Multiparametric MRI analysis of Glioblastoma Multiforme tissues using Support Vector Machines

By

J. Gabriel Heredia

A dissertation in partial fulfillment of the requirement for the degree of

Doctor of Philosophy

(Medical Physics)

At the

University of Wisconsin-Madison

2013

Date of final oral examination: 5/28/2013

The dissertation is approved by the following members of the Final Oral Committee:

M. Elizabeth Meyerand, Professor, Medical Physics

Soren M. Bentzen, Professor, Medical Physics

Patrick A. Turski, Professor, Radiology

Rasmus Birn, Assistant Professor, Psychiatry

Dana Tudorascu, Assistant Professor, Biostatistics

To my tutor, my friend …my wife, Athena
(in that order)

Acknowledgments

Firstly, I would like to thank God for blessing me with family, friends, health, and love. I sincerely thank Dr. M. Elizabeth Meyerand for being the most supportive and encouraging professor I have ever had. It has truly been a joy to work for you. I would never have accomplished this goal without your kind words of encouragement and assurance during the many difficult moments.

I would also like to thank Drs. Elