the data using a robust registration-based brain extraction method (animaAtlasBasedBrainExtraction, available at anima.irisa.fr, RRID:SCR_017017 and RRID:SCR_01707). Thirdly, baseline and follow-up T1w, T2w, and FLAIR scans are rigidly registered on the FLAIR baseline using a block matching registration method [animaPyramidalBMRegistration (19)]. We used the FLAIR baseline scan as reference for the registration as this is the one generally used by experts in clinical practice. Nevertheless, we did not observe any notable difference in results when modifying the choice of the reference (neither in training, nor in testing). Fourth, images are all cropped using the FLAIR baseline as a mask in order to reduce pointless data. Fifth, bias due to spatial inhomogeneity is estimated using the N4 algorithm (20) and removed from the data (animaN4BiasCorrection). Finally, for each pair of baseline and follow-up images (e.g., FLAIR baseline and FLAIR follow-up) voxel intensities are jointly corrected using a 2 fold procedure: (i) first, their joint histogram is linearly rescaled so that it best fits the y = x line in a least square sense, (ii) second, a Nyul standardization (21) on an in-house multisequence template is applied independently on each acquisition (animaNyulStandardization).

Deep Neural Network Architecture and Learning

The core of the segmentation module consists of a fully convolutional neural network that was trained to segment new lesions from a pair of preprocessed FLAIR, T1w, and T2w acquisitions. Specifically, we adopted the nnU-net framework proposed by Isensee et al. [(22), github.com/MIC-DKFZ/nnUNet] that enables training of a 3D U-Net (23) while automating the choice of the hyperparameter values. This framework has been shown to outperform a number of deep learning-based methods on a variety of segmentation tasks. Precisely, our 3D U-Net has 6 input channels (one for each sequence and each time point) of size [160, 192, 64]. To fit this frame, each input image is first resampled to size [0.5, 0.5, 1.1 mm] (median training image resolution) and then each set of 6 images (3 sequences for each of the 2 time points) is split into patches of such a size. Finally, each such 6× [160, 192, 64] patch is processed independently and aggregated to others to form the final softmax outputs map. Data augmentation included: (i) isotropic rescaling, (ii) 3D rotation, (iii) mirroring in the sagittal plane, (iv) smooth elastic deformations and (v) intensity enhancements and attenuations on lesion voxels (modeling the diversity of signal change due to lesions). This network was trained to minimize the sum of Cross-Entropy and Dice loss over the training dataset and included a drop-out based regularization (with probability = 0.2). Training was performed using a stochastic gradient descent run over 1,000 epochs, each of them consisting of 250 minibatches. Learning was conducted on a GPU NVIDIA Quadro P6000, 24 GB and lasted 10 days. Prediction for a given patient lasted about 6 min (including pre and post-processing) on the same hardware.

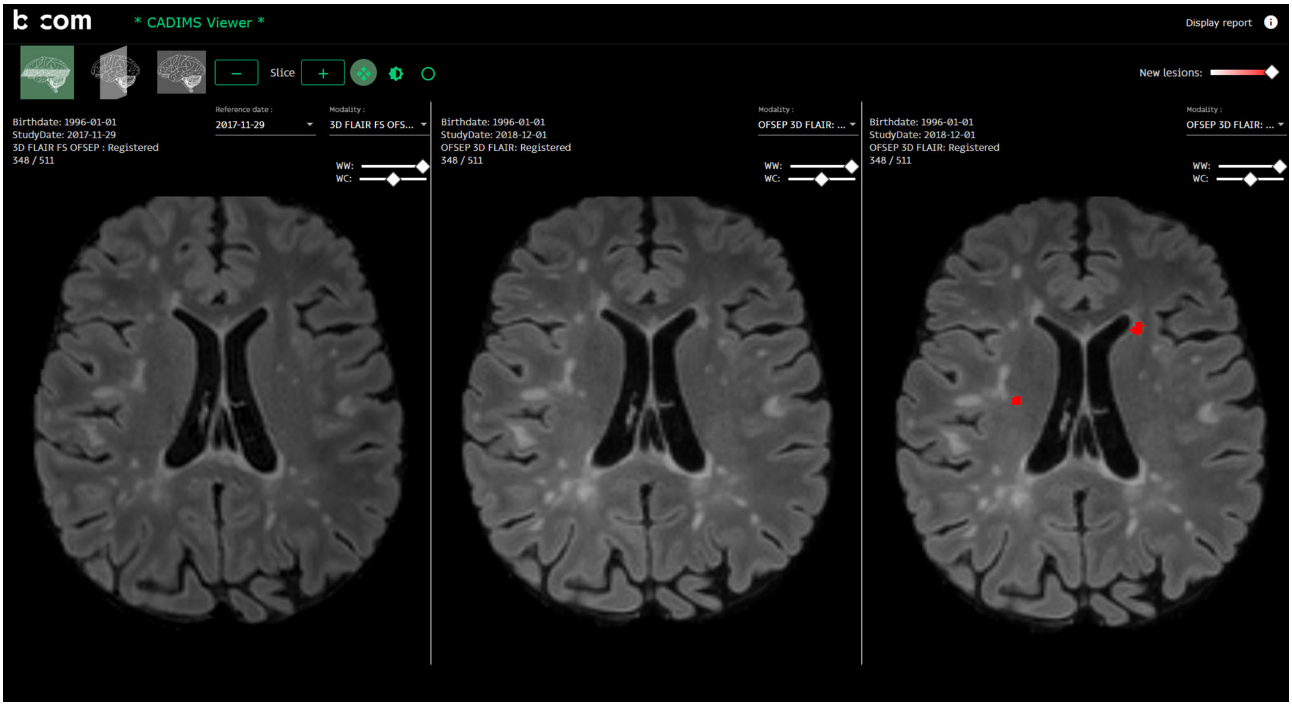
Data Post-processing

Once the neural network evaluated for a given pair of acquisitions, a binary segmentation map was obtained from the network softmax outputs using the following empirical procedure. First, the softmax outputs map is binarized using a threshold of 0.01. Second, connected components (26-connectivity) were extracted from the resulting binary map. Third, only connected components with volume larger than 12 mm^3 and including at least one voxel with softmax value >0.1 were selected as new lesions in the final output mask. Last, preprocessed data and the corresponding segmentation mask were resampled to the original baseline FLAIR image slab and resolution.

The Visualization Module: Efficient and Adapted Reporting

The CADIMS software has been designed to allow a fast and intuitive access to the preprocessed volumes and new lesions segmentation masks (**Figure 3**, as well as Video available in **Supplementary Material**). It was built in collaboration with a neurologist and a neuroradiologist following MS patients to meet their clinical needs. It consists of a MRI viewer usable from a standard web browser. It has been developed using the AMI framework (https://github.com/FNNDSC/ami) for the visualization of medical images and integrated in an Angular application. It allows the visualization of DICOM images that are directly retrieved from the hospital PACS using DICOMWeb^{TM}, the DICOM Standard for web-based medical imaging. As explained in section The transfer and storage modules: Servers interoperability and data access, the processed scans and the segmentation maps are transferred back via the telemedicine platform and are directly available in the viewer. Moreover, images are still stored durably in the centralized backed up PACS and are also available to any clinicians connected to the MUSIC network.

From a practical perspective, the viewer is composed of three synchronized views where three registered images are



Birthdate: 1996-01-01
StudyDate: 2017-11-29
3D FLAIR FS OFSEP : Registered
348 / 511

Reference date : 2017-11-29
Modality : 3D FLAIR FS OFS...

Birthdate: 1996-01-01
StudyDate: 2018-12-01
OFSEP 3D FLAIR: Registered
348 / 511

Modality : OFSEP 3D FLAIR: ...

Birthdate: 1996-01-01
StudyDate: 2018-12-01
OFSEP 3D FLAIR: Registered
348 / 511

Modality : OFSEP 3D FLAIR: ...

FIGURE 3 | The CADIMS viewer: The CADIMS MRI viewer is usable from a standard web browser. It consists of three synchronized views displaying from left to right (i) the baseline scan, (ii) the follow-up scan and (iii) the follow-up scan with segmented new lesions highlighted in red.

visualized simultaneously (from left-to-right: initial image, follow-up image, follow-up image with new lesions mask). The viewer displays the FLAIR images in the axial plane at startup. The other sequences (T1-w, T2-w) and other planes (sagittal, coronal) can be visualized by selecting them on dedicated menus. If more than one follow-up MRIs has been acquired for the same patient, previous acquisitions are also accessible. The viewer also integrates the following basic navigation functionalities: padding, zooming, and intensity windowing, all accessible from the computer-mouse.

Evaluation of the MUSIC Workflow

In this section, we present the datasets used and the two sets of experiments conducted to assess the added value of the MUSIC workflow on routine clinical practice.

Data Sets

Patients from 5 MRI centers were prospectively included in the MUSIC project. All patients were informed and written consents were obtained. All patients were included in the OFSEP ("Observatoire Français de la Sclérose en Plaques") cohort, registered on clinicaltrials.gov (NCT02889965) and compliant with French data confidentiality regulations. The study was approved by the relevant ethics committee.

Inclusion criteria were chosen to target a population with a substantial number of active patients. They included (i) a diagnosis of MS according to 2017 Mc Donald criteria (2); (ii) a disease duration <10 years; (iii) an Expanded Disability Status Scale (EDSS) score <4; (iv) a follow-up MRI available 10–16 months after the first MRI.

Our evaluation dataset consists of 54 pairs (baseline and follow-up) of 2D or 3D FLAIR, T1w, and T2w scans acquired on 9 different 3T MR scanners from Siemens, Philips and General Electrics. Thirty out of the 54 studied MS patients had a follow-up scan on a different MR scanner than the first scan, and 22 out of these 30 on a MR scanner from a different manufacturer. The overall allocation between scanners is depicted in **Table 1**. All data were acquired according to the OFSEP recommendations (6). The median and range FLAIR, T1w, and T2w spatial resolutions (in mm) were, respectively [1, 1, 1] (range [0.7, 0.7, 0.6]; [1, 1, 1]), [1, 1, 1] (range [0.5, 0.5, 2]; [1, 1, 1]) and [0.7, 0.7, 3] (range [0.5, 0.5, 1]; [1, 1, 3]). Data were not preselected according to quality criteria and a few acquisitions were of lower quality (example in **Figure 4**). Patients main characteristics were: mean age 35 yo (SD = 10), mean EDSS 1.1 (SD = 1.3), disease duration 3.7 years (SD = 1.3), percent of women = 67%.

Experimental Setting

We conducted two experiments involving three experts with different levels of experience: a senior neuroradiologist with 15 years of experience (named "expert 1" below), a senior neurologist with 8 years of experience (named "expert 2" below) and a junior radiologist (named "expert 3" below). Each of these two experiments are detailed below and consisted of the visual analysis of a set of pairs of acquisitions in two different conditions.

TABLE 1 | Repartition of patients in the different scanners (each row is a different scanner).

Manufacturer	Version	Number of sessions (overall 108)
Phillips	Ingenia	47
Siemens	Prisma	23
Siemens	Verio	17
Phillips	Ingenia	9
Phillips	Ingenia	5
Siemens	Aera	3
General Electrics	SIGNA Explorer	2
Siemens	Aera	1
General Electrics	SIGNA Explorer	1

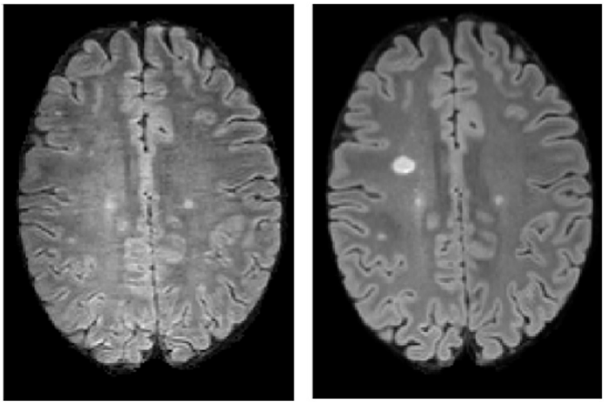


FIGURE 4 | A pair of FLAIR acquisitions from a patient of the evaluation dataset experiencing a low quality baseline scan. **(Left)** Baseline FLAIR axial slice. **(Right)** Follow-up FLAIR axial slice.

Impact of the Segmentation Module on Expert Performances

In this first experiment, we assessed the added value of the segmentation module on the ability of each expert to detect new lesions arising between the two time points. This experiment was conducted on 48 patients out of the 54. It consists of a 2-fold procedure. In its first phase, each expert was asked to annotate all notable new lesions—by simply drawing a point near the center of the lesion—from the pre-registered FLAIR, T1w, and T2w volumes for the two time points of interest. Then, in a second phase 2 weeks later, each expert was asked to perform the same exercise with an additional input: the segmentation mask provided by the segmentation module. Annotated lesions as well as time to perform each segmentation were recorded.

This experiment was performed on a dedicated reading system allowing MR volume annotation built from the fsleyes software (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLeyes). The time measurement was automated. All experts were asked to conduct this experiment in situations as similar as possible to clinical practice. In particular, experts were explicitly instructed not to spend more time reading MRIs than they would have in clinical routine. However, they had to be in a quiet environment so as not to be interrupted during their reading. A few days prior to the first phase, each expert experienced a short session to experiment with the reading system.

Impact of the MUSIC Workflow on Routine Clinical Practice

In this second experiment, we explored the added value of the overall MUSIC workflow in clinical practice. This experiment was conducted on 6 patients. Again, it was a 2-fold procedure. In the first phase, each expert was asked to visualize the MRI data and write a radiological report using the fully manual procedure currently in use. Hence, the images were viewed directly from the PACS and MRI for the two time points were manually roughly registered. The presence of new lesions was visually assessed and annotated in the radiological report, without any computer-aided tool. In the second phase of the experiment, 2 weeks later, each expert was asked to repeat the exercise via the MUSIC workflow (i.e., from a user perspective, using the new lesion segmentation mask and realigned data in the dedicated web MRI viewer). The experts measured the time needed to load data, read the MRI and write the report in the two phases. As in the first experiment, experts were explicitly instructed not to spend more time reading MRIs than they would have in clinical routine. They again had to be in a quiet environment so as not to be interrupted during their reading.

Statistical Analysis

Impact of the Segmentation Module on Expert Performances

First, for each expert and during each phase of this first experiment, detected lesions were colocalized using an automated analysis and manual intervention when necessary. This stage allows us to produce a mapping between each detected lesion, the names of the experts who detected it and the phase (phase 1 or/and phase 2) in which it was detected. Second, each lesion that has been reported, regardless of the phase of the experiment, was labeled as a true positive or a false positive via a consensus reading of all lesions from the two most experienced experts. Finally, we computed:

- The number of lesions detected by each expert as well as the overall number of individual lesions (i.e., counted only once for all experts) detected, for each phase.
- The inter-expert differences on detected lesions within each phase reported as ratio, pairwise Cohen's kappa statistics and multi-rater Fleiss' kappa statistic and associated 95% confidence intervals (CI).
- The number of lesions detected in phase 1 and not in phase 2 and conversely.
- The averaged patient-wise number of lesions detected by experts in each phase, that is compared between phases using a paired student test.
- The number of patients reported with at least one notable lesion by each expert and in each phase, as well as the associated pairwise Cohen's kappa statistics and multi-rater Fleiss' kappa statistic and associated 95% CIs. The overall sensitivity and specificity associated to this categorization (i.e., at least one new lesion vs. no new lesion) was then computed

TABLE 2 | Inter-expert heterogeneity during phase 1.

	Expert 1	Expert 2	Expert 3
Expert 1	–	23/90 κ =0.38 [0.22, 0.53]	17/90 κ = 0.62 [0.48, 0.75]
Expert 2	16/83	–	19/83 κ = 0.44 [0.29, 0.59]
Expert 3	8/83	18/83	–

The first figures give the number of lesions detected by the expert in row not detected by the expert in column. For example, over its 90 new lesions detected, expert 1 detected 23 lesions that were not detected by expert 2 and 17 that were not detected by expert 3. The second figures (only once per expert combination) give the Cohen's Kappa coefficient and its associated 95% CI.

for each phase and tested for equality between phase 1 and phase 2 using a logistic regression including a patient and an expert random effect.

- The pooled inter-expert standard deviation associated to the number of lesions detected in each phase, that is compared between the phases.
- The individual sensitivity together with its 95% CI for each expert and each phase. Moreover, for each expert, sensitivity is tested for equality between phase 1 and 2 using a logistic regression including a patient random effect. Associated odds ratio, p-values for odds ratio = 1 and associated 95% CI are reported.

Finally, mean time elapsed for each expert and each phase was estimated and tested for equality between phases using a paired student test.

Impact of the MUSIC Workflow on Routine Clinical Practice

First, radiological reports from this second experiment were gathered. Then for each expert and each setting (i.e., using the full MUSIC workflow or using the current manual approach), patients were categorized according to the report as: "no activity," "1 lesion" or "> 1 lesion."

Second, the time spent to perform radiological readings for each of the three experts and each of the two settings were summarized and the mean times elapsed in the two settings were tested for equality using a paired t-test.

RESULTS

Impact of the Segmentation Module on Expert Performances

Detection of New Lesions Without the Segmentation Mask

During the first phase, overall 113 lesions were detected. The three experts, respectively, detected 90, 83, and 83 new lesions. **Table 2** reports the difference of lesions detected from one expert to another as well as the inter-rater Cohen's Kappas, illustrating the high inter-rater variability on detected lesions. Moreover, the overall Fleiss's Kappa coefficient was 0.47 with 95% CI = [0.38, 0.57]. **Figure 5** gives an example of a notable lesion detected by only one of the three experts.

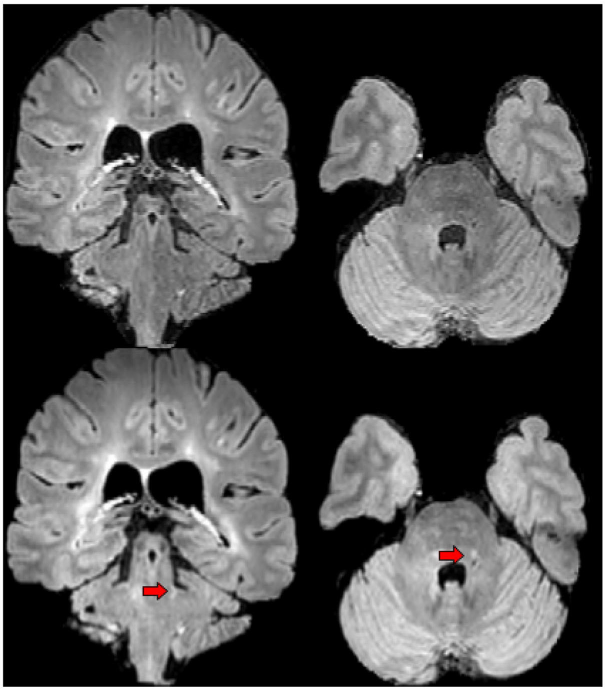


FIGURE 5 | An example of lesion detected by expert 1 in the first phase of the experiment but not by experts 2 and 3. First row shows the baseline FLAIR scan (from left-to-right: coronal and axial view), second row shows the FLAIR scan 1 year later (from left-to-right: coronal and axial view). Red arrows designate the lesion of interest.

At the patient scale, depending on the experts, 19, 19, and 20 patients out of 48 were reported to have at least one sign of MRI disease activity. When combining the different expert segmentations, this number increased to 22. The inter-rater Cohen's Kappa coefficients associated with these patients classifications were: for Expert 1-Expert 2: 0.83 [0.66, 0.99], for Expert 2-Expert 3: 0.78 [0.74, 1], and for Expert 1-Expert 3: 0.96 [0.87, 1]. The overall Fleiss's Kappa coefficient was 0.86 [0.75, 0.97].

Detection of New Lesions With the Segmentation Mask

During the second phase (i.e., when segmentation masks provided by the segmentation module were used as supplemental information), the three experts, respectively, detected 114, 111, and 104 lesions. Overall 125 lesions were detected. **Table 3** reports the difference of lesions detected from one expert to another in this second phase. The overall Fleiss's Kappa coefficient was 0.59 [0.49, 0.69]. **Table 4** details the number of lesions from the segmentation module accepted and rejected by the experts as well as the number of supplemental lesions added. Overall, a large majority of the 121 candidate lesions detected by the segmentation module were accepted by the experts (between 103 and 107 depending on the expert). Eleven lesions out of these 121 were rejected by each of the three experts. After the consensus reading, one supplemental lesion proposed by the

TABLE 3 | Inter-expert disparity during phase 2.

	Expert 1	Expert 2	Expert 3
Expert 1	–	10/114 $\kappa = 0.44$ [0.29, 0.59]	16/114 $\kappa = 0.51$ [0.34, 0.69]
Expert 2	7/111	–	11/111 $\kappa = 0.69$ [0.53, 0.83]
Expert 3	6/104	4/104	–

The first figures give the number of lesions detected by the expert in row not detected by the expert in column. For example, over its 114 detected new lesions, expert 1 detected 10 lesions that were not detected by expert 2 and 16 that were not detected by expert 3. The second figures (only once per expert combination) give the Cohen's Kappa coefficient and its associated 95% CI.

TABLE 4 | Relevance of the segmentation masks produced by the segmentation module.

	Expert 1	Expert 2	Expert 3
Accepted lesions	105	107	103
Rejected lesions	16	14	18
Supplemental lesions	9	4	1

Number of lesions accepted, rejected and added by the different experts when using the segmentation mask as supplemental information.

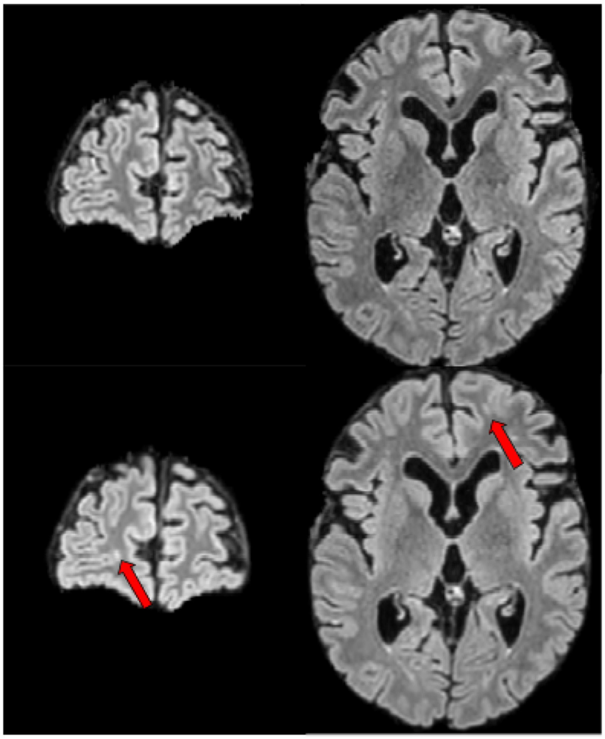


FIGURE 6 | Example of a lesion detected by none of the experts in the first phase of the experiment, detected by the segmentation module and accepted by all experts in the second phase of the experiment. First row shows the baseline FLAIR scan (from left-to-right: coronal and axial view), second row shows the FLAIR scan 1 year later (from left-to-right: coronal and axial view). Red arrows designate the lesion of interest.

segmentation module was rejected, leading to a total of 12 false positive lesions distributed among 8 patients (10% rejection rate) for the segmentation module. At the patient scale, depending on the experts, 24, 23, and 23 patients were reported to have at least one sign of disease activity. When combining the different expert segmentations, this number rises to 25. The inter-rater Cohen's Kappa coefficients associated with these patients classifications were: for Expert 1-Expert 2: 0.96 [0.88, 1], for Expert 2-Expert 3: 0.92 [0.80, 1], and for Expert 1-Expert 3: 0.88 [0.74, 1]. The overall Fleiss's Kappa coefficient was 0.92 [0.83, 1].

Consensus Lesions Reading and Patient Characteristics

Overall, 138 individual lesions were reported by the experts during the two phases. Two of these 138 lesions were then discarded during the concerted reading (one was reported in phase 1 and the other one in phase 2). The patient-wise repartition of lesions is given in **Supplementary Figure 1**. Briefly, the median lesion number was 1, ranging from 0 to 18. Twenty-two patients (about 46%) did not develop new lesions.

Comparison of New Lesions Detection With and Without the Segmentation Mask at the Lesion Scale

By comparing lesions detected in the two phases (and excluding the two false positive lesions), we identified 103 cases of lesions that were not detected by an expert in the first phase but were detected by this expert in the second phase. **Figure 6** displays an example of lesion that was detected by the segmentation module and accepted by the three experts in the second phase of the experiment but that was reported by none of the three experts during the first phase of the experiment. Conversely, we identified only 30 cases of lesions that were first detected by an expert in the first phase but not detected in the second phase.

Table 5 reports the statistics on lesion detection averaged over patients and highlights the added value of the segmentation module to increase expert performance. Similarly, **Table 6** reports increased ability of each expert to detect new lesions using the segmentation module. Finally, **Table 7** reports the statistics on elapsed time for each of two phases for the three experts and highlights the gain in expert processing time brought by the use of the segmentation mask.

Comparison of New Lesions Detection With and Without the Segmentation Mask at the Patient Scale

Table 8 provides a contingency table summarizing the numbers of patients that were identified as having no lesion, one lesion or more than one lesion during the two phases of the experiment. Moreover, 20 patients (by adding those identified by each expert) were wrongly identified as having no new lesion in the first phase, against only 8 patients in the second phase of the experiment. The overall sensitivity at the patient scale (i.e., no new lesion vs. at least one new lesion) was 0.74 in the first phase, and 0.90 in the second phase of the experiment (p-value for unit odds

TABLE 5 | Statistics on lesion detections averaged over patients and differences between phase 1 (lesions detection only using patient acquisitions) and phase 2 (lesions detection using patient acquisitions and segmentation mask produced by the segmentation module).

	Phase 1	Phase 2	Phase 2 to Phase 1 differences value, [95%CI] and *p*-value for no difference
Mean number of detected lesion per patient and per expert	1.8 Lesion	2.3 Lesion	Mean difference = 0.5 [−0.78, −0.23] $p = 5.10^{-4}$
Pooled standard deviation from interexpert variability	0.76 Lesion	0.55 Lesion	Mean difference = 0.09 [−0.02, 0.29] $p = 0.12$

First row: averaged patient-wise number of lesions detected in phase 1 (column 1) and 2 (column 2) and mean difference, associated 95% CI and p-value for a null difference between the two phases (column 3). Second row: pooled inter-expert standard deviation associated with lesions number detected in phase 1 (column 1) and 2 (column 2) and mean difference, associated 95% CI and p-value for a null difference between the two phases (column 3).

TABLE 6 | Ability of each expert to detect a new lesion during the two phases.

	Phase 1 sensitivity [95%CI]	Phase 2 sensitivity [95%CI]	Phase 2 to Phase 1 differences odds ratio, [95%CI] and *p*-value
Expert 1	0.66 [0.58, 0.74]	0.84 [0.76, 0.90]	2.77 [1.55, 5.15] $p = 7.10^{-4}$
Expert 2	0.60 [0.51, 0.68]	0.82 [0.74, 0.88]	3.35 [1.84, 6.29] $p = 8.10^{-5}$
Expert 3	0.61 [0.52, 0.69]	0.75 [0.68, 0.83]	2.31 [1.33, 4.10] $p = 3.10^{-3}$

For each expert (in row): the ratio of new lesions detected over the overall 136 lesions (the overall number of lesions detected on all patients, by all experts during the two phases and confirmed during the concerted reading) as well its 95% CI for phase 1 (first column) and phase 2 (second column) and difference between phase 1 and phase 2 (third column).

TABLE 7 | Statistics on time elapsed for each of two phases for the three expert and comparison between the two phases.

	Phase 1 duration (mean, [range])	Phase 2 duration (mean, [range])	Phase 1 to Phase 2 difference [mean, (sd), *p*-value]
Expert 1	317s [144, 807]	232 s [91, 603]	85 s (137 s) $p = 10^{-5}$
Expert 2	283s [125, 847]	204s [93, 511]	78 s (126 s) $p = 10^{-5}$
Expert 3	272 s [146, 525]	160 s [82, 287]	112 s (79 s) $p = 10^{-13}$

First column: Mean time and associated range associated with the processing of the 48 patients in Phase 1. Second column: same elements for phase 2. Third column: Mean time reduction from Phase 2 to Phase 1, associated standard deviation and p-value for a null time reduction.

ratio = 0.003). Moreover, for each expert and each phase, the patient-wise specificity was equal to 1.

TABLE 8 | Contingency table of numbers of patients reported with no (0), one (1), or more than one (>1) lesion in the two phases of the experiment.

		Phase 2		
		0	1	>1
Phase 1	0	71 (23, 23, 25)	13 (5, 5, 3)	2 (1, 1, 0)
	1	3 (1, 2, 0)	10 (3, 3, 4)	2 (0, 0, 2)
	>1	0 (0, 0, 0)	2 (1, 0, 1)	41 (14, 14, 13)

In each cell, the top figure indicates the overall number of reported patients while the three bottom figures give, respectively, these figures for expert 1, expert 2, and expert 3.

When assessing the reported new lesions as detected by the segmentation module (i.e., with no adjustment by an expert), we computed a sensitivity of 0.90 and specificity of 0.84 at the patient scale.

Impact of the MUSIC Workflow on Routine Clinical Practice

Supplementary Table 1 gives the main elements of reporting for each expert and each patient when using a standard manual examination of data from the clinical PACS (phase 1) and when using the MUSIC workflow (phase 2). In particular, expert 2 and expert 3 reported two patients without activity in phase 1 (patients 2 and 3) while they reported a notable new lesion for these same patients in the second phase. Finally, **Table 9** gives the mean time elapsed by the three experts in the two settings. Mean times elapsed in the two settings differ significantly [mean difference $= 2^{'}45^{''}$ (SD $= 2^{'}00^{''}$), $p = 10^{-4}$].

DISCUSSION

While there is a growing number of methodological works addressing the question of automating the detection of new MS lesions from one acquisition to another using deep learning techniques [e.g., (7, 9, 10)], the integration of such tools in clinical

TABLE 9 | Mean (standard deviation) time elapsed to perform radiological readings of the patients of interest for each of the three experts and each of two phases of the second experiment.

	Expert 1 mean (standard deviation)	**Expert 2 mean (standard deviation)**	**Expert 3 mean (standard deviation)**
Phase 1	4′45″ (1′30″)	7′00″ (3′00)	5′15″ (1′30″)
Phase 2	3′15″ (0′30″)	3′00″ (0′30″)	3′45″ (0′30″)

practice as an aid to clinicians and the associated added-value on the resulting radiological reports have not been fully evaluated. This work aims at providing elements to document these two points. In particular, we described a fully-integrated workflow and showed that the proposed workflow increases MRI reader performance to detect new MS lesions on longitudinal MRI scans while decreasing MRI comparison time. Beyond the number of lesions detected, our workflow has an impact on the number of MS patients classified as stable or active based on their MRI, even by the most experienced neuroradiologist. It may therefore have substantial consequences on the therapeutic management of MS patients.

Visual Detection of New MS Lesions Is a Complex Task

First, as previously reported, we observed a high inter-expert variability in the detection of new FLAIR lesions (24, 25). In practice, a significant part of this variability is not due to differences of MR signal interpretations but related to the difficulty to visually notice them within the whole 3D volumes of interest. Indeed, while we did not investigate the intra-expert variability, a previous study reported a mean intraobserver kappa score for new lesions detection at 0.72 (25). As expected, in the present study, the expert with the highest level of experience (neuroradiologist with 15 years of experience) detected a higher number of new lesions than the other clinicians.

Automated New Lesion Segmentation Tools Provide a Relevant and Valuable Aid for Clinicians

Second, we observed that the use of lesion masks produced by the lesion detection module significantly increases the number of lesions detected regardless of the level of expert's experience (more than 15% more lesions with the MUSIC workflow than without). This observation is in line with recent studies not involving deep learning based segmentation (26–28). In parallel, while not significant, we also observed a natural reduction of the inter-expert variability when using the segmentation masks.

It is also interesting to note that we deliberately put ourselves in difficult conditions by including longitudinal data acquired on different scanners in 56% of the cases. These conditions are representative of the follow-up conditions in clinical practice where patients may be followed in different centers and on different scanners. Moreover, we did not discard lower quality acquisitions from the study. Despite these heterogeneities the segmentation module provides valuable aid to clinicians. In particular, we did not observe evidence of mean differences in sensitivity of the segmentation module depending on whether baseline and follow-up data come from the same scanner or from two different scanners/brands (mean difference = 0.10, p = 0.44 for "same scanner vs. different scanners," mean difference = 0.04, p = 0.77 for "same brand vs. different brands"). This point must however be mitigated by our sample size that may be too low to evidence subtle mean differences. Meanwhile, the rejection rate, i.e., the percentage of candidate lesions detected by the segmentation module that were rejected by the experts was moderate (about 10%) and most segmentation masks (about 80%) did not present any false positive lesions. Overall, this rate must be considered in light of our methodological choices. Indeed, in this work, we chose to accept the presence of a reasonable amount of false positives (favoring sensitivity over specificity). Optimizing the balance between the number of false negative lesions (increasing experts' acceptance and comfort) and the number of true positive lesions (decreasing the probability to miss a new lesion) may consist of interesting future directions. In particular the segmentation module, and especially the post-processing rules that drive most of this balance, could be modified for this purpose. This optimization could also depend on acquisition characteristics (e.g., acquisition signal-to-noise ratio, scanner brand) to reduce potential effect of these factors on performance.

It is worth noting that we do not think these results are intrinsically related to our segmentation module. Indeed, while being built on state-of-the-art solutions and exhibiting satisfying performances, it may be replaced by other recent methods of the literature [e.g., (7–10)]. Our aim is not to show the superiority of our segmentation module but to evidence the potential impact of using state-of-the-art segmentation methods on MS clinical practice.

Using a New Lesion Segmentation Mask Was Well-Received by the Experts

Importantly, all three experts reported a satisfying and comfortable reading experience when using the segmentation mask as an aid, especially with the full workflow (Experiment 2). Additionally, for each of them, the time spent to analyze the images was significantly reduced in the second phase of the experiments.

More specifically, the three experts were satisfied by the information provided by the segmentation masks and reported that the segmentation module offered very good performances. While this result is satisfactory, it also raises issues related to the confidence to place in these segmentation masks, especially regarding their potential lack of sensitivity. As an example, in the second phase of our first experiment, expert 3 only added 1 supplemental lesion to those proposed by the segmentation module, while being the expert exhibiting the highest gain of processing time between Phase 1 and Phase 2. While on average, the performances of expert 3 were notably superior with the

MUSIC workflow than without, this observation also suggests a risk for the experts to place too much trust into the automated outputs. We think that ways to mitigate such risk, such as confidence intervals or uncertainty estimation (10, 29), have to be considered in future methodological developments.

Automated New Lesion Segmentation Tools May Have Substantial Consequences on the Therapeutic Management of MS Patients

Beyond lesion-wise statistics, our results suggest that the use of segmentation masks has also consequences at the patient level. Indeed, in the first experiment, it allows each expert to identify three to five supplementary active MS patients (i.e., with at least one new lesion on MRI). This result, which may seem important, should be interpreted in relation to our dataset, including particularly active patients. This high activity rate is well-explained by our inclusion criteria, selecting patients at the early stage of the disease.

Moreover, it would be interesting to evaluate the consequences on patient management by the clinician (in particular with respect to potential treatment changes). Indeed, the appearance of new lesions under treatment is recognized as being prognostic of an increased risk of clinical relapse and of disability progression. It consequently often leads to a change of treatment in clinical practice (30, 31). This point could be evaluated in a future study including the neurologists in charge of these patients. In the longer term, the objectives would be to evaluate the impact of such a tool on the evolution of disability in patients and on the costs of managing the disease. Finally, it would be interesting to evaluate this workflow from the patient's point of view. There is indeed a potential added-value of a straightforward visualization enhancing new lesions to facilitate the clinician-patient dialogue, especially to argue for a change of DMT.

Limits and Perspectives

Our study has several limitations that need to be discussed. First, our evaluation must be interpreted in light of our population, which exhibits a high prevalence of new lesions due to our inclusion criteria. Indeed, we voluntarily put ourselves in a setting where the inter-expert and intra-expert variabilities are exacerbated and, as a consequence, where a computer-guided aid is likely to offer a high added-value. If the number of active patients had been lower, we can reasonably assume that average expert performances without the computer-guided aid would have been better and that the resulting added-value of our workflow would have been less pronounced.

Secondly, all FLAIR, T2-w, and T1-w images were used as input to the automatic lesion detection module. These 3 sequences correspond to those currently recommended in the OFSEP protocol in France (6) and are mostly performed in clinical routine for the follow-up of MS patients. However, in some cases, due to time constraints, some of these sequences are not acquired. Our segmentation module therefore needs to be adapted and evaluated to deal with this configuration.

Thirdly, while it is consistent with that used in other studies (26, 32), our evaluation sample size (54 patients) is limited. It will be interesting to evaluate our workflow and confirm our results on a larger sample from all centers involved in the follow-up of MS patients in our region. Moreover, some MRI scanners are under-represented in our sample (as GE scanners) and the size of our cohort did not allow us to analyze the performance of our tool by subgroup, e.g., according to the type of MRI scanners used. Despite these limitations, overall, the added value of our segmentation module compared to a standard radiological reading appears clearly significant, both on the number of lesions detected and on the time to perform this task.

Fourth, the fact that all readings were firstly performed without assistance (phase 1) and secondly using the segmentation mask as an aid (phase 2) may have introduced a bias that would have been reduced by using a dedicated design. However, we are confident about the lack of such substantial bias. Indeed, a 2-week period was included between the two phases and this period consisted, for each expert, of a dense clinical and radiological activity. Moreover, the number of data analyzed was consequent and the order of analysis of the patients was different between the two readings.

Fifth, our segmentation module could be improved following recent methodological advances. In particular, a two-path encoder that extracts hierarchical features for each time-point separately, while allowing for an exchange of information at certain levels of abstraction, might be explored in the future (33). In parallel, the design of methods using both a joint analysis of the baseline and follow-up acquisitions (as in the present work) and an analysis of each cross sectional segmentation probability maps, obtained from dedicated algorithms, could maximize the use of the information available in the different annotated databases (34, 35). In particular, these latter segmentation probability maps could be obtained estimating the confidence maps associated with the presence of lesions (10), that have already shown their interests to detect new MS lesions.

Sixth, our experiments were limited to follow-up with two time points and did not include settings with more time points. Our workflow can actually deal with such settings by processing the data sequentially, using the first baseline images as reference target for registration, and performing segmentations independently for each consecutive pair of acquisition sessions.

Finally, in our study, we mainly evaluated our workflow at the lesion scale. Evaluating the impact of such workflow at the patient scale, and in particular its consequences on patient's management (continuation or change of treatment, effect on disability progression for example) is a final objective that we did not fully address in this study and constitutes the future directions of our work.

CONCLUSION

The workflow proposed in this paper consists of a fully-integrated and user-friendly computer-aided MRI reading system, potentially accessible to all neurologists and radiologists in a given area. Importantly, the aid provided by the segmentation module significantly improved both the number of new FLAIR lesions detected by MRI-readers, including highly experienced ones, the number of patients classified as having active disease, and the time spent interpreting follow-up MRIs. These results should make us think about how to widely disseminate such workflows, to allow an optimized follow-up for all MS patients wherever they are followed and whatever the level of expertise of their clinicians.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by CPP Ille-de-France VI (POCADIMS protocol). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

BC, AK, GP, FG, J-CF, and EC: writing paper. AK, GE, J-CF, OC, PL'H, EC, and GE: designing framework. BC, GP, BL, FG, OC, and EC: designing framework modules. BC, AK, J-CF, GP, and GE: designing experiments. BC and AK: analysis experiments. AK, J-CF, NE, and RC: expert segmentation. All authors contributed to the article and approved the submitted version.

FUNDING

This study received funding from Biogen, Teva Santé, Novartis Pharma, Roche SAS, Merck Serono, Sanofi. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication. As a part of the MUSIC project, the academic POCADIMS study (NCT03205280) was funded by the Programme Hospitalier de Recherche Clinique-Interregional (PHRC-I) 2016, Groupement Interrégional de Recherche Clinique et d'Innovation Grand Ouest. The study was supported by a grant from the Institut des Neurosciences Cliniques de Rennes (INCR).

ACKNOWLEDGMENTS

Authors thank the Observatoire Français de la Sclérose en Plaques (OFSEP) which was supported by a grant provided by the French State and handled by the Agence Nationale de la Recherche within the framework of the Investments for the Future program, under the reference ANR-10-COHO-002, by the Eugène Devic EDMUS Foundation against multiple sclerosis and by the ARSEP Foundation. Authors thank Damien Le Port for his help to gather and document the dataset used in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.740248/full#supplementary-material

Supplementary Figure 1 | Number of patients for each new lesion count (from 0 to 18).

Supplementary Video | The CADIMS viewer in action: an example of data visualization in the CADIMS viewer.

REFERENCES

1. McGinley MP, Goldschmidt CH, Rae-Grant AD. Diagnosis and treatment of multiple sclerosis: a review. *JAMA*. (2021) 325:765–79. doi: 10.1001/jama.2020.26858
2. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol*. (2018) 17:162–73. doi: 10.1016/S1474-4422(17)30470-2
3. Fisniku LK, Brex PA, Altmann DR, Miszkiel KA, Benton CE, Lanyon R, et al. Disability and T2 MRI lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis. *Brain*. (2008) 131(Pt 3):808–17. doi: 10.1093/brain/awm329
4. Rotstein DL, Healy BC, Malik MT, Chitnis T, Weiner HL. Evaluation of no evidence of disease activity in a 7-year longitudinal multiple sclerosis cohort. *JAMA Neurol*. (2015) 72:152–8. doi: 10.1001/jamaneurol.2014.3537
5. Wattjes MP, Ciccarelli O, Reich DS, Banwell B, de Stefano N, Enzinger C, et al. 2021 MAGNIMS-CMSC-NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis. *Lancet Neurol*. (2021) 20:653–70. doi: 10.1016/S1474-4422(21)00095-8
6. Brisset J-C, Kremer S, Hannoun S, Bonneville F, Durand-Dubief F, Tourdias T, et al. New OFSEP recommendations for MRI assessment of multiple sclerosis patients: special consideration for gadolinium deposition and frequent acquisitions. *J Neuroradiol*. (2020) 47:250–8. doi: 10.1016/j.neurad.2020.01.083
7. Krüger J, Opfer R, Gessert N, Ostwaldt A-C, Manogaran P, Kitzler HH, et al. Fully automated longitudinal segmentation of new or enlarged multiple sclerosis lesions using 3D convolutional neural networks. *Neuroimage Clin*. (2020) 28:102445. doi: 10.1016/j.nicl.2020.102445
8. Lladó X, Ganiler O, Oliver A, Martí R, Freixenet J, Valls L, et al. Automated detection of multiple sclerosis lesions in serial brain MRI. *Neuroradiology*. (2012) 54:787–807. doi: 10.1007/s00234-011-0992-6
9. Salem M, Valverde S, Cabezas M, Pareto D, Oliver A, Salvi J, et al. A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. *Neuroimage Clin*. (2020) 25:102149. doi: 10.1016/j.nicl.2019.102149
10. McKinley R, Wepfer R, Grunder L, Aschwanden F, Fischer T, Friedli C, et al. Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence. *Neuroimage*. (2020) 25:102104. doi: 10.1016/j.nicl.2019.102104
11. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol*. (2021) 31:3797–804. doi: 10.1007/s00330-021-07892-z
12. Zeng C, Gu L, Liu Z, Zhao S. Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI. *Front Neuroinform*. (2020) 14:610967. doi: 10.3389/fninf.2020.610967
13. Salem M, Cabezas M, Valverde S, Pareto D, Oliver A, Salvi J, et al. A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. *Neuroimage Clin*. (2018) 17:607–15. doi: 10.1016/j.nicl.2017.11.015
14. Jain S, Ribbens A, Sima DM, Cambron M, De Keyser J, Wang C, et al. Two time point MS lesion segmentation in brain MRI: an expectation-maximization framework.

Front Neurosci. (2016) 10:576. doi: 10.3389/fnins.2016.00576

15. Cabezas M, Corral JF, Oliver A, Díez Y, Tintoré M, Auger C, et al. Improved automatic detection of new T2 lesions in multiple sclerosis using deformation fields. *AJNR Am J Neuroradiol*. (2016) 37:1816–23. doi: 10.3174/ajnr.A4829
16. Ganiler O, Oliver A, Diez Y, Freixenet J, Vilanova JC, Beltran B, et al. A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology*. (2014) 56:363–74. doi: 10.1007/s00234-014-1343-1
17. Sweeney EM, Shinohara RT, Shea CD, Reich DS, Crainiceanu CM. Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI. *AJNR Am J Neuroradiol*. (2013) 34:68–73. doi: 10.3174/ajnr.A3172
18. Fartaria MJ, Kober T, Granziera C, Bach Cuadra M. Longitudinal analysis of white matter and cortical lesions in multiple sclerosis. *Neuroimage Clin*. (2019) 23:101938. doi: 10.1016/j.nicl.2019.101938
19. Commowick O, Wiest-Daesslé N, Prima S. Block-matching strategies for rigid registration of multimodal medical images. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. Barcelona: IEEE (2012). p. 700–3. doi: 10.1109/ISBI.2012.6235644
20. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. (2010) 29:1310–20. doi: 10.1109/TMI.2010.2046908
21. Nyúl LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. *IEEE Trans Med Imaging*. (2000) 19:143–50. doi: 10.1109/42.836373
22. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. (2021) 18:203–11. doi: 10.1038/s41592-020-01008-z
23. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Munich: Springer International Publishing (2015). p. 234–41. doi: 10.1007/978-3-319-24574-4_28
24. Erbayat Altay E, Fisher E, Jones SE, Hara-Cleaver C, Lee J-C, Rudick RA. Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. *JAMA Neurol*. (2013) 70:338–44. doi: 10.1001/2013.jamaneurol.211
25. Molyneux PD, Miller DH, Filippi M, Yousry TA, Radü EW, Adèr HJ, et al. Visual analysis of serial T2-weighted MRI in multiple sclerosis: intra- and interobserver reproducibility. *Neuroradiology*. (1999) 41:882–8. doi: 10.1007/s002340050860
26. Bilello M, Arkuszewski M, Nucifora P, Nasrallah I, Melhem ER, Cirillo L, et al. Multiple sclerosis: identification of temporal changes in brain lesions with computer-assisted detection software. *Neuroradiol J*. (2013) 26:143–50. doi: 10.1177/197140091302600202
27. Galletto Pregliasco A, Collin A, Guéguen A, Metten MA, Aboab J, Deschamps R, et al. Improved detection of new MS lesions during follow-up using an automated MR coregistration-fusion method. *AJNR Am J Neuroradiol*. (2018) 39:1226–32. doi: 10.3174/ajnr.A5690
28. Dahan A, Pereira R, Malpas CB, Kalincik T, Gaillard F. PACS integration of semiautomated imaging software improves Day-to-Day MS disease activity detection. *AJNR Am J Neuroradiol*. (2019) 40:1624–9. doi: 10.3174/ajnr.A6195
29. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf Fusion*. (2021) 76:243–97. doi: 10.1016/j.inffus.2021.05.008
30. Sormani MP, Rio J, Tintorè M, Signori A, Li D, Cornelisse P, et al. Scoring treatment response in patients with relapsing multiple sclerosis. *Mult Scler*. (2013) 19:605–12. doi: 10.1177/1352458512460605
31. Gasperini C, Prosperini L, Tintoré M, Sormani MP, Filippi M, Rio J, et al. Unraveling treatment response in multiple sclerosis: a clinical and MRI challenge. *Neurology*. (2019) 92:180–92. doi: 10.1212/WNL.0000000000006810
32. Zopfs D, Laukamp KR, Paquet S, Lennartz S, Pinto Dos Santos D, Kabbasch C, et al. Follow-up MRI in multiple sclerosis patients: automated co-registration and lesion color-coding improves diagnostic accuracy and reduces reading time. *Eur Radiol*. (2019) 29:7047–54. doi: 10.1007/s00330-019-06273-x
33. Gessert N, Krüger J, Opfer R, Ostwaldt A-C, Manogaran P, Kitzler HH, et al. Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs. *Comput Med Imaging Graph*. (2020) 84:101772. doi: 10.1016/j.compmedimag.2020.101772
34. Carass A, Roy S, Jog A, Cuzzocreo JL, Magrath E, Gherman A, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *Neuroimage*. (2017) 148:77–102. doi: 10.1016/j.neuroimage.2016.12.064
35. Commowick O, Istace A, Kain M, Laurent B, Leray F, Simon M, et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci Rep*. (2018) 8:13650. doi: 10.1038/s41598-018-31911-7

Conflict of Interest: GP and EC were employed by b.com. PL'H was employed by Biotrial.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare that this study received funding from Biogen, Teva Santé, Novartis Pharma, Roche SAS, Merck Serono and Sanofi. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Medicine

ORIGINAL RESEARCH
published: 25 November 2021
doi: 10.3389/fmed.2021.681183

Segmentation of Cerebral Small Vessel Diseases-White Matter Hyperintensities Based on a Deep Learning System

Wei Shan[1,2,3†], *Yunyun Duan*[1†], *Yu Zheng*[2], *Zhenzhou Wu*[2], *Shang Wei Chan*[2], *Qun Wang*[1,2,3], *Peiyi Gao*[2], *Yaou Liu*[2*], *Kunlun He*[4,5*] and *Yongjun Wang*[1,2*]

[1] Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China, [2] National Center for Clinical Medicine of Neurological Diseases, Beijing, China, [3] Beijing Institute for Brain Disorders, Beijing, China, [4] Laboratory of Translational Medicine, Chinese PLA General Hospital, Beijing, China, [5] Key Laboratory of Ministry of Industry and Information Technology of Biomedical Engineering and Translational Medicine, Chinese PLA General Hospital, Beijing, China

OPEN ACCESS

Edited by:
Raphael Meier,
Bern University Hospital, Switzerland

Reviewed by:
Wonsang You,
Sun Moon University, South Korea
Chandana Nagaraj,
National Institute of Mental Health and Neurosciences (NIMHANS), India

****Correspondence:***
Kunlun He
kunlunhe@plagh.org
Yongjun Wang
weishantiantan@gmail.com
Yaou Liu
yaou08@163.com

† These authors share first authorship

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *16 March 2021*
Accepted: *28 September 2021*
Published: *25 November 2021*

Citation:
Shan W, Duan Y, Zheng Y, Wu Z, Chan SW, Wang Q, Gao P, Liu Y, He K and Wang Y (2021) Segmentation of Cerebral Small Vessel Diseases-White Matter Hyperintensities Based on a Deep Learning System. Front. Med. 8:681183. doi: 10.3389/fmed.2021.681183

Objective: Reliable quantification of white matter hyperintensities (WHMs) resulting from cerebral small vessel diseases (CSVD) is essential for understanding their clinical impact. We aim to develop and clinically validate a deep learning system for automatic segmentation of CSVD-WMH from fluid-attenuated inversion recovery (FLAIR) imaging using large multicenter data.

Method: A FLAIR imaging dataset of 1,156 patients diagnosed with CSVD associated WMH (median age, 54 years; 653 males) obtained between September 2018 and September 2019 from Beijing Tiantan Hospital was retrospectively analyzed in this study. Locations of CSVD-WMH on the FLAIR scans were manually marked by two experienced neurologists. Using the manually labeled data of 996 patients (development set), a U-shaped novel 2D convolutional neural network (CNN) architecture was trained for automatic segmentation of CSVD-WMH. The segmentation performance of the network was evaluated with per pixel and lesion level dice scores using an independent internal test set ($n = 160$) and a multi-center external test set ($n = 90$, three medical centers). The clinical suitability of the segmentation results, classified as acceptable, acceptable with minor revision, acceptable with major revision, and not acceptable, was analyzed by three independent neuroradiologists. The inter-neuroradiologists agreement rate was assessed by the Kendall-W test.

Results: On the internal and external test sets, the proposed CNN architecture achieved per pixel and lesion level dice scores of 0.72 (external test set), and they were significantly better than the state-of-the-art deep learning architectures proposed for WMH segmentation. In the clinical evaluation, neuroradiologists observed the segmentation results for 95% of the patients were acceptable or acceptable with a minor revision.

Conclusions: A deep learning system can be used for automated, objective, and clinically meaningful segmentation of CSVD-WMH with high accuracy.

Keywords: masking white matter hyperintensities, deep learning, neural network, segmentation, clinical evaluation

INTRODUCTION

White matter accounts for approximately half of the adult cerebral hemisphere volume, and it primarily contains myelinated axons that connect various gray matter areas of the cerebral cortex and subcortical regions with each other (1). White matter lesions damage this connectivity, leading to an interruption in communication between different functional areas, which ultimately manifests in a form of various neurobehavioral disorders (1, 2).

White matter hyperintensity (WMH), or leukoaraiosis are characteristic lesions of the white matter that appear as hyperintense regions on the fluid-attenuated inversion recovery (FLAIR) magnetic resonance images (MRI) (3–6). Clinically, WMHs can be caused by many conditions, such as plaque accumulation in the white matter small vessels, small vessel inflammation, toxicity after medication use, genetic white matter diseases, infections, demyelinating diseases, metabolic diseases, tumors, brain trauma, and persistent chronic damage in white matter small vessels (4). Matsue and others considered that these imaging findings correspond to a series of histological changes. For example, histological analysis revealed that the ventricle's high signal corresponded to a pale myelin sheath, perivascular proliferation, a discontinuous inner layer of ependyma, and increased subependymal glia. The hyperintensity in the deep and subcortical white matter has been primarily observed as a result of the pale myelin sheath and perivascular hyperplasia. Perivascular hyperplasia has been mainly found in the frontal and/or apical subcortical white matter (4, 7–13). The diameter of hyperplastic areas was usually <3 mm and had an obvious boundary. The diffused white matter lesions (WMLs) in Binswanger's disease are characterized by a pale myelin sheath and tissue thinning due to the loss of myelin sheaths and axons. All of the above WMLs show different degrees of arteriosclerosis (12, 13).

Although WMLs are closely related to cerebrovascular diseases and vascular risk factors, their pathogenesis remains largely unclear and they can be caused by multiple factors (14). WMHs have been observed to be the main manifestation of cerebral small vessel disease (SVD) and they are important factors in the indication of stroke, dementia, and aging (7–13). Additionally, WMHs have been observed to be prevalent in aged people (15).

At present, the Age-related White Matter Changes (ARWMC), Fazekas, modified Scholten's, and Ylikoski scales are widely used in clinical practice (16–18). Existing quantitative methods are time-consuming, laborious, and subjective. Currently, deep convolutional neural networks (CNNs) have been shown to be useful and effective in medical applications. Thus, a highly accurate system for automatic segmentation of WMH aid neuroradiologists in timely quantitative assessment of WMH and significantly reduce the time required for diagnoses (4, 19–22).

In this work, we propose a deep learning system (DLS) for efficient, objective, and automatic prediction of WMH from the FLAIR images. We compare the proposed DLS with the state-of-the-art deep learning architectures and validate its performance using two independent multi-center test datasets. Finally, to analyze the clinical utility of the proposed DLS and check its acceptance by clinicians, we perform a qualitative analysis whereby three clinical neuroradiologists assess the accuracy and quality of the WMH segmentation on four levels, viz: acceptable, acceptable with minor revision, acceptable with major revision, and not acceptable.

MATERIALS AND METHODS

The study was approved by the Ethics Committee of the Beijing Tiantan Hospital in accordance with the Helsinki Declaration. Written informed consent from the participants was not required for participation in this study.

Study Design and Participants

This study retrospectively analyzed the data from 1,156 patients diagnosed with the CSVD associated WMH admitted to the Beijing Tiantan Hospital between September 2018 and September 2019. The patients with a mention of WMH in their electronic health records (EHRs) were reviewed by clinicians for the presence of WMH and the patients with confirmed WMH were included in this analysis. Patients with poor FLAIR image quality were excluded from the analysis. The included patients were randomly divided into a development dataset ($n = 996$, ∼85% of the data) and an independent internal test dataset ($n = 160$). Furthermore, for external validation of the segmentation performance, 90 randomly selected patients with clinically diagnosed WMH from the Third China National Stroke Registry (CNSR-III) study were included in the analysis as an external test dataset.

Data Distribution

MRI Acquisition

All the patients were reviewed for the availability of good quality FLAIR images. The scans were acquired from multiple different scanners with a field strength of either 1.5T or 3T according to the clinically used FLAIR collection protocol. The analyzed images had an axial thickness between 0.55 and 1.2 mm and the sagittal and coronal view spacings were between 0.43 and 0.9 (equal along both the planes).

Manual Annotation of the WMH

In total, we included 34,228 T2-FLAIR images from 1,156 patients from Beijing Tiantan Hospital with labeled segmented WMHs. In this data set, we labeled 12,087 small leukoencephalopathies (<20 plex*spacing), 14,759 medium leukoencephalopathies (between 20 and 150 plex*spacing) and 4,003 large leukoencephalopathies (over 150 plex*spacing).

For clinical evaluation data set included 90 patients' T2-FLAIR images from three other hospitals across China, which were included in The Third China National Stroke Registry (CNSR-III). Additional detailed information about the lesion sizes can be found in **Table 1**. Each volumetric MRI had a vertical spacing between 0.55 and 1.2 mm. For each image, the spacing along the x- and y-directions varied from 0.43*0.43 to 0.9*0.9 mm^2

TABLE 1 | Data distribution in the manuscript.

		Positive number
DLS development	Traning set	870 patients
	Validation set	126 patients
	Inner test set[#]	160 patients
	Summary	1,156 patients
Clinical evaluation	Test data set	90 patients
	Summary	1,246 patients

[#] *inner test set used for the code optimizzatio program only.*
Test data set used for the clinical evaluation only.

TABLE 2 | Validation Test 1.

Data set	Lesion size	Percentage correctly labeled (*n*,%)	Dice
Data set 1			
Small	<20 (plex*spacing)	462 (64.71%)	
Medium	20 ~ 150 (plex*spacing)	(80.09%)	
Large	>150 (plex*spacing)	39 (96.12%)	0.722
Data set 2			
Small	<20 (plex*spacing)	601 (68.37%)	
Medium	20 ~ 150 (plex*spacing)	909 (82.86%)	
Large	>150 (plex*spacing)	325 (96.73%)	0.776
Date set 3			
Small	<20 (plex*spacing)	361 (50.14%)	
Medium	20 ~ 150 (plex*spacing)	425 (68.77%)	
Large	>150 (plex*spacing)	234 (92.49%)	0.722

between consecutive pixels. The distribution of pixel spacings for each data set is shown in **Table 2**.

Development of Deep Learning System for WMH Segmentation

For automatic segmentation of the WMH, we developed a deep learning system using the data from the training dataset along with manual annotations (**Figure 1**). The deep learning system consisted of a four layered modified U-Net architecture which is presented in Figures. The architecture was trained using 996 patients' data from the development dataset. The model was designed to predict a 2D lesion mask using 2D axial slices of FLAIR images. The FLAIR images were first preprocessed by scaling the global (3D) image intensities to follow a standard normal distribution (mean of 0, and standard deviation of 1). Next, images were zero-padded to obtain square-shaped images in the axial plane. The images were next transformed to have uniform axial dimensions of 384 × 384 pixels either using bilinear interpolation (for images with dimensions smaller than 384 × 384 pixels) or using the center crop technique (for images with dimensions larger than 384 × 384 pixels). The decision to center crop the larger images was taken to preserve the spatial resolution of the image which was observed to crucial in the detection of small lesions.

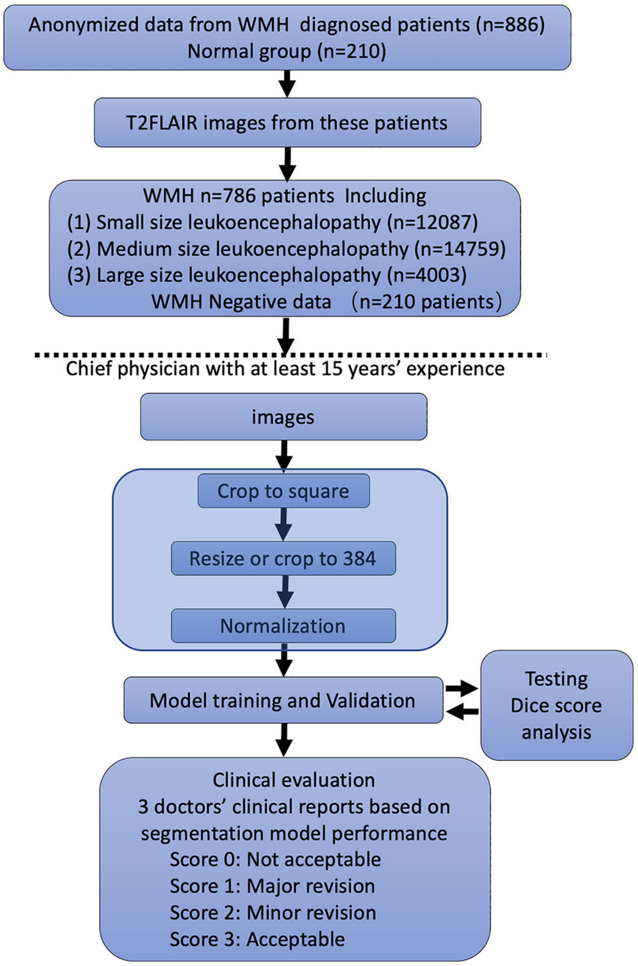

Anonymized data from WMH diagnosed patients (n=886)
Normal group (n=210)
T2FLAIR images from these patients
WMH n=786 patients Including
(1) Small size leukoencephalopathy (n=12087)
(2) Medium size leukoencephalopathy (n=14759)
(3) Large size leukoencephalopathy (n=4003)
WMH Negative data (n=210 patients)
Chief physician with at least 15 years' experience
images
Crop to square
Resize or crop to 384
Normalization
Model training and Validation
Testing
Dice score
analysis
Clinical evaluation
3 doctors' clinical reports based on
segmentation model performance
Score 0: Not acceptable
Score 1: Major revision
Score 2: Minor revision
Score 3: Acceptable

FIGURE 1 | Flowchart of the distribution of patients in the training and clinical evaluation steps. The distribution and classification of all samples in each step was used for the model training and clinical evaluation steps.

The model was trained using the above preprocessed 2D axial slices of T2 Flair scans (input shape: 384 × 384 × 1) with an Adam optimizer for 200 epochs using a cross-entropy loss and a batch size of 32. The initial learning rate was set to 3×10^{-4}. To increase the generalizability of the model, data augmentation strategies including vertical flip, horizontal flip, rotation, contrast enhancement, scaling, translation, and addition of Gaussian noise were randomly applied to the images during the training process. The learning rate was modulated based on the dice score on the reserved validation set (n = 126 patients from the development dataset). The learning rate was reduced by 10% if the validation set dice score did not improve for 30 consecutive epochs. To avoid model overfitting, the training was stopped if the validation set dice did not improve for 60 consecutive epochs. After the completion of training, the model with the highest dice score on the validation set was selected as a final segmentation model. This model was then used for automatic segmentation of WMH in the external and internal test datasets. The complete 3D WMH mask for each patient was computed by concatenating the 2D WMH masks from all the axial slices.

Performance Evaluation of the DLS

The segmentation performance of the proposed DLS was assessed using per-pixel precision, recall, dice score, and accuracy. The precision was defined as the total number of correctly predicted WMH pixels divided by the total number of pixels predicted to be of WMH. The recall was defined as the total number of correctly predicted WMH pixels divided by the total number of WMH pixels in the ground truth segmentation. The dice score was calculated as 2*precision*recall/(precision + recall). Also, based on the precision and recall, the receiver operating characteristics (ROC) curves were constructed and the area under the ROC was calculated. All these metrics were calculated for each patient and final results on the entire dataset were calculated as the arithmetic mean of the per-patient value. Also, the dice score was independently calculated for small, medium, and large lesions.

Also, the dice score is biased toward the correct prediction of large WMH and by correctly segmenting one large WMH the model can have a high dice score despite it missing multiple small WMH. Therefore, considering the importance of correct segmentation of small WMH, we also employed a lesion-wise precision, recall, dice score as a performance measure. In the lesion-wise analysis, a lesion was said to be correctly identified if at least 40% of the lesioned pixels were correctly marked by the prediction model. In this manner, by counting the correctly identified lesions, and missed lesions, the lesion dice score, precision, and recall were calculated.

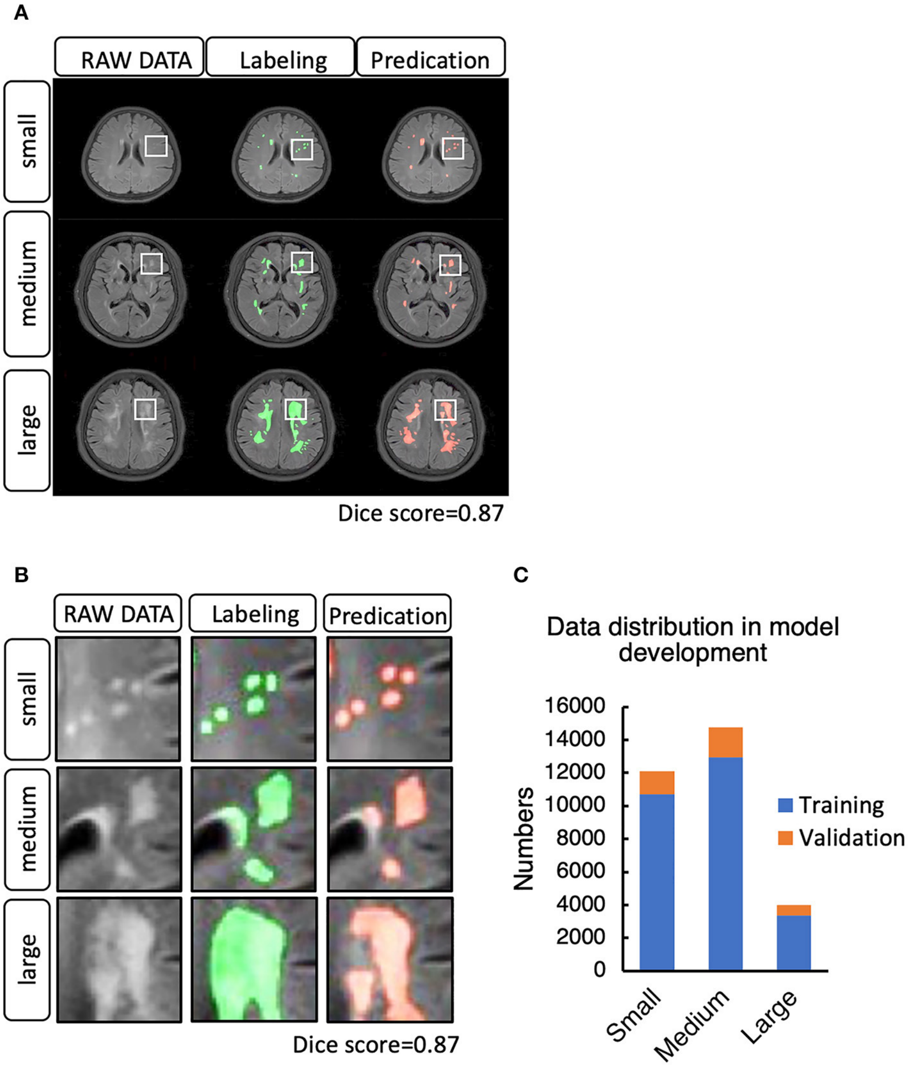


FIGURE 2 | (A) Example cases of white matter hyperintensities (WMHs) labeled manually and by the DLS system. **(B)** WMH lesion distribution in the training and validation step. **(C)** Data distribution in the model development for training and validation.

Lastly, using the above evaluation metrics, we compared the segmentation performance of the proposed DLS with the state-of-the-art WMH segmentation architectures named U-Resnet and 3D-unet. The architectures were constructed according to the best settings proposed by the respective authors and were trained using the same training data as that of the proposed DLS.

Clinical Evaluation of the Proposed DLS

To analyze the clinical utility of the proposed DLS and assess its acceptance by clinicians, we performed a qualitative clinical analysis. In this analysis, three expert neuroradiologists with more than 7 years of experience independently assessed the WMH segmentation results of the proposed DLS for the 90 patients from the external test set. Each neurologist was instructed to rate the segmentation quality of the proposed DLS into four grades, with each of them being defined as:

Grade I (perfectly acceptable, score 4): no missed lesions and <5% mismatch between the predicted and the ground-truth lesions.

Grade II (acceptable with minor revision, score 3): small lesions: 1-4 missed lesions and <10% mismatch for predicted lesions; medium lesions: <2 missed lesions and <5% mismatch; large lesions: no missed lesions.

Grade III (acceptable with major revision, score 2): small lesions: more than four missed lesions and <50% mismatch; medium lesions: more than two missed lesions. Large lesions: more than 30% mismatch.

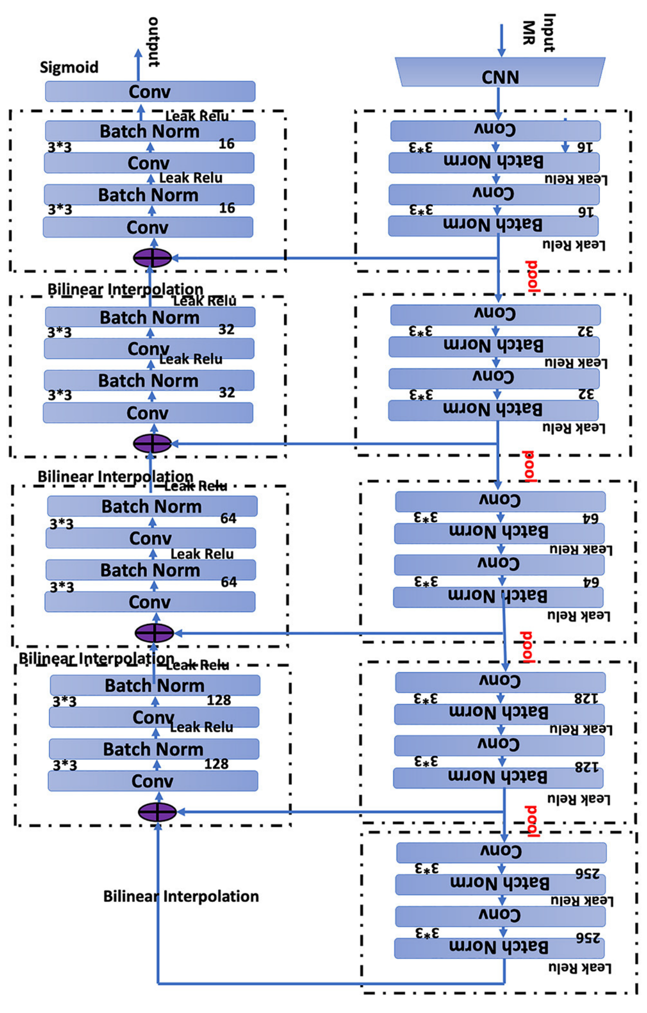


FIGURE 3 | Network architecture of the proposed two-dimensional (2D) convolutional neural network (CNN). The network has 19 layers integrating nine Convolution blocks. Bilinear interpolating arrows indicate up sampling operations to make predictions for the segmentation task. The pool arrow indicates the down sampling operation to gradually increasing the receptive field for the segmentation task. Concatenate connections are used to fuse Multi-scale features in the network. Batch normalization is a linear transformation of the features performed to reduce the covariance shift, thus speeding up the training procedure. Convolution bars indicate the convolution operation, which computes the features. The number 16, 32, 64, 128, 256 indicates the number of channels in that layer, and 3·3·3·3·3·3 denotes the size of the 2D CNN kernels.

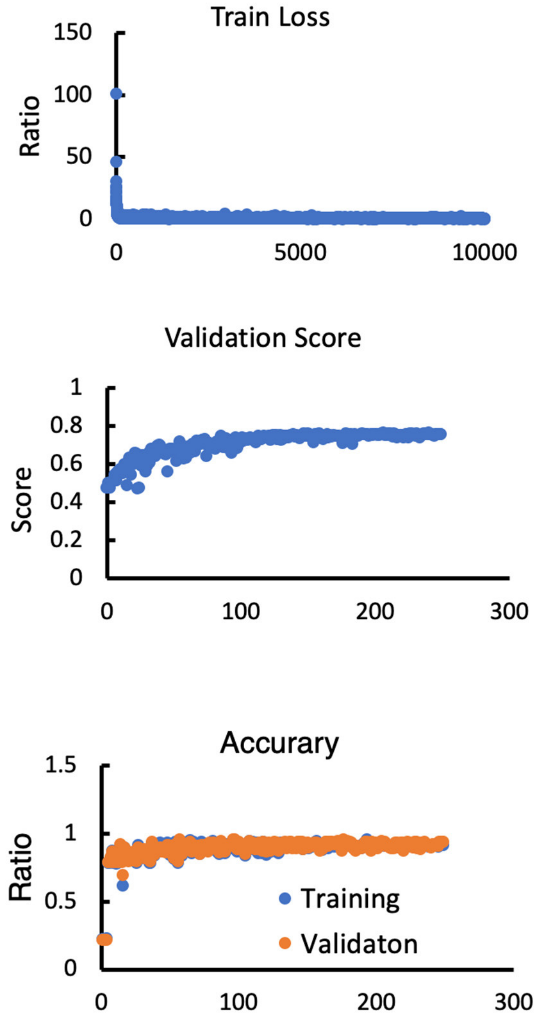


FIGURE 4 | Model performance in terms of the training loss, validation score, training accuracy and validation accuracy.

TABLE 3 | Models head-to head analysis (Data set 2)/Correct labled ratio.

Models	Small lesions (879)	Medium lesions (1,097)	Large lesions (336)
U-Resnet	551, 62.68%	863, 78.66%	321, 95.54%
3D-unet	365, 41.52%	778, 70.92%	328, 97.62%
Our model	601, 68.37%	909, 82.86%	325, 96.73%

TABLE 4 | Models head-to head analysis.

	Our model	Our model no preprocess	U-Resnet	3D-unet
ACC.	0.97	0.906	0.97	0.93
Sensitivity	0.7244	0.5706	0.6024	0.6499
Specificity	0.9989	0.9998	0.9998	0.9997
AUC	0.9959	0.9944	0.9958	0.9896

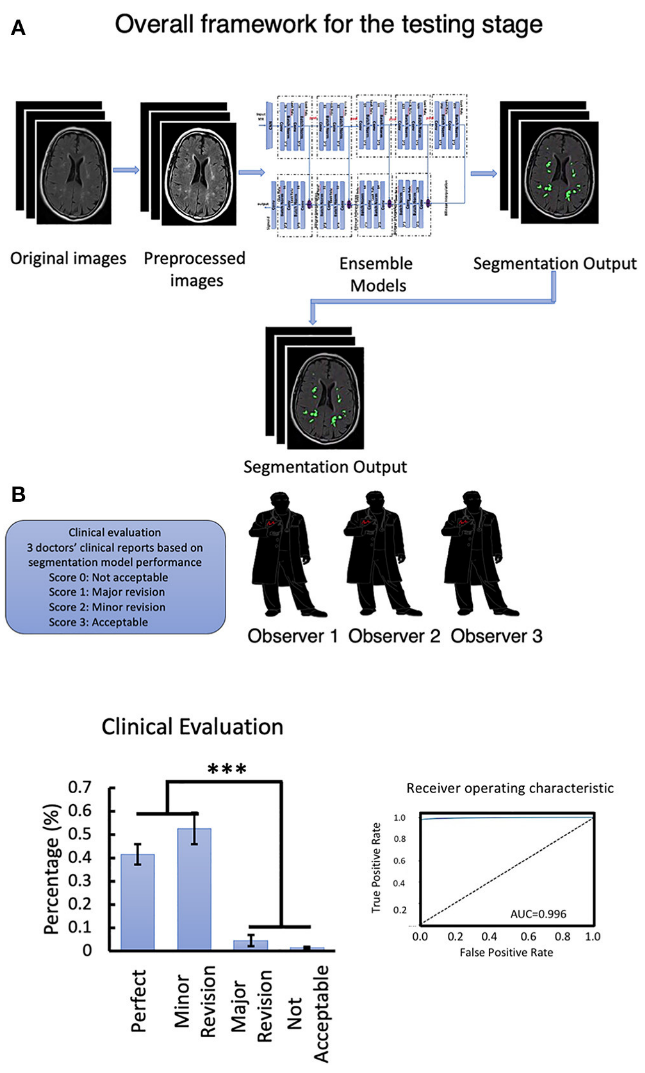


FIGURE 5 | (A) Overall framework for the testing stage. **(B)** Clinical evaluation of the testing data set and Segmentation model ROC-curve and AUC score analysis. Number of neuroradiologists are 3. $^{***}P < 0.001$.

Grade III (not acceptable, score 1): small lesions: more than eight missed and more than 50% mismatch; medium lesions: more than two missed; Large lesions: more than 30% missed and more than 30% mismatch.

Statistical Analysis

The inter-radiologist agreement rate and the Kendall W statistic were calculated for each validation using SPSS software (version 20.0). One-way ANOVA with *post hoc* Tukey's test was applied to assess the differences between each group. Statistical significance was considered at $p < 0.05$. ROC curve and AUC score are performed for the segmentation analysis (https://www.kaggle.com/kmader/use-roc-curves-to-evaluate-segmentation-methods).

RESULTS

Baseline Imaging Characteristics

The FLAIR images from the 1,156 patients contained a total of 34,228 2D axial slices. In these slices, following manual annotations, a total of 12,087 small, 14,759 medium, and 4,003 large WMH lesions were identified (**Figure 1**). The distribution of the lesion size was observed to be consistent across the development, internal test, and external test datasets.

Segmentation Performance of the DSL

To set up the DLS, the images were first labeled manually. In summary, we manually labeled ~12,087 small lesions, 14,759 medium lesions, and 4,003 large lesions for training and validation (**Figures 2A-C**; **Table 2**). The network architecture of the proposed 2D convolutional neural network is shown in **Figure 3**. The model quality control parameters could be fond in **Figure 4**. More detailed information on the network can be found in the Network Architecture portion of the Methods section. After training and validation, the DLS was tested with the testing data set. The accuracy of the DLS-generated masking is represented in **Figures 2A,B**, with a Dice score of 0.87.

In the segmentation of WMH lesions, the proposed DLS achieved average pixel-wise dice score, precision, and recall of 0.711, 0.789, and 0.647 on the external test set. The lesion wise dice score, precision, recall, and accuracy achieved by the model were 0.735, 0.725, and 0.653 on the external test set. Also, the dice score of the model in segmentation of small, medium, and large WMH was 0.53, 0.82, and 0.96, respectively. Furthermore, in the lesion level analysis on the external test set, the model could correctly identify 61.07, 77.24, and 95.11% of the small, medium, and large lesions, respectively, and the detailed results of this analysis are presented in **Table 2**. A few examples of WMH segmentation using the proposed system are presented in

TABLE 5 | Clinical evaluation 1.

	Physician 1	Physician 2	Physician 3
Perfect (score 3)	34	33	45
Minor revision (score 2)	54	53	35
Major revision (score 1)	1	3	8
Not acceptable (score 0)	1	1	2

Figure 2. Also, in the segmentation of the WMH, the proposed DLS achieved a mean AUC of 0.9959 on the external test dataset (**Table 3**).

Lastly, the average pixel-wise dice score achieved by the UresUnet and 3D-unet networks on the external datasets were 0.584, and 0.623, respectively, and these were worse than the performance of the proposed DLS. Beside the preprocess is also import in the DLS development. For more detail information about the models head-to-head analysis in **Table 4**.

All the testing data are summarized in **Tables 2**, **3**, representing the relabeling results between the DLS tool and the experts (percentage correctly labled rato). From the table, we can see that the manual image labeling is precise and perfectly matches the contouring with the true signaling. This is because the labeling tool and pixels could not be well-controlled when manually drawing the labeling. Thus, the Dice score does not perfectly reflect the DLS segmentation result. These data can only support DLS training and validation. Visually, we checked all the data and found a strong concordance between our DLS and human experts for lesion contouring but, as mentioned above, with low Dice scores.

Clinical Assessment of the DLS Segmentation

The workflow of clinical evaluation (**Figure 5**) and results of the clinical acceptability analysis of the DLS are presented in **Table 5**. In this analysis, the majority [85 of 90 (94.0%)] of the DLS-generated segmentations were deemed satisfactory by the experts (no revision required, $n = 37$; minor revision, $n = 47$) (**Table 3**). Only four patients were assessed to require major revision, with two patients having clinically unacceptable segmentation results. In the assessment of the interrater agreement between the three neuroradiologists for the 90 test patients, the Kendall W test produced a score of 0.006 ($p = 0.605$) indicating a good inter-rater agreement.

DISCUSSION

In this paper, based on a large dataset of FLAIR images from more than 1,000 patients and with more than 50,000 lesions, we trained a DLS for automatic, and objective segmentation of WMHs. The proposed system was evaluated using pixel-wise and lesion-wise dice scores on internal and external test datasets. The results indicated that the proposed DLS achieved a consistent performance across both the test datasets, indicating good generalizability in the segmentation of WMH from different data sources. Furthermore, in the clinical acceptance analysis, with the 95% acceptance rate by the neuroradiologists, the segmentation results produced by the proposed DLS were observed to have a high clinical acceptance rate. These results collectively indicate that the proposed system can be deployed in clinical practice to quantitatively assess the WMH load in an end-to-end manner with high accuracy and in significantly reduced analysis time. Such a system can aid clinicians in fast and accurate assessment of WMH of the CSVD origin.

Limitation

This retrospective study analyzed the data from multiple different scanners which could result in a more robust and better generalizable model. However, our analysis did not exhaustively include the data from all the scanners and associated FLAIR image collection protocols, and hence, more extensive testing of the model, in prospective studies is necessary before its adaptation for clinical use. Second, our DLS system for segmentation of WMHs is solely based on MRI-FLAIR imaging features and it does not include complementary information that can be provided by other MRI sequences. Therefore, the possibility of better WMH segmentation using multiple imaging modalities should be explored in future studies.

CONCLUSION AND CONTRIBUTIONS

This study presented a DLS for the segmentation of WMH. Our findings indicate that the DLS can segment the WMHs with good accuracy and significantly smaller analysis time, minimizing the need for the physicians to perform repetitive tasks associated with segmentation. Additionally, the DLS model can reduce intra- and inter neuroradiologists' variation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Beijing Tiantan Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WS and YD wrote the initial draft of the manuscript, provided both figures and made preliminary revisions. YZ and ZW contributed to DLS development and medical test organization. YD, QW, ZW, PG, YW, and KH made preliminary revisions to the manuscript. YL, KH, and YW made crucial revisions to the manuscript. YZ main contributor for the model development. All authors planned the manuscript, critically revised the initial draft, and made final improvements prior to submission.

FUNDING

This study was funded by the China National Neurological Clinical Research Centre, Beijing Postdoctoral Research Foundation (ZZ 2019-09), and China Postdoctoral Science Foundation (No. 2019M660719).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.681183/full#supplementary-material

REFERENCES

1. Filley CM, Douglas Fields R. White matter and cognition: making the connection. J Neurophysiol. (2016) 116:2093-104. doi: 10.1152/jn.00221.2016
2. Mota B, Dos Santos SE, Ventura-Antunes L, Jardim-Messeder D, Neves K, Kazu RS, et al. White matter volume and white/gray matter ratio in mammalian species as a consequence of the universal scaling of cortical folding. *Proc Natl Acad Sci.* (2019) 116:15253-61. doi: 10.1073/pnas.1716956116
3. Narayana PA. White matter changes in patients with mild traumatic brain injury: MRI perspective. *Concussion*. (2017) 2:CNC35. doi: 10.2217/cnc-2016-0028
4. Duan Y, Shan W, Liu L, Wang Q, Wu Z, Liu PY, et al. Primary categorizing and masking cerebral small vessel disease based on "Deep Learning System." *Front Neuroinform.* (2020) 14:17. doi: 10.3389/fninf.2020.00017
5. Kim JW, Byun MS, Yi D, Lee JH, Ko K, Jeon SY, et al. Association of moderate alcohol intake with *in vivo* amyloid-beta deposition in human brain: a cross-sectional study. *PLoS Med.* (2020) 17: 1003022. doi: 10.1371/journal.pmed.1003022
6. Joutel A, Chabriat H. Pathogenesis of white matter changes in cerebral small vessel diseases: beyond vessel-intrinsic mechanisms. *Clin Sci.* (2017) 131:635-51. doi: 10.1042/CS20160380
7. Kazutaka S, Shoichiro S, Fumiaki N, Yusuke S, Kunihiro N, Yoshihiro M, et al. The Japan stroke data bank investigators Abstract WP178: impact of white matter hyperintensity on severity and outcome of acute ischemic and hemorrhagic stroke: the Japan Stroke Data Bank. *Stroke.* (2018) 49(Suppl_1):AWP178. doi: 10.1161/str.49.suppl_1.WP178
8. Shinoda T, Nakashita S, Hamada M, Hirono K, Ito M, Kashihara K, et al. Multi-center observational study of personality and impulse control disorders in Japanese patients with Parkinson's disease. *Parkinsonism Relat Disord.* (2018) 46:e68. doi: 10.1016/j.parkreldis.2017.11.232
9. Scarpelli M, Salvolini U, Diamanti L, Montironi R, Chiaromoni L, Maricotti M. MRI and pathological examination of post-mortem brains: the problem of white matter high signal areas. *Neuroradiology.* (1994) 36:393-8. doi: 10.1007/BF00612126
10. Rezaie P, Dean A. Periventricular leukomalacia, inflammation and white matter lesions within the developing nervous system. *Neuropathology.* (2002) 22:106-32. doi: 10.1046/j.1440-1789.2002.00438.x
11. van der Knaap MS, Valk J. *Subcortical Arteriosclerotic Encephalopathy. Magnetic Resonance of Myelin, Myelination, Myelin Disorders.* Berlin, Heidelberg: Springer (1995). p. 391-7.
12. Rocca MA, Nusbaum AO, Absinta M, Rapalino O, Fung K-M, Filippi M. White matter diseases and inherited metabolic disorders. *Magnet Resonan Imaging Brain Spine.* (2009) 4:365.
13. Urbach H, Tschampa H, Flacke S, Thal D. MRI of vascular dementia and differential diagnoses. *Clin Neuroradiol.* (2007) 17:88-97. doi: 10.1007/s00062-007-7009-1
14. Dalby RB, Chakravarty MM, Ahdidan J, Sørensen L, Frandsen J, Jonsdottir KY, et al. Localization of white-matter lesions and effect of vascular risk factors in late-onset major depression. *Psychol Med.* (2010) 40:1389. doi: 10.1017/S0033291709991656
15. Prins ND, Scheltens P. White matter hyperintensities, cognitive impairment and dementia: an update. *Nat Rev Neurol.* (2015) 11:157-65. doi: 10.1038/nrneurol.2015.10
16. Wahlund LO, Barkhof F, Fazekas F, Bronge L, Augustin M, Sjogren M, et al. A new rating scale for age-related white matter changes applicable to MRI and CT. *Stroke.* (2001) 32:1318-22. doi: 10.1161/01.STR.32.6.1318
17. Wardlaw JM, Smith EE, Biessels GJ, Cordonnier C, Fazekas F, Frayne R, et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol.* (2013) 12:822-38. doi: 10.1016/S1474-4422(13)70152-2
18. Pohjasvaara T, Mäntylä R, Ylikoski R, Kaste M, Erkinjuntti T. Clinical features of MRI-defined subcortical vascular disease. *Alzheimer Dis Assoc Disord.* (2003) 17:236-42. doi: 10.1097/00002093-200310000-00007
19. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. *J Med Syst.* (2018) 42:226. doi: 10.1007/s10916-018-1088-1
20. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: full training or fine tuning?. *IEEE Trans Med Imaging.* (2016) 35:1299-312. doi: 10.1109/TMI.2016.2535302
21. Milletari F, Navab N, Ahmadi S-A. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE (2016).
22. Li Q, Cai W, Wang X, Zhou Y, Feng DD, Chen M. Medical image classification with convolutional neural network. In: *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*. IEEE (2014).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Neuroscience

ORIGINAL RESEARCH
published: 30 December 2021
doi: 10.3389/fnins.2021.752780

Robust, Primitive, and Unsupervised Quality Estimation for Segmentation Ensembles

Florian Kofler [1,2,3]*, Ivan Ezhov [1,3], Lucas Fidon [4], Carolin M. Pirkl [1], Johannes C. Paetzold [1,3], Egon Burian [2], Sarthak Pati [1,5,6,7], Malek El Husseini [1,2], Fernando Navarro [1,3,8], Suprosanna Shit [1,3], Jan Kirschke [2], Spyridon Bakas [5,6,7], Claus Zimmer [2], Benedikt Wiestler [2†] and Bjoern H. Menze [1,9†]

[1] Department of Informatics, Technical University Munich, Munich, Germany, [2] Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany, [3] TranslaTUM - Central Institute for Translational Cancer Research, Technical University of Munich, Munich, Germany, [4] School of Biomedical Engineering & Imaging Sciences, King's College London, London, United Kingdom, [5] Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Pennsylvania, PA, United States, [6] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Pennsylvania, PA, United States, [7] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Pennsylvania, PA, United States, [8] Department of Radio Oncology and Radiation Therapy, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany, [9] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

OPEN ACCESS

Edited by:
*Gabriele Lohmann,
Max Planck Institute for Biological Cybernetics, Germany*

Reviewed by:
*Xue Feng,
University of Virginia, United States
Qian Wang,
Shanghai Jiao Tong University, China*

***Correspondence:**
*Florian Kofler
florian.kofler@tum.de*

† These authors share senior authorship

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *03 August 2021*
Accepted: *26 October 2021*
Published: *30 December 2021*

Citation:
Kofler F, Ezhov I, Fidon L, Pirkl CM, Paetzold JC, Burian E, Pati S, El Husseini M, Navarro F, Shit S, Kirschke J, Bakas S, Zimmer C, Wiestler B and Menze BH (2021) Robust, Primitive, and Unsupervised Quality Estimation for Segmentation Ensembles. Front. Neurosci. 15:752780. doi: 10.3389/fnins.2021.752780

A multitude of image-based machine learning segmentation and classification algorithms has recently been proposed, offering diagnostic decision support for the identification and characterization of glioma, Covid-19 and many other diseases. Even though these algorithms often outperform human experts in segmentation tasks, their limited reliability, and in particular the inability to detect failure cases, has hindered translation into clinical practice. To address this major shortcoming, we propose an unsupervised quality estimation method for segmentation ensembles. Our primitive solution examines discord in binary segmentation maps to automatically flag segmentation results that are particularly error-prone and therefore require special assessment by human readers. We validate our method both on segmentation of brain glioma in multi-modal magnetic resonance - and of lung lesions in computer tomography images. Additionally, our method provides an adaptive prioritization mechanism to maximize efficacy in use of human expert time by enabling radiologists to focus on the most difficult, yet important cases while maintaining full diagnostic autonomy. Our method offers an intuitive and reliable uncertainty estimation from segmentation ensembles and thereby closes an important gap toward successful translation of automatic segmentation into clinical routine.

Keywords: quality estimation, failure prediction, anomaly detection, ensembling, fusion, OOD, CT, MR

1. INTRODUCTION

Advances in deep learning for segmentation have facilitated the automated assessment of a variety of anatomies and pathologies in medical imaging. In particular for glioma, automatic segmentation has shown great promise as a basis for objective assessment of tumor response (Kickingereder et al., 2019). In segmentation challenges such as BraTS (Menze et al., 2015), VerSe (Sekuboyina et al., 2021) and LiTS (Bilic and et al., 2019) virtually all top-performing solutions are based on ensembling. Recent efforts such as *HD-GLIO* (Kickingereder et al., 2019; Isensee et al., 2021), *GaNDLF* (Pati et al., 2021), and *BraTS Toolkit* (Kofler et al., 2020) have paved the way to apply state-of-the-art deep-learning ensembles in clinical practice. Even though algorithms often outperform human readers (Kofler et al., 2021), algorithmic reliability remains a major obstacle toward safe implementation of automated segmentation (and hence volumetry) into clinical routine (D'Amour et al., 2020). Researchers in the field of Out-of-Distribution (OOD) detection try to address this shortcoming by discovering systematic patterns within convolutional neural networks (CNN) (Schölkopf et al., 2001; Jungo et al., 2018; Mehrtash et al., 2020; Berger et al., 2021; Ruff et al., 2021). These sophisticated anomaly detection methods have the disadvantage of being limited to CNNs, often specific CNN architectures.

In contrast, we present a primitive, and therefore more applicable, solution exploiting discord in binary segmentation maps to estimate segmentation quality in an unsupervised fashion. We evaluate our method on segmentation of brain glioma in multi-modal magnetic resonance (MR)—and of lung lesions in computer tomography (CT) images. Our method allows detecting error-prone segmentation results, which require special assessment by human readers. Working only on binary segmentation maps enables our method to analyze the segmentations of human readers, classical machine learning, and modern deep learning approaches interchangeably. As segmentations are the basis for objective disease assessment as well as subsequent image analysis, our method addresses an urgent need for improving the trustworthiness of automatic segmentation methods. Furthermore, by implementing our method healthcare providers can streamline efficient use of human workforce, arguably the most persistent and major bottleneck in healthcare service worldwide (Krengli et al., 2020; Starace et al., 2020).

2. METHODS

2.1. Unsupervised Quality Estimation

Figure 1 depicts the quality estimation procedure. By aggregating and comparing multiple candidate segmentations, cases with large discordance, therefore a high chance of failure, can be rapidly identified. In more detail, our method consists of the following steps:

1. We obtain candidate segmentations from all methods in an ensemble, and then compute a fusion from the candidate segmentations.
2. We calculate similarity metrics between the fused segmentation result and the individual candidate segmentations.
3. We obtain the threshold for setting an alarm value by subtracting the *median absolute deviation (mad)* of the similarity metric times the tunable parameter α from its *median* value. This happens individually for each candidate image. We prefer the *median* based statistics for their better robustness toward statistical outliers. For metrics that are negatively correlated with segmentation performance, such as Hausdorff distance, we propose to use the additive inverse.
4. We set an alarm flag if the individual similarity metric is below the computed threshold. For *infinite* (or *Nan*) values, which can for instance happen for distance-based metrics such as Hausdorff distance, alarm flags are raised too.
5. Finally, we accumulate the alarm flags to obtain risk scores and therefore quality estimation for each image.

The results of this procedure are illustrated in **Figure 4**. We hypothesize that a higher count of alarm flags is associated with worse segmentation quality, here measured by lower volumetric Dice performance.

2.2. MR Experiment: Multi-Modal Brain Tumor Segmentation

To test the validity of our approach we use BraTS Toolkit *(btk)* (Kofler et al., 2020) to create a segmentation ensemble for brain glioma in multi-modal magnetic resonance (MR) images. Therefore, we incorporate five segmentation algorithms (Feng et al., 2019; Isensee et al., 2019; McKinley et al., 2019, 2020; Zhao et al., 2019) developed within the scope of the BraTS challenge (Menze et al., 2015; Bakas et al., 2017a,b,c, 2018). We compute alarms according to the above procedure based on Dice similarity and Hausdorff distances.

2.2.1. Fusions and Segmentation Metrics

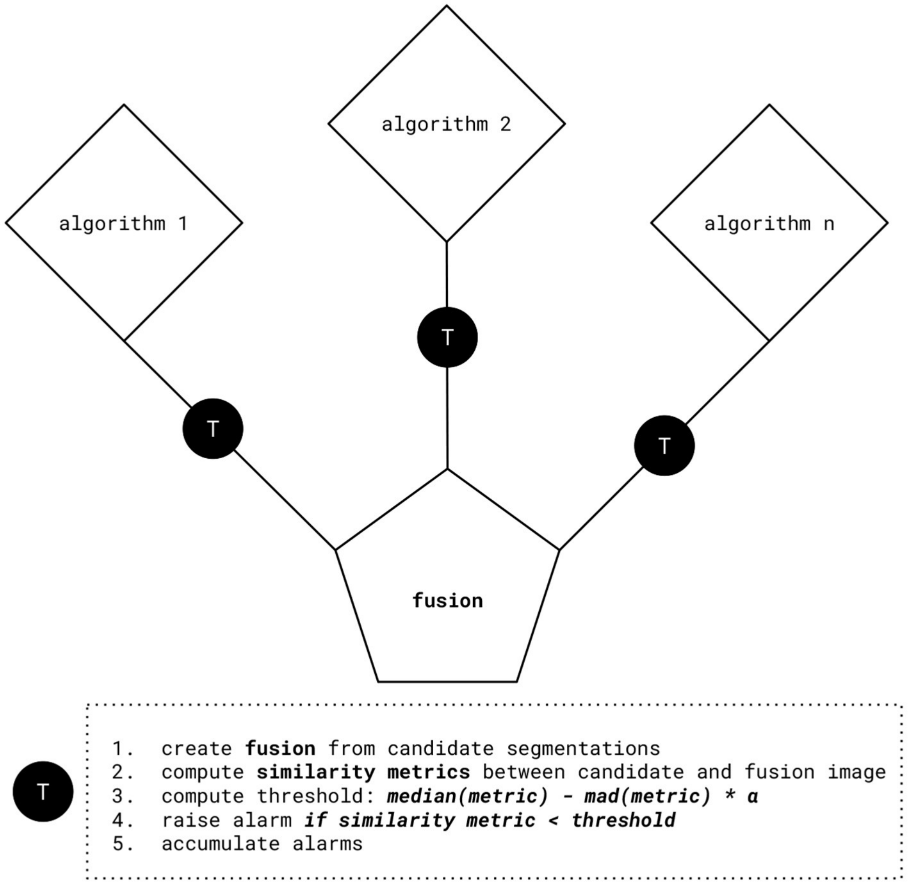
We fuse the segmentations with an equally weighted majority voting using *btk* (Kofler et al., 2020) and compute segmentation quality metrics with *pymia* (Jungo et al., 2021). **Figure 2** illustrates fusions and individual segmentations with an example exam.

2.2.2. Data

We evaluate on a dataset of 68 cases capturing the wide diversity in glioma imaging. Our dataset consists of 15 high-grade glioma (HGG) from the publicly available Rembrandt dataset (Gusev et al., 2018), as well as another 25 HGG from TUM university hospital (MRI TUM). Furthermore, we evaluate 13 low-grade glioma (LGG) from Rembrandt and 15 from MRI TUM. Two expert radiologists generated the ground truth segmentations using *ITK-SNAP* (Yushkevich et al., 2006) and corrected each other's tumor delineations.

2.3. CT Experiment: COVID-19 Lung CT Lesion Segmentation

For further validation, we compose an ensemble based on the MONAI challenge baseline (MONAI CORE Team, 2020)



```
1.  create fusion from candidate segmentations
2.  compute similarity metrics between candidate and fusion image
3.  compute threshold: median(metric) - mad(metric) * α
4.  raise alarm if similarity metric < threshold
5.  accumulate alarms
```

FIGURE 1 | Quality estimation procedure. After computing fusion from the candidate segmentations, similarity metrics between the fused and the candidate segmentations are evaluated. Using this information, we obtain threshold values by subtracting the median absolute deviation (mad) of similarity metrics times the tunable parameter α from their median value. We set an alarm flag if the individual similarity metric is below the computed threshold, for example: $median(Dice) - mad(Dice) * \alpha$.

developed for the *COVID-19 Lung CT Lesion Segmentation Challenge - 2020* (Clark et al., 2013). To segment lung lesions in computer tomography (CT) images, the code implements a 3d-Unet inspired by Falk et al. (2019). q2a1 We first train the original baseline for 500 epochs. Then we generate a small ensemble of three networks by warmstarting the training with the baseline's model weights and replacing the following parameters for the respective model for training another 500 epochs:

To obtain our first model (ADA) we swap the baseline's original Adam optimizer to *AdamW* (Loshchilov and Hutter, 2019). In a similar fashion, the second model (RAN) utilizes Ranger (Wright, 2019) to make use of Gradient Centralization (Yong et al., 2020). Our third model (AUG) adds an augmentation pipeline powered by batchgenerators (Isensee et al., 2020), torchio (Pérez-García et al., 2020), and native MONAI augmentations. In addition we switch the optimizer to stochastic gradient descent (*SGD*) with momentum (momentum = 0.95).

Our metric for training progress is the volumetric Dice coefficient. All networks are trained with an equally weighted Dice plus binary cross-entropy loss. The training is stopped once we observe no further improvements for the validation set. We conduct model selection by choosing the respective model with the best volume Dice score on the validation set. The code for the CNN trainings is publicly available via GitHub (***censored to maintain the double blind review process***).

2.3.1. Fusions and Segmentation Metrics

To unify the individual outputs of our ensembles' components to a segmentation mask we choose SIMPLE (Langerak et al., 2010) fusion. SIMPLE is an iterative fusion method introduced by Langerak et al., which tends to outperform generic majority voting across various segmentation problems. An example segmentation for one exam is illustrated in **Figure 3**. We generate SIMPLE fusions using BraTS Toolkit (Kofler et al., 2020) and generate alarms for Dice scores calculated with *pymia* (Jungo et al., 2021). Segmentation quality metrics, in particular volumetric Dice coefficient and Hausdorff distances, for the test set are obtained through the challenge portal (COVID Challenge Team, 2021).

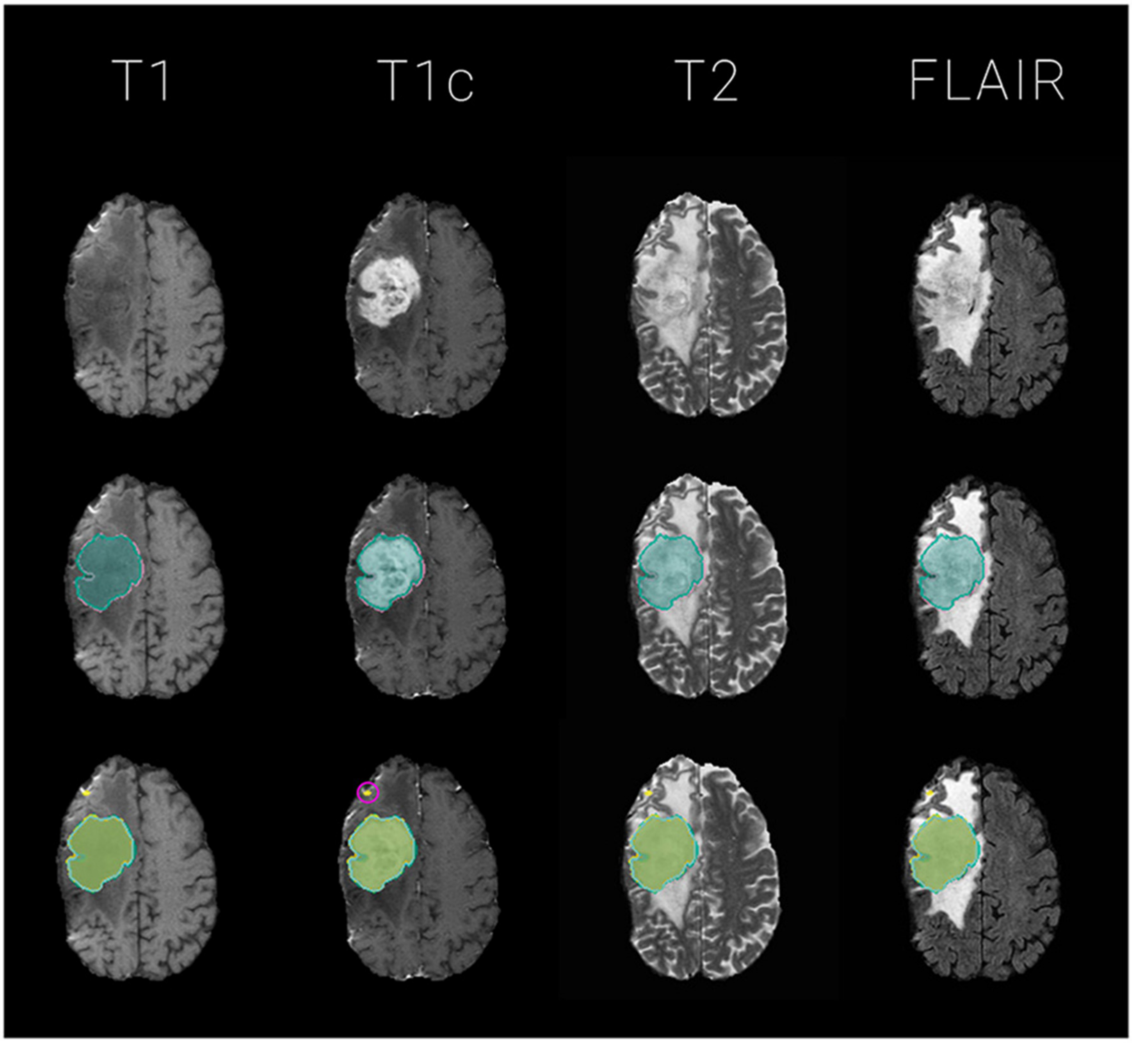


FIGURE 2 | Exemplary glioma segmentation exam with multi-modal MR. Segmentations are overlayed on T1, T1c, T2, FLAIR images for the tumor's center of mass, defined by the *tumor core* (*necrosis* and *enhancing tumor*) of the ground truth label. The segmentation outlines represent the *tumor core* labels, meaning the sum of *enhancing tumor* and *necrosis* labels. **Top**: the four input images without segmentation overlay; **Middle**: ground truth segmentation (*GT*) in *reddish purple* vs. majority voting fusion (*mav*) in *bluish green*; **Bottom**: *mav* fusion in *bluish green* vs. individual segmentation algorithms in various colors. Notice the small outliers encircled in pink on the frontal lobe which probably contribute to the raise of 3 Dice - and 4 Hausdorff distance based alarms for this particular exam with a mediocre volumetric Dice similarity coefficient with the *ground truth* data of *0.66*.

2.3.2. Data

We run our experiments on the public dataset of the COVID-19 Lung CT Lesion Segmentation Challenge - 2020 (COVID Challenge Team, 2021), supported by the Cancer Imaging Archive (TCIA) (Clark et al., 2013).

2.4. Calibration of Alpha (*α*)

The α parameter can be fine-tuned to account for different optimization targets and adjusted dynamically depending on workload, e.g., in an extreme triage scenario, an alarm flag could only be raised for the strongest outliers, hence a high α should be chosen. Once the situation has been amended, α can be reset to a smaller value, resulting in a more sensitive failure prediction.

With the default value $\alpha = 0$ the threshold is set to the median. Therefore, approximately half of the cases will trigger an alarm for each metric. Alternatively, alpha can be automatically adjusted to maximize the Pearson correlation coefficient with a segmentation quality metric or entropy, or combinations thereof. **Tables 1**, **2** illustrate how the distributions of alarm counts correlate with Dice performance and the resulting entropy in response to variations in α.

Note that α can also be adjusted for each segmentation target class, as well as, each of the ensemble's components, and for each similarity metric on an individual basis to fine-tune the quality estimation toward specific needs. For instance, hence the *enhancing tumor* label is of higher clinical relevance for glioma

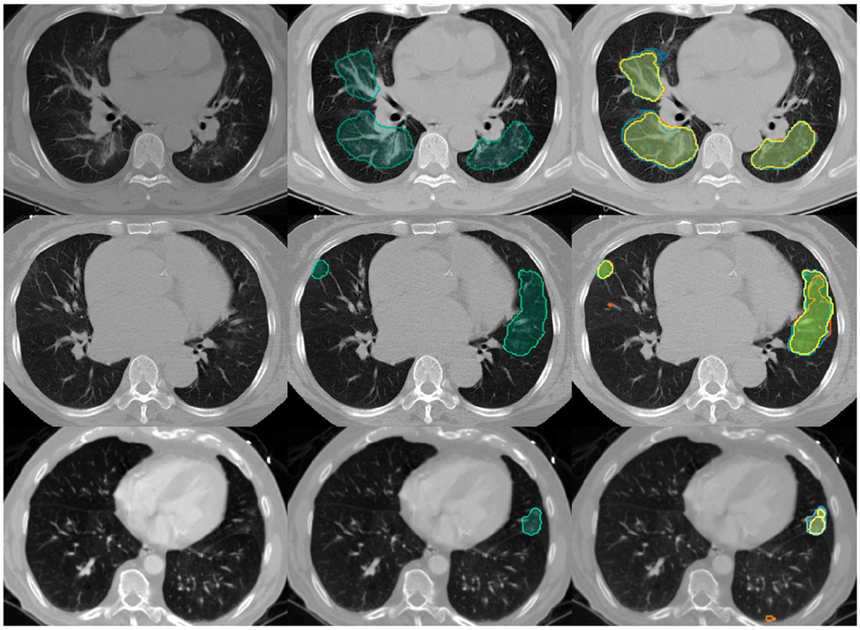


FIGURE 3 | Example Covid-19 lung lesion segmentation exams with CT images. Segmentations are overlayed for the lesions' center of mass, defined by the slice with most lesion voxels: **Left**: the empty input images; **Middle**: SIMPLE segmentation fusion (simple) in *bluish green*; **Right**: SIMPLE fusion in *bluish green* vs. individual segmentation algorithms in various colors. The volumetric Dice similarity coefficients with the *ground truth* and respective alarm counts are as following: Top row: *0.81, 0*; Middle row: *0.58, 2*; Last row: *0.14, 3*.

TABLE 1 | Distribution of alarm counts depending on α for the MR experiment: The table illustrates the number of images classified in the individual alarm count categories *(a)* from *0* to *10*; for different values of α.

Alpha	Entropy	r:dice	r:hd	0a	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a
−3.00	−0.00	NA	NA	0	0	0	0	0	0	0	0	0	0	68
−2.00	0.22	NA	0.04	0	0	0	0	0	0	0	0	0	4	64
−1.00	1.28	−0.27	−0.2	0	0	0	1	0	2	2	5	5	13	40
−0.75	1.80	−0.55	−0.27	0	0	1	3	4	2	10	5	4	14	25
−0.50	2.02	−0.63	−0.3	0	6	1	3	5	4	11	1	10	8	19
−0.25	2.33	−0.7	−0.38	3	5	4	5	7	4	6	7	8	7	12
−0.10	2.37	−0.73	−0.41	7	4	4	6	4	7	7	8	7	6	8
0.00	2.35	−0.76	−0.45	9	5	7	4	4	6	6	8	8	3	8
0.10	2.30	−0.77	−0.46	9	6	10	3	6	7	2	9	5	3	8
0.25	2.28	−0.77	−0.51	11	7	12	3	2	7	3	8	5	5	5
0.50	2.23	−0.78	−0.59	15	11	8	3	2	4	5	8	4	4	4
0.75	2.06	−0.73	−0.59	18	13	7	3	1	5	6	7	2	6	0
1.00	1.97	−0.72	−0.58	23	12	3	3	2	6	8	6	3	2	0
2.00	1.71	−0.66	−0.55	30	10	6	4	3	8	2	5	0	0	0
3.00	1.40	−0.65	−0.52	37	11	4	1	3	10	1	1	0	0	0

Additionally, we depict the Pearson correlation coefficients for the Dice (r:dice) *- and Hausdorff distance* (r:hd) *based alarm counts with volumetric Dice segmentation performance, as well as the respective alarm count distribution's entropy. The selected value for* α *of 0.1 is highlighted in pink The resulting computed thresholds are depicted in* ***Table 3****.*

(Weller et al., 2014), one might consider setting the associated thresholds to more conservative values using a smaller *alpha*.

For simplicity, we set parameter α to *0.1* for each class, component and metric in our analysis. This results in a slightly less conservative failure prediction compared to the default.

3. RESULTS

Our method accurately predicts the segmentation performance in both experiments and is able to capture segmentation failures. Even though our code is not optimized for speed, the

TABLE 2 | Distribution of alarm counts depending on α for the CT experiment: The table illustrates the number of images classified in the individual alarm count categories *(a)* from *0* to *3*; for different values of α.

Alpha	Entropy	r:dice	0a	1a	2a	3a
−3.00	−0.00	NA	0	0	0	46
−2.00	−0.00	NA	0	0	0	46
−1.00	0.58	−0.45	0	3	5	38
−0.75	0.88	−0.56	5	2	6	33
−0.50	1.19	−0.67	6	7	8	25
−0.25	1.32	−0.64	10	7	10	19
−0.10	1.36	−0.73	12	8	11	15
0.00	1.37	−0.7	13	8	14	11
0.10	1.37	−0.7	15	10	11	10
0.25	1.33	−0.62	18	9	11	8
0.50	1.20	−0.61	23	6	12	5
0.75	1.17	−0.69	25	9	8	4
1.00	1.13	−0.71	26	10	6	4
2.00	0.86	−0.67	33	8	2	3
3.00	0.66	−0.62	37	6	1	2

Additionally, we depict the Pearson correlation coefficients for the Dice (r:dice) *based alarm counts with volumetric Dice segmentation performance, as well as the respective alarm count distribution's entropy. The selected value for* α *of 0.1 is highlighted in pink. The resulting computed Dice similarity thresholds are as following: ADA: 0.9489; RAN: 0.9446; AUG: 0.9024.*

computation of the fused segmentation masks, similarity metrics and resulting alarm counts is a matter of seconds. Quantitative metrics for the MR and CT experiment are summarized in **Figure 4**.

3.1. MR Experiment

Setting α to *0.1* leads to an even distribution across alarm count groups, (see **Tables 1**, **3**). **Figure 4A** plots the average Dice coefficients across the tumors labels: *enhancing tumor, necrosis and edema* against the alarm count. We observe a strong negative correlation between segmentation performance and increasing alarm count: *Pearson's* $r = -0.72$, $p = 3.874e\text{-}12$. This is also reflected in the Hausdorff distance, (see **Figure 4B**).

3.2. CT Experiment

Choosing an α of *0.1* leads to an even distribution across alarm count groups, (see **Table 2**). **Figure 4C** plots Dice coefficients[1] on the challenge test set against alarm count. As for the MR experiment, we find a strong negative correlation between segmentation performance and increasing alarm count: *Pearson's* $r = -0.70$, *p-value* $= 4.785e\text{-}08$. As observed before, this effect is mirrored by the Hausdorff distance, (see **Figure 4D**).

4. DISCUSSION

It is important to note that, the validity of our method is closely tied to the chosen evaluation metrics' representation of segmentation performance (Kofler et al., 2021). For our experiments, we evaluate the volumetric Dice score and Hausdorff distance. Based on this fundamental assumption, we provide an unsupervised quality estimation for segmentation ensembles that does not perform any background diagnostic decisions and fully maintains the radiologists' diagnostic autonomy.

We demonstrate efficacy for two different use cases, namely multi-modal glioma segmentation in brain MR and Covid-19 lesion segmentation in lung CT images. The sensitivity of our method can be fine-tuned to specific requirements by adjusting α for ensemble components, classes, and segmentation quality metrics. Additionally, the low computational requirements make it easy to integrate into existing pipelines as computing the alarms takes only seconds and creates very little overhead.

Even though there are various efforts, such as the *BraTS algorithmic repository*[2], to facilitate clinical translation of state-of-the-art segmentation algorithms, quality estimation mechanisms represent a currently unmet, yet important milestone on the road toward reliably deploying deep learning segmentation pipelines in clinical practice. The proposed solution can assist clinicians in navigating the plethora of exams, which have to be reviewed daily. It provides a neat prioritization mechanism, maximizing the efficient use of human expert time, by enabling focus on the most difficult, yet important cases.

It is important to note further limitations of our method. First of all, it can only be applied to model ensembles and not to single algorithms. However, as most top-performing segmentation solutions employ ensembling techniques there is a broad field of potential application. Second, the computation of alarms relies on discordance in the ensemble. If all components of the ensemble converge to predicting the same errors they cannot be detected. Notably, we did not observe such a case in our experiments, even though our CT segmentation ensemble featured only three models employing the same architecture and little variation in training parameters. As our method profits from bigger ensembles and more variations in the network training, one could argue that our experiment is probably more difficult than most real-world scenarios. Along these lines, Roy et al. (2019) activated dropout during inference and Fort et al. (2020) demonstrated that it might be enough to choose different random initialization to achieve variance in network outputs. Third, even though the default value of α, *0* and *0.1*, which we chose for demonstration purposes, performed well in our experiments, there might be segmentation problems for which α needs to be manually fine-tuned.

Future research could investigate whether α how global thresholding, instead of the proposed individual thresholding per algorithm, affects the results. It should also be explored whether the methodology can be improved by including further

[1] Our basic ensemble reaches a median volumetric Dice score of *0.67*. We observe a wide performance distribution with a minimum of *0*, a maximum of *0.93* and a standard deviation of *0.25* around a mean of *0.61*, as displayed in **Figure 4C**. With regard to volumetric Dice coefficients mainly low-performing outliers separate our method from the top-performing methods in the challenge.

[2] https://www.med.upenn.edu/sbia/brats2017/algorithms.html

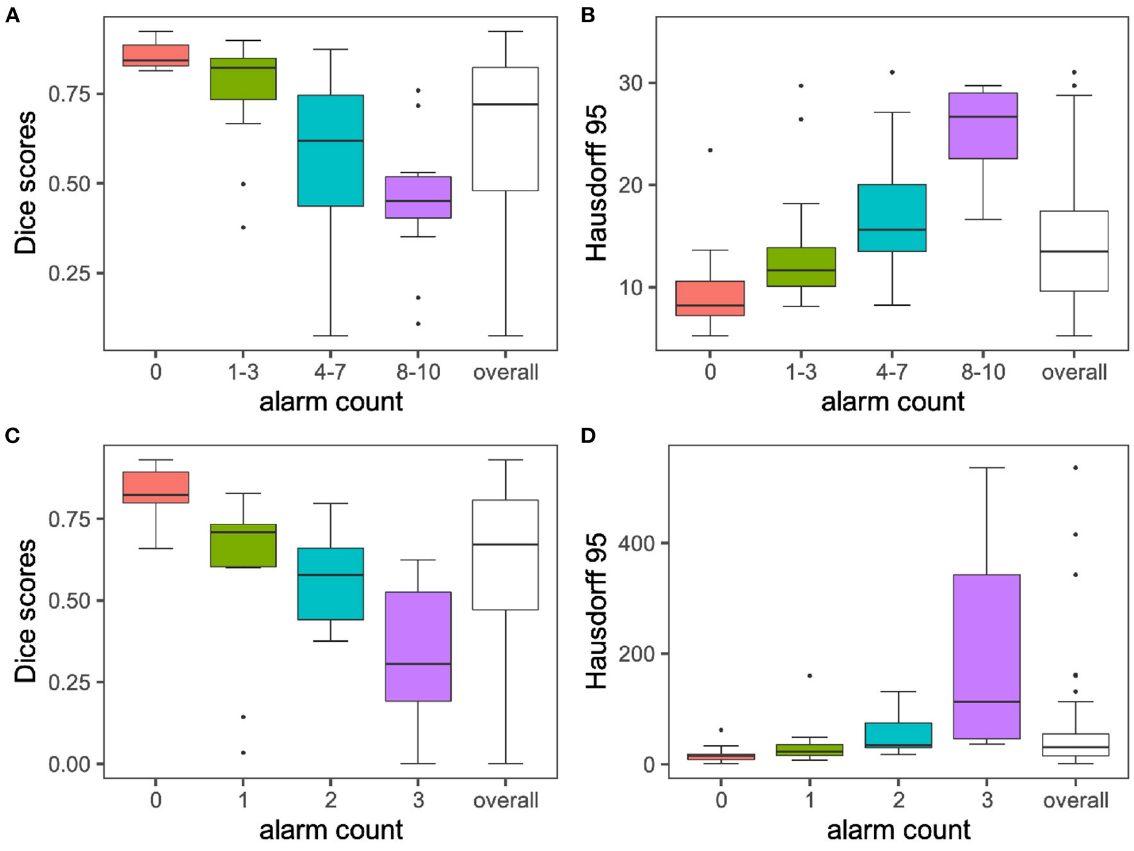


FIGURE 4 | Segmentation performances vs. alarm counts. The group means are illustrated with horizontal black lines. For display purposes only the 0–95 percent quantile is displayed for Hausdorff distances on the y-axis. In line with the performance of the volumetric Dice coefficient, Hausdorff distances increase with increasing alarm count. Infinite values for Hausdorff distances, which can happen when ground truth or prediction are empty, are excluded from the plot. Subplots **(A)** + **(B)** illustrate findings for the MR experiment, while subplots **(C)** + **(D)** depict results for the CT experiment.

TABLE 3 | Thresholds computed with $\alpha = 0.1$ for the MR experiment per algorithm: The columns *Dice* and *Hausdorff* depict, the respective volumetric Dice and Hausdorff distance based thresholds for the alarm computation for each of the segmentation algorithms.

Algorithm	Citation	Dice	Hausdorff
micdkfz	Isensee et al., 2019	0.9055	10.2277
xfeng	Feng et al., 2019	0.9092	8.9835
scan2019	McKinley et al., 2020	0.9147	8.8292
scan	McKinley et al., 2019	0.9084	10.4850
zyx	Zhao et al., 2019	0.9293	8.4451

segmentation metrics and to which extend it generalizes to other segmentation problems.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The CT data can be found here: https://covid-segmentation.grand-challenge.org/data/. The MR data will be published at: https://neuronflow.github.io/btk_evaluation/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

FK, IE, and LF contributed to conception and design of the study. FK, IE, CP, and JP wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

The annotation of the dataset was made possible through the joint work of Children's National Hospital, NVIDIA, and

National Institutes of Health for the COVID-19-20 Lung CT Lesion Segmentation Grand Challenge. BM, BW, and FK are supported through the SFB 824, subproject B12. Supported by Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81. LF, SS, and IE are supported by the Translational Brain Imaging Training Network (TRABIT) under the European Union's Horizon 2020 research & innovation program (Grant agreement ID: 765148). With the support of the Technical University of Munich - Institute for Advanced Study, funded by the German Excellence Initiative. JP and SS are supported by the Graduate School of Bioengineering, Technical University of Munich. Research reported in this publication was partly supported by the National Cancer Institute (NCI) and the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health (NIH), under award numbers NCI:U01CA242871 and NINDS:R01NS042645. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH. Research reported in this publication was partly supported by AIME GPU cloud services.

ACKNOWLEDGMENTS

We would like to thank the YouTube channel Primitive Technology for inspiration regarding the manuscript's title and methodology.

REFERENCES

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017a). *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM Collection*. The Cancer Imaging Archive.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al. (2017b). *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG Collection*. The Cancer Imaging Archive.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., et al. (2017c). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* 4:170117. doi: 10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., et al. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*.

Berger, C., Paschali, M., Glocker, B., and Kamnitsas, K. (2021). Confidence-based out-of-distribution detection: a comparative study and analysis. *arXiv preprint arXiv:2107.02568*. doi: 10.1007/978-3-030-87735-4_12

Bilic, P., Christ, P. F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., et al. (2019). The liver tumor segmentation benchmark (LiTS). *arXiv preprint arXiv:1901.04056*.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., et al. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* 26, 1045–1057. doi: 10.1007/s10278-013-9622-7

COVID Challenge Team (2021). *COVID Challenge*. Available online at: https://covid-segmentation.grand-challenge.org/Data/

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., et al. (2019). U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* 16, 67–70. doi: 10.1038/s41592-018-0261-2

Feng, X., Tustison, N., and Meyer, C. (2019). "Brain tumor segmentation using an ensemble of 3D U-Nets and overall survival prediction using radiomic features," in *International MICCAI Brainlesion Workshop*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer), 279–288. doi: 10.1007/978-3-030-11726-9_25

Fort, S., Hu, H., and Lakshminarayanan, B. (2020). Deep ensembles: a loss landscape perspective. *arXiv preprint arXiv:1912.02757*.

Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J.-C., Fine, H., and Madhavan, S. (2018). The rembrandt study, a large collection of genomic data from brain cancer patients. *Sci. Data* 5:180158. doi: 10.1038/sdata.2018.158

Isensee, F., and et al. (2019). "No new-net," in *International MICCAI Brainlesion Workshop* (Cham: Springer), 234–244. doi: 10.1007/978-3-030-11726-9_21

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). NNU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211. doi: 10.1038/s41592-020-01008-z

Isensee, F., Jager, P., Wasserthal, J., Zimmerer, D., Petersen, J., Kohl, S., et al. (2020). *Batchgenerators - A Python Framework for Data Augmentation*. doi: 10.5281/zenodo.3632567

Jungo, A., Meier, R., Ermis, E., Herrmann, E., and Reyes, M. (2018). Uncertainty-driven sanity check: application to postoperative brain tumor cavity segmentation. *arXiv preprint arXiv:1806.03106*.

Jungo, A., Scheidegger, O., Reyes, M., and Balsiger, F. (2021). pymia: a python package for data handling and evaluation in deep learning-based medical image analysis. *Comput. Methods Prog. Biomed*. 198:105796. doi: 10.1016/j.cmpb.2020.105796

Kickingereder, P., Isensee, F., Tursunova, I., Petersen, J., Neuberger, U., Bonekamp, D., et al. (2019). Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol*. 20, 728–740. doi: 10.1016/S1470-2045(19)30098-1

Kofler, F., Berger, C., Waldmannstetter, D., Lipkova, J., Ezhov, I., Tetteh, G., et al. (2020). Brats toolkit: translating brats brain tumor segmentation algorithms into clinical and scientific practice. *Front. Neurosci*. 14:125. doi: 10.3389/fnins.2020.00125

Kofler, F., Ezhov, I., Isensee, F., Balsiger, F., Berger, C., Koerner, M., et al. (2021). Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for cnn training beyond rolling the dice coefficient. *arXiv preprint arXiv:2103.06205*.

Krengli, M., Ferrara, E., Mastroleo, F., Brambilla, M., and Ricardi, U. (2020). Running a radiation oncology department at the time of coronavirus: an Italian experience. *Adv. Radiat. Oncol*. 5, 527–530. doi: 10.1016/j.adro.2020.03.003

Langerak, T. R., van der Heide, U. A., Kotte, A. N., Viergever, M. A., Van Vulpen, M., and Pluim, J. P. (2010). Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). *IEEE Trans. Med. Imaging* 29, 2000–2008. doi: 10.1109/TMI.2010.2057442

Loshchilov, I., and Hutter, F. (2019). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

McKinley, R., Meier, R., and Wiest, R. (2019). "Ensembles of densely-connected cnns with label-uncertainty for brain tumor segmentation," in *International MICCAI Brainlesion Workshop*, eds A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum (Cham: Springer), 456–465. doi: 10.1007/978-3-030-11726-9_40

McKinley, R., Rebsamen, M., Meier, R., and Wiest, R. (2020). "Triplanar ensemble of 3D-to-2D CNNs with label-uncertainty for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi and S. Bakas (Cham: Springer International Publishing), 379–387. doi: 10.1007/978-3-030-46640-4_36

Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., and Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Med. Imaging* 39, 3868–3878. doi: 10.1109/TMI.2020.3006437

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi: 10.1109/TMI.2014.2377694

MONAI CORE Team (2020). *MONAI*. doi: 10.5281/zenodo.4323059

Pati, S., Thakur, S. P., Bhalerao, M., Baid, U., Grenko, C., Edwards, B., et al. (2021). Gandlf: A generally nuanced deep learning framework for scalable end-to-end clinical workflows in medical imaging. *arXiv preprint arXiv:2103.01006*.

Pérez-García, F., Sparks, R., and Ourselin, S. (2020). TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *arXiv preprint arXiv:2003.04696*. doi: 10.1016/j.cmpb.2021.106236

Roy, A. G., Conjeti, S., Navab, N., and Wachinger, C. (2019). Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control. *Neuroimage* 195, 11–22. doi: 10.1016/j.neuroimage.2019.03.042

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., et al. (2021). A unifying review of deep and shallow anomaly detection. *Proc. IEEE* 109, 756–795. doi: 10.1109/JPROC.2021.3052449

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Comput*. 13, 1443–1471. doi: 10.1162/089976601750264965

Sekuboyina, A., Husseini, M. E., Bayat, A., Loffler, M., Liebl, H., Li, H., et al. (2021). Verse: a vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical Image Anal*. 2021:102166. doi: 10.1016/j.media.2021.102166

Starace, V., Brambati, M., Battista, M., Capone, L., Gorgoni, F., Cavalleri, M., et al. (2020). A lesson not to be forgotten. Ophthalmologists in Northern Italy become internists during the SARS-CoV-2 pandemic. *Am. J. Ophthalmol*. 220, 219–220. doi: 10.1016/j.ajo.2020.04.044

Weller, M., van den Bent, M., Hopkins, K., Tonn, J. C., Stupp, R., Falini, A., et al. (2014). Eano guideline for the diagnosis and treatment of anaplastic gliomas and glioblastoma. *Lancet Oncol*. 15, e395-e403. doi: 10.1016/S1470-2045(14)70011-7

Wright, L. (2019). Ranger - a synergistic optimizer. *GitHub Repos*. Available online at: https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer

Yong, H., Huang, J., Hua, X., and Zhang, L. (2020). Gradient centralization: a new optimization technique for deep neural networks. *arXiv preprint arXiv:2004.01461*. doi: 10.1007/978-3-030-58452-8_37

Yushkevich, P. A., Piven, J., Cody Hazlett, H., Gimpel Smith, R., Ho, S., Gee, J. C., et al. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31, 1116–1128. doi: 10.1016/j.neuroimage.2006.01.015

Zhao, Y.-X., Zhang, Y. M., Song, M., and Liu, C. L. (2019). “Multi-view semi-supervised 3D whole brain segmentation with a self-ensemble network,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Cham: Springer), 256–265. doi: 10.1007/978-3-030-32248-9_29

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 13 January 2022
doi: 10.3389/fmed.2021.774632

The Patterns of Morphological Change During Intracerebral Hemorrhage Expansion: A Multicenter Retrospective Cohort Study

Chang Jianbo[1†], Xiao Ting[2,3†], Chen Yihao[1], Wang Xiaoning[2], Shang Hong[2], Zhang Qinghua[4], Ye Zeju[5], Wang Xingong[6], Tian Fengxuan[7], Chai Jianjun[8], Ma Wenbin[1], Wei Junji[1], Feng Ming[1*], Jianhua Yao[2*] and Wang Renzhi[1*]

[1] Department of Neurosurgery, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China, [2] Tencent AI Lab, Shenzhen, China, [3] Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [4] Department of Neurosurgery, Shenzhen Nanshan Hospital, Shenzhen, China, [5] Department of Neurosurgery, Dongguan People's Hospital, Dongguan, China, [6] Department of Neurosurgery, Linyi People Hospital, Linyi, China, [7] Department of Neurosurgery, Qinghai Provincial People's Hospital, Xining, China, [8] Department of Neurosurgery, Zhangqiu People Hospital, Jinan, China

OPEN ACCESS

Edited by:
Susanne Wegener,
University of Zurich, Switzerland

Reviewed by:
Wenyue Su,
Western University of Health Sciences, United States
Archana Hinduja,
The Ohio State University, United States

***Correspondence:**
Feng Ming
jackietz@163.com
Jianhua Yao
Jianhua.yao@gmail.com
Wang Renzhi
wangrz@126.com

† These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: 12 September 2021
Accepted: 17 December 2021
Published: 13 January 2022

Citation:
Jianbo C, Ting X, Yihao C, Xiaoning W, Hong S, Qinghua Z, Zeju Y, Xingong W, Fengxuan T, Jianjun C, Wenbin M, Junji W, Ming F, Yao J and Renzhi W (2022) The Patterns of Morphological Change During Intracerebral Hemorrhage Expansion: A Multicenter Retrospective Cohort Study. Front. Med. 8:774632. doi: 10.3389/fmed.2021.774632

Objectives: Hemorrhage expansion (HE) is a common and serious condition in patients with intracerebral hemorrhage (ICH). In contrast to the volume changes, little is known about the morphological changes that occur during HE. We developed a novel method to explore the patterns of morphological change and investigate the clinical significance of this change in ICH patients.

Methods: The morphological changes in the hematomas of ICH patients with available paired non-contrast CT data were described in quantitative terms, including the diameters of each hematoma in three dimensions, the longitudinal axis type, the surface regularity (SR) index, the length and direction changes of the diameters, and the distance and direction of movement of the center of the hematoma. The patterns were explored by descriptive analysis and difference analysis in subgroups. We also established a prognostic nomogram model for poor outcomes in ICH patients using both morphological changes and clinical parameters.

Results: A total of 1,094 eligible patients from four medical centers met the inclusion criteria. In 266 (24.3%) cases, the hematomas enlarged; the median absolute increase in volume was 14.0 [interquartile range (IQR), 17.9] mL. The initial hematomas tended to have a more irregular shape, reflected by a larger surface regularity index, than the developed hematomas. In subtentorial and deep supratentorial hematomas, the center moved in the direction of gravity. The distance of center movement and the length changes of the diameters were small, with median values of less than 4 mm. The most common longitudinal axis type was anterior–posterior (64.7%), and the axis type did not change between initial and repeat imaging in most patients (95.2%). A prognostic nomogram model including lateral expansion, a parameter of morphological change, showed good performance in predicting poor clinical outcomes in ICH patients.

Conclusions: The present study provides a morphological perspective on HE using a novel automatic approach. We identified certain patterns of morphological change in HE, and we believe that some morphological change parameters could help physicians predict the prognosis of ICH patients.

Keywords: intracerebral hemorrhage, hemorrhage expansion, anatomy, shape, stroke

INTRODUCTION

Spontaneous intracerebral hemorrhage (sICH) produces mortality or disability in approximately 50% of cases (1, 2), imposing a severe burden (3, 4). Hemorrhage expansion (HE) occurs in approximately one-quarter of sICH patients (2, 5) and is a major determinant of deterioration and death (5–7). Thus, it is important to explore the changes associated with HE (8, 9).

In contrast to the volume changes (5, 10), little is known about the morphological changes that occur in HE (11). Some studies have shown that hemorrhagic lesions expand asymmetrically and non-uniformly, especially in the hyperacute phase (12, 13). However, the patterns of morphological change have not been explored. Additionally, many studies have found that the initial shape of a hematoma was associated with the quality of outcomes (14, 15). However, few studies have focused on the relationship between morphological changes and patient prognosis.

In the current study, we developed and applied a novel approach to explore the patterns of morphological change during HE, which provided a new perspective on hematoma expansion and might help physicians predict the prognosis of ICH patients.

METHODS

Subjects

All data were obtained from the Chinese Intracranial Hemorrhage Image Database (CICHID), which was initiated by Peking Union Medical College Hospital (PUMCH) in February 2019 and supported by the Group of Medical Data, Chinese Medical Doctor Association (16). As of October 2020, the database contained approximately twenty-eight thousand scans from eight thousand patients at twenty-two centers located in Mainland China. All medical records and CT images were anonymized. The CT scans were in Digital Imaging and Communications in Medicine (DICOM) format.

The inclusion criteria were as follows: 1. The cohort from each center included more than 100 patients; 2. the medical records were searchable through the case retrieval system in each center; 3. the patients were adults diagnosed with spontaneous intracerebral hemorrhage (ICH); 4. the patients had one initial and at least one repeat non-contrast computed tomography (NCCT) scan not preceded by surgery; and 5. the initial CT scan was taken within 24 h after symptom onset, and the repeat scan was taken more than 8 but less than 72 h after the initial scan. The exclusion criteria were as follows: 1. The patients were diagnosed with secondary intracranial hemorrhage, such as epidural hemorrhage, subdural hemorrhage, traumatic brain injury, brain tumor, or hemorrhagic transformation of ischemic infarction; 2. the medical records were not available; 3. the hematoma volume in the repeat CT scan was < 3 mL or the volume had decreased by more than 3 mL; and 4. the scans were low-quality images or failed to be registered to the atlas.

The boundary of each hematoma was determined on CT axial slices by a semiautomatic method, in which research assistants independently used the software platform ITK-SNAP 3.6 (17) to correct the boundary drawn by the laboratory's in-house automatic hematoma segmentation software (18). The following clinical characteristics were collected: age, sex, symptom onset time, Glasgow Coma Scale (GCS) score, Glasgow Outcome Scale (GOS) score, initial hematoma volume, location, intraventricular hemorrhage (IVH), and expansion. Both the hematoma boundary and the clinical characteristics were assessed by researchers (WXN, CYH and SH), and any disagreement was reviewed by a neurosurgeon (CJB). Absolute change and relative change were used to describe the change in hematoma volume, and HE was defined as an increase of at least 6 mL or 33% (5, 10).

Measurement of Changes in Hematoma Morphology

The shape irregularity of each hematoma was measured by the surface regularity (SR) index (19, 20), calculated as follows: SR index $= \pi^{1/3}(6V)^{2/3}/A$, where V represents the volume and "A" represents the surface area of the hematoma. The SR index ranges on a continuous scale from 0 (very irregular shape) to 1 (perfectly regular sphere) (21).

Hematoma morphology was characterized in the initial CT by three diameters determined on the slices with the maximum hematoma area in the planes parallel to the coordinate system; these diameters are presented as length (anterior–posterior, AP), width (left–right, LR) and height (superior–inferior, SI) (**Supplementary Figure 1**) (22). The longitudinal axis of each initial hematoma was categorized into one or four types: AP, LR, SI, or no longitudinal axis (NL), determined by which diameter was the longest. The NL group was defined by pairwise ratios ranging between 0.850 and 1.176 for all pairs of diameters, which means that all three diameters were similar (**Supplementary Figure 2**).

The changes in diameters between the initial and repeat scans were described by the length change and direction change. The length change was calculated as the difference between the diameters of the hematoma on the initial and repeat scans. The direction change of the diameters was defined by the axis that showed the largest absolute change in length, categorized as AP, LR, SI or no direction change (**Figure 2**). "No direction change" was defined by similar length changes in two directions or in all three directions.

The geometric center was defined as the centroid of the largest connected region of the hematoma in 3D space. The distance of center movement was defined as the spatial distance between the geometric center locations of the hematoma on the initial scan and the repeat scan. The direction of center movement is presented as the projection of the vector's direction in the standard planes (axial, coronal and sagittal). The directions of center movement from all cases in the same anatomical region were synthesized into one arrow and visualized in an atlas (ICBM 2009c Non-linear Symmetric template, MNI152) (23) (**Supplementary Figure 3**).

Image Processing

To compare morphological changes among different cases, the paired CT scans were registered to an atlas. After the skull was stripped away and the brain was extracted using BET (24), each pair of initial and repeat CT scans from the same patient was spatially registered to a common atlas using the MNI152 template (23, 25). The registration pipeline consists of two sequential linear registrations and two sequential non-linear registrations, where both registration tools were provided by Advanced Normalization Tools (ANTs), and the non-linear registration was based on a B-spline function (26). The Dice coefficient between the registered CT and the template was calculated after each registration. If the Dice value did not reach the predetermined threshold (0.93), the registration was considered a failure, and the pair of scans was excluded. To ensure the registration quality, a predetermined threshold was applied by visually checking the registered CT quality. After registration to a common atlas, all CTs and their hemorrhage masks were in the same template space. The synthesis to determine the direction of center movement was performed by the package Mayavi (27). The surface area, volume, geometric center and diameters were calculated by the image processing package Skimage (28).

Prediction Model for Poor Outcomes (GOS $\leq$ 3) at Discharge

To explore the clinical significance of changes in hematoma morphology, we conducted multivariate logistic regression incorporating clinical parameters and hematoma morphological change parameters. Multiple logistic regression was used to select the most useful predictive variables for poor outcomes (GOS $\leq$ 3) at discharge. All useful predictors, defined as those with $P < 0.05$, were used to develop the final multivariate logistic regression, and a nomogram was then built to predict which ICH patients could have poor outcomes (29). Discrimination was evaluated by the area under the curve (AUC) value of the receiver operating characteristic curve (ROC), and calibration was measured by the calibration curve (30).

Statistical Analysis

Baseline characteristics were summarized as counts [percentages (%)] for categorical variables and the mean [standard deviation (SD)] or median [interquartile range (IQR)] for continuous variables. Both clinical and morphological change parameters were investigated by descriptive analysis and difference analysis in subgroups (expansion and longitudinal axis type). A two-sided Pearson's chi-squared test or Fisher's exact test was used for categorical variables, and Student's *t*-test, ANOVA or the Mann-Whitney U test was used for continuous variables. The threshold for statistical significance was set to 0.05. Statistical analyses were conducted with SPSS Statistics (version 21.0.0, IBM, Armonk, New York) and R (version 3.6.3, R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

The database contained 3,231 ICH patients who underwent repeat CT between Jan. 1, 2016, and Aug. 30, 2020, at 4 medical centers; 1,094 of these patients met the inclusion criteria to be analyzed (**Figure 1**, **Supplementary Table 1**). The baseline characteristics of the patients are summarized in **Table 1**. Most included patients were elderly males with ICH in deep supratentorial brain regions. HE occurred in 24.3% of patients, with median absolute and relative increases of 14.0 (IQR 17.9) mL and 53.4 (IQR 97.3) %, respectively. The initial volume in the expansion group was larger than that in the non-expansion group (25.9 vs. 18.7 mL, $P < 0.05$). To compare the morphological changes, the CTs were registered to an atlas, as shown in **Supplementary Figure 4**.

The morphological characteristics of HE are shown in **Table 2**. The hematomas became more irregular in repeat CTs in both the expansion and non-expansion groups, with the median SR index decreasing from 0.542 to 0.515. The median change in the SR index was−0.027, and the change in this index was significantly larger in the expansion group (-0.049) than in the non-expansion group (-0.020, $P < 0.05$).

The most common type of longitudinal axis was the AP direction (64.7%), followed by the SI direction (19.5%). A total of 10.3% of hematomas had an approximately spherical shape and were categorized as the NL type in this study. There was no significant difference in the longitudinal axis types between the expansion and non-expansion groups (**Table 2**). The length change of the diameters was small, and the largest change was in the AP direction, with an increase of 4.8 mm in the expansion group (**Figure 2**). The distance of center movement was small, with a median of 3.5 mm for all patients and 6.1 mm in the expansion group.

We investigated the patterns of change in the diameters and geometric center for different longitudinal axis types (**Table 3**). Regardless of the longitudinal axis type, the diameter direction change was mostly AP [43.1% (471/1094) in total, ranging from 38.3 to 45.2%], followed by LR (ranging from 23.9 to 31.0%). The length change of the diameters was < 3.3 mm in most cases.

The distance of center movement was small, ranging from 2.3 to 4.1 mm. As **Figure 3** shows, the direction of center movement in deep supratentorial hematomas was in the direction of gravity as patients lay in a supine position. A similar pattern was also observed in subtentorial hematoma; however, it did not exist in

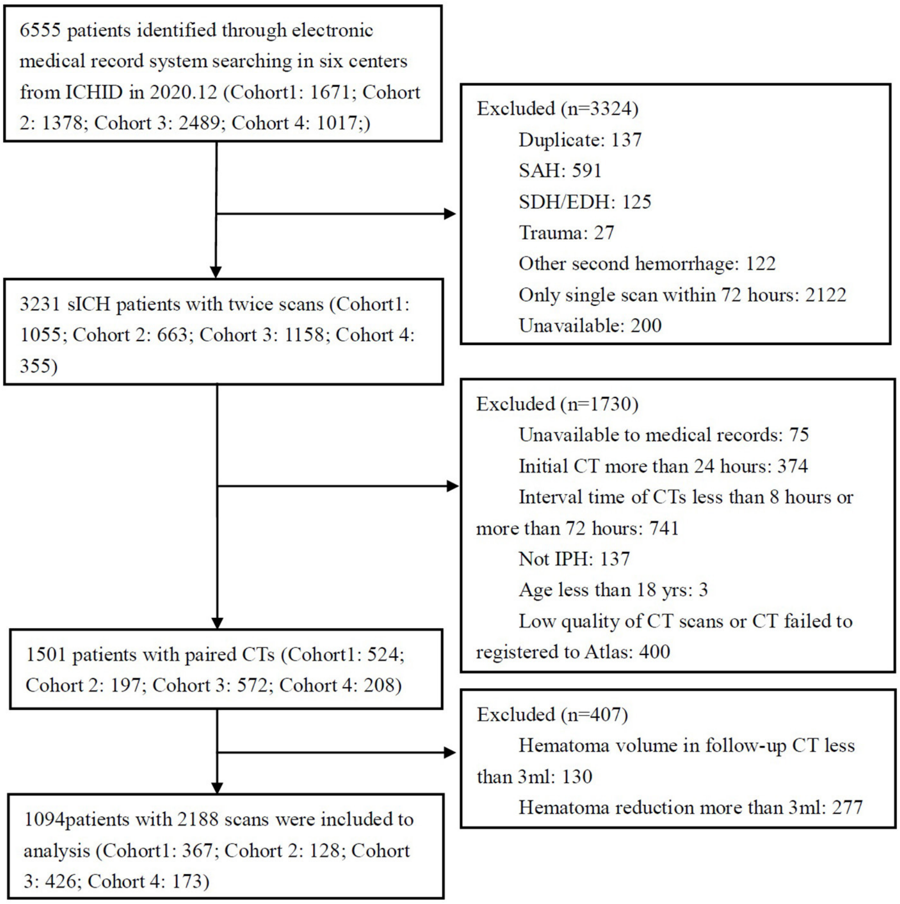

6555 patients identified through electronic medical record system searching in six centers from ICHID in 2020.12 (Cohort1: 1671; Cohort 2: 1378; Cohort 3: 2489; Cohort 4: 1017;)
Excluded (n=3324)
Duplicate: 137
SAH: 591
SDH/EDH: 125
Trauma: 27
Other second hemorrhage: 122
Only single scan within 72 hours: 2122
Unavailable: 200
3231 sICH patients with twice scans (Cohort1: 1055; Cohort 2: 663; Cohort 3: 1158; Cohort 4: 355)
Excluded (n=1730)
Unavailable to medical records: 75
Initial CT more than 24 hours: 374
Interval time of CTs less than 8 hours or more than 72 hours: 741
Not IPH: 137
Age less than 18 yrs: 3
Low quality of CT scans or CT failed to registered to Atlas: 400
1501 patients with paired CTs (Cohort1: 524; Cohort 2: 197; Cohort 3: 572; Cohort 4: 208)
Excluded (n=407)
Hematoma volume in follow-up CT less than 3ml: 130
Hematoma reduction more than 3ml: 277
1094patients with 2188 scans were included to analysis (Cohort1: 367; Cohort 2: 128; Cohort 3: 426; Cohort 4: 173)

FIGURE 1 | Flowchart of patient selection.

TABLE 1 | Baseline characteristics of ICH patients.

Variable	**Total (*N* = 1,094)**	**Expansion group (*N* = 266)[a]**	**Non-expansion group (*N* = 828)**	***P*-value**
Male, *n* (%)	698 (63.8)	195 (73.3)	503 (60.7)	*
Age, median (IQR), y	61.0 (18.0)	59.0 (19.0)	61.0 (18.0)	
Onset to CT, median (IQR), hr	3.0 (4.0)	3.0 (3.0)	3.0 (5.0)	*
Time interval between CT scans, median (IQR), hr	22.8 (19.8)	22.4 (20.2)	22.9 (20.0)	
GCS score, median (IQR)	14 (4)	13 (5)	14 (3)	*
GOS score, median (IQR)	3 (1)	3 (1)	3 (1)	*
Initial hematoma volume, median (IQR), mL	20.2 (25.4)	25.9 (33.2)	18.7 (23.0)	*
IVH, *n* (%)	374 (34.2)	91 (34.2)	283 (34.2)	
Hematoma location				*
Deep, *n* (%)	763 (69.8)	176 (66.2)	587 (71.0)	
Lobar, *n* (%)	236 (21.6)	75 (28.2)	161 (19.5)	
Subtentorial, *n* (%)	94 (8.6)	15 (5.6)	79 (9.5)	
Absolute change in hematoma volume, median (IQR), mL	0.9 (5.5)	14.0 (17.9)	0.4 (2.4)	*
Percentage change in hematoma volume, median (IQR), %	4.5 (24.4)	53.4 (97.3)	6.8 (12.4)	*

[a] *Expansion was defined as a volume change ≥ 6 mL or 33%. *P < 0.05. IQR, interquartile range; CT, computed tomography; GCS, glasgow coma scale; GOS, glasgow outcome scale; IVH, intraventricular hemorrhage.*

TABLE 2 | Morphological characteristics of hematoma expansion.

	Total ($N = 1,094$)	Expansion ($n = 266$)	Non-expansion ($n = 828$)	*P*-value
SR index on admission, mean (SD)	0.542 (0.104)	0.536 (0.106)	0.544 (0.104)	
SR index on follow-up, mean (SD)	0.515 (0.101)	0.487 (0.098)	0.524 (0.101)	
SR index change, mean (SD)	−0.027 (0.073)	−0.049 (0.090)	−0.020 (0.065)	*
Longitudinal axis type of hematoma on admission, *n* (%)				
AP	708 (64.7)	175 (65.8)	533 (64.4)	
LR	60 (5.5)	9 (3.4)	51 (6.2)	
SI	213 (19.5)	55 (20.7)	158 (19.1)	
NL	113 (10.3)	27 (10.2)	86 (10.4)	
Diameters of hematoma on admission, mean (SD), mm				
Length (AP)	60.54 (23.3)	64.1 (25.0)	59.4 (22.6)	*
Width (LR)	43.9 (15.0)	46.9 (17.2)	43.0 (14.1)	*
Height (SI)	55.3 (15.8)	58.8 (16.6)	54.2 (15.4)	*
Length change of hematoma diameters, mean (SD), mm				
AP	2.8 (9.8)	4.8 (13.0)	2.1 (8.5)	*
LR	1.9 (9.0)	2.8 (11.3)	1.6 (8.1)	*
SI	−0.5 (9.2)	−1.2 (16.8)	−0.4 (4.7)	*
Distance of center movement, mean (SD), mm	3.5 (5.4)	6.1 (8.2)	2.7 (3.7)	*

$^*P < 0.05$. AP, anterior-posterior; LR, left-right; SI, superior-inferior; NL, no longitudinal axis; SD, standard deviation; SR, surface regularity.

some supratentorial lobar hematomas, such as those in the frontal lobe, parietal lobe and occipital lobe (**Figure 3**, Attachment 1).

Although most changes in diameter length were small, there were 139 cases with obvious changes, where the changes in diameter length and the distance of center movement were both greater than 10 mm, or the direction change was inconsistent with the longitudinal type. However, among these cases with obvious changes, only 53 (38.1%) patients' longitudinal axis types were changed, accounting for 4.84% of all patients (53/1,094).

To explore the clinical significance of the morphological change in HE, we established a prognostic nomogram to predict poor outcomes (GOS $\leq$ 3). Eight potential predictors (age, volume, location, GCS, hematoma expansion, initial SR index, hematoma diameter length, and length change of the LR diameter) were selected from 19 collected variables by using multivariate logistic regression (**Supplementary Table 2**). Then, the logistic regression analysis was visualized as a nomogram (**Figure 4A**), which was preliminarily built to predict the probability of poor outcome in ICH patients. ROC curve analysis indicated that the nomogram performed well in prognostic prediction, with an AUC of 0.824 (95% CI 0.800, 0.846). The calibration plot also showed excellent agreement between the nomogram predictions and actual observations in ICH patients with GOS $\leq$ 3 (**Figure 4**). In particular, the length change of the LR diameter (the lateral expansion) was a morphological change factor that contributed strongly to the model, with an odds ratio of 1.1386 (95% CI: 1.0216, 1.2691). All these findings suggested that our prediction model including morphological change parameters had good performance in predicting poor clinical outcomes in ICH patients. This scale should remind physicians to pay attention to lateral expansion, especially in ICH patients who are predicted to have GOS $\leq$ 3.

DISCUSSION

In this study, we investigated the patterns of morphological change in sICH based on a large cohort. We found that the initial hematoma tended to be more irregularly shaped, with a larger SR index, than the developed hematoma. In deep supratentorial hematomas and subtentorial hematomas, the direction of center movement was toward the pull of gravity. Most hematomas had their longitudinal axis in the AP direction (64.7%), and the direction of the diameter change was AP in approximately 40% of patients. The length change of the diameters and the distance of center movement were < 4 mm. The longitudinal axis type did not change between the initial and repeat CT scans in most patients. In addition, one morphological change parameter, the length change of the diameter in the LR direction (lateral expansion), was found to be associated with poor prognosis in ICH patients. The prediction model including lateral expansion for ICH patients with poor prognosis showed good performance. The results of our analysis provide a new perspective on hematoma expansion in terms of morphological changes. Physicians should take these results as a reminder not to ignore ICH with lateral expansion.

The SR index quantifies irregularity by the ratio of surface area to volume (20, 21). As in previous studies (21), the SR index decreased in the repeat hematoma group, especially in the expansion group, which indicated that the hematomas tended to become more irregular as they developed. Previous studies showed that the efficacy of hematoma evacuation surgery decreased in irregular hematoma (31), and our findings reminded the surgeon to consider the tendency of hematoma to be more irregular before making surgical decisions.

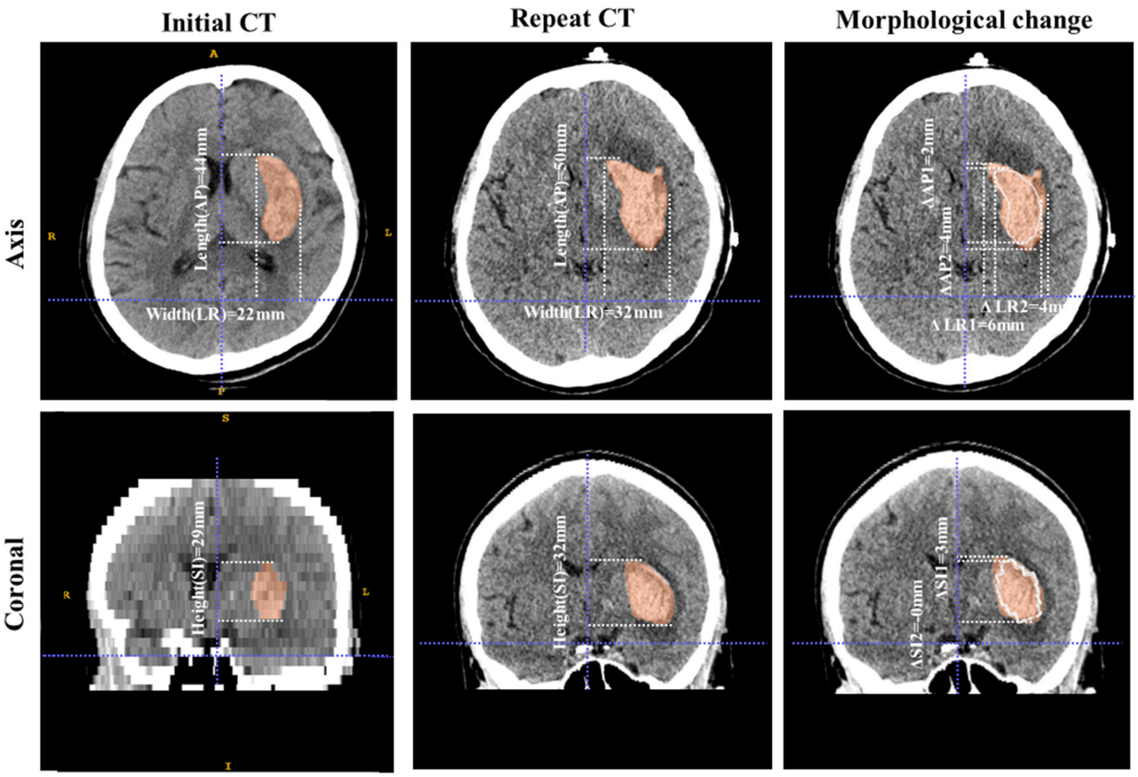


FIGURE 2 | Example of morphological change in a hematoma. These images are from a 73-year-old male patient with ICH. The volume of the hematoma was 15.6 mL on initial CT and 23.5 mL on repeat CT. The first column shows the shape characteristics of the initial hematoma, including its three diameters (length 44 mm, width 22 mm, and height 29 mm). The longitudinal axis is of the AP type and the SR index is 0.628. The second column shows the shape characteristics of the hematoma on the repeat scan, including its three diameters (length 50 mm, width 32 mm, and height 32 mm). The longitudinal axis is of the AP type and the SR index is 0.552. The third column shows the morphological change. The white line is the contour of the initial hematoma and the red area is the hematoma as of the repeat scan. The length changes of the hematoma diameters are 6 mm, 10 mm, and 3 mm in the AP, LR, and SI directions, respectively. The direction change of the hematoma diameters is LR and the longitudinal axis type does not change (AP). The SR index decreases, which means that the hematoma becomes more irregular from initial CT to repeat CT.

TABLE 3 | Morphological changes by longitudinal axis type.

	Hematoma longitudinal axis type			
	AP (*n* = 708)	**LR (*n* = 60)**	**SI (*n* = 213)**	**NL (*n* = 113)**
Direction change of hematoma diameters, *n* (%)				
AP	320 (45.2)	23 (38.3)	83 (39.0)	45 (39.8)
LR	197 (27.8)	16 (26.7)	66 (31.0)	27 (23.9)
SI	88 (12.4)	12 (20.0)	21 (9.9)	21 (18.6)
No direction change	103 (14.5)	9 (15.0)	43 (20.2)	20 (17.7)
Length change of hematoma diameters, mean (SD), mm				
AP	3.0 (10.5)	1.8 (6.7)	2.2 (7.4)	3.3 (10.9)
LR	1.8 (8.2)	0.4 (4.0)	2.9 (13.2)	0.9 (5.1)
SI	−0.3 (6.9)	−0.8 (4.4)	−0.6 (6.4)	−1.7 (21.0)
Distance of center movement, mean (SD), mm	3.6 (3.5)	2.3 (2.0)	3.4 (7.0)	4.1 (10.4)

AP, anterior-posterior; LR, left-right; SI, superior-inferior; NL, no longitudinal axis; SD, standard deviation.

Similar to previous studies (11–13, 22), our study showed that hematoma growth was asymmetric and that the geometric center moved in HE. Furthermore, we found a pattern regarding center movement. In subtentorial and deep supratentorial hematomas, the center tended to move in the direction of gravity. Although the center moved only a short distance, researchers should take

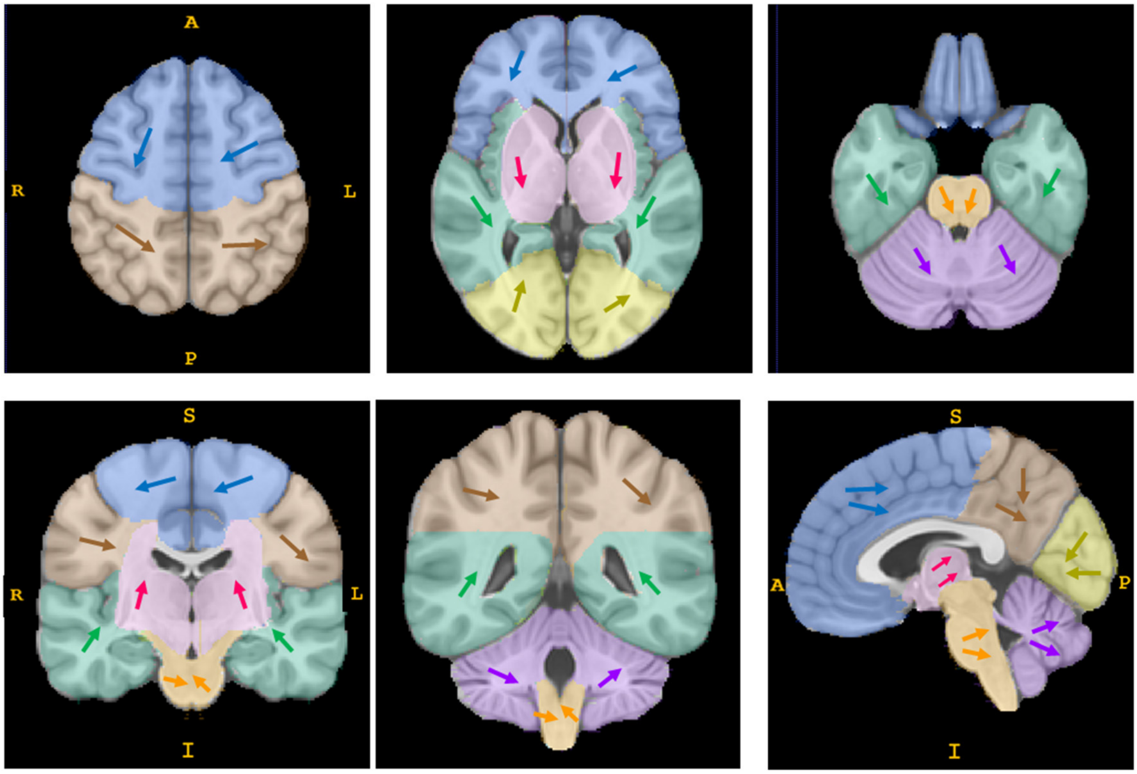


FIGURE 3 | The direction of movement of the geometric center of hematomas. Each arrow of a different color represents the synthesized direction of center movement in a different anatomical region (the frontal lobe is shown in blue, the parietal lobe in brown, the insula and temporal lobe in green, the occipital lobe in yellow, the basal ganglia/thalamus area in pink, the brain stem in orange, and the cerebellum in purple). The direction of center movement in deep supratentorial regions (basal ganglia/thalamus area) and subtentorial regions (brain stem and cerebellum) was in the direction of gravity as patients lay in a supine position; some supratentorial lobar hematomas showed no such pattern. This is a schematic diagram; the details are shown in attachment 1.

this movement as a reminder to consider the effect of gravity when studying the pathophysiological mechanism of acute-phase hematoma formation and expansion.

In this study, we found that the longitudinal axis type did not change in 90% of HE cases. Although the direction change of the diameters did not always align with the longitudinal axis, the length changes of the diameters were ordinarily < 4 mm, which is insufficient to change the longitudinal axis type. The median distance of center movement was only 3.5 mm, such that the change would not influence the drainage trajectory traversing the epicenter of the hematoma (31, 32). These findings alleviate the concern as to whether the longitudinal direction or geometric center would change after drainage surgery for HE (33, 34).

Another interesting finding in our study was that lateral expansion (the length change of the diameter in the LR direction) was associated with poor outcomes. As important factors in HE and prognosis, the shape features of the initial hematoma have been described by various methods and demonstrated to be associated with outcomes. These studies have included qualitative analytical variables, such as Barras grade, island sign and satellite sign (8, 14, 20), and quantitative analytical variables, such as the SR index, compactness, Fourier factor and fractal dimension (15, 21, 35). The present study is the first to relate changes in hematoma morphology to the patient's prognosis. The prediction model showed that the risk of a poor outcome increased by a factor of 1.139 for every 1 mm of lateral expansion. This finding should remind physicians not to ignore lateral expansion, especially in ICH patients who are predicted to have poor outcomes.

This study has several limitations. For the purposes of maintaining analytic rigor, we excluded patients who had undergone any invasive operation before repeat CT, as these interventions directly affect hematoma shape. We also excluded patients with time intervals of < 8 h or more than 72 h between scans, as the shape of a hematoma tends to be most stable from 8 to 72 h. The volume of the lesion may significantly increase in the first 8 h, and absorption starts after 72 h (34). These criteria may bias our sample toward populations less severely affected by their hemorrhage. In addition, the absolute value of the length change was affected by the accuracy of the registration process. Although we strictly excluded failed registration (36) and manually checked each case, there might still have been some registration error. Considering that the actual length change was small, this registration error may have an adverse effect on the accuracy of the calculated absolute value of the length change. Moreover, the longitudinal axis in our study was defined as the

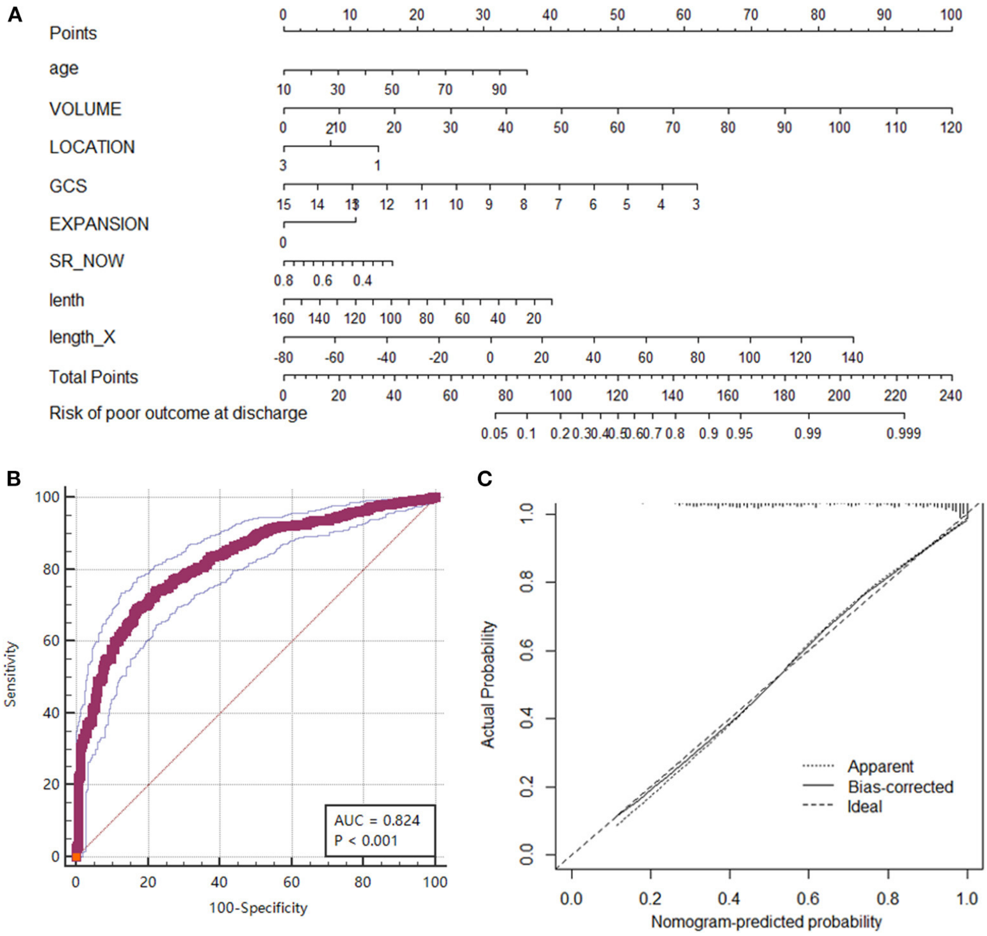


FIGURE 4 | Nomogram of the prognostic model for predicting poor outcomes (GOS ≤ 3) at discharge. **(A)** The nomogram was developed from a multivariable logistic regression model based on age, volume, location, GCS, hematoma expansion, initial SR index, diameter lengths, and length change in the LR direction. **(B)** ROC curve of the nomogram representing the discrimination performance of the model. **(C)** Calibration curve of nomogram. A calibration curve depicts the calibration of a model in terms of the agreement between the predicted risk of a poor outcome and the outcome actually observed. The Y-axis represents the actual poor-outcome rate. The X-axis represents the predicted poor-outcome risk. The diagonal dotted line represents a perfect prediction by an ideal model. The solid black line represents the performance of the nomogram, where a closer fit to the diagonal dotted line represents a better prediction.

longest of the three diameters. These diameters were measured parallel to the coordinate system to compare the direction changes across different hematomas (**Supplementary Figure 1**). Thus, the longitudinal axis as defined here is merely the projection of the actual longitudinal axis onto the coordinate axis that best approximates its direction. This imperfect definition might limit the interpretability of the findings. Finally, while our study included more than one thousand patients who underwent repeat imaging, all the patients were from a single country, China. The retrospective nature of this multicenter analysis is another limitation. Further validation must be carried out with independent data sets to ensure generalizability.

In conclusion, the present study provides a morphological perspective on hematoma expansion by a novel approach. We identified certain patterns of morphological change in HE. As hematomas enlarged, they shifted in the direction of gravity and tended to be more irregular. The most common longitudinal axis type of hematoma was AP, which did not change during HE. Based on our findings, we used morphological change parameters to establish a novel, promising prognostic nomogram model for the individualized prediction of poor outcomes in ICH patients. This nomogram requires further validation in other centers.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board of Peking Union Medical College Hospital. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

CJB and XT were the project lead for the current study and contributed equally to the study. CY, WXia, ZQ, YZ, WXin, TF, and CJB collected and extracted data. CJB, XT, and SH conducted data and statistical analysis. CJB wrote the manuscript. FM, MW, and WJ provided significant advice about the manuscript. WR and JY reviewed and revised the manuscript. WR, FM, and JY are co-corresponding authors. All authors read and approved the final manuscript.

FUNDING

This study has received funding by National Key R&D Program of China (2018YFA0108603), Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Science (2020-I2M-C&T-B-031 and 2020-I2M-C&T-B-028), Beijing Tianjin Hebei basic research cooperation project [19JCZDJC64600 (Z)], and Huazhong University of Science and Technology Union Shenzhen Hospital Fund (NS202001).

ACKNOWLEDGMENTS

We thank the Chinese Intracranial Hemorrhage Image Database (CICHID) and the Group of Medical Data, Chinese Medical Doctor Association for data support and assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2021.774632/full#supplementary-material

REFERENCES

1. Qureshi AI, Palesch YY, Barsan WG, Hanley DF, Hsu CY, Martin RL, et al. Intensive blood-pressure lowering in patients with acute cerebral hemorrhage. *N Engl J Med.* (2016) 375:1033–43. doi: 10.1056/NEJMoa1603460
2. Anderson CS, Heeley E, Huang Y, Wang J, Stapf C, Delcourt C, et al. Rapid blood-pressure lowering in patients with acute intracerebral hemorrhage. *N Engl J Med.* (2013) 368:2355–65. doi: 10.1056/NEJMoa1214609
3. Global, regional, and national burden of stroke, 1990-2016: a systematic analysis for the global burden of disease study 2016. *Lancet Neurol.* (2019) 18:439–58. doi: 10.1016/S1474-4422(19)30034-1
4. Feigin VL, Krishnamurthi RV, Parmar P, Norrving B, Mensah GA, Bennett DA, et al. Update on the global burden of ischemic and hemorrhagic stroke in 1990-2013: the GBD 2013 study. *Neuroepidemiology.* (2015) 45:161–76. doi: 10.1159/000441085
5. Dowlatshahi D, Demchuk AM, Flaherty ML, Ali M, Lyden PL, Smith EE. Defining hematoma expansion in intracerebral hemorrhage: relationship with patient outcomes. *Neurology.* (2011) 76:1238–44. doi: 10.1212/WNL.0b013e3182143317
6. Davis SM, Broderick J, Hennerici M, Brun NC, Diringer MN, Mayer SA, et al. Hematoma growth is a determinant of mortality and poor outcome after intracerebral hemorrhage. *Neurology.* (2006) 66:1175–81. doi: 10.1212/01.wnl.0000208408.98482.99
7. Li Z, You M, Long C, Bi R, Xu H, He Q, et al. Hematoma expansion in intracerebral hemorrhage: an update on prediction and treatment. *Front Neurol.* (2020) 11:702. doi: 10.3389/fneur.2020.00702
8. Morotti A, Boulouis G, Dowlatshahi D, Li Q, Barras CD, Delcourt C, et al. Standards for detecting, interpreting, and reporting noncontrast computed tomographic markers of intracerebral hemorrhage expansion. *Ann Neurol.* (2019) 86:480–92. doi: 10.1002/ana.25563
9. Hemphill JC 3rd, Greenberg SM, Anderson CS, Becker K, Bendok BR, Cushman M, et al. Guidelines for the management of spontaneous intracerebral hemorrhage: a guideline for healthcare professionals from the American heart association/American stroke association. *Stroke.* (2015) 46:2032–60. doi: 10.1161/STR.0000000000000069
10. Yogendrakumar V, Ramsay T, Fergusson DA, Demchuk AM, Aviv RI, Rodriguez-Luna D, et al. Redefining hematoma expansion with the inclusion of intraventricular hemorrhage growth. *Stroke.* (2020) 51:1120–7. doi: 10.1161/STROKEAHA.119.027451
11. Schlunk F, Greenberg SM. The pathophysiology of intracerebral hemorrhage formation and expansion. *Transl Stroke Res.* (2015) 6:257–63. doi: 10.1007/s12975-015-0410-1
12. Boulouis G, Dumas A, Betensky RA, Brouwers HB, Fotiadis P, Vashkevich A, et al. Anatomic pattern of intracerebral hemorrhage expansion: relation to CT angiography spot sign and hematoma center. *Stroke.* (2014) 45:1154–6. doi: 10.1161/STROKEAHA.114.004844
13. Edlow BL, Bove RM, Viswanathan A, Greenberg SM, Silverman SB. The pattern and pace of hyperacute hemorrhage expansion. *Neurocrit Care.* (2012) 17:250–4. doi: 10.1007/s12028-012-9738-5
14. Morotti A, Arba F, Boulouis G, Charidimou A. Noncontrast CT markers of intracerebral hemorrhage expansion and poor outcome: a meta-analysis. *Neurology.* (2020) 95:632–43. doi: 10.1212/WNL.0000000000010660
15. Kliś KM, Krzyzewski RM, Kwinta BM, Stachura K, Popiela TJ, Gasowski J, et al. Relation of intracerebral hemorrhage descriptors with clinical factors. *Brain sciences.* (2020) 10:252. doi: 10.3390/brainsci10040252
16. Wang RenZhi, Chang JianBo, Ming F. Prospects for precious diagnosis, assessment, prediction and treatment of hemorrhagic stroke. *Zhongguo Xian Dai Shen Jing Ji Bing Za Zhi.* (2019) 19:618–21. doi: 10.3969/j.issn.1672?6731.2019.09.003
17. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage.* (2006) 31:1116–28. doi: 10.1016/j.neuroimage.2006.01.015
18. Çicek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016*. Athens: Springer, Cham (2016).
19. Pérez-Beteta J, Molina-García D, Ortiz-Alhambra JA, Fernández-Romero A, Luque B, Arregui E, et al. Tumor surface regularity at MR imaging predicts survival and response to surgery in patients with glioblastoma. *Radiology.* (2018) 288:218–25. doi: 10.1148/radiol.2018171051
20. Oge DD, Topcuoglu MA, Gocmen R, Arsava EM. The dynamics of hematoma surface regularity and hematoma expansion in acute intracerebral hemorrhage. *J Clin Neurosci.* (2020) 74:160–3. doi: 10.1016/j.jocn.2020.01.081
21. Salazar P, Di Napoli M, Jafari M, Jafarli A, Ziai W, Petersen A, et al. Exploration of multiparameter hematoma 3D image analysis for predicting outcome after intracerebral hemorrhage. *Neurocrit Care.* (2020) 32:539–49. doi: 10.1007/s12028-019-00783-8
22. Liu R, Huynh TJ, Huang Y, Ramsay D, Hynynen K, Aviv RI. Modeling the pattern of contrast extravasation in acute intracerebral hemorrhage using dynamic contrast-enhanced MR. *Neurocrit Care.* (2015) 22:320–4. doi: 10.1007/s12028-014-0071-z

23. Fonov VS, Evans AC, McKinstry RC, Almli CR, Collins DL. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage.* (2009) 47:S102. doi: 10.1016/S1053-8119(09)70884-5
24. Akkus Z, Kostandy P, Philbrick KA, Erickson BJ. Robust brain extraction tool for CT head images. *Neurocomputing.* (2020) 392:189–95. doi: 10.1016/j.neucom.2018.12.085
25. Fonov V, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*. (2011) 54:313–27. doi: 10.1016/j.neuroimage.2010.07.033
26. Tustison NJ, Cook PA, Klein A, Song G, Das SR, Duda JT, et al. Large-scale evaluation of ANTs and freesurfer cortical thickness measurements. *NeuroImage*. (2014) 99:166-—79. doi: 10.1016/j.neuroimage.2014.05.044
27. Ramachandran P, Varoquaux G. Mayavi: 3D visualization of scientific data. *Comput Sci Eng.* (2011) 13:40–51. doi: 10.1109/MCSE.2011.35
28. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in python. *PeerJ.* (2014) 2:e453. doi: 10.7717/peerj.453
29. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat Med.* (2007) 26:5512–28. doi: 10.1002/sim.3148
30. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA*. (2017) 318:1377–84. doi: 10.1001/jama.2017.12126
31. Awad IA, Polster SP, Carrión-Penagos J, Thompson RE, Cao Y, Stadnik A, et al. Surgical performance determines functional outcome benefit in the minimally invasive surgery plus recombinant tissue plasminogen activator for intracerebral hemorrhage evacuation (MISTIE) procedure. *Neurosurgery*. (2019) 84:1157–68. doi: 10.1093/neuros/nyz077
32. Scaggiante J, Zhang X, Mocco J, Kellner CP. Minimally invasive surgery for intracerebral hemorrhage. *Stroke.* (2018) 49:2612–20. doi: 10.1161/STROKEAHA.118.020688
33. de Oliveira Manoel AL. Surgery for spontaneous intracerebral hemorrhage. *Critical Care*. (2020) 24:45. doi: 10.1186/s13054-020-2749-2
34. Garg R, Biller J. Recent advances in spontaneous intracerebral hemorrhage. F1000Res. (2019) 8:F1000. doi: 10.12688/f1000research.16357.1
35. Wang CW, Liu YJ, Lee YH, Hueng DY, Fan HC, Yang FC, et al. Hematoma shape, hematoma size, Glasgow coma scale score and ICH score: which predicts the 30-day mortality better for intracerebral hematoma? *PLoS ONE.* (2014) 9:e102326. doi: 10.1371/journal.pone.0102326
36. Rister B, Horowitz MA, Rubin DL. Volumetric image registration from invariant keypoints. *IEEE Trans Image Process.* (2017) 26:4900–10. doi: 10.1109/TIP.2017.2722689

Conflict of Interest: JY, XT, SH, and WXia are employee of Tencent AI Lab.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 17 March 2022
doi: 10.3389/fmed.2022.797586

Clinically Deployed Computational Assessment of Multiple Sclerosis Lesions

Siddhesh P. Thakur[1,2,3], *Matthew K. Schindler*[4], *Michel Bilello*[1,2*†] *and Spyridon Bakas*[1,2,3*†]

[1] Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA, United States, [2] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [3] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [4] Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

Multiple Sclerosis (MS) is a demyelinating disease of the central nervous system that affects nearly 1 million adults in the United States. Magnetic Resonance Imaging (MRI) plays a vital role in diagnosis and treatment monitoring in MS patients. In particular, follow-up MRI with T2-FLAIR images of the brain, depicting white matter lesions, is the mainstay for monitoring disease activity and making treatment decisions. In this article, we present a computational approach that has been deployed and integrated into a real-world routine clinical workflow, focusing on two tasks: (a) detecting new disease activity in MS patients, and (b) determining the necessity for injecting Gadolinium Based Contract Agents (GBCAs). This computer-aided detection (CAD) software has been utilized for the former task on more than 19, 000 patients over the course of 10 years, while its added function of identifying patients who need GBCA injection, has been operative for the past 3 years, with $>85\%$ sensitivity. The benefits of this approach are summarized in: (1) offering a reproducible and accurate clinical assessment of MS lesion patients, (2) reducing the adverse effects of GBCAs (and the deposition of GBCAs to the patient's brain) by identifying the patients who may benefit from injection, and (3) reducing healthcare costs, patients' discomfort, and caregivers' workload.

Keywords: Multiple Sclerosis, clinical setting, machine learning, deep learning, gadolinium

OPEN ACCESS

Edited by:
Aaron Carass,
Johns Hopkins University,
United States

Reviewed by:
Blake Dewey,
Johns Hopkins University,
United States
Pouria Tohidi,
Johns Hopkins University,
United States

Correspondence:
Spyridon Bakas
sbakas@upenn.edu
Michel Bilello
mbilello@upenn.edu

†These authors share senior authorship

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *18 October 2021*
Accepted: *17 February 2022*
Published: *17 March 2022*

Citation:
Thakur SP, Schindler MK, Bilello M and Bakas S (2022) Clinically Deployed Computational Assessment of Multiple Sclerosis Lesions. Front. Med. 9:797586. doi: 10.3389/fmed.2022.797586

1. INTRODUCTION

Multiple Sclerosis (MS) is a chronic immune-mediated disease that affects the central nervous system (CNS) with a complex pathophysiology. The prevalence of MS in the United States of America is reported as approximately 1 million adults (1), with several more million patients recorded worldwide.

MRI is an essential tool in the diagnosis and treatment monitoring of MS. Patients with MS typically undergo annual follow up MRI scanning that commonly includes T1 post-contrast images to assess for subclinical disease, i.e., formation of new focal demyelinating lesions in the absence of clinical symptoms. Previous work has shown that in the absence of a new T2-weighted Fluid-Attenuated-Inversion-Recovery (T2-FLAIR) lesion, contrast does not usually add additional clinical information to the interpretation of the scan (2). A common challenge in the clinical assessment of MS is relying on visual interpretation of images, particularly in the case of high lesion burden, to determine if new lesions developed. Automated techniques could aid clinicians in their visualization of new MS lesions improving efficiency and confidence in clinical decisions.

Here, we present a computer-aided-detection (CAD) approach that uses machine learning techniques to detect changes in white matter brain lesions on MRI scans of patients with MS. The presented system has been deployed and fully integrated into the routine clinical workflow for 10 years. Although it was initially designed and used solely as a neuroradiological aid in detecting MS lesion burden changes, this CAD approach is also contributing to reducing the number of gadolinium injections in the MS patient population. Specifically, our CAD system focuses on two targets: (i) assessing MS lesion burden changes from a given previous time-point; and (ii) reducing the administration of GBCAs, based on the detection of new/growing lesions. In a nutshell, the development of our CAD system initiated in 2009 and following its offline evaluation it is clinically translated and integrated in the routine clinical workflow since 2012 focusing solely on its first target. In 2019, the same CAD system was further successfully evaluated and integrated to the clinical workflow for the reduction of GBCA administration.

2. LITERATURE REVIEW

According to related literature (3), computational methods for the assessment of MS, can be divided into two categories: (1) lesion detection, and (2) lesion-change detection. As shown in **Figure 1**, the lesion detection approach detects both static and dynamic MS lesions on a given single-time MRI volume. These segmentation methods are usually supervised and rely on distinguishing hyperintense lesions from normal appearing white matter tissue in the brain. The lesion-change detection is a longitudinal analysis of volumes taken at different time-points, and a lesion quantification approach is required to see the lesion changes quantitatively (5). A lesion-change occurs as a result of "*tissue transformation*" or even "*tissue deformation*" (6). "*Tissue transformation*" in the context of MS lesions refers to the change in signal intensity within a MS lesion (after accounting for acquisition differences), whereas "*tissue deformation*" refers to surrounding tissue changes as a result of the lesion's expansion or contraction. Neurologists referring patients for a follow-up MRI, want to know if new lesions have formed since the previous timepoint. This information may prompt neurologists to modify the treatment regimen, in order to avoid future recurrences.

The current clinical routine to detect new white matter lesions is based on the visual observation and longitudinal comparison of T2-FLAIR MRI brain scans, by neuroradiologists, from current and previous sessions. However, the typical acquisition protocol for MS patients includes high-resolution 3-D MRI scans, which render this manual reviewing process a tedious and time-consuming task. The current clinical practise based on visual observation can be inaccurate if there are large angulation differences between the two studies or at times, when particular 3-D protocols are not followed and 2-D scans of large slice thickness are acquired instead. These constraints suggest that the utilization of CAD tools could contribute and software-based interventions to speeding up the whole procedure, while at the same time improving the accuracy of quantification. Taking into consideration the optimal patient care, a semi-automatic "*human-in-the-loop*" approach (where the neuroradiologist removes potential false positive detections generated by the computational tool) may be the best solution.

There are multiple ways for MS lesion detection, and some of them are (i) intensity-based approaches, which depend on detecting the changes of intensity (7, 8), (ii) deformation-based approaches, which analyse the deformation of brain tissue (9, 10), (iii) segmentation-based approaches, which segment white matter hyper-intensities from the acquired scans (11, 12), and (iv) subtraction-based approaches, which depend on subtracting two longitudinal scans (7).

In the intensity-based approaches, a voxelwise comparison is made between MRI scans of different time-points to detect and segment new MS lesions (7, 8). In the deformation-based approaches, the new lesions detected in a T2-FLAIR scan are identified by analyzing the deformation fields between the different MRI scans, obtained through non-rigid registration (9, 10). The non-rigid registration method between the two timepoints has shown to improve the detection of the new T2-w MS lesions in longitudinal studies (10, 13). These deformation fields can be generated through non-rigid registration approaches, either based on optimization (14) or newer learning-based approaches (15). Typically, both the "*tissue transformation*" (*via* intensity change) and the "*tissue deformation*" occur, and as such the mass effect of the particular lesion needs to be taken into account for a precise assessment of the lesion's evolution status.

Furthermore, numerous strategies that combine intensity-based and deformation-based approaches have been proposed. Cabezas et al. (10) modified Ganiler et al. (7)'s subtraction pipeline by merging subtraction and Deformation Field (DF) operators to reduce the amount of false positive lesions found by the subtraction pipeline. Registration is characterized as an optimization issue that must be solved for each volume pair of longitudinal scans using a similarity metric, while enforcing smoothness requirements on the mapping in these approaches. Because solving this optimization is generally computationally costly (16–19), it is exceedingly slow in practise. Various GPU-based accelerated methodologies have been presented to improve the efficiency and speed up the optimization (20–22).

Currently, Convolutional Neural networks (CNNs) have shown superior performance in brain imaging, particularly for segmenting tissues, (23, 24), brain extraction (25–27), brain tumors (27–33), and white matter lesions (34, 35). During training, learning-based registration techniques learn a parameterized registration function from a set of images. Some proposed methods (36, 37) use a precomputed DF as the ground truth (GT), while others depend solely on image registration or segmentation masks, without comparing the predicted DF to a precomputed DF (38, 39). Balakrishnan et al. (15) developed a new CNN approach that computes the deformation between two images by training the network using a similarity metric and a regularization term similar to traditional registration methods, yielding results that are comparable to current state-of-the-art approaches.

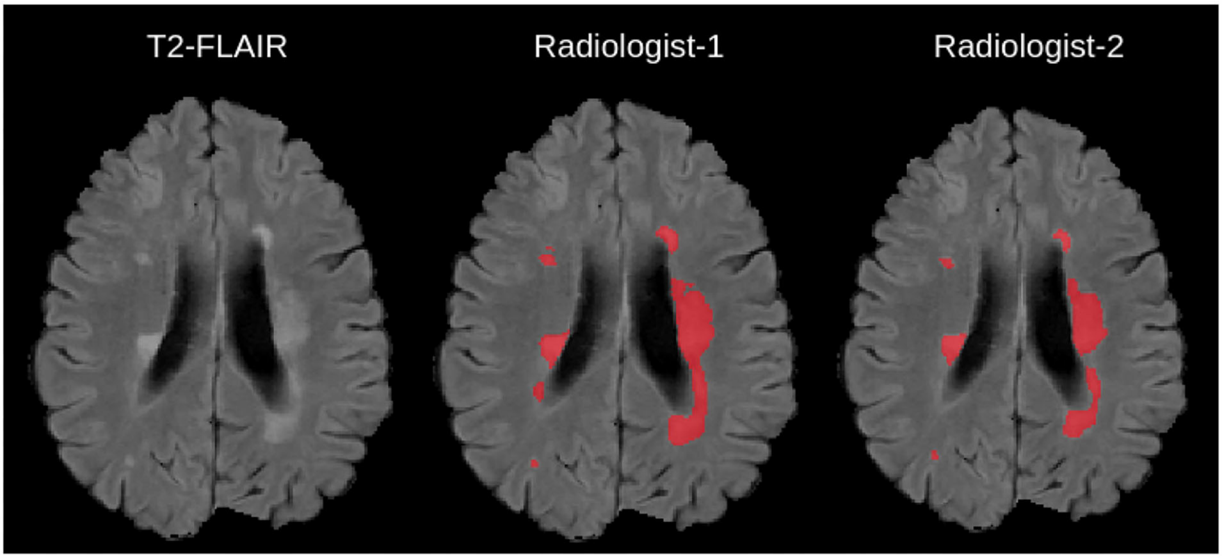


FIGURE 1 | Example Illustration of Multiple Sclerosis lesions overlaid on a 2-D T2-FLAIR scan, together with manual delineations from independent experts taken from (4).

3. MATERIALS AND METHODS

3.1. Data

The routine MRI acquisition protocol for MS patients in the University of Pennsylvania Health System (UPHS) network includes (i) 3-D T2-FLAIR (ii) high resolution isotropic or near-isotropic T1 pre-contrast [3-D magnetization-prepared 180 degrees radio-frequency pulses and rapid gradient-echo (MPRAGE)], as well as (iii) 2D T2-weighted images, and (iv) 30-direction Diffusion Tensor Imaging (DTI). Additionally, 3-D T1 post-contrast images are optional and acquired only if new lesions are detected from the CAD results. All MRI sequences described here are acquired within the UPHS, at multiple satellite sites. However, the vast majority of MS patients get their MRI scans at the main site of the Hospital of the University of Pennsylvania (HUP). The scanner magnetic field strength of the equipment used to acquire these MRI scans was either 1.5 or 3 Tesla, with the HUP scanners being exclusively at 3 Tesla.

The CAD approach presented here uses only the 3-D T2-FLAIR, which is the first acquired sequence in the UPHS acquisition protocol for MS patients. In some rare cases, the prior T2-FLAIR sequence is part of an outside study that has been uploaded to PACS.

Since the acquisition parameters vary across sites, we briefly cover them as majority of scans are performed at 3/1.5 Tesla, with the sagittal 3D T2/FLAIR acquired using the following parameters: TR/TE/TI = 5,000/395/1,800 ms, FOV 250 × 250 × 160 mm, matrix of 256 × 256 × 160, near isotropic $1mm^3$ voxel size. Outside studies with only available 2D T2-FLAIR scans with slice thickness larger than or equal to 5mm are not considered useful and as such our CAD system is not applied to them. However, outside studies with 2D T2-FLAIR scans with <5mm thickness are still used by resampling the higher resolution scans to match the lower resolution images. Notably, the proportion of patients with 2D T2-FLAIR scans from outside studies have been rare.

3.2. CAD System Overview

The functionality of the CAD approach is described in the following sections and visually summarized in **Figure 2**. The overview of the CAD method can be explained as follows: after acquiring the 3D T2-FLAIR scan at Timepoint-2, the CAD system is executed. The registration of Timepoint-2 to a Timepoint-1 3D T2-FLAIR scan occurs, followed by brain extraction and bias field correction for both time points. Then subtraction and false positive reduction methods are applied to identify new lesions, as well as resolve false positives generated by the CAD system. While the CAD system is running, for the routine protocol, we acquire the T1 precontrast, 2D T2-weighted, and DTI images. After the CAD system points out whether new lesions are present, the decision to inject GBCAs is delivered to the MRI technologists.

The hereby presented computational approach, integrated and routinely used in clinical practice since April 2012, has been applied to the assessment of MS lesion scans more than 19, 000 times and is currently assessing more than 200 MS patients per month. **Figure 3** provides a visual representation of the CAD's lifecycle to-date. The "3-D lab" (i.e., a UPHS team of technologists, trained for executing specific software, e.g., for post-processing cardiac CT, MR arteriograms, and more) currently runs and monitors the CAD approach for every scanned MS patient.

3.3. Pre-processing

Prior to any image processing specific to the CAD targets, a set of pre-processing steps are considered essential toward defining the search space to assess lesion burden changes.

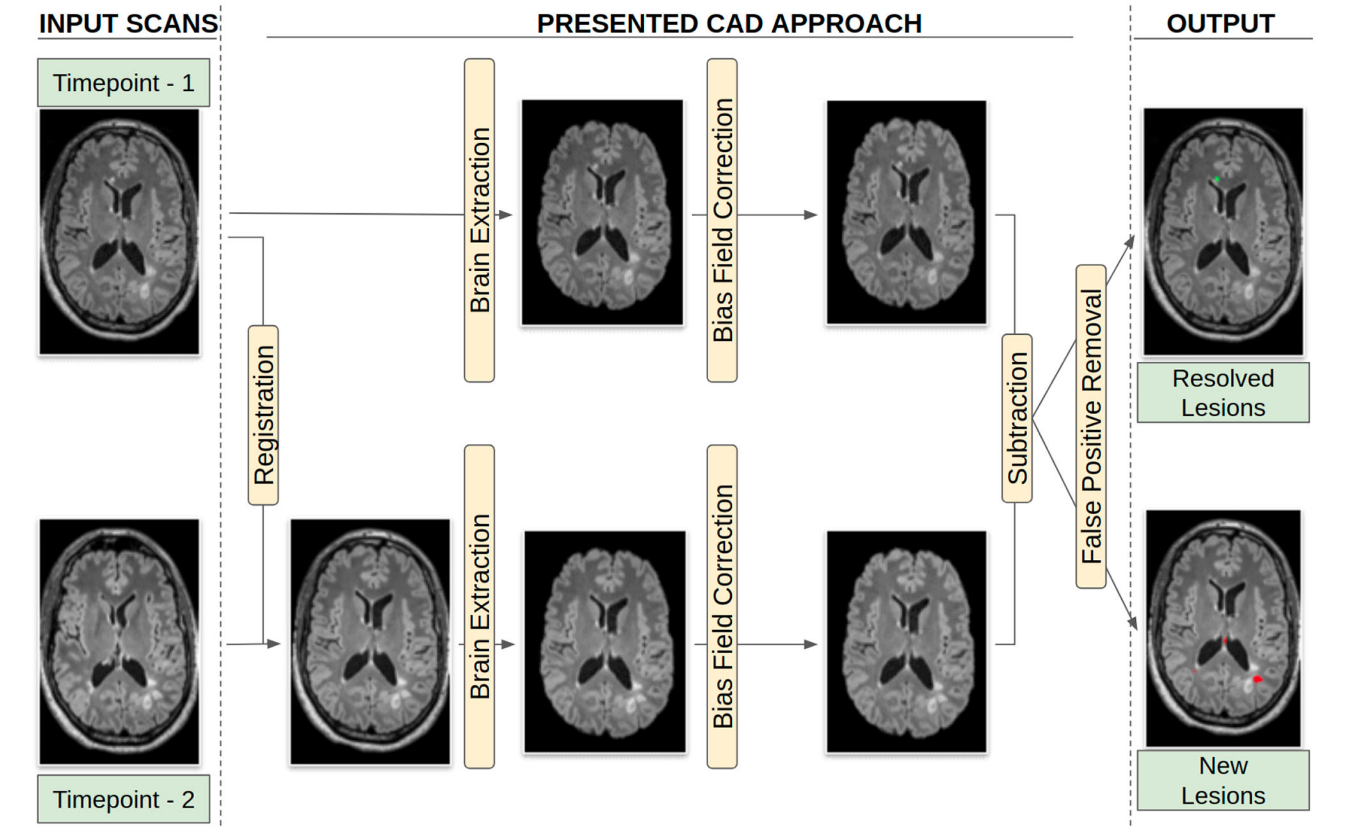


FIGURE 2 | Workflow followed for generating the lesion maps.

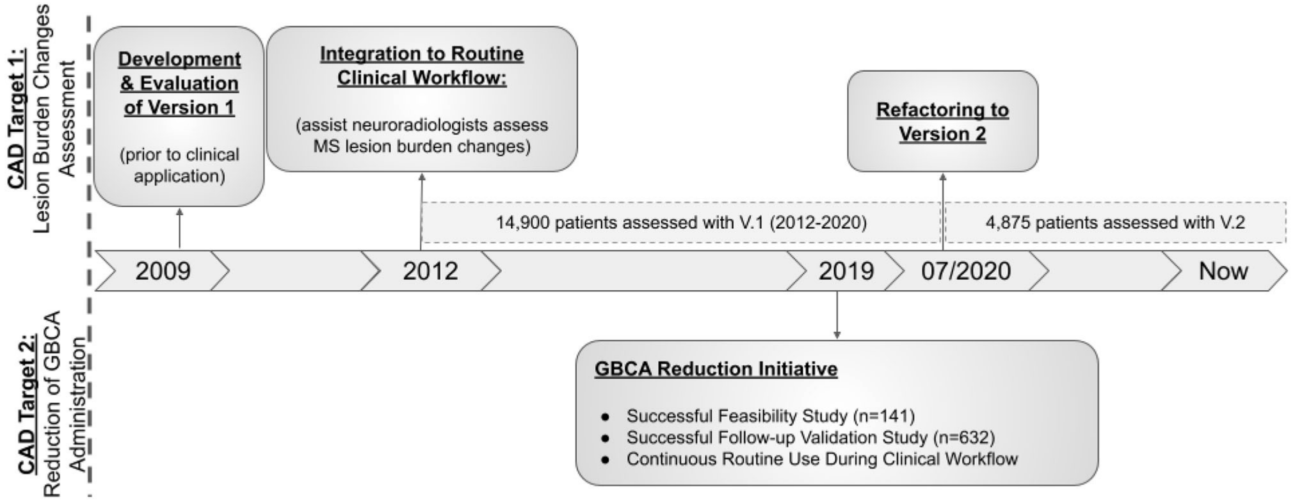

CAD Target 1: Lesion Burden Changes Assessment
Development & Evaluation of Version 1
(prior to clinical application)
Integration to Routine Clinical Workflow:
(assist neuroradiologists assess MS lesion burden changes)
Refactoring to Version 2
14,900 patients assessed with V.1 (2012-2020)
4,875 patients assessed with V.2
2009
2012
2019
07/2020
Now
CAD Target 2: Reduction of GBCA Administration
GBCA Reduction Initiative
Successful Feasibility Study (n=141)
Successful Follow-up Validation Study (n=632)
Continuous Routine Use During Clinical Workflow

FIGURE 3 | Visual representation of the CAD's lifecycle to-date.

All acquired MRI scans are stored in the SECTRA Picture Archiving and Communications System (PACS)[1], as DICOM file sequences. Once a scan is retrieved from the PACS, the CAD system first converts the DICOM file sequences into the Neuroimaging Informatics Technology Initiative (NIfTI) (40) file format to facilitate subsequent image processing steps.

[1] https://medical.sectra.com/

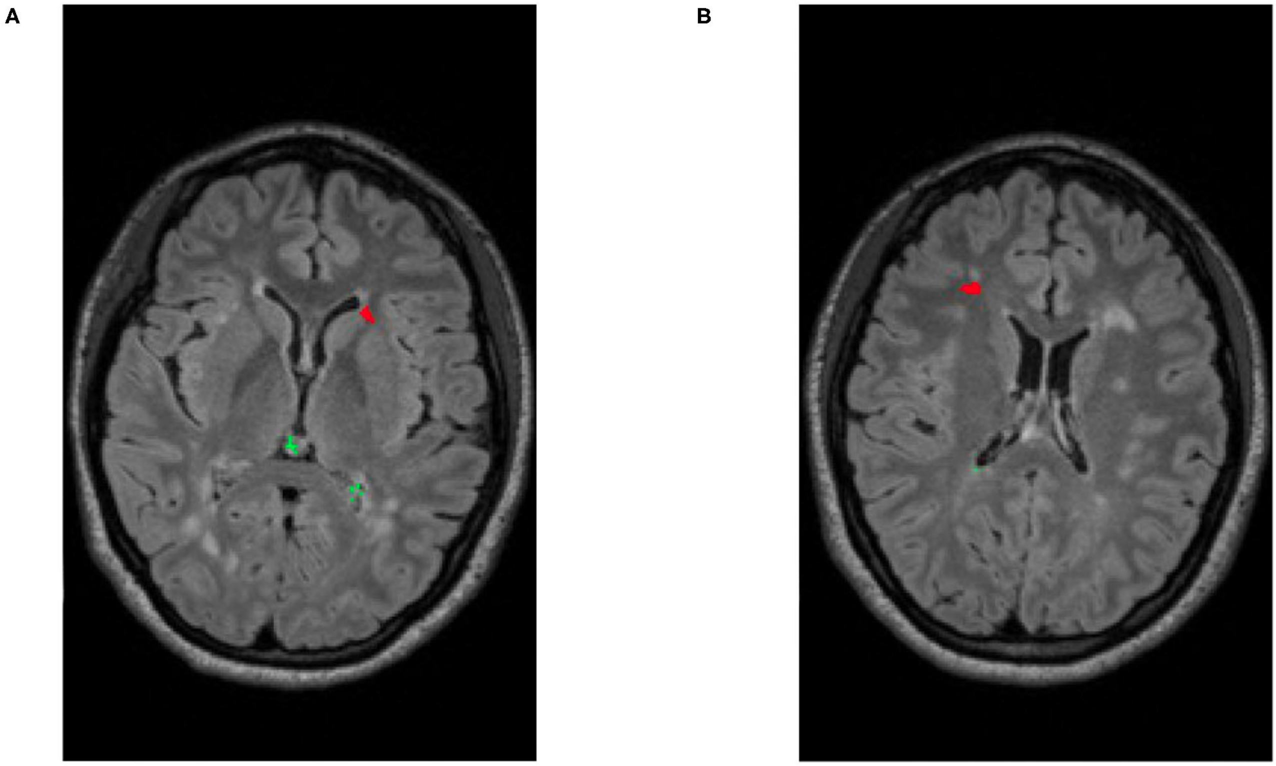


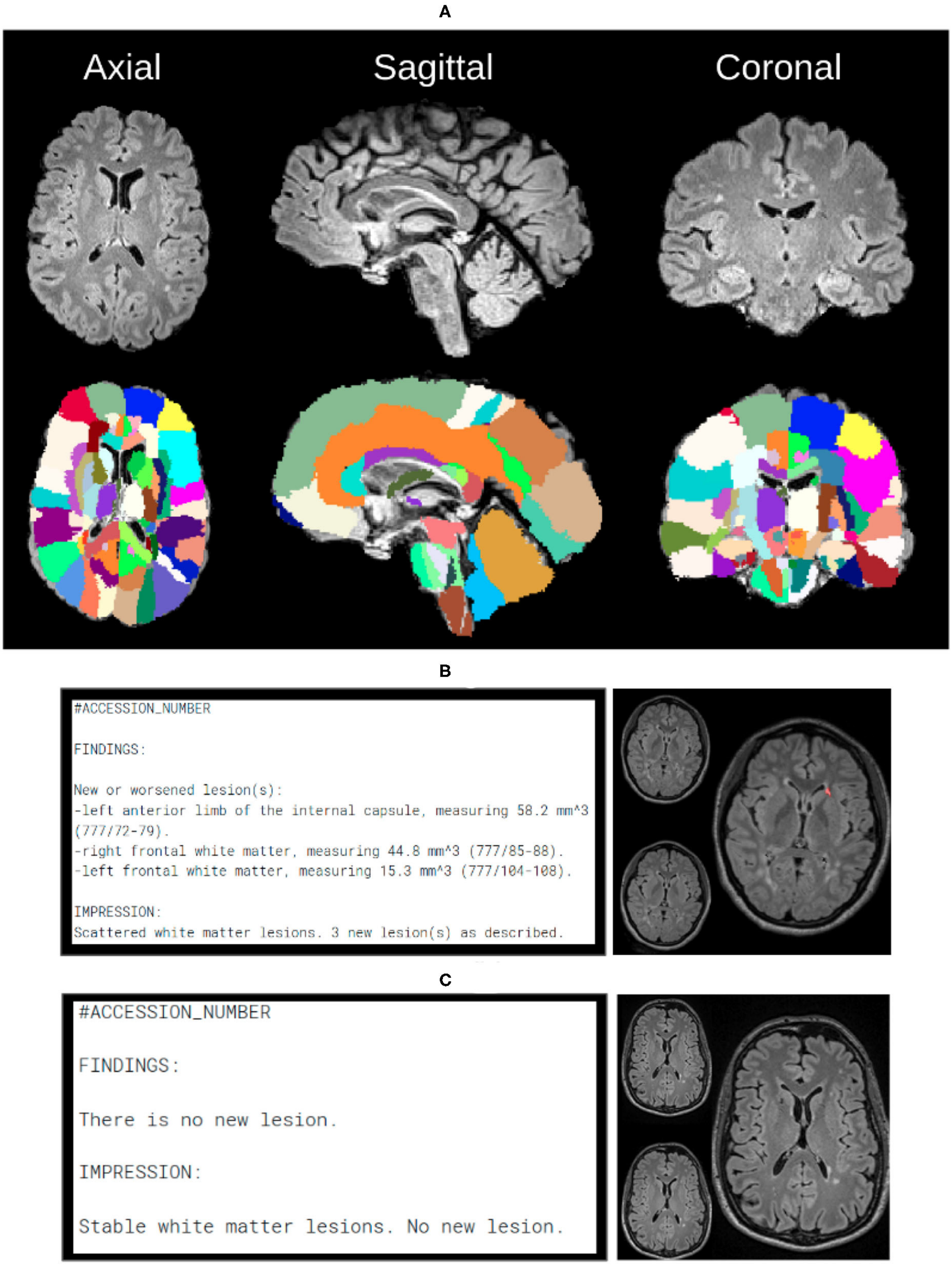
FIGURE 4 | Illustrative examples of detecting false positives and new lesions. **(A)** An example of patient with two identified false positives (depicted in green) and a new lesion (depicted in red). **(B)** Example of a different patient with a single identified false positive (green) and a single new lesion (red).

Since the CAD approach is intended for use across multiple UPHS sites, some harmonization needs to be considered in the imaging space to account for any heterogeneity in the acquisition protocol, thereby ensuring consistent interpretation of the input scans. The most typical harmonization considered here is on the normalization of the scanning resolution, since the acquired MRI scans can range from good quality $1mm^3$ isotropic resolution (or near-isotropic) to a scan of much higher slice thickness, e.g., $5mm^3$. Specifically, when the resolution of the scans across the two time-points differs, the higher resolution scans are downsampled to match the lower resolution scans. The reason to select the lower resolution target is to reduce interpolation artifacts that would have been generated when going from lower resolution to higher resolution. The software can work with 2D images, but the low out-of-plane resolution (slice thickness) limits results accuracy.

Following this resolution normalization, all apparent non-brain tissue has to be removed from the MRI scans to facilitate optimal downstream analyses, by removing parts of skull and to keep the region of interest focused to the brain. Firstly, each patient's Timepoint-2 T2-FLAIR MRI scan has to be rigidly registered to the patient's Timepoint-1 T2-FLAIR anatomical space. Then, the step of brain extraction (also known as skull-stripping) is performed, in order to reduce false positives that may be generated during the downstream analysis, by including portions that do not belong to the brain tissue.

3.4. Intensity Processing

After identifying the complete search space comprising of the brain tissue apparent in the acquired scans, certain intensity processing needs to take place, to facilitate further analyses. Firstly, the magnetic field strength inhomogeneties observed in the acquired scans are corrected by applying the N4 bias field correction (41), available through the ANTs Toolkit (42). We then apply histogram matching to normalize intensities between the "Timepoint-1" and "Timepoint-2" scans, prior to the subtraction process. This step allows us to take into account any contrast differences that are not related to the lesion appearances. Subsequently, an intensity subtraction takes place between the different time-point MRI scans (Timepoint-1 & Timepoint-2). This subtraction operation identifies new lesions (i.e., through their post-subtraction higher intensity appearance). Thirdly, and importantly, a false positive reduction routine is applied to compensate for false positive "artifacts" occurred due to potential misregistrations. This routine leverages the temporal intensity information between the T2-FLAIR scans of Timepoint-1 and Timepoint-2. Specifically, it assesses the pixel intensity of the lesion's center of gravity, and if it is higher in Timepoint-1 the identified lesion is considered as a false positive, but otherwise a true positive. If this criterion is not met, the identified lesion is considered as a false positive. However, in order to maintain high sensitivity, false positive detections from the CAD are tolerated, and are eventually discarded by human evaluation. An example of a true lesion and a false positive detection is shown in **Figure 4**.



```
#ACCESSION_NUMBER

FINDINGS:

New or worsened lesion(s):
-left anterior limb of the internal capsule, measuring 58.2 mm^3
(777/72-79).
-right frontal white matter, measuring 44.8 mm^3 (777/85-88).
-left frontal white matter, measuring 15.3 mm^3 (777/104-108).

IMPRESSION:
Scattered white matter lesions. 3 new lesion(s) as described.
```

```
#ACCESSION_NUMBER

FINDINGS:

There is no new lesion.

IMPRESSION:

Stable white matter lesions. No new lesion.
```

FIGURE 5 | In order to create automated reports, we need anatomical atlas, which can be seen through the subfigures. **(A)** Over 130 anatomical regions of jacob atlas identified and overlayed which allows the software to detect the exact location of new lesions in the brain. **(B)** Example of the automatically generated report, indicating new found lesions. **(C)** Example of the automatically generated report, indicating lack of no newly identified lesions.

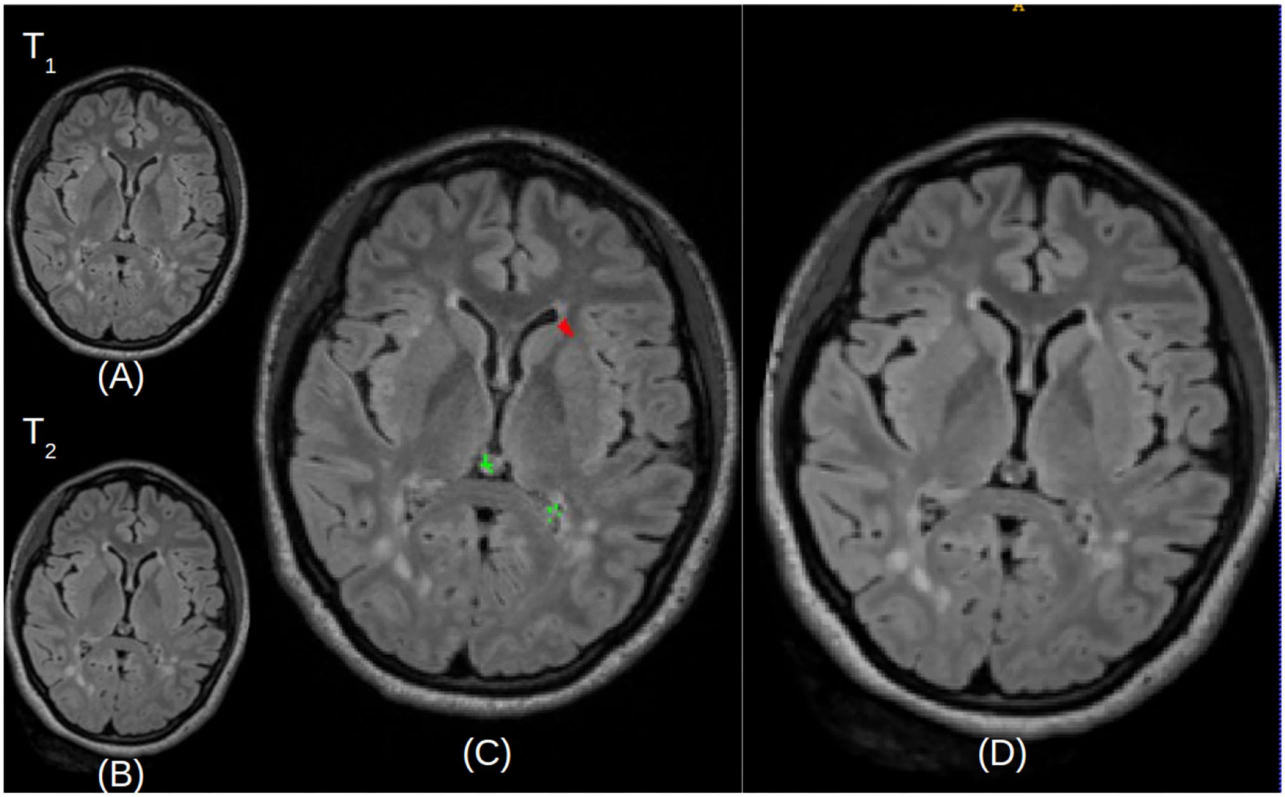


FIGURE 6 | Illustrative examples of resulted images stored in the DICOM file format. **(A,B)** Represent the Timepoint-1 and Timepoint-2 scans of a given patient, respectively. **(C)** Describes Timepoint-2 **(B)**, with superimposed annotations for detected new lesions (depicted in red) and false positives (depicted in green). **(D)** Is a larger version of Timepoint-2 **(B)** without annotations, for visualization purposes.

3.5. Atlas Mapping and Lesion Quantification

After any new or changed lesions are detected, the lesion space is affinely registered to the Jacob Atlas Map (43) (**Figure 5**), to identify an approximate anatomical location of the lesion. Additionally, the size of new lesions is also returned in mm^3. These measurements are included in a draft radiology report available to the neuroradiologist for editing, thereby improving patient care with quantification information, and improving radiology workflow efficiency. **Figures 5B,C** depict representative positive and negative report examples, respectively.

Finally, the CAD system generates DICOM files from the lesion map images, with the temporal pair of images adjacent to assist radiologists re-verify the images manually, without the need of opening up images from every Timepoint separately. **Figure 6** shows the scans of Timepoint-1 (**Figure 6A**) and Timepoint-2 (**Figure 6B**) on the image and the predicted lesions (**Figure 6C**) in the larger size to identify lesions in the subject. These DICOM files are sent to the patient's folder in PACS, so they are available to the neuroradiologist that reads the case.

3.6. Gadolinium Reduction Initiative

A new initiative was started in 2019, namely to use the CAD results not only to assist neuroradiologists, but also to determine in real-time which MS patients would benefit from T1 post-contrast imaging. Since the discovery that gadolinium deposition can be detected in the brain of patients that undergo serial MRI studies with GBCAs, many people involved in the domain of Radiology have started initiatives to decrease the numbers of unnecessary contrast injection when performing MRI scans (44). This is one example of application of the principle of precaution, as the long term effects of this deposition in the brain, and probably other organs of the body, are yet unknown and to be determined. The ability of the presented CAD approach to determine accurately and in real time the patients that would benefit from the injection of GBCAs and those that would not, has decreased the rate of injections substantially. This not only has a positive impact on patients by decreasing their exposure to unnecessary gadolinium, but also benefits the caring healthcare institution.

The principle of this initiative is based on the routine use of the CAD system and only in the case of newly detected white matter lesions to intravenously inject GBCAs, enabling the acquisition of post-contrast T1 imaging. Specifically, once the patient is placed in the scanner, the MR technologist initiates the acquisition of the 3D T2-FLAIR sequence. Once this acquisition is complete, the MR technologist contacts the 3D lab technologist to runs the CAD tool. While the CAD tool is being executed, the MR technologist continues with the acquisition of the remaining non-contrast sequences. Once the CAD 3D Lab technologist assesses

TABLE 1 | GBCA reduction initiative: results from the 2 month feasibility study.

	New brain lesion	No new brain lesion	Total
Gad given	14	3	17
Gad not given	4	120	124
Total	18	123	141

TABLE 2 | GBCA reduction initiative: results from the 3 month follow up validation study.

	New brain lesion	No new brain lesion	Total
Gad given	119	146	265
Gad not given	17	350	367
Total	136	496	632

the CAD results, informs the MR technologist if there is a need for an intravenous injection (e.g., butterfly) and an acquisition of a T1 post-contrast sequence, subject to a new lesion being identified by the CAD system.

To assess the value and the potential clinical relevance of this gadolinium reduction initiative, we conducted a initial 2-month feasibility study involving 141 patients. During this feasibility study, the CAD was already integrated in the clinical workflow for the assessment of lesion burden changes, and hence the feasibility study was directly conducted. After the successful conclusion of this initial feasibility study, we conducted a follow up validation study, including 632 MS patients over the course of 3 months. The purpose of the follow-up study was to confirm the success of the approach in reducing the use of GBCAs, in a larger patient population. The metrics of sensitivity and specificity were calculated for patients who received GBCAs when new brain lesions were found (Equation 1). Following the successful conclusion of both studies, we started using the CAD system as part of our clinical routine for reducing the unnecessary use of GBCAs.

$$sensitivity = \frac{\text{GBCA given for new brain lesion}}{\text{GBCA given for new brain lesion + GBCA not given for new brain lesion}} \tag{1}$$

Inclusion/exclusion of patients in our studies was based on informed consent of participation. For any included patient, the choice to inject GBCA was based on the detection of new lesions as communicated from the CAD operator to the MR technician. The monitored outcome was the performance of correctly identifying patients in need of GBCAs with the use of our CAD system. The sensitivity and specificity calculated here assess the results of the feasibility study (**Table 1**), as well as the follow up validation study (**Table 2**) of giving Gadolinium to MS patients when new lesions are detected.

3.7. CAD's Lifecycle To-Date

Almost 20,000 patients have been assessed with the clinically deployed CAD system to-date. However, the presented CAD tool has undergone a code refactoring during its lifecycle, to improve performance in terms of execution time and sensitivity, resulting in a second version (**Figure 3**). Specifically, the initial development and evaluation of the presented CAD tool has successfully concluded in 2012, resulting in its original deployed v.1.0. This version was integrated to the routine clinical workflow for MS patients across the UPHS network. Ever since, we have been monitoring technological developments and methodological advancements that could improve the performance of the deployed tool. Taking into consideration its high throughput application we have only performed some basic code refactoring in 2020, when we observed that the numbers of assessed scans were lowered due to pandemic-related cancellations. The number of patients assessed during the clinical use of v.1.0 (2012-07/2020) and v.2.0 (07/2020-Now) were 14,900 and 4,875, respectively.

The algorithmic differences between the CAD tool's v.1.0 and v.2.0 relate to the steps of rigid registration and brain extraction. Further modifications have also been considered that are unrelated to any methodological components and are associated with the optimization of graphical elements of the tool according to feedback from the technologists in the "3D-lab". For the rigid registration step, we specifically substituted the FSL's FLIRT (45) algorithm that was used in v.1, with "Greedy" (https://github.com/pyushkevich/greedy) (46) to optimize for the total execution time. "Greedy" is a CPU-based C++ implementation of the greedy diffeomorphic registration algorithm (47) and was designed and developed for rapid registration of radiologic scans. "Greedy" shares the Symmetric Normalization (SyN) of the ANTs registration approach (42). Greedy, on the other hand, is non-symmetric, which makes it quicker (in applications like multi-atlas segmentation, where symmetric property is not required). For the brain extraction step, the initial version of the CAD system (v.1.0) used the "Brain Extraction Tool" (BET) (48). During the CAD's refactoring to its second version, in 2020, BET was substituted by an in-house deep learning based method (26) developed explicitly for brain MRI scans including pathologies, with the intention of improving the execution time, as well as the brain extraction quality.

4. RESULTS

Figure 7 depicts the number of clinical cases assessed by the CAD software, since its integration into the routine clinical workflow. The increase over the years reflects both the growth of the patient population at the UPHS MS clinic and the greater application of the CAD software across the UPHS network, i.e., at satellite locations. We should note the drop in the patients evaluated in 2020 due to the COVID-19 pandemic, when patients cancelled or postponed their followup MRI examinations.

For the assessment of lesion burden changes, we have not conducted an explicit quantitative performance evaluation of the two versions of the software. However, the "3D-lab" technologists have internally reported the sensitivity of the initial version in detecting new lesions as 88%, whereas the sensitivity of v.2.0 being greater than 95%, following the aforementioned

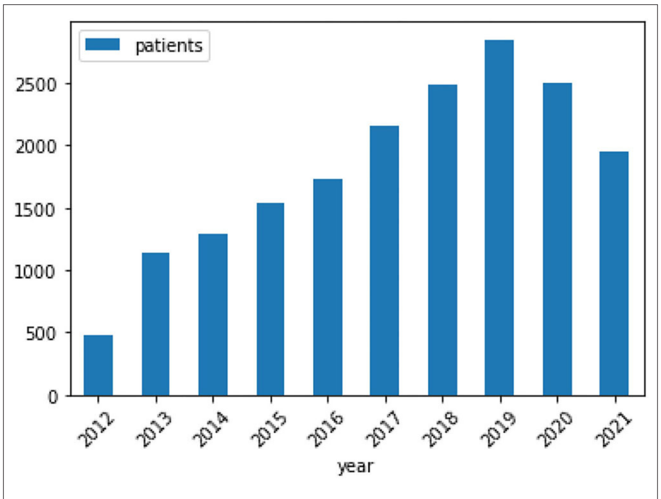


FIGURE 7 | Annual use of the software displayed per year for the number of patients assisted.

TABLE 3 | GBCA reduction initiative: quantitative performance evaluation from the feasibility and the follow up validation studies.

Metric	Feasibility study	Follow-up study
Sensitivity	0.78	0.88
Specificity	0.98	0.71
Precision	0.82	0.45
Recall	0.78	0.88
Positive predictive value	0.82	0.45
Negative predictive value	0.97	0.95
False positive rate	0.02	0.30
False negative rate	0.22	0.13
Accuracy	0.95	0.75
F1 score	0.80	0.59

algorithmic modifications. In terms of execution time, the "3-D lab" technologists require minimal manual intervention to execute the tool (approximately 1 min), and the total time that the approach takes to perform a single patient assessment, on an 4-core CPU (Intel Xeon W-2123 3.60 GHz), has been reported as 11 min for v.1.0 and 5.5 min for v.2.0, on average. The false positives are relatively easy to discard by humans because they tend to occur outside of the white matter of the brain, and in specific areas of the images that are inherently noisy.

For the initial testing phase of gadolinium reduction initiative, we conducted a 2 month feasibility study during which a total of 141 patients participated. Following **Table 1**, we note an 88% reduction in the overall use of GBCAs. We further note a 98% reduction of GBCAs use, when considering only patients with new lesions. Following a performance evaluation, we maintained a high specificity (shown in **Table 3**), while keeping high sensitivity of 78%. This feasibility study showed promising results, and the new protocol, with contrast imaging conditioned upon finding new disease activity on CAD results, became the current standard.

Further evaluation of the gadolinium reduction initiative described a 3 months analysis for 632 additional MS patients, as a follow up validation. In this analysis, we note a reduction of 58% of GBCAs' use, which is lower when compared to the 88% reduction observed in the feasibility study. We further note a 71% reduction of GBCAs use on only patients with existing lesions. This study yielded an increase in sensitivity to 88% (from 78% of the feasibility study), while a reduction occurred in specificity from 98% of the feasibility study to 71%.

These metrics are being passively tracked by the "3-D lab", and the current estimate is about 85% reduction in GBCAs use, as the protocol continues to be utilized across nearly all UPHS sites.

5. DISCUSSION AND FUTURE WORK

In this study we have presented a CAD based method deployed and integrated to the routine clinical workflow for (i) assisting neuroradiologists assess MS lesion burden changes, and (ii) reducing the need for use of GBCAs. We demonstrate the successful evaluation of this computational approach in both the initial evaluation studies and during its routine clinical use, following its complete integration to the clinical workflow since 2012. The findings of this study support our claims that CAD based systems built around clinical settings for MS can contribute in improving patient care and assist radiologists in making better informed decisions.

Temporal changes in patients with existing diagnosed MS lesions were identified better through the presented computational approach using a 3-D T2-FLAIR MRI sequence, than the routine clinical interpretation based on visual observation. The computational approach assisted in improving the sensitivity and false-positive ratio in identifying patients with new (or growing) lesions compared to manual interpretation. Cases can be run in real time (during the patient scanning session, and in $<$ 10 min) within the clinical workflow due to the processing time being so short. The approach has been tuned toward producing the highest possible sensitivity of 90% on a patient basis, where GBCAs are given only when necessary, but still with a low rate of false positive of 30%, allowing for efficient temporal change assessment (**Table 3**).

Non-enhancing new lesions are also of great interest from a clinical standpoint. In fact, as the number of treatment options for MS patients grows, neurologists caring for them are more interested than ever in knowing if new lesions have emerged from previous scans, regardless of their enhancing status. The presented computational approach is not intended to detect lesions that are enhancing. However, we do not believe that these are clinically significant, and the detection of enhancing lesions "manually" (i.e., by visual observation) is relatively simple.

The presented computational approach helps answer the essential clinical question that neurologists are asking when ordering a follow up MRI scan: "Are there new white matter lesions from the prior scan?". This is still one of the most relevant metric for assessing the performance of a therapeutic regimen. The high sensitivity (90%) of this approach in detecting new focal MS lesions allows neurologists to determine if a patient's current

Disease Modifying Therapy (DMT) is appropriately controlling their disease (i.e., no new MS lesions) and may be continued, or if it is not controlling their disease (i.e., new lesions are detected) and a change in DMT may need to be considered. This is more relevant nowadays that the number of available therapeutic options has increased in the past several years, including higher efficacy drugs that also carry the potential for more adverse effects.

Although the presented approach has a clear benefit to clinical practice, it also has its limitations. One of them is that the different time-point MRI scans have to be acquired at the same institution, or more specifically to have a record stored under the institutional PACS. This is required for the approach to produce appropriate results shared with the attending clinician through the platform used typically for the assessment of MS patients. Use of multi-institutional data with medical record number varying across patients and scanning sessions has not been utilized yet, as it was out of scope of this 10-year analyses. Another limitation is the assessment of the rarely observed spinal cord lesions that are not taken into consideration. Any new lesions formed around the spinal cord are currently not considered/processed, and can be potentially missed through the computational approach, since we have primarily focused only on the brain. Limitations also occur when a lower resolution space is used as the reference space to avoid the interpolation artifacts that are generated going from a lower resolution space to a higher resolution space.

The presented approach could also be further utilized in a clinical research setting, such as drug trials, when the ability of the approach to detect temporal changes consistently and reliably, with high sensitivity 90%, is critical. Although the approach presented here does not calculate either the exact volume of each lesion, or the total change in lesion load, this quantification capability belongs to the immediate future work incorporating multi-institutional pilot projects.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because since the dataset collected contains patient health information, this data cannot be made public. Requests to access the datasets should be directed to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board of the University of Pennsylvania. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MB and SB: study conception and design. MB, ST, and SB: software design and development. ST, MB, MS, and SB: data analysis and interpretation and reviewed and edited. ST, MS, and MB: wrote the initial manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

Research reported in this publication was partly supported by the National Cancer Institute (NCI) and the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health (NIH), under Award Numbers NCI:U01CA242871, NCI:U24CA189523, and NINDS:R01NS042645.

REFERENCES

1. Wallin MT, Culpepper WJ, Campbell JD, Nelson LM, Langer-Gould A, Marrie RA, et al. The prevalence of MS in the United States. *Neurology*. (2019) 92:e1029–40. doi: 10.1212/WNL.0000000000007035
2. Mattay RR, Davtyan K, Bilello M, Mamourian AC. Do all patients with multiple sclerosis benefit from the use of contrast on serial follow-up MR imaging? A retrospective analysis. *Am J Neuroradiol*. (2018) 39:2001–6. doi: 10.3174/ajnr.A5828
3. Lladó X, Ganiler O, Oliver A, Martí R, Freixenet J, Valls L, et al. Automated detection of multiple sclerosis lesions in serial brain MRI. *Neuroradiology*. (2012) 54:787–807. doi: 10.1007/s00234-011-0992-6
4. Carass A, Roy S, Jog A, Cuzzocreo JL, Magrath E, Gherman A, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *Neuroimage*. (2017) 148:77–2. doi: 10.1016/j.neuroimage.2016.12.064
5. Kohler C, Wahl H, Ziemssen T, Linn J, Kitzler HH. Exploring individual multiple sclerosis lesion volume change over time: development of an algorithm for the analyses of longitudinal quantitative MRI measures. *Neuroimage Clin*. (2019) 21:101623. doi: 10.1016/j.nicl.2018.101623
6. Thirion JP. Image matching as a diffusion process: an analogy with Maxwell's demons. *Med Image Anal*. (1998) 2:243–60. doi: 10.1016/S1361-8415(98)80022-4
7. Ganiler O, Oliver A, Diez Y, Freixenet J, Vilanova JC, Beltran B, et al. A subtraction pipeline for automatic detection of new appearing multiple sclerosis lesions in longitudinal studies. *Neuroradiology*. (2014) 56:363–74. doi: 10.1007/s00234-014-1343-1
8. Elliott C, Arnold DL, Collins DL, Arbel T. Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. *IEEE Trans Med Imaging*. (2013) 32:1490–503. doi: 10.1109/TMI.2013.2258403
9. Rey D, Subsol G, Delingette H, Ayache N. Automatic detection and segmentation of evolving processes in 3D medical images: application to multiple sclerosis. *Med Image Anal*. (2002) 6:163–79. doi: 10.1016/S1361-8415(02)00056-7
10. Cabezas M, Corral JF, Oliver A, Diez Y, Tintoré M, Auger C, et al. Improved automatic detection of new T2 lesions in multiple sclerosis using deformation fields. *Am J Neuroradiol*. (2016) 37:1816–23. doi: 10.3174/ajnr.A4829
11. Zhang H, Oguz I. Multiple sclerosis lesion segmentation-a survey of supervised CNN-based methods. *arXiv [Preprint] arXiv*:2012.08317 (2020). doi: 10.1007/978-3-030-72084-1_2
12. Fartaria MJ, Kober T, Granziera C, Bach Cuadra M. Longitudinal analysis of white matter and cortical lesions in multiple sclerosis. *Neuroimage Clin*. (2019) 23:101938. doi: 10.1016/j.nicl.2019.101938
13. Salem M, Cabezas M, Valverde S, Pareto D, Oliver A, Salvi J, et al. A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. *Neuroimage Clin*. (2018) 17:607–15. doi: 10.1016/j.nicl.2017.11.015
14. Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey. *IEEE Trans Med Imaging*. (2013) 32:1153–90. doi: 10.1109/TMI.2013.2265603

15. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Trans Med Imaging*. (2019) 38:1788–800. doi: 10.1109/TMI.2019.2897538
16. Beg MF, Miller MI, Trouvé A, Younes L. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int J Comput Vis*. (2005) 61:139–57. doi: 10.1023/B:VISI.0000043755.93987.aa
17. Ashburner J. A fast diffeomorphic image registration algorithm. *Neuroimage*. (2007) 38:95–113. doi: 10.1016/j.neuroimage.2007.07.007
18. Glocker B, Komodakis N, Tziritas G, Navab N, Paragios N. Dense image registration through MRFs and efficient linear programming. *Med Image Anal*. (2008) 12:731–41. doi: 10.1016/j.media.2008.03.006
19. Dalca AV, Bobu A, Rost NS, Golland P. Patch-based discrete registration of clinical brain images. In: Wu G, Coupé P, Zhan Y, Munsell BC, Rueckert D, editors. *Patch-Based Techniques in Medical Imaging*. Cham: Springer International Publishing (2016). p. 60–7. doi: 10.1007/978-3-319-47118-1_8
20. Punithakumar K, Boulanger P, Noga M. A GPU-accelerated deformable image registration algorithm with applications to right ventricular segmentation. *IEEE Access*. (2017) 5:20374–82. doi: 10.1109/ACCESS.2017.2755863
21. Wu J, dan Li D, yao Li J, cong Yin Y, chang Li P, Qiu L, et al. Identification of microRNA-mRNA networks involved in cisplatin-induced renal tubular epithelial cells injury. *Eur J Pharmacol*. (2019) 851:1–12. doi: 10.1016/j.ejphar.2019.02.015
22. Han X, Hibbard LS, Willcut V. GPU-accelerated, gradient-free MI deformable registration for atlas-based MR brain image segmentation. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Miami, FL: IEEE (2009). p. 141–8.
23. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *Neuroimage*. (2015). 108:214–24. doi: 10.1016/j.neuroimage.2014.12.061
24. Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJNL, Išgum I. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans Med Imaging*. (2016) 35:1252–61. doi: 10.1109/TMI.2016.2548501
25. Thakur SP, Doshi J, Pati S, Ha SM, Sako C, Talbar S, et al. Skull-stripping of glioblastoma MRI scans using 3D deep learning. *Brainlesion*. (2019) 11992:57–68. doi: 10.1007/978-3-030-46640-4_6
26. Thakur S, Doshi J, Pati S, Rathore S, Sako C, Bilello M, et al. Brain extraction on MRI scans in presence of diffuse glioma: multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. *Neuroimage*. (2020) 220:117081. doi: 10.1016/j.neuroimage.2020.117081
27. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Hum Brain Mapp*. (2019) 40:4952–64. doi: 10.1002/hbm.24750
28. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*. (2017) 4:170117. doi: 10.1038/sdata.2017.117
29. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM Collection*. The Cancer Imaging Archive (2017).
30. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. *Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection*. The Cancer Imaging Archive (2017).
31. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv [Preprint] arXiv:*181102629 (2018). doi: 10.48550/arXiv.1811.02629
32. Baid U, Ghodasara S, Bilello M, Mohan S, Calabrese E, Colak E, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv [Preprint] arXiv:*210702314 (2021). doi: 10.48550/arXiv.2107.02314
33. Pati S, Thakur SP, Bhalerao M, Baid U, Grenko C, Edwards B, et al. Gandlf: a generally nuanced deep learning framework for scalable end-to-end clinical workflows in medical imaging. *arXiv [Preprint] arXiv:*210301006 (2021). doi: 10.48550/arXiv.2103.01006
34. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. *Med Image Anal*. (2017) 35:18–31. doi: 10.1016/j.media.2016.05.004
35. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. (2017) 36:61–78. doi: 10.1016/j.media.2016.10.004
36. Sokooti H, de Vos B, Berendsen F, Lelieveldt BPF, Išgum I, Staring M. Nonrigid image registration using multi-scale 3D convolutional neural networks. In: *Medical Image Computing and Computer Assisted Intervention, MICCAI 2017*. Cham: Springer International Publishing (2017). p. 232–9. doi: 10.1007/978-3-319-66182-7_27
37. Yang X, Kwitt R, Styner M, Niethammer M. Quicksilver: fast predictive image registration–a deep learning approach. *Neuroimage*. (2017) 158:378–96. doi: 10.1016/j.neuroimage.2017.07.008
38. Li H, Fan Y. Non-rigid image registration using self-supervised fully convolutional networks without training data. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. (2018). p. 1075–8. doi: 10.1109/ISBI.2018.8363757
39. de Vos BD, Berendsen FF, Viergever MA, Staring M, Išgum I. End-to-end unsupervised deformable image registration with a convolutional neural network. In: Cardoso MJ, Arbel T, Carneiro G, Syeda-Mahmood T, Tavares JMRS, Moradi M, editors. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Québec City, QC: Springer International Publishing (2017). p. 204–12. doi: 10.1007/978-3-319-67558-9_24
40. Cox RW, Ashburner J, Breman H, Fissell K, Haselgrove C, Holmes CJ, et al. A (sort of) new image data format standard: NiFTI-1. In: *10th Annual Meeting of the Organization for Human Brain Mapping*. (2004).
41. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. (2010) 29:1310–20. doi: 10.1109/TMI.2010.2046908
42. Tustison NJ, Cook PA, Klein A, Song G, Das SR, Duda JT, et al. Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage*. (2014) 99:166–79. doi: 10.1016/j.neuroimage.2014.05.044
43. Kabani NJ, MacDonald D, Holmes CJ, Evans AC. 3D anatomical atlas of the human brain. *Neuroimage*. (1998) 7:S717. doi: 10.1016/S1053-8119(18)31550-7
44. Rudie JD, Mattay RR, Schindler M, Steingall S, Cook TS, Loevner LA, et al. An initiative to reduce unnecessary gadolinium-based contrast in multiple sclerosis patients. *J Am Coll Radiol*. (2019) 16:1158–64. doi: 10.1016/j.jacr.2019.04.005
45. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*. (2002) 17:825–41. doi: 10.1006/nimg.2002.1132
46. Yushkevich PA, Pluta J, Wang H, Wisse LEM, Das S, Wolk D. IC-P-174: fast automatic segmentation of hippocampal subfields and medial temporal lobe subregions in 3 Tesla and 7 Tesla T2-weighted MRI. *Alzheimers Dement*. (2016) 12:P126–7. doi: 10.1016/j.jalz.2016.06.205
47. Joshi S, Davis B, Jomier M, Gerig G. Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage*. (2004) 23:S151–60. doi: 10.1016/j.neuroimage.2004.07.068
48. Smith SM. Fast robust automated brain extraction. *Hum Brain Mapp*. (2002) 17:143–55. doi: 10.1002/hbm.10062

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 24 March 2022
doi: 10.3389/fmed.2022.807443

Cerebral Microbleed Automatic Detection System Based on the "Deep Learning"

Pingping Fan[1,2,3†], Wei Shan[2,4,5†], Huajun Yang[2,4], Yu Zheng[2], Zhenzhou Wu[2], Shang Wei Chan[2], Qun Wang[2,3,5], Peiyi Gao[1], Yaou Liu[1,2,4*], Kunlun He[6,7*] and Binbin Sui[2,3*]*

[1] Department of Radiology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China, [2] National Clinical Research Center for Neurological Diseases, Beijing, China, [3] Tiantan Neuroimaging Center of Excellence, Beijing, China, [4] Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China, [5] Beijing Institute for Brain Disorders, Beijing, China, [6] Laboratory of Translational Medicine, Chinese PLA General Hospital, Beijing, China, [7] Key Laboratory of Ministry of Biomedical Engineering and Translational Medicine, People's Liberation Army General Hospital, Beijing, China

OPEN ACCESS

Edited by:
Richard McKinley,
Bern University Hospital, Switzerland

Reviewed by:
Qian Ren,
Hebei Medical University, China
Zhichao Lai,
Peking Union Medical College Hospital (CAMS), China

****Correspondence:***
Wei Shan
weishanns@gmail.com
Yaou Liu
yaou80@163.com
Kunlun He
kunlunhe@plagh.org
Binbin Sui
reneesui@163.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Translational Medicine, a section of the journal Frontiers in Medicine

Received: *02 November 2021*
Accepted: *22 February 2022*
Published: *24 March 2022*

Citation:
Fan P, Shan W, Yang H, Zheng Y, Wu Z, Chan SW, Wang Q, Gao P, Liu Y, He K and Sui B (2022) Cerebral Microbleed Automatic Detection System Based on the "Deep Learning". Front. Med. 9:807443. doi: 10.3389/fmed.2022.807443

Objective: To validate the reliability and efficiency of clinical diagnosis in practice based on a well-established system for the automatic segmentation of cerebral microbleeds (CMBs).

Method: This is a retrospective study based on Magnetic Resonance Imaging-Susceptibility Weighted Imaging (MRI-SWI) datasets from 1,615 patients (median age, 56 years; 1,115 males, 500 females) obtained between September 2018 and September 2019. All patients had been diagnosed with cerebral small vessel disease (CSVD) with clear cerebral microbleeds (CMBs) on MRI-SWI. The patients were divided into training and validation cohorts of 1,285 and 330 patients, respectively, and another 30 patients were used for internal testing. The model training and validation data were labeled layer by layer and rechecked by two neuroradiologists with 15 years of work experience. Afterward, a three-dimensional convolutional neural network (CNN) was applied to the MRI data from the training and validation cohorts to construct a deep learning system (DLS) that was tested with the 72 patients, independent of the aforementioned MRI cohort. The DLS tool was used as a segmentation program for these 72 patients. These results were evaluated and revised by five neuroradiologists and subjected to an output analysis divided into the missed label, incorrect label, and correct label. The interneuroradiologists DLS agreement rate, which was assessed using the interrater agreement kappas test, was used for the quality analysis.

Results: In the detection and segmentation of the CMBs, the DLS achieved a Dice coefficient of 0.72. In the evaluation of the independent clinical data, the neuroradiologists reported that more than 90% of the lesions were directly detected and less than 10% of lesions were incorrectly labeled or the label was missed by our DLS. The kappa value for interneuroradiologist DLS agreement reached 0.79 on average.

Conclusion: Based on the results, the automatic detection and segmentation of CMBs are feasible. The proposed well-trained DLS system might represent a trusted tool for the segmentation and detection of CMB lesions.

Keywords: cerebral microbleed, deep learning, neural network, segmentation, clinical evaluation

INTRODUCTION

Cerebral microbleeds (CMBs) are radiological constructs that were first observed and defined on MRI (1). T2*-weighted gradient-recalled echo (GRE) and susceptibility-weighted imaging (SWI) are commonly used to detect CMB in clinical practice (2). On GRE images or SWI, a CMB is a small elliptical or circular lesion of 2–5 mm but sometimes up to 10 mm (3). According to previous studies, SWI is usually the main modality recommended for quantifying numbers of CMBs, as it shows higher sensitivity and reliability for CMB detection than GRE imaging. The pathophysiology of CMB has not yet been fully elucidated. Histopathologically, microbleeds represent the perivascular focal collection of hemosiderin deposits (1–5). Vitreous degeneration of small vessels and vascular amyloidosis are considered to be the two main pathological mechanisms. They might damage the small vascular wall and cause the destruction of the blood-brain barrier. The focal remnant deposits of hemosiderin are most likely secondary to such arteriolar and capillary damage caused by multiple mechanisms, which result in blood product leakage in the perivascular space (6). A group of risk factors for CMB has been reported, including age, hypertension, cholesterol, diabetes mellitus, and smoking (7–9). CMB is associated with an increased risk of several diseases and conditions. CMBs increase the risk of subsequent ischemic stroke and intracranial hemorrhage (ICH) (2, 3, 10, 11). CMBs are associated with small vascular disease (SVD) and are thus more likely to accompany strokes with lacunar infarction than infarction caused by cardioembolism or atherosclerosis (12). CMB is also expected to cause ICH (8, 13). Therefore, CMB was considered a predictor of future stroke and hemorrhage in patients receiving thrombolytic therapy or long-term antithrombotic treatment for ischemic stroke (3).

Cerebral microbleeds (CMBs) are also associated with an increased risk of cognitive impairment and dementia in patients with normal cognitive function, mild cognitive impairment, and dementias such as Alzheimer's disease (4, 14–16). In addition, CMBs may be present in individuals with some genetic diseases, such as cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) or Moyamoya disease (17, 18). The risk and extent of CMB, which is considered a biomarker of SVD, has been used as an index for evaluating the status of underlying diseases and might influence the management of these diseases (12). Thus, a systemic and quantitative evaluation of CMB with high accuracy and efficiency is essential in assessing disease prognosis. At present, visual scoring systems are used in CMB evaluations, including the Microbleed Anatomical Rating Scale (MARS) (19) and the Brain Observer MicroBleed Scale (BOMBS) (20). However, the reliability of these methods in assessing the number and location of CMBs is relatively low without the use of evaluation tools. In recent years, automated or semiautomated brain imaging analysis methods have been applied to evaluate CMBs (21–23). A deep learning system (DLS) for automatic CMB detection was developed and analyzed in terms of reliability to support clinical work with consistent and efficient CMB identification and simplify the clinical workflow of CMB marking. We invited five clinical neuroradiologists to assess the performance of the proposed DLS, especially the number and location of CMBs based on SWI sequences. This study aimed to validate an appropriately trained DLS that could be trusted by a neuroradiologist with sufficient experience.

MATERIALS AND METHODS

Standard Protocol Approvals, and Patient Consent

This study was approved by the ethics committee of Beijing Tiantan Hospital and fulfilled the Declaration of Helsinki.

Image Dataset

We retrospectively obtained MRI-SWI data with good SWI image quality from 1,615 patients, and all the data were obtained from Beijing Tiantan Hospital. According to the SWI acquisition protocol used clinically, scans were obtained using multiple different scanners with a field strength of 1.5T or 3T. In this dataset, we labeled 10,525 lesions, including 9,387 small size lesions ranging in size from 2 to 5 mm and 1,138 large lesions ranging in size from 5 to 10 mm. The basic information of the patients and manufacturers is provided in **Table 1**, and more detailed information about lesion sizes is presented in **Table 2**. The clinical evaluation dataset (test data) included MRI-SWI images from 72 patients with CMBs in the Third China National Stroke Registry (CNSR-III), a nationwide registry of ischemic stroke or transient ischemic attack (TIA) in China based on etiology, imaging, and biological markers that recruit consecutive patients with ischemic stroke or TIA from 201 hospitals that cover 22 provinces and four municipalities in China. This dataset is independent of the previous 1,615 patients.

Data Quality Control

Minor artifacts or mildly reduced signal-noise ratios with no effects on diagnosis or no artifacts and optimal artifacts were selected for the evaluation of image quality in this retrospective study. Diagnostic Screening: All patient electric health records (EHRs) were reviewed and reanalyzed by medical doctors before preprocessing the images, labeling, and generating ensemble models. The segmentation labels with CMBs were based on a manual slice by slice analysis of the MRI-SWI data. After labeling all images with CMBs, all data were rechecked and endorsed by two radiologists with 15 years of clinical experience, which were used for DLS training and validation.

Network Architecture

The SWI image is a three-dimensional (3D) axial slice, and the image size is Z*X*Y. Z represents the number of slices and X*Y represents the length and width of each slice. Microbleeding is a disease with contextual information. The normal network used two-dimensional (2D) U-net and 3D U-net for prediction, but the number of slices of the SWI was quite different. If 3D U-net and resampling are used, some slice information will be missing, and a simple 2D U-net will lack the upper and lower slice information.

TABLE 1 | Basic information of the patients, manufacturers, and parameters of scanners.

Patients characteristics (training/validation dataset)	Patient (images) metric
Number of patients	1285/330
Female to male ratio	405:880/95:235
Different macufacturers (numbers in the training/validation datasets)	
GE	287/72
Siemens	356/88
Philips	642/170
Field strength (numbers in the training/validation datasets)	
1.5T	174/34
3T	1111/296
Scanner model (numbers in the training/validation datasets)	
Verio	89/25
Ingenia	429/116
Achieva	77/13
Trio Tim	110/33
Signa HDxt	71/10
DiSCOVERY MR750	205/70
Ingenia CX	133/33
Skyra	16/2
Avanto	100/18
Aera	43/6
SIGNA Explorer	12/2
Prisma	0/2
Resolution (numbers in the training/validation datasets)	
512 × 384	96/15
432 × 432	459/157
512 × 512	326/60
256 × 192	183/47
256 × 232	62/15
480 × 480	21/7
768 × 624	9/3
256 × 224	42/9
224 × 256	10/2
320 × 320	9/1
384 × 264	3/0
320 × 260	18/3
640 × 520	13/2
310 × 320	1/0
352 × 352	9/1
256 × 256	21/8
260 × 320	1/0
560 × 560	2/0

TABLE 2 | Data distribution.

	patients	Small lesions	Large lesions
Training dataset	1,285 (79.6%)	7,461 (79.5%)	927 (81.5%)
Validation dataset	330 (20.4%)	1,926 (20.5%)	211 (18.5%)
Summary	1,615	9,387	1,138

Therefore, in this article, we adopted a new approach. In the training process, three consecutive layers of slices were used as input. The size was X*Y*3, and the output was X*Y. The middle slice of the positive sample had microbleeds, and the middle slice of the negative sample had no microbleeds. Bleeding was not set for the analysis of whether microbleeds were present in the upper and lower layers. Here, the ratio of positive to negative was 1:1. During the test, the entire image was input into the network in sequence according to the scanning order.

We implemented a 3D CNN to extract representative features for complicated CMBs based on the MRI-SWI sequences. Specifically, we designed a full CNN architecture composed of encoder and decoder paths to conduct the segmentation task. More specifically, our network was based on the widely-used modified 3D U-Net architecture with 3 layers. The detailed network architecture is shown in **Figure 1**, which provides a detailed description of the segmented network used to detect CMBs.

Establishing the Deep Learning Algorithm

Before feeding the model, all MRI-SWI data were preprocessed by scaling the global (3D) image intensities and were standardized across the acquisition parameters to increase the convergence rate of network training. We performed the normalization and alignment based on the histogram peaks to the white matter content in the MRI. All images were cropped into squares according to the shortest side, and the size was cropped or resized to 384 × 384 pixels. According to the histogram, we deleted fewer points (less than 1e-4), and the window width was determined and then max-min normalized. Lesions with sizes other than 2–10 mm were deleted.

The training set and validation set consisted of CMB data (n = 1,285 positive volumetric scans and 330 positive volumetric scans, respectively). The model was trained using 3D axial SWI slices. The SWI data from all patients were preprocessed, resized, and normalized to have a uniform size of 384 × 384 × 3 pixels and pixel intensities in the range of 0–1. Using these data, the network was trained using binary cross-entropy loss and the Adam optimizer with an initial learning rate of 10^{-3}. During training, model training progress was monitored using a validation set Dice score. The learning rate was reduced by a factor of 0.1. An early stopping criterion was implemented if the validation Dice did not improve for 30 consecutive epochs to avoid model overfitting. Training stopped if the validation Dice score did not improve for 60 consecutive epochs. At the end of the training, the model with the highest Dice score for the validation set was retrieved. Its performance was evaluated on the test dataset. The training was stopped when the training loss was less than 10 and the validation scores reached 0.8, as we presumed that the DLS reached the optimal performance at this time. The dataset was augmented in the training process, including image vertical, rotation, translation, contrast changes and other parameters, to increase the robustness of the model. Thus, it forms a more diverse dataset with slight differences.

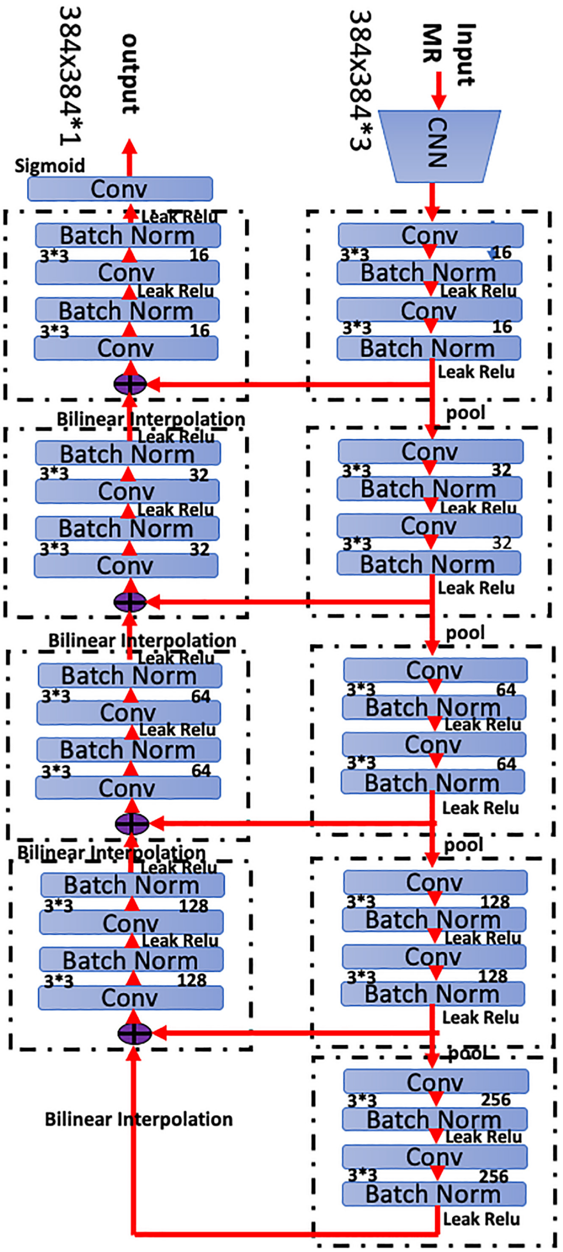


FIGURE 1 | Network architecture of the proposed three-dimensional (3D) convolutional neural network (CNN). The network has 28 layers integrating six residual blocks. Bilinear interpolating arrows indicate upsampling operations to provide dense predictions for the segmentation task. Skip connections are used to fuse low- and high-level features in the network. Batch normalization is a linear transformation of the features to reduce covariance shift and accelerate the training process. The convolution bar represents the convolution operation that computes features. The number *64* indicates the number of channels in that layer, and *3 3 3 3 3* denotes the size of the 3D CNN kernels.

If the prediction mask and the reference mask intersected, a value greater than the threshold value was considered predicted correctly; otherwise, prediction error was considered. Similarly, if the reference mask and forecast masks intersect, a value greater than the threshold value was considered correctly predicted; otherwise, the prediction miss was considered. The threshold was set to 0.4 obtained from the optimized model results. Computed precision and recall were analyzed using this method. After the establishment of the model, 90 healthy patients without CMBs were used as the control to test the model, and no false-positive CMBs were detected. Then, the model entered the evaluation phase.

Evaluation Dataset and Reference Standard

Commonly used metrics known as the Dice score, precision, recall, and accuracy were used to evaluate the performance of the proposed segmentation networks (24). Pixel level dice score was the primary model performance criteria and it was calculated as follows:

$$\begin{aligned}\mathrm{TP}_{\mathrm{pixel}} &= \left|\text{Pixels correctly predicted as positive}\right| \\ &= |\mathit{Predicted\ Mask}\ \cap \mathit{Ground\ truth\ mask}|\end{aligned}$$

$$\mathrm{FP}_{\mathrm{pixel}} = \left|\text{Pixels wrongly predicted as positive}\right|$$

$$\mathrm{FN}_{\mathrm{pixel}} = |\text{Pixels wrongly predicted as negative}|$$

$$Dice_{pixel} = \frac{2 * TP_{pixel}}{2 * TP_{pixel} + FP_{pixel} + FN_{pixel}}$$

Along with the dice score, the pixel level precision and recall were also calculated as follows:

$$Presicion_{pixel} = \frac{TP_{pixel}}{TP_{pixel} + FP_{pixel}}$$

$$Recall_{pixel} = \frac{TP_{pixel}}{TP_{pixel} + FN_{pixel}}$$

Also, along with the pixel-level computations, to understand how good the model is in identifying isolated lesions, the lesion level precision and recall were computed. For this computation, first, the individual lesions were identified as a set of continuous positive pixels from both the predicted and ground truth mask. Next, the overlap between the predicted lesions and the ground truth lesions was computed and the lesions were termed as TP lesions if this overlap was greater than 40% of the true lesion. Following this, the lesion-level precision and recall were computed as follows:

$$\mathrm{TP}_{\mathrm{lesion}} = \left|\text{Correctly predicted as lesions}\right|$$

$$\mathrm{Predicted}_{\mathrm{lesion}} = \left|\text{Predicted lesions}\right|$$

$$\mathrm{True}_{\mathrm{lesion}} = \left|\text{Ground truth lesions}\right|$$

$$Presicion_{lesion} = \frac{TP_{lesion}}{\text{Predicted}_{lesion}}$$

$$Recall_{lesion} = \frac{TP_{lesion}}{True_{lesion}}$$

We count the data level and patient level at the same time, and the data level is calculated on the entire data set. The patient level is calculated for each patient first and then averaged.

Furthermore, based on the condition of whether the model was able to identify at least one correct lesion, the patients were also classified into the TP, TN, FP, and FN categories. Based on this data, the patient-level FP rate (FPR), FN rate (FNR), and TP rate (TPR) were calculated as follows:

$$FPR = \frac{FP}{TN + FP}$$

$$FNR = \frac{FN}{TP + FN}$$

$$TPR = \frac{TP}{TP + FN}$$

The FP rate indicates the model's tendency for wrongly identifying a patient as having infarction (Type I error rate) whereas, FNR indicates the possibility of the model missing a patient with infarction (Type II error rate). Based on the FPR and TPR, the receiver operating characteristics (ROC) curve was constructed: the abscissa was FPR and the ordinate was TPR. Then the area under the ROC (AUC) was calculated. Pixel-level ROC and lesion-level ROC were defined as follows:

Pixel-level ROC: taking each pixel as a sample, the ROC curve is calculated from the pixel prediction probability and ground truth.

Lesion-level ROC: taking each lesion as a sample, the average pixel probability of each lesion is counted, and the ROC curve is calculated from the average probability and ground truth.

Compared with pixel-level ROC, lesion-level ROC is more clinically relevant. So we constructed the lesion-level ROC to evaluate the performance of the system.

We generated predictions for 72 patients randomly chosen by a doctor among patients who had SWI sequences in their records from the CNSR-III research group to evaluate the performance of the model and assess whether it would meet the clinical requirements. The clinical diagnosis must meet the inclusion criteria, and all of these patients are independent of the previous training and validation datasets.

The clinical doctors included in this study are top experts neuroradiologists with at least 15 years of clinical experience. After DLS prediction, we asked them to categorize the prediction results into three subgroups: correct label, missed label, and incorrect label. Each of these terms was defined as follows:

Correct label: the label was accurate compared with the ground truth.

Missed label: compared with the ground truth, the model did not produce the corresponding label.

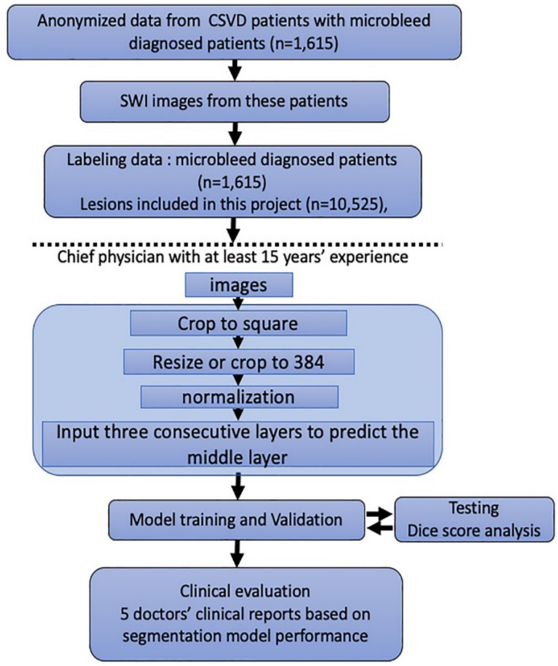

Anonymized data from CSVD patients with microbleed diagnosed patients (n=1,615)
SWI images from these patients
Labeling data : microbleed diagnosed patients (n=1,615)
Lesions included in this project (n=10,525),
Chief physician with at least 15 years' experience
images
Crop to square
Resize or crop to 384
normalization
Input three consecutive layers to predict the middle layer
Model training and Validation
Testing
Dice score analysis
Clinical evaluation
5 doctors' clinical reports based on segmentation model performance

FIGURE 2 | Flowchart of the patients' distribution in training and clinical evaluation sets. Model training and clinical evaluation steps use the distribution and classification of all samples in each step.

Incorrect label: the model assigned additional labels that were not in the ground truth label during the test.

Ground truth: two different chief physicians double confirmed the ground truth label in the test dataset (72 patients).

Doctors were requested to perform the segmentation to the best of their abilities, without any constraint on time or duration to ensure that they evaluated the data in its best state. They revised the prediction results obtained from DLS when the prediction results were missing or incorrect.

Statistical Analysis

The clinical evaluation was performed by five clinical medical doctors to assess the deep learning segmentation results. The interradiologist agreement test was performed for each validation case using the SPSS software (version 20.0) (IBM, Armonk, NY, United States). The statistical significance was set to $P < 0.05$, and a kappa value >0.21, based on the ground truth.

RESULTS

Patient Demographic Characteristics

In **Figure 2**, we present the entire method for the DLS setup. A total of 1,615 MRI examinations with l0,525 lesions identified in 1,615 patients were included. We randomly distributed these data into a training cohort and a validation cohort; thus, no significant differences in sex or age were observed.

Additionally, our dataset comprised an adequate number of CMBs, and the distribution of the lesion data had no

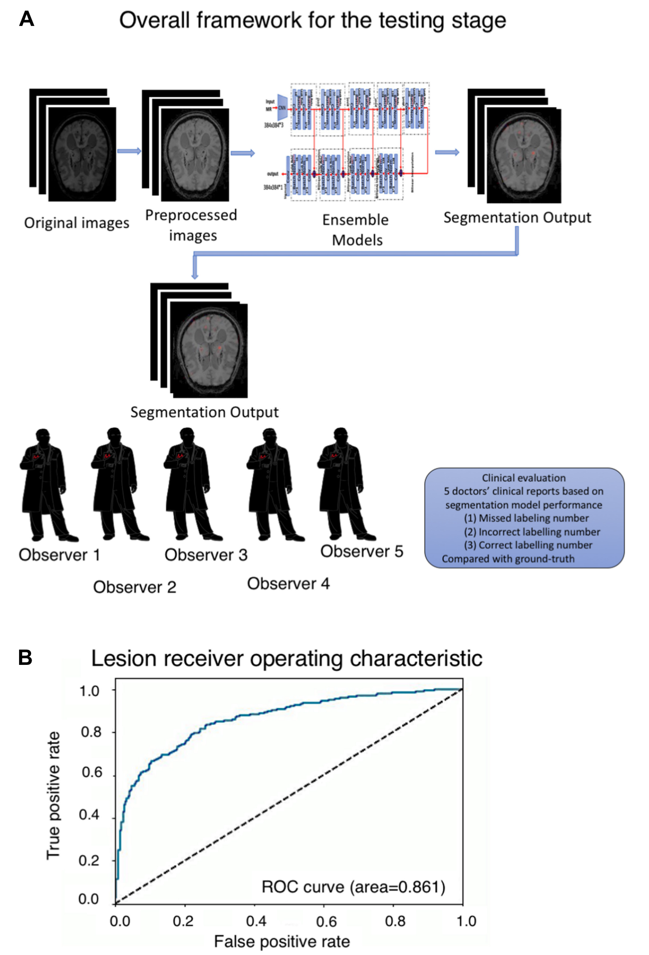

A
Overall framework for the testing stage
Original images
Preprocessed images
Ensemble Models
Segmentation Output
Segmentation Output
Observer 1
Observer 2
Observer 3
Observer 4
Observer 5
Clinical evaluation
5 doctors' clinical reports based on segmentation model performance
(1) Missed labeling number
(2) Incorrect labelling number
(3) Correct labelling number
Compared with ground-truth
B
Lesion receiver operating characteristic
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area=0.861)

FIGURE 3 | (A) Overall framework for the testing stage. **(B)** Segmentation receiver operating characteristic (ROC) curve and area under the curve (AUC) score in the lesion level analysis.

bias. The parameters of these data were obtained from a similar investigator and scanner. We confirmed the scanner parameters of pixel and thinness. The brightness and contrast were normalized before being input into the DLS system.

DLS Set-Up and Performance of the DLS Contouring Method

A total of 10,525 lesions were manually labeled to establish the DLS. Briefly, we manually labeled approximately 9,387 small size lesions (2–5 mm, 7,461 lesions for the training set and 1,926 lesions for the validation set) and 1,138 large lesions (5–10 mm, 927 lesions for the training set, and 211 lesions for the validation set) for training and validation (**Table 1**). The network architecture of the proposed 3-dimensional convolutional neural network is shown in **Figure 1**, and more detailed information about the network is presented in the methods section. After training and validation, the DLS was tested using the testing dataset. The average pixelwise DSC, precision, and recall of the proposed DLS reached 0.72, 0.718, and 0.765, respectively. The 3D lesionwise precision and recall reached 0.751 and 0.852, respectively. A lesion level analysis was performed on the independent test set, and the results showed that the sensitivities of detecting the small and large lesions were 84.4 and 93.51%, respectively. Additionally, in the lesion level analysis, the AUC score of the proposed DLS system was 0.861, and the ROC curve is shown in **Figure 3B**. The patient level analysis was also performed with the FP, FN, TN, and TP of 10, 2, 84, and 93, respectively. And the FP rate and FN rate were 0.106 and 0.021, respectively. The detailed results of the model are presented in **Table 3**.

TABLE 3 | Model performance obtained from the testing dataset.

	DSC	Precision	Recall	Sensitivity	Specificity
Small lesions	0.71	0.707	0.762	84.4%	78.07%
Large lesions	0.73	0.729	0.768	93.51%	83.72%
In average	0.72	0.718	0.765	/	/

TABLE 4 | Clinical evaluation.

	Observer 1	Observer 2	Observer 3	Observer 4	Observer 5	Average
Correct label	787 (94.4%)	770 (92.8%)	790 (94.6%)	784 (93.3%)	761 (91.2%)	778.4 (93.3%)
Incorrect label	27 (3.2%)	44 (5.3%)	24 (2.9%)	30 (3.6%)	53 (6.4%)	35.6 (4.3%)
Missed label	20 (2.4%)	16 (1.9%)	21 (2.5%)	26 (3.1%)	20 (2.4%)	20.6 (2.5%)

The data show the relabeling results after a comparison between the DLS tool and expert labeling results. From the data, we found that the labels attained from the model were accurate and perfectly matched the contour of the real signal. However, the labeling tools and pixels did not adequately control the manual labeling, and the Dice score did not adequately reflect the DLS segmentation results. These data could only support DLS training and validation. Visually, we examined all the data and found that our DLS and human experts had strong consistency in the lesion contour, but the Dice score was low, as described above.

Assessment of DLS-Generated Contours by Human Experts

Figure 3A presents the overall framework for data prediction and the clinical evaluation process. Based on the labeling sensitivity and Dice score, the sensitivity of labeling small lesions was approximately 84.47% and that of large lesions was approximately 93.51%, with an average Dice score of approximately 0.72. The specificity of labeling was approximately 78.07% for small lesions and 83.72% for large lesions (**Table 3**).

However, in the clinical evaluation, the doctors evaluated the output labeling results and revised the labeling results to "missed label," "incorrect label," and "correct label." The missed label group included approximately 20.6 lesions on

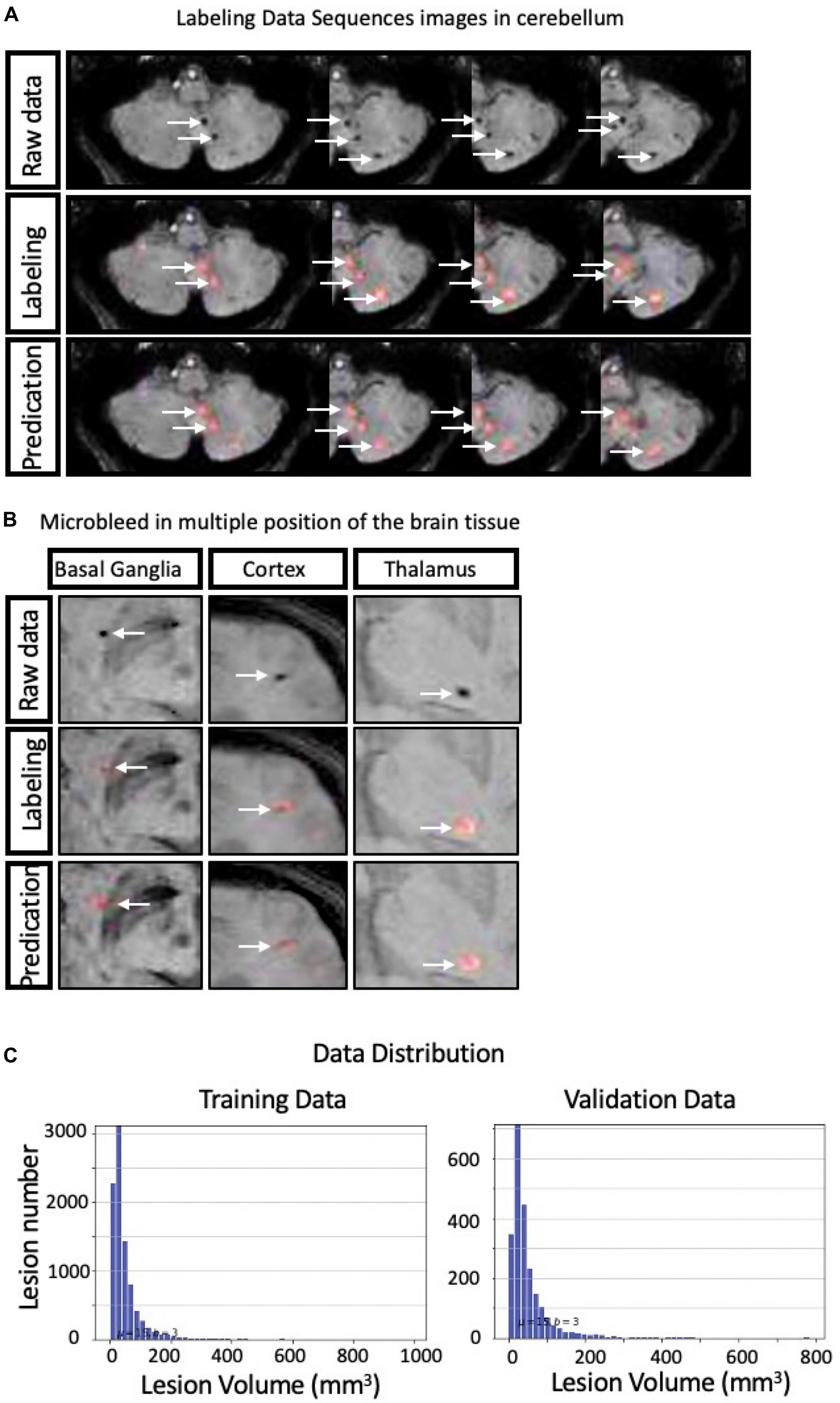


FIGURE 4 | Representative cases of manual cerebral microbleeding (CMB) labeling and labeling with the deep learning system (DLS) system **(A,B)** along with the data distribution **(C)**.

average and 2.5% in total, and the incorrect-label group included approximately 35.6 lesions on average and 4.3% in total. We concluded that using DLS as a contouring accuracy evaluation criterion is reliable and provides accurate lesion quantification. The average kappa value for the internal agreement between observers and DLS prediction was 0.79. Detailed information in the clinical evaluation is presented in **Table 4**. Several examples obtained from the DLS are shown in **Figure 4** and compared with those obtained manually.

DISCUSSION

Cerebral microbleed (CMB) is closely related to many diseases, including SVD, AD, and CADASIL. A previous study has shown that in addition to the location of the CMB, the number is also an independent predictor of the severity of cognitive impairment and dementia in multiple fields (25). Therefore, systematic and accurate quantification of CMB is of great clinical significance.

Based on the SWI data from 1,615 patients with a total of 10,525 lesions, we established a DLS that automatically and objectively segmented CBMs. Compared with other studies (26–28) that automatically recognized CMBs using deep convolutional networks, our DLS was trained with a larger dataset, and the sensitivity and specificity of the model were high, suggesting that it was reliable and would better serve clinicians. Previous studies usually adopted 2D CNNs to construct automatic detection systems, but they lacked upper and lower slice information. In our study, the FN rate of the DLS was low, and we used the widely-used modified 3D U-Net architecture. It made full use of the spatial information of biomarkers and accelerated the computing speed. In addition to studies that used 2D/3D CNNs to detect CMBs on MRI-SWI, Chesebro et al. (29) presented an algorithm for microbleed automated detection using geometric identification criteria (MAGIC) to detect CMBs automatically. It has reasonable precision on both T2*-weighted GRE images and SWI and had high sensitivity in longitudinal identification, with 50% of longitudinal microbleeds correctly labeled. Limited to the algorithm, this study was unable to discriminate between edge artifacts and true positives better than other studies using deep convolutional networks.

We evaluated the DLS performance based on the Dice score, which allows for minor uncertainties in the neighborhood of a few pixels, and the region-wise F1 score, which may not be a suitable indicator for success in evaluating lesions. The AUC of our DLS was 0.861, revealing the excellent performance of the system. Due to the high AUC and low FP rate and FN rate, we propose that it accurately quantified CMBs. Based on a manual data recheck and the variation in lesion marking by individual neuroradiologists, we performed a clinical evaluation based on a multicenter analysis with a score scale. Our DLS performed favorably according to the evaluation by neuroradiologists with an average accuracy of 93.3%. Our marking results were directly or clinically accepted, with most DLS-identified CMBs agreed upon by expert specialists. Moreover, these processes were performed much faster than the manual evaluation process [DLS 2.8 s/case vs. doctors 146 s/case (on average)], which is time-consuming and produces systemic and quantified results, significantly minimizing heterogeneity among neuroradiologists in the delineation of lesions. The results in this study showed that our model was used for the diagnosis and evaluation of CMBs and was more reliable than manual evaluation performed by specialists.

Limitations

First, several different MRI devices with varying scan parameters produced all the images evaluated in this study. The use of these images might increase the data diversity in training the algorithm and testing interpretation subjectivity. However, we were unable to include all the different devices or their corresponding parameter sets for each patient. Therefore, the further clinical application of our system may be challenging due to this limitation. Second, due to the heterogeneity in different neuroradiologists' clinical backgrounds, the accurate recognition and consistent interpretation of the number and location of CMBs by all of these clinicians was challenging. Notably, all annotations made in the dataset have been endorsed by associate chief physicians with at least 15 years of experience. As our model was trained on these data, the limitation of clinical experience in these doctors might affect their evaluation of CMBs and subsequently affect the training process. Improving data quality using more experienced doctors and rigorous training of study protocols might optimize the reliability of training our DLS model. Third, the well-trained DLS has advantages in overcoming the heterogeneity of individual human interpretations with good consistency based on the training features (30). This automated procedure is independent of clinical experience, overcoming limitations imposed by an individual physician's visual sensitivity and clinical experience. The results from the DLS report are produced instantly by the graphical processing unit after input with the output scanning results, which is helpful for neuropathologists to perform the interpretation process faster. Prospective clinical studies are needed to determine whether this hypothesis is valid, and the interpretation should also be modified by performing a post-DLS analysis to match the equipment and the DLS. In addition, our DLS system for CMBs is based only on MRI-SWI and does not include other useful clinical diagnostic information, such as natural history and other imaging performance in the resulting output. Thus, the information is limited in producing a powerful and clinically significant prediction, and differential diagnosis, such as calcification and normal vascular fluid voids, is sometimes needed. Currently, our DLS only serves as a method for assisting neuroradiologists. Future studies involving more comprehensive clinical information are necessary. Another limitation of this study is that it is based on local and regional data. All the data were collected in China and thus do not include data from other countries and regions.

CONCLUSION AND CONTRIBUTIONS

In summary, we developed a DLS tool to perform the CMB lesion segmentation. Our results show that DLS can significantly and quickly masks CMBs in less time to reduce physicians' repetitive labor. Additionally, based on the DLS model, variation within and between neuroradiologists might be reduced. The resulting output produced by the system will be more subjective.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Beijing Tiantan Ethics Committee. According to the requirements of national legislation and institutions, written informed consent for participation was not required for this study.

AUTHOR CONTRIBUTIONS

PF, WS, and HY wrote the initial draft of the manuscript. WS prepared figures and made preliminary revisions. YZ and ZW contributed to DLS development and medical test organizations. QW, ZW, SWC, and KH performed the preliminary revision. YL, KH, and BS performed crucial revisions. All authors together planned the manuscript, critically revised the initial draft, and made final improvements prior to submission.

FUNDING

This study was funded by the China National Neurological Clinical Research Centre, Beijing Postdoctoral Research Foundation (No. ZZ 2019-09), and China Postdoctoral Science Foundation (No. 2019M660719).

REFERENCES

1. Petrault M, Casolla B, Ouk T, Cordonnier C, Berezowski V. Cerebral microbleeds: beyond the macroscope. *Int J Stroke.* (2019) 14:468–75. doi: 10.1177/1747493019830594
2. Wilson D, Ambler G, Lee KJ, Lim JS, Shiozawa M, Koga M, et al. Cerebral microbleeds and stroke risk after ischaemic stroke or transient ischaemic attack: a pooled analysis of individual patient data from cohort studies. *Lancet Neurol.* (2019) 18:653–65. doi: 10.1016/S1474-4422(19)30197-8
3. Shuaib A, Akhtar N, Kamran S, Camicioli R. Management of cerebral microbleeds in clinical practice. *Transl Stroke Res.* (2019) 10:449–57. doi: 10.1007/s12975-018-0678-z
4. Lee J, Sohn EH, Oh E, Lee AY. Characteristics of cerebral microbleeds. *Dement Neurocogn Disord.* (2018) 17:73–82.
5. Greenberg SM, Vernooij MW, Cordonnier C, Viswanathan A, Al-Shahi Salman R, Warach S, et al. Cerebral microbleeds: a guide to detection and interpretation. *Lancet Neurol.* (2009) 8:165–74. doi: 10.1016/S1474-4422(09)70013-4
6. Shoamanesh A, Kwok CS, Benavente O. Cerebral microbleeds: histopathological correlation of neuroimaging. *Cerebrovasc Dis.* (2011) 32:528–34. doi: 10.1159/000331466
7. Granger JP. An emerging role for inflammatory cytokines in hypertension. *Am J Physiol Heart Circ Physiol.* (2006) 290:H923–4. doi: 10.1152/ajpheart.01278.2005
8. Jeon SB, Kang DW, Cho AH, Lee EM, Choi CG, Kwon SU, et al. Initial microbleeds at MR imaging can predict recurrent intracerebral hemorrhage. *J Neurol.* (2007) 254:508–12. doi: 10.1007/s00415-006-0406-6
9. Vergouwen MD, de Haan RJ, Vermeulen M, Roos YB. Statin treatment and the occurrence of hemorrhagic stroke in patients with a history of cerebrovascular disease. *Stroke.* (2008) 39:497–502. doi: 10.1161/STROKEAHA.107.488791
10. Wilson D, Charidimou A, Ambler G, Fox ZV, Gregoire S, Rayson P, et al. Recurrent stroke risk and cerebral microbleed burden in ischemic stroke and TIA: a meta-analysis. *Neurology.* (2016) 87:1501–10. doi: 10.1212/WNL.0000000000003183
11. Fisher M. Cerebral microbleeds and thrombolysis: clinical consequences and mechanistic implications. *JAMA Neurol.* (2016) 73:632–5. doi: 10.1001/jamaneurol.2016.0576
12. Wardlaw JM, Smith EE, Biessels GJ, Cordonnier C, Fazekas F, Frayne R, et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol.* (2013) 12:822–38. doi: 10.1016/S1474-4422(13)70124-8
13. van Etten ES, Auriel E, Haley KE, Ayres AM, Vashkevich A, Schwab KM, et al. Incidence of symptomatic hemorrhage in patients with lobar microbleeds. *Stroke.* (2014) 45:2280–5. doi: 10.1161/STROKEAHA.114.005151
14. Kono Y, Wakabayashi T, Kobayashi M, Ohashi T, Eto Y, Ida H, et al. Characteristics of cerebral microbleeds in patients with fabry disease. *J Stroke Cerebrovasc Dis.* (2016) 25:1320–5. doi: 10.1016/j.jstrokecerebrovasdis.2016.02.019
15. Charidimou A, Boulouis G, Pasi M, Auriel E, van Etten ES, Haley K, et al. MRI-visible perivascular spaces in cerebral amyloid angiopathy and hypertensive arteriopathy. *Neurology.* (2017) 88:1157–64. doi: 10.1212/WNL.0000000000003746
16. Yates PA, Villemagne VL, Ellis KA, Desmond PM, Masters CL, Rowe CC. Cerebral microbleeds: a review of clinical, genetic, and neuroimaging associations. *Front Neurol.* (2014) 4:205. doi: 10.3389/fneur.2013.00205
17. Yakushiji Y, Wilson D, Ambler G, Charidimou A, Beiser A, van Buchem MA, et al. Distribution of cerebral microbleeds in the East and West: individual participant meta-analysis. *Neurology.* (2019) 92:e1086–97. doi: 10.1212/wnl.0000000000007039
18. Haller S, Vernooij MW, Kuijer JPA, Larsson EM, Jager HR, Barkhof F. Cerebral microbleeds: imaging and clinical significance. *Radiology.* (2018) 287:11–28. doi: 10.1148/radiol.2018170803
19. Gregoire SM, Chaudhary UJ, Brown MM, Yousry TA, Kallis C, Jager HR, et al. The microbleed anatomical rating scale (MARS): reliability of a tool to map brain microbleeds. *Neurology.* (2009) 73:1759–66. doi: 10.1212/WNL.0b013e3181c34a7d
20. Cordonnier C, Potter GM, Jackson CA, Doubal F, Keir S, Sudlow CL, et al. improving interrater agreement about brain microbleeds: development of the Brain Observer MicroBleed Scale (BOMBS). *Stroke.* (2009) 40:94–9. doi: 10.1161/STROKEAHA.108.526996
21. Al-Masni MA, Kim WR, Kim EY, Noh Y, Kim DH. Automated detection of cerebral microbleeds in MR images: a two-stage deep learning approach. *Neuroimage Clin.* (2020) 28:102464. doi: 10.1016/j.nicl.2020.102464
22. Cuadrado-Godia E, Dwivedi P, Sharma S, Ois Santiago A, Roquer Gonzalez J, Balcells M, et al. Cerebral small vessel disease: a review focusing on pathophysiology, biomarkers, and machine learning strategies. *J Stroke.* (2018) 20:302–20. doi: 10.5853/jos.2017.02922
23. Zivadinov R, Ramasamy DP, Benedict RR, Polak P, Hagemeier J, Magnano C, et al. Cerebral microbleeds in multiple sclerosis evaluated on susceptibility-weighted images and quantitative susceptibility maps: a case-control study. *Radiology.* (2016) 281:884–95. doi: 10.1148/radiol.2016160060
24. Sudre CH, Li W, Vercauteren T, Ourselin S, Cardoso MJ. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and*

7th International Workshop, ML-CDS 2017. (Vol. 10553), Cham: Springer (2017). p. 240–48. doi: 10.1007/978-3-319-67558-9_28

25. Seo SW, Hwa Lee B, Kim EJ, Chin J, Sun Cho Y, Yoon U, et al. Clinical significance of microbleeds in subcortical vascular dementia. *Stroke*. (2007) 38:1949–51. doi: 10.1161/STROKEAHA.106.477315
26. Chen Y, Villanueva-Meyer JE, Morrison MA, Lupo JM. Toward automatic detection of radiation-induced cerebral microbleeds using a 3D deep residual network. *J Digit Imag*. (2019) 32:766–72.
27. Qi D, Hao C, Lequan Y, Lei Z, Jing Q, Defeng W, et al. Automatic detection of cerebral microbleeds from MR images *via* 3D convolutional neural networks. *IEEE Transact Med Imag*. (2016) 35:1182–95. doi: 10.1109/TMI.2016.2528129
28. Chen H, Yu L, Dou Q, Shi L, Mok V, Heng PA. Automatic detection of cerebral microbleeds *via* deep learning based 3D feature representation. In: *Proceedings of the 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. Brooklyn, NY: IEEE (2015). p. 764–7.
29. Chesebro AG, Amarante E, Lao PJ, Meier IB, Mayeux R, Brickman AM. Automated detection of cerebral microbleeds on T2*-weighted MRI. *Sci Rep*. (2021) 11:4004. doi: 10.1038/s41598-021-83607-0
30. Duan Y, Shan W, Liu L, Wang Q, Wu Z, Liu P, et al. Primary categorizing and masking cerebral small vessel disease based on "deep learning system". *Front Neuroinform*. (2020) 14:17. doi: 10.3389/fninf.2020.00017

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Neuroimaging

ORIGINAL RESEARCH
published: 25 April 2022
doi: 10.3389/fnimg.2022.832512

Weakly Supervised Skull Stripping of Magnetic Resonance Imaging of Brain Tumor Patients

Sara Ranjbar[1], Kyle W. Singleton[1], Lee Curtin[1], Cassandra R. Rickertsen[1], Lisa E. Paulson[1], Leland S. Hu[1,2], Joseph Ross Mitchell[3,4]† and Kristin R. Swanson[1]†*

[1] Mathematical NeuroOncology Lab, Department of Neurosurgery, Mayo Clinic, Phoenix, AZ, United States, [2] Department of Diagnostic Imaging and Interventional Radiology, Mayo Clinic, Phoenix, AZ, United States, [3] Department of Medicine, Faculty of Medicine & Dentistry and the Alberta Machine Intelligence Institute, University of Alberta, Edmonton, AB, Canada, [4] Provincial Clinical Excellence Portfolio, Alberta Health Services, Edmonton, AB, Canada

Automatic brain tumor segmentation is particularly challenging on magnetic resonance imaging (MRI) with marked pathologies, such as brain tumors, which usually cause large displacement, abnormal appearance, and deformation of brain tissue. Despite an abundance of previous literature on learning-based methodologies for MRI segmentation, few works have focused on tackling MRI skull stripping of brain tumor patient data. This gap in literature can be associated with the lack of publicly available data (due to concerns about patient identification) and the labor-intensive nature of generating ground truth labels for model training. In this retrospective study, we assessed the performance of Dense-Vnet in skull stripping brain tumor patient MRI trained on our large multi-institutional brain tumor patient dataset. Our data included pretreatment MRI of 668 patients from our in-house institutional review board–approved multi-institutional brain tumor repository. Because of the absence of ground truth, we used imperfect automatically generated training labels using SPM12 software. We trained the network using common MRI sequences in oncology: T1-weighted with gadolinium contrast, T2-weighted fluid-attenuated inversion recovery, or both. We measured model performance against 30 independent brain tumor test cases with available manual brain masks. All images were harmonized for voxel spacing and volumetric dimensions before model training. Model training was performed using the modularly structured deep learning platform NiftyNet that is tailored toward simplifying medical image analysis. Our proposed approach showed the success of a weakly supervised deep learning approach in MRI brain extraction even in the presence of pathology. Our best model achieved an average Dice score, sensitivity, and specificity of, respectively, 94.5, 96.4, and 98.5% on the multi-institutional independent brain tumor test set. To further contextualize our results within existing literature on healthy brain segmentation, we tested the model against healthy subjects from the benchmark LBPA40 dataset. For this dataset, the model achieved an average Dice score, sensitivity, and specificity of 96.2, 96.6, and 99.2%, which are, although comparable to other publications, slightly lower than the performance of models trained on healthy patients. We associate this drop in performance with the use of brain tumor data for model training and its influence on brain appearance.

Keywords: MRI, brain tumors, brain extraction, skull stripping, deep learning, weakly supervised learning

OPEN ACCESS

Edited by:
Susanne Wegener,
University of Zurich, Switzerland

Reviewed by:
Sidong Liu,
Macquarie University, Australia
Juan Miguel Valverde,
University of Eastern Finland, Finland

***Correspondence:**
Sara Ranjbar
ranjbar.sara@mayo.edu

†These authors share senior authorship

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroimaging

***Received:** 09 December 2021*
***Accepted:** 21 February 2022*
***Published:** 25 April 2022*

Citation:
Ranjbar S, Singleton KW, Curtin L, Rickertsen CR, Paulson LE, Hu LS, Mitchell JR and Swanson KR (2022) Weakly Supervised Skull Stripping of Magnetic Resonance Imaging of Brain Tumor Patients. Front. Neuroimaging 1:832512. doi: 10.3389/fnimg.2022.832512

INTRODUCTION

Magnetic resonance imaging (MRI) has a pivotal role in noninvasive diagnosis and monitoring of many neurological diseases (Fox and Schott, 2004; Bauer et al., 2013). The large amount of data produced in routine patient care has prompted the birth of many studies aiming to automate image analysis tasks relevant to patient care including volumetric analyses (Filipek et al., 1997; Shattuck et al., 2001), tissue classification (Hu et al., 2015, 2017; Kickingereder et al., 2016; Ramkumar et al., 2017), disease staging (Chaddad et al., 2018; Ranjbar et al., 2019b), and localization of pathology (Fox and Schott, 2004; Bauer et al., 2013). To successfully characterize both normal baseline and pathological deviation (Kalavathi and Prasath, 2016) on MRI, non-brain tissues such as fat, skull, eyeballs, eyes, and teeth need to be removed from images, as well as cerebrospinal fluid (CSF) surrounding the brain. As manual annotation of brain tissue in a volumetric MRI is excruciatingly labor intensive, many automatic “whole brain extraction” or “skull stripping” techniques have been introduced in the literature to tackle this need. Separating brain and non-brain tissue has been achieved using edge-based (Somasundaram and Kalaiselvi, 2011; Speier et al., 2011), intensity-based (Ashburner and Friston, 2000; Hahn and Peitgen, 2000), and deformable surface-based methods (Smith, 2002; Jenkinson et al., 2005; Zhuang et al., 2006; Galdames et al., 2012). Atlas-based (Leung et al., 2011) and patch-based (Eskildsen et al., 2012; Roy et al., 2017) methods define the boundaries of the brain by registering images to one or many atlases either on the entire image or on nonlocal image patches. Hybrid methods (Segonne et al., 2001; Rehm et al., 2004) that integrate several of the above approaches have been found (Boesen et al., 2004; Iglesias et al., 2011) superior to any individual method in accuracy at the expense of time efficiency.

However, these methods offer fluctuating accuracies with heterogeneous datasets with varying levels of image resolutions, noise, and artifacts (Kalavathi and Prasath, 2016), and as they are designed for healthy brains, they fail in the presence of pathological conditions on images (Speier et al., 2011). Glioblastoma (GBM), a brain tumor known for its diffuse infiltration, creates serious challenges for most skull stripping methods because of large regions of edema or administration of contrast agents during the examination (Speier et al., 2011). Moreover, GBMs are often cortically localized with abnormalities extending to the edge of the brain and deformities in MRI known as brain shift, which can throw off morphological skull stripping approaches that have rigid assumptions about brain appearance.

Recent success of deep learning has made a lasting impact in computer vision and by extension in biomedical image analysis. Deep convolutional neural networks (CNNs) have shown success in several neuroimaging applications such as MR sequence classification (Ranjbar et al., 2019a), prediction of genetic mutation using MRI (Chang et al., 2018; Yogananda et al., 2019), and tumor segmentation (Işin et al., 2016; Pereira et al., 2016). Naturally, several works have explored the utility of deep learning approaches in MRI skull stripping (Kleesiek et al., 2016; Mohseni Salehi et al., 2017) and have reported high performance on publicly available datasets of normal brains. Given the level of variability that we routinely observe in brain tumor data with respect to image quality as well as shape, size, and the location of abnormalities, rule-based approaches might not be well-suited for skull stripping MRI data in oncology, and there is a need for learning-based approaches for skull stripping MRI of patients with brain tumors. However, labeled training data are scarce in this case as whole-brain labels require substantial time to obtain and have no immediate clinical utility. In the absence of fully ground truth labels, weakly supervised learning, where imperfect and inexact labels are used for model training, offers a more approachable alternative and has previously shown success in segmentation of brain structures on MRI (Bontempi et al., 2020). In this work, we assessed the performance of a weakly supervised three-dimensional (3D) skull stripping approach to generate brain masks for multi-institutional brain tumor data when training data were also brain tumor data. To the best of our knowledge, our work is the first of its kind as no previous study has explored the use of both imperfect labels and pathological MRIs to train a skull stripping model.

The contributions of our work are therefore (1) training a 3D CNN for brain extraction leveraging a diverse set of multi-institutional brain tumor data for model training, (2) use of imperfect automatically generated labels for ground truth, (3) comparison of results across two clinically standard MRI sequences (T1-weighted post injection of gadolinium contrast ([T1Gd] or fluid-attenuated inversion recovery [FLAIR]) used in oncology, and (4) assessing the performance of a skull stripping model trained on brain tumor data on a dataset of healthy subjects.

MATERIALS AND METHODS

Data

Brain Tumor Images

Our in-house institutional review board (IRB)–approved repository [described in our previous work; Ranjbar et al., 2019a), which contains more than 70,000 serial structural MR studies of 2,500+ unique brain tumor patients acquired across 20+ institutions, was used as the source of brain tumor data. We included paired pretreatment T1Gd and FLAIR series of 668 adult brain tumor image series. The vast majority of this dataset consists of one imaging time point per patient with available T1Gd and FLAIR series, with the exception of one patient with two time points and another with three, which were also acquired at different institutions. We used patients with paired imaging available to compare model performance across different input combinations without concerns about dataset differences influencing the results. We also excluded post-treatment images from the cohort as brain tumor treatment typically including surgery, radiation, and chemotherapy can have varying effects on the appearance of MRI. Because of the retrospective nature of our database, various anatomical and quantitative MRI sequences were available for our patients, and the availability of a certain sequence was dependent on the decision of the patient’s clinical team. We chose to include only T1Gd and FLAIR sequences because of their common use in clinical practice and their prevalence in our database. These

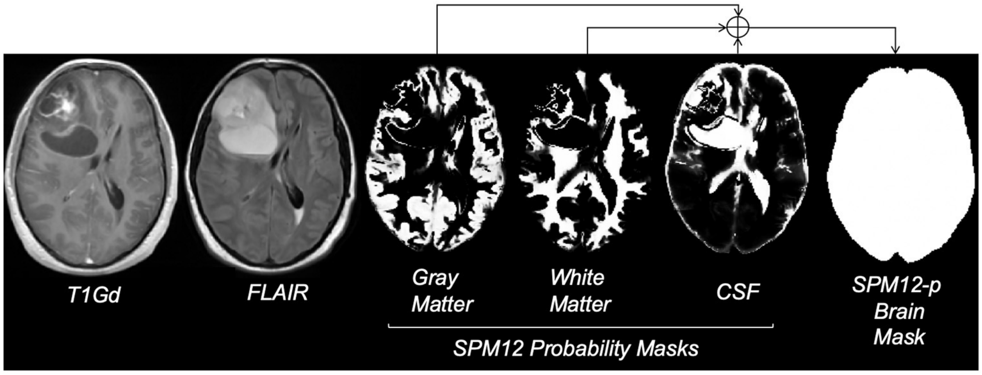


FIGURE 1 | Steps for creating the SPM12-p brain masks; images reflect the MRI of a 29-year-old male brain tumor patient with a diagnosis of GBM. FLAIR refers to fluid-attenuated inversion recovery MRI and T1Gd refers to T1-weighted MRI with gadolinium contrast enhancement. Gray matter, white matter, and CSF probability masks were generated using the SPM12 software. Bright voxels in these masks reflect higher probability. The final brain mask was generated by combining probability masks, using a threshold of 0.7, and minimal post-processing.

series were randomly assigned to 586 training, 52 validation, and 30 test cases. Imaging time points from the same patient were placed in the same data split. As creating ground truth labels for the entire brain on volumetric MRI is very cumbersome and time-consuming, the number of test cases were limited to only 30.

As the data were acquired between 1990 and 2016, many factors varied among samples including field strength and acquisition parameters. We used a number of preprocessing steps to harmonize the data including noise reduction with nonlinear curvature-flow noise reduction (Sethian, 1999), radiofrequency non-uniformity correction reduced using the N4 algorithm (Tustison et al., 2010), resizing to a common matrix size of 240 × 240 × 64 voxels and a voxel resolution of 1 × 1 × 2 mm. The SimpleElastix framework (Marstal et al., 2016) was used to rigidly coregister the FLAIR image to the T1Gd image within each study to enable a comparative experiment of model training on both sequences simultaneously.

Brain Tumor Labels

Given the large size of our cohort and the time-consuming nature of manual segmentation, we devised an automatic approach to substitute manual delineation of brain masks for model training. We used the Statistical Parameter Mapping (Penny et al., 2011) software SPM12, which contains tools for processing many neuroimaging modalities including structural MRI. SPM12 software generated probability maps for gray matter, white matter, and CSF from all T1Gd MRIs. For each case, the maps were combined into a single map and binarized using 0.7 probability (empirically decided) to generate a brain mask. In some cases, the presence of tumor necrosis resulted in occasional missing areas inside the combined mask, which we accounted for by performing minimal morphological operations erosion followed by dilation to fill in the gaps. The final post-processed result for each brain (referred to as SPM12-p) was stored as a label for model training and validation (**Figure 1**). SPM12 was run in MATLAB version 2018a, and postprocessing steps were executed in Python 3.6.6. This process was also conducted on test cases to allow for comparison of labels with manual ground truth.

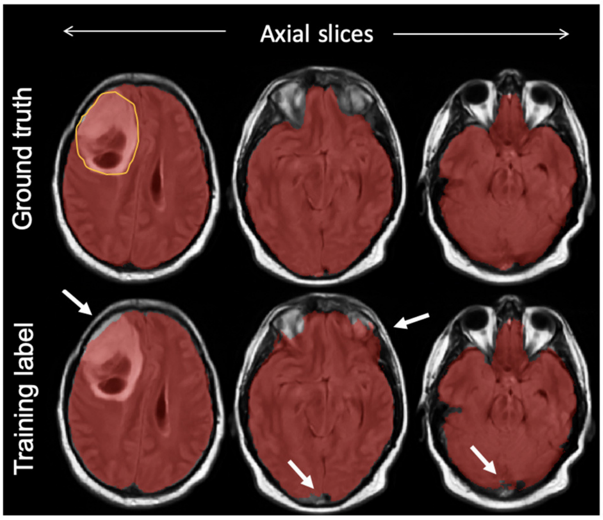


FIGURE 2 | An example of a final training label compared with ground truth; semiautomatically generated training labels were created using SPM12 software. As highlighted with arrows, compared with ground truth delineated manually, the training label included some undersegmentation and oversegmentation particularly around the edges of the brain, but included the bulk of the tumor (outlined on top left slice).

On the test set, we manually segmented brain regions to establish ground truth for estimating model performance. The intracranial volume was defined as the combination of gray matter, white matter, subarachnoid CSF, ventricles (lateral, third, fourth), and cerebellum as suggested by a previous work in the literature (Roy et al., 2017). Manual segmentation was initiated by one of two trained individuals with experience in MRI tumor segmentation using our in-house semiautomatic software. The results were further loaded into the ITK-SNAP

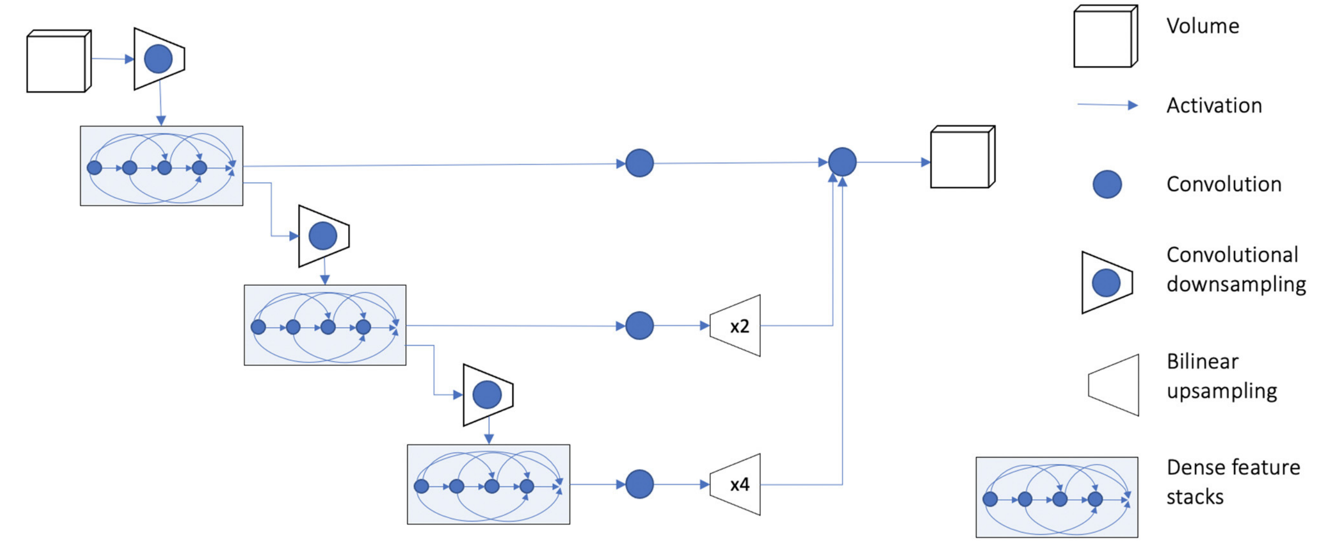


FIGURE 3 | Overview of model architecture. Detailed description of the model architecture is available in Gibson et al. (2018a). The output of the model is resized to the original input image dimension during postprocessing. An implementation of the model is available in the NiftyNet platform (http://niftynet.io) in code repositories.

(Yushkevich et al., 2006) software version 3.8.0 and corrected manually by a third individual as needed. **Figure 2** compares the manual mask and SPM12-p label for one of the test cases.

To further enable comparison with existing atlas-based skull stripping methods in the literature, we generated a third set of labels for the test cases using the Multi-cONtrast brain STRipping method (MONSTR; Roy et al., 2017), a patch-based multiatlas skull stripping method. Although not extensively tested on brain tumor patient data, MONSTR is a benchmark skull stripping approach that was advertised for having success in brain extraction of pathological MRI including patients with traumatic brain injuries and tumors. We refer to these brain masks as MONSTR masks hereon. MONSTR masks were generated using both T1Gd and FLAIR contrasts as inputs.

Healthy Subjects Data

The publicly available LONI Probabilistic Brain Atlas Project (LBPA40) (Shattuck et al., 2009) consisting of T1-weighted MRI of 40 healthy subjects was used for evaluation of the model against publicly available benchmarks. The corresponding manually delineated brain masks included in this dataset were used as ground truth. Although training data for this work were entirely brain tumor patients, using this dataset will allow us to contextualize our work within the existing skull stripping literature that have evaluated their approach on MRI of healthy subjects.

Model Training and Convolutional Neural Network

We used TensorFlow (version 1.12.0) and the medical imaging deep learning platform NiftyNet (Li et al., 2017; Gibson et al., 2018b; version 0.6.0) for implementation of all experiments. NiftyNet is a modularly structured deep learning platform tailored toward medical image analysis applications with modules for preprocessing, network training, evaluation, and inference. Minimal coding is required from the user using this platform, and the specific settings related to preprocessing images, training, and testing can be communicated via a configuration file. We used the 3D fully CNN (Long et al., 2015) architecture known as dense V-network (Dense-Vnet) that has previously demonstrated success in establishing voxel-to-voxel connections between input and output images in multiorgan segmentation of abdominal computed tomography images (Gibson et al., 2018a). The architecture of the model is shown in **Figure 3**, and it only differs from the original model in the size of input image (in our case, 240 × 240 × 64) and the lack of priors. The encoder block of the segmentation network generates three different sized sets of feature maps using dense feature stacks (Huang et al., 2017). The outputs are upsampled using the decoder block so that the smaller feature maps match the original input size. The final output is the concatenated version of all outputs after a single convolution in the skip connection. It should be noted that the Dense-Vnet architecture is designed to work with a smaller version of the original image to constrain memory usage (i.e., the first convolutional downsampling layer in **Figure 3**), and the final output is resized to the original image size during postprocessing. An implementation of the model and post-processing is available in the NiftyNet platform (http://niftynet.io).

Hyperparameters included learning rate, optimizer, and augmentation, which were selected using the validation set. Training was conducted using He weight initialization (He et al., 2015), whitening (scaling image intensities to 0–1), adam (Kingma and Ba, 2014) optimizer with a batch size of 6, and the Dice coefficient as the loss criteria (Milletari et al., 2016). We trained the model for a maximum of 300 iterations, and the model that performed best on the validation set was used as the final model. It should be added that the results reported

TABLE 1 | Comparison of model performance across input type on the test set.

Model input	Dice score	Sensitivity	Specificity	Hausdorff distance
T1Gd	93.09 (1.78)	96.14 (3.81)	97.92 (1.28)	3.69 (0.55)
FLAIR	**94.54** (1.09)	**96.39** (2.34)	98.48 (1.05)	**3.39 (0.44)**
T1Gd + FLAIR	94.47 (1.61)	94.80 (3.49)	**98.84** (0.79)	3.44 (0.49)

Values indicate mean and standard deviation. Best result is highlighted in bold font.

TABLE 2 | Comparison of performance between model and non-learning methods on the test set.

Method	Dice score	Sensitivity	Specificity	Hausdorff distance
MONSTR	91.34 (6.76)	88.22 (7.44)	**98.91 (2.22)**	3.67 (0.75)
SPM12-p	93.36 (3.75)	93.39 (6.59)	98.76 (1.05)	3.44 (0.80)
Our approach	**94.54 (1.09)**	**96.39 (2.34)**	98.48 (1.05)	**3.39 (0.44)**

Values indicate mean and standard deviation. Best result is highlighted in bold font.

here were generated without the use of any augmentation as data augmentation (including rotation, scaling, and flipping images on the x-axis) did not improve model performance on the validation set. All experiments were conducted on an Ubuntu 17.10 system with a single Nvidia TITAN V GPU. The source code for NiftyNet platform along with instructions on how to call the platform via terminal is available at: https://github.com/NifTK/NiftyNet.

Our trained models along with the complete list of parameters utilized for model training are available at: https://github.com/SARARANJBAR/skullstripping_niftynet.

Experiments

Using only brain tumor data, we evaluated the performance of the network across MRI contrasts by repeating model training three times: first using only T1Gd MRIs, second using only FLAIR MRIs, and finally using both series as inputs. When both T1Gd and FLAIR sequences were provided to the network as input, the two images were simultaneously provided to the model. Apart from input image type, all other training parameters were identical between different runs. We evaluated model performance using Dice similarity coefficient (Kingma and Ba, 2014), sensitivity, specificity, and Hausdorff distance (Kingma and Ba, 2014), comparing predicted labels with manual brain masks. Sensitivity measures the detection rate of brain tissue, and specificity measures how much non-brain tissue is correctly identified, whereas Dice score evaluates the trade-off between sensitivity and specificity, measuring the overlap of predictions and ground truths. Hausdorff distance measures the Euclidean distance between the farthest contours of the ground truth and predictions and is relevant to this work to assess accuracy of predictions at the edge of the brain.

In addition to brain tumor data, we used the healthy subject data from LBPA40 (Gibson et al., 2018b) dataset to evaluate the performance of trained models on a publicly available benchmark. Other deep-learning skull stripping methods in the literature (Chang et al., 2009; Kleesiek et al., 2016; Mohseni Salehi et al., 2017; Lucena et al., 2019) have used this data collection to evaluate their model. Although our model was not trained on healthy subjects, we believe addition of this experiment will help place our work within existing literature. Average Dice score was used as the performance measure. The Dice scores of previous approaches were acquired from their publications.

RESULTS

Table 1 compares the performance of model training on brain tumor data across input types on previously unseen test cases with available ground truth. We found the model trained on FLAIR to achieve the highest Dice score and sensitivity, and the model trained on both sequences was superior to single input models in specificity (98.84%). Our FLAIR-only model achieved a mean Dice score of 94.54%, a sensitivity of 96.39%, and specificity of 98.48% on the test set with available ground truth. The average Dice score for the FLAIR-only model was not significantly higher than that of the model trained on both sequences ($p = 0.83$, t-test) but was significantly higher than that of the T1Gd-only model ($p = 0.00042$), which was also significantly outperformed by the model trained on both ($p = 0.0027$). The model trained on both modalities achieved a slightly higher but non-significant mean specificity than the FLAIR-only model ($p = 0.14$), with the FLAIR model significantly outperforming the model trained on both in mean sensitivity ($p = 0.043$). The T1Gd model was significantly lower in mean specificity than the model trained on both modalities ($p = 0.0016$) and lower than the model trained only on FLAIR; this result was not significant ($p = 0.068$). The T1Gd-only model had a slightly lower mean sensitivity than the FLAIR-only model ($p = 0.7612$). The average Hausdorff distance between the predictions of the FLAIR model and ground truth was also superior to that of T1Gd-only ($p = 0.023$) and dual input (T1Gd + FLAIR) models ($p = 0.71$). **Table 2** compares the performance of our model with non-learning methods MONSTR and SPM12. While MONSTR did not fail to include the regions occupied by tumors into the segmentation, its performance was much worse in identifying the boundaries of the brain in other regions, and oversegmentation and undersegmentation were observed at the top and bottom slices. In comparison, SPM12-p showed a much improved sensitivity. Our model was superior in Dice score, Hausdorff distance, and sensitivity compared with both non-learning approaches. An example of predicted brain mask and comparison with MONSTR and SPM12 is presented in **Figure 4**. Using the same machine for training, generating an SPM12-p mask required an average of 2–3 min compared with 10–20 min for MONSTR, and 2–3 s for the model. Longer runtime is expected for MONSTR as atlas-based methods tend to take longer than other approaches.

Figure 5 shows two examples of a model prediction (red), ground truth (blue), and overlap (purple) (left). This prediction achieved a relatively low Dice score of 92.4%, with areas of both underprediction and overprediction. In this case, the model more commonly underpredicted the anterior and posterior regions of the brain, while overpredicting the superior and inferior regions. This prediction achieved a relatively high Dice score of 96.6%, primarily underpredicting the superior region and

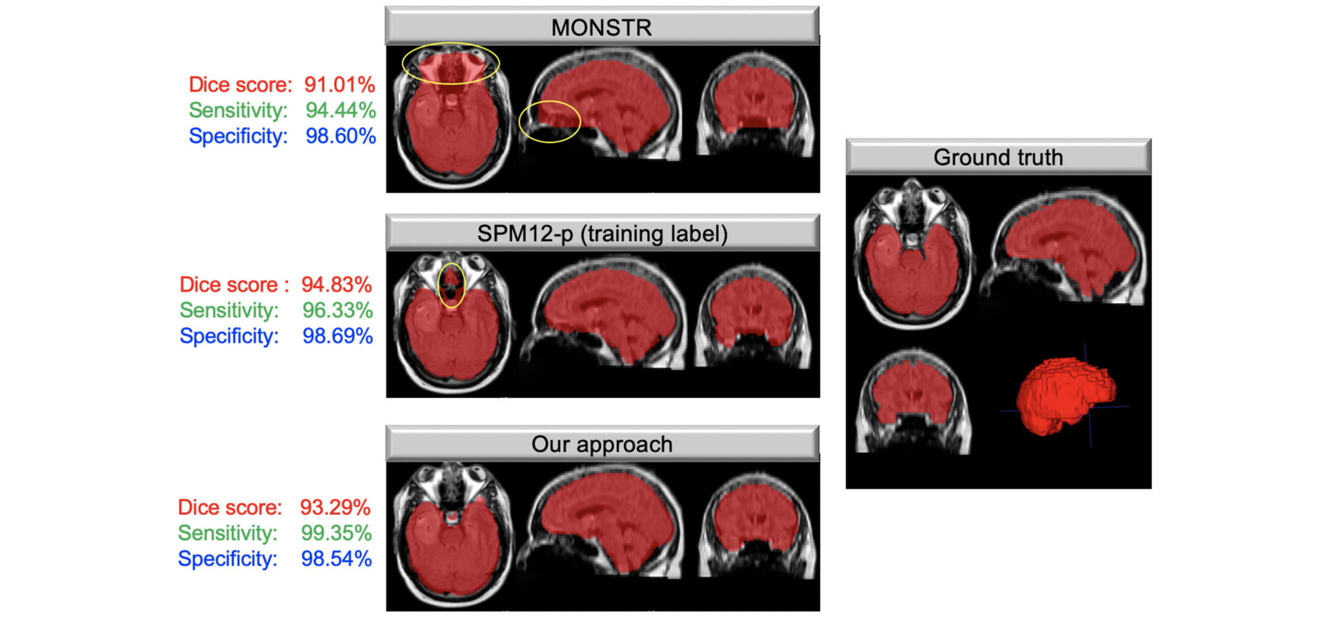


FIGURE 4 | Masks overlaid on brain tumor MRIs; images on the left show the brain masks created using MONSTR, SPM12-p, and our model in different anatomical views. Areas highlighted in yellow show errors in results. The right image shows the ground truth manual segmentation. Our approach performed very well and much better than the other two methods. The Dice coefficient, sensitivity, and specificity, calculated based on the ground truth for this case, are shown to the left of each image.

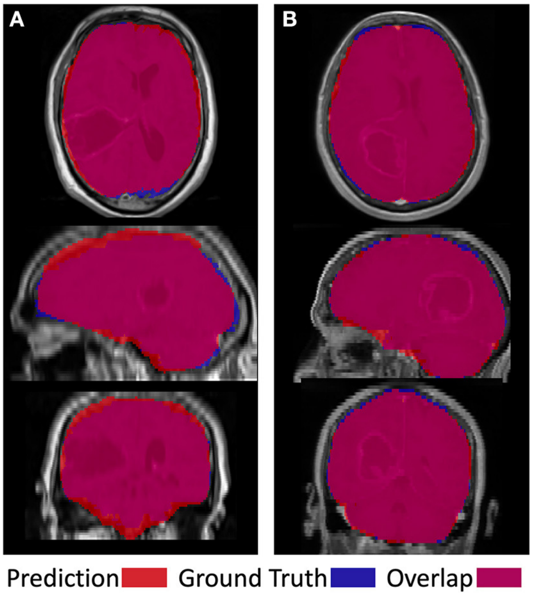


FIGURE 5 | Visual examples of a less successful case **(A)** and a more successful case **(B)**. Prediction is shown in red, ground truth in blue, and overlap in purple.

overpredicting the inferior regions. Importantly, there is no evidence that the net suffered from the presence of tumor abnormalities in either case.

Table 3 presents the performance of our model on healthy subjects. On average, our model achieved a Dice score of 96.2%, sensitivity of 96.6%, and specificity of 99.2% on the LBPA40 dataset. Overall, our results were within the range of those reported by others in similar applications. However, our Dice score and sensitivity were on the lower end of scores. We believe this is expected given that, unlike others, we trained our model using brain-tumor patient data that divert from the normal brain due to imaging patterns resulting from pathology.

TABLE 3 | Comparison of performance with previous literature on healthy brains from the LBPA40 dataset.

Method	Dice score	Sensitivity	Specificity
CONSNet (Milletari et al., 2016)	97.35 (0.003)	97.26 (0.007)	**99.54 (0.001)**
Auto-U-Net (Mohseni Salehi et al., 2017)	**97.73 (0.003)**	**98.31 (0.006)**	99.48 (0.001)
U-Net (Mohseni Salehi et al., 2017)	96.79 (0.004)	97.22 (0.016)	99.34 (0.002)
3D CNN (Kleesiek et al., 2016)	96.96 (0.010)	97.46 (0.010)	99.41 (0.003)
Our approach	96.17 (0.220)	96.60 (0.080)	99.22 (0.090)

Performance measures of others' works are extracted from their publication. Values in bold font indicate the best result.

DISCUSSION

Despite the large body of existing literature on automatic skull striping methods on MRI, few have reported robustness in the presence of a pathology (Thakur et al., 2019). The closest work to ours is the modality-agnostic 3D CNN created by

Thakur et al., Lucena et al. (2019), which was tested on brain tumor data from three different institutions compared with ours with 20+ institutions. Authors trained their network with pretreatment images of glioma patients using T1-weighted, T1Gd, T2-weighted, and FLAIR sequences. Their model achieved an average Dice coefficient of 97.8% on images from the training institution and 95.6, 91.6, and 96.9% on datasets of other institutions. Another learning-based skull stripping approach is the work of Kleesiek et al. (2016), in which authors created a modality-agnostic fully convolutional CNN model with similar input channels as Thakur et al. and achieved an average Dice of 95.2% and a sensitivity of 96.25% on a cohort of 53 brain tumor patients from training institution. Our work differs from these works (Kleesiek et al., 2016; Thakur et al., 2019) in a number of ways. First, our approach is considered weakly supervised, as the network was trained using automatically generated labels with known imperfections (Malone et al., 2015) compared with accurate ground truth delineated by neuroradiologists. The data used in this work were collected at 20+ institutions from 1990 to 2016 using a variety of imaging devices that has been shown to impact the outcome of skull stripping (Rex et al., 2004; Fennema-Notestine et al., 2006). However, we argue that an advantage of this type of data heterogeneity is that it better approximates the data found in clinical practice and therefore can serve as a realistic benchmark for estimating model performance in clinical practice. The fact that our result is within the range of reported performance in Thakur et al. (Lucena et al., 2019) on data from other institutions is a good indicator for this argument. Given that the CSF is dark on both FLAIR and T1Gd images, and brain tissue is brighter than CSF on both images, the major visual difference between the two images is the high intensity of skull on T1Gd and its low intensity the FLAIR image. This can result in a sharper edge at the boundary of the brain on the FLAIR images, which we associate with the improved performance of the FLAIR model. That said, given the small size of our test set and similarly promising results of our other models, we urge the reader not to discount models trained only on T1Gd or a combination of images. One limitation of our work is that we did not train a sequence-agnostic model. In our results, the FLAIR model yielded the highest Dice and sensitivity, and the addition of T1Gd slightly improved specificity. Given the heterogeneity of data types across institutions, a sequence-agnostic approach is beneficial for ensuring utility across data found in clinical practice, and we intend to adopt a similar approach in future work.

Because of the size of our cohort and the labor-intensive nature of manual segmentation, we needed an automatic method to create brain masks for training. We selected SPM12 because of its reported comparable performance with manual delineation in segmenting total intracranial volume on MRI even in the presence of neurodegenerative pathology (Malone et al., 2015). Compared with ground truth, the SPM12-p labels achieved a Dice of 93.34% on the test set. Visualization of model output against ground truth showed the net was not hindered by the presence of tumor abnormalities; rather, the differences in Dice score were related to the overall brain shape. Despite the reported high performance of MONSTR in skull stripping brain tumor data, we found its performance worse than SPM12, demonstrated by comparing the Dice score of generated masks with ground truth (**Table 2**). As a result of this finding, we decided to proceed with model training with SPM12. However, no single automatic method for generating labels can outperform consensus methods that combine different skull-stripping methods through a meta-algorithm and allow for combining the strength of different approaches. In the work of Lucena et al. (Milletari et al., 2016), the authors generated silver standard labels for training using the STAPLE (Warfield et al., 2004) method combining eight different segmentation approaches into a probabilistic consensus mask, and achieved a Dice score of 97.3% and sensitivity of 97.2% on healthy subjects. In comparison, our approach could be considered a "bronze standard" given that our labels were acquired using one segmentation method. In future work, we aim to repeat our analysis using a silver standard.

Among the non–learning-based skull stripping approaches in the literature, the MONSTR algorithm (Roy et al., 2017) was reported to outperform other methods on a small cohort of five brain tumor cases with an average Dice agreement of 96.95% with ground truth. MONSTR achieved a moderate Dice score of 91.34% on the test set. In comparison, SPM12-p outperformed MONSTR, particularly with respect to sensitivity (93.39 vs. 88.22%), as well as average runtime for creating masks (2–3 vs. 10–20 min on the machine used for model training). Discrepancy between the results here and the reported performance in the original paper could also be related to our use of T1Gd and FLAIR inputs for creating MONSTR masks, as opposed to T1Gd and T2W images that were used in the original results (Roy et al., 2017). The worse performance by MONSTR could also be associated with the atlas-based nature of the algorithm, which can result in inaccuracies when images deviate from healthy brain MRIs. The performance of our model on healthy subjects was decidedly on the lower end of reported results for deep learning–based skull stripping models in the literature. Mohseni Salehi et al. (2017) compared the performance of a voxel-wise approach using three convolutional pathways for each anatomical plane and a fully convolutional U-Net (Ronneberger et al., 2015) architecture and achieved Dice coefficients of 97.7 and 96.8% on two publicly available datasets of normal brains. Although the authors used the U-Net architecture, which might be considered dated in today's deep learning context, their approach achieved a higher performance than ours because of their use of different convolutional pathways for each anatomical plane. Kleesiek et al. (2016) used a 3D input-agnostic fully convolutional network and compared its performance to six other skull stripping methods on publicly available datasets. Whereas, Kleesiek et al. (2016) reported the performance of their model on merged public datasets, others (Lucena et al., 2019) reported their performance on the LBPA40 dataset alone to be an average Dice score of 97.0% and sensitivity of 97.4%. Lucena et al. (Milletari et al., 2016) adopted a brain extraction model consisting of three parallel, fully convolutional networks using the U-Net architecture and achieved a Dice score of 97.3% and sensitivity of 97.2%. Here again, the authors utilized parallel pathways to achieve high performance. Our approach did not yield the same level of Dice score on the LBPA40 dataset. We believe this is expected given that unlike others we trained our network using only brain-tumor MRI and did not use manually delineate or consensus methods for training labels. In future work, we intend to adopt a consensus method for creating training labels. To maximize generalizability and utility of this tool, we will supplement brain tumor data with healthy subjects to improve model performance on healthy subjects as well as to stay relevant for utility in clinical settings. In addition to using pathological MRI for model training with suboptimal labels, we adopted a straightforward volumetric training approach with no pathway parallelization for different anatomical planes. This could also explain the drop in our model performance compared with others.

In summary, we assessed the performance of a deep learning model in MRI brain extraction of a diverse multi-institutional brain tumor patient dataset using weak labels. On previously unseen brain tumor cases, our approach reached comparable performance to previous literature. The model underperformed compared with state-of-the-art models in the literature on healthy subjects, which can be attributed to the absence of healthy patients in our training set and our rather simplistic model training approach. The shortcomings can be addressed by fine tuning the model on healthy subjects, leveraging a consensus approach to generating training labels, and allocating training pathways within the model for different anatomical planes. Despite the shortcomings, we believe that our approach can be a practical choice for skull stripping MRI data in repositories of brain tumor patients given its turnaround time and simplicity. In future work, we intend to extend this work to perform skull striping on post-treatment MRIs.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: MR imaging data of brain tumor patients used in this study was acquired from our in-house IRB-approved repository which contains patient information and therefore is subject to HIPAA regulations. Due to the proprietary nature of patient data and patient information, we are not at liberty to freely share data with readers. However, data may be available for sharing upon the request of qualified parties if patient privacy and intellectual property interests of our institution are not compromised. Typically, data access will occur through collaboration and may require interested parties to obtain an affiliate appointment with our institution prior to data access. Requests to access these datasets should be directed to https://mathematicalneurooncology.org. Healthy subject data used in this work were acquired from the publicly available LBP40A dataset. Transforms from delineation and native radiological spaces are available on The LONI Probabilistic Brain Atlas Project webpage at: https://resource.loni.usc.edu/resources/atlases-downloads/.

ETHICS STATEMENT

All procedures performed in the studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Our de-identified data repository of patients with brain cancer includes retrospective data collected from medical records and prospective data collection. Research on the data repository was reviewed and approved by Mayo Clinic Institutional Review Board. Prior to collection of retrospective data, informed consent was waived for those participants by the Mayo Clinic Institutional Review Board (IRB# 15-002337). Written informed consent was obtained for all prospectively enrolled participants as approved by Mayo Clinic Institutional Review Board (IRB# 17-009682). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SR, KWS, JRM, and KRS contributed to study design. SR led model generation and data processing as well as writing the first draft of the manuscript. KWS and KRS created the infrastructure necessary for conducting the study. SR, KWS, JRM, LC, CRR, and LEP contributed to data collection and data preprocessing. LSH was the clinical lead of the study and reviewed the accuracy of the ground truth brain masks. JRM and KRS share senior authorship. All authors have substantially contributed to conducting this research and drafting the manuscript. All authors have edited the manuscript and have approved the contents.

FUNDING

This publication would not have been possible without the support of the James S. McDonnell Foundation, the Ivy Foundation, the Mayo Clinic, the Zicarelli Foundation, and the NIH (R01 NS060752, R01 CA164371, U54 CA210180, U54 CA143970, U54 CA193489, U01 CA220378, and U01 CA250481-02).

REFERENCES

Ashburner, J., and Friston, K. J. (2000). Voxel-based morphometry-the methods. *Neuroimage* 11, 805–821. doi: 10.1006/nimg.2000.0582

Bauer, S., Wiest, R., Nolte, L.-P., and Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumor studies. *Phys. Med. Biol.* 58, R97–R129. doi: 10.1088/0031-9155/58/13/R97

Boesen, K., Rehm, K., Schaper, K., Stoltzner, S., Woods, R., Lüders, E., et al. (2004). Quantitative comparison of four brain extraction algorithms. *Neuroimage* 22, 1255–1261. doi: 10.1016/j.neuroimage.2004.03.010

Bontempi, D., Benini, S., Signoroni, A., Svanera, M., and Muckli, L. (2020). CEREBRUM: a fast and fully-volumetric Convolutional Encoder-decodeR for weakly-supervised sEgmentation of BRain strUctures from out-of-the-scanner MRI. *Med Image Anal.* 62:101688. doi: 10.1016/j.media.2020.101688

Chaddad, A., Desrosiers, C., and Niazi, T. (2018). Deep radiomic analysis of MRI related to Alzheimer's disease. *IEEE Access* 6, 58213–58221. doi: 10.1109/ACCESS.2018.2871977

Chang, H.-H., Zhuang, A. H., Valentino, D. J., and Chu, W.-C. (2009). Performance measure characterization for evaluating neuroimage segmentation algorithms. *Neuroimage* 47, 122–135. doi: 10.1016/j.neuroimage.2009.03.068

Chang, P., Grinband, J., Weinberg, B. D., Bardis, M., Khy, M., Cadena, G., et al. (2018). Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. *Am. J. Neuroradiol.* 39, 1201–1207. doi: 10.3174/ajnr.A5667

Eskildsen, S. F., Coupé, P., Fonov, V., Manjón, J. V., Leung, K. K., Guizard, N., et al. (2012). BEaST: brain extraction based on nonlocal segmentation technique. *Neuroimage* 59, 2362–2373. doi: 10.1016/j.neuroimage.2011.09.012

Fennema-Notestine, C., Burak Ozyurt, I., Clark, C. P., Morris, S., Bischoff-Grethe, A., Bondi, M. W., et al. (2006). Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location. *Hum. Brain Mapp.* 27, 99–113. doi: 10.1002/hbm.20161

Filipek, P. A., Semrud-Clikeman, M., Steingard, R. J., Renshaw, P. F., Kennedy, D. N., and Biederman, J. (1997). Volumetric MRI analysis comparing subjects having attention-deficit hyperactivity disorder with normal controls. *Neurology* 48, 589–601. doi: 10.1212/WNL.48.3.589

Fox, N. C., and Schott, J. M. (2004). Imaging cerebral atrophy: normal ageing to Alzheimer's disease. *Lancet* 363, 392–394. doi: 10.1016/S0140-6736(04)15441-X

Galdames, F. J., Jaillet, F., and Perez, C. A. (2012). An accurate skull stripping method based on simplex meshes and histogram analysis for magnetic resonance images. *J. Neurosci. Methods* 206, 103–119. doi: 10.1016/j.jneumeth.2012.02.017

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., et al. (2018a). Automatic multi-organ segmentation on abdominal CT with dense V-networks. *IEEE Trans. Med. Imaging* 37, 1822–1834. doi: 10.1109/TMI.2018.2806309

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., et al. (2018b). NiftyNet: a deep-learning platform for medical imaging. *Comput. Methods Prog. Biomed.* 158, 113–122. doi: 10.1016/j.cmpb.2018.01.025

Hahn, H. K., and Peitgen, H.-O. (2000). "The skull stripping problem in MRI solved by a single 3D watershed transform," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2000* (Berlin; Heidelberg: Springer), 134–143. doi: 10.1007/978-3-540-40899-4_14

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago). doi: 10.1109/ICCV.2015.123

Hu, L. S., Ning, S., Eschbacher, J. M., Baxter, L. C., Gaw, N., Ranjbar, S., et al. (2017). Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. *Neuro Oncol.* 19, 128–137. doi: 10.1093/neuonc/now135

Hu, L. S., Ning, S., Eschbacher, J. M., Gaw, N., Dueck, A. C., Smith, K. A., et al. (2015). Multi-parametric MRI and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma. *PLoS ONE* 10:e0141506. doi: 10.1371/journal.pone.0141506

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 4700–4708. doi: 10.1109/CVPR.2017.243

Iglesias, J. E., Liu, C.-Y., Thompson, P. M., and Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans. Med. Imaging* 30, 1617–1634. doi: 10.1109/TMI.2011.2138152

Işin, A., Direkoglu, C., and Sah, M. (2016). Review of MRI-based brain tumor image segmentation using deep learning methods. *Proc. Comput. Sci.* 102, 317–324. doi: 10.1016/j.procs.2016.09.407

Jenkinson, M., Pechaud, M., and Smith, S. (2005). "Others. BET2: MR-based estimation of brain, skull and scalp surfaces," in *Eleventh Annual Meeting of the Organization for Human Brain Mapping* (Toronto), 167.

Kalavathi, P., and Prasath, V. B. S. (2016). Methods on skull stripping of MRI head scan images-a review. *J. Digit. Imaging* 29, 365–379. doi: 10.1007/s10278-015-9847-8

Kickingereder, P., Bonekamp, D., Nowosielski, M., Kratz, A., Sill, M., Burth, S., et al. (2016). Radiogenomics of glioblastoma: machine learning-based classification of molecular characteristics by using multiparametric and multiregional MR imaging features. *Radiology* 281, 907–918. doi: 10.1148/radiol.2016161382

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arxiv.org/abs/1412.6980.* doi: 10.48550/arXiv.1412.6980

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., et al. (2016). Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *Neuroimage* 129, 460–469. doi: 10.1016/j.neuroimage.2016.01.024

Leung, K. K., Barnes, J., Modat, M., Ridgway, G. R., Bartlett, J. W., Fox, N. C., et al. (2011). Brain MAPS: an automated, accurate and robust brain extraction technique using a template library. *Neuroimage* 55, 1091–1108. doi: 10.1016/j.neuroimage.2010.12.067

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M. J., and Vercauteren, T. (2017). "On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task," in *Information Processing in Medical Imaging* (Boone, NC: Springer International Publishing), 348–360. doi: 10.1007/978-3-319-59050-9_28

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3431–3440. doi: 10.1109/CVPR.2015.7298965

Lucena, O., Souza, R., Rittner, L., Frayne, R., and Lotufo, R. (2019). "Convolutional neural networks for skull-stripping in brain MR imaging using silver standard masks," in *Artificial Intelligence in Medicine* (Poznan), 48–58. doi: 10.1016/j.artmed.2019.06.008

Malone, I. B., Leung, K. K., Clegg, S., Barnes, J., Whitwell, J. L., Ashburner, J., et al. (2015). Accurate automatic estimation of total intracranial volume: a nuisance variable with less nuisance. *Neuroimage* 104, 366–372. doi: 10.1016/j.neuroimage.2014.09.034

Marstal, K., Berendsen, F., Staring, M., and Klein, S. (2016). "SimpleElastix: a user-friendly, multi-lingual library for medical image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Las Vegas, NV), 134–142. doi: 10.1109/CVPRW.2016.78

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford, CA). doi: 10.1109/3DV.2016.79

Mohseni Salehi, S. S., Erdogmus, D., and Gholipour, A. (2017). Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging. *IEEE Trans. Med. Imaging* 36, 2319–2330. doi: 10.1109/TMI.2017.2721362

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. (2011). *Statistical Parametric Mapping*: *The Analysis of Functional Brain Images*. Elsevier; Academic Press.

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). "Deep convolutional neural networks for the segmentation of gliomas in multi-sequence MRI," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, eds A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels (Munich: Springer International Publishing), 131–143. doi: 10.1007/978-3-319-30858-6_12

Ramkumar, S., Ranjbar, S., Ning, S., Lal, D., Zwart, C. M., Wood, C. P., et al. (2017). MRI-based texture analysis to differentiate sinonasal squamous cell carcinoma from inverted papilloma. *AJNR Am. J. Neuroradiol.* 38, 1019–1025. doi: 10.3174/ajnr.A5106

Ranjbar, S., Singleton, K. W., Jackson, P. R., Rickertsen, C. R., Whitmire, S. A., Clark-Swanson, K. R., et al. (2019a). Deep convolutional neural network for annotation of magnetic resonance imaging sequence type. *J. Digit. Imaging* 33, 439–446. doi: 10.1007/s10278-019-00282-4

Ranjbar, S., Velgos, S. N., Dueck, A. C., Geda, Y. E., Mitchell, J. R., and Alzheimer's Disease Neuroimaging Initiative (2019b). Brain MR radiomics to differentiate cognitive disorders. *J. Neuropsychiatry Clin. Neurosci.* 31, 210–219. doi: 10.1176/appi.neuropsych.17120366

Rehm, K., Schaper, K., Anderson, J., Woods, R., Stoltzner, S., and Rottenberg, D. (2004). Putting our heads together: a consensus approach to brain/non-brain segmentation in T1-weighted MR volumes. *Neuroimage* 22, 1262–1270. doi: 10.1016/j.neuroimage.2004.03.011

Rex, D. E., Shattuck, D. W., Woods, R. P., Narr, K. L., Luders, E., Rehm, K., et al. (2004). A meta-algorithm for brain extraction in MRI. *Neuroimage* 23, 625–637. doi: 10.1016/j.neuroimage.2004.06.019

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015* (Munich: Springer International Publishing), 234–241. doi: 10.1007/978-3-319-24574-4_28

Roy, S., Butman, J. A., Pham, D. L., and Alzheimers Disease Neuroimaging, Initiative (2017). Robust skull stripping using multiple MR image contrasts insensitive to pathology. *Neuroimage* 146, 132–147. doi: 10.1016/j.neuroimage.2016.11.017

Segonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., et al. (2001). A hybrid approach to the Skull Stripping problem in MRI. *NeuroImage* 22, 1060–1075. doi: 10.1016/S1053-8119(01)91584-8

Sethian, J. A. (1999). *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. Cambridge University Press.

Shattuck, D. W., Prasad, G., Mirza, M., Narr, K. L., and Toga, A. W. (2009). Online resource for validation of brain segmentation methods. *Neuroimage* 45, 431–439. doi: 10.1016/j.neuroimage.2008.10.066

Shattuck, D. W., Sandor-Leahy, S. R., Schaper, K. A., Rottenberg, D. A., and Leahy, R. M. (2001). Magnetic resonance image tissue classification using a partial volume model. *Neuroimage* 13, 856–876. doi: 10.1006/nimg.2000.0730

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp*. 17, 143–155. doi: 10.1002/hbm.10062

Somasundaram, K., and Kalaiselvi, T. (2011). Automatic brain extraction methods for T1 magnetic resonance images using region labeling and morphological operations. *Comput. Biol. Med*. 41, 716–725. doi: 10.1016/j.compbiomed.2011.06.008

Speier, W., Iglesias, J. E., El-Kara, L., Tu, Z., and Arnold, C. (2011). Robust skull stripping of clinical glioblastoma multiforme data. *Med. Image Comput. Comput. Assist. Interv*. 14(Pt 3), 659–666. doi: 10.1007/978-3-642-23626-6_81

Thakur, S., Doshi, J., Min Ha, S., and Shukla, G. (2019). NIMG-40. Robust modality-agnostic skull-stripping in presence of diffuse glioma: a multi-institutional study. *Neuro-Oncology*. 21(Suppl. 6), vi170. doi: 10.1093/neuonc/noz175.710

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Warfield, S. K., Zou, K. H., and Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans. Med. Imaging* 23, 903–921. doi: 10.1109/TMI.2004.828354

Yogananda, C. G. B., Shah, B. R., Vejdani-Jahromi, M., Nalawade, S. S., Murugesan, G. K., Yu, F. F., et al. (2019). A novel fully automated mri-based deep learning method for classification of idh mutation status in brain gliomas. *Neuro Oncol.* 22, 402–411. doi: 10.1093/neuonc/noz199

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., et al. (2006). User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* 31, 1116–1128. doi: 10.1016/j.neuroimage.2006.01.015

Zhuang, A. H., Valentino, D. J., and Toga, A. W. (2006). Skull-stripping magnetic resonance brain images using a model-based level set. *Neuroimage* 32, 79–92. doi: 10.1016/j.neuroimage.2006.03.019

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership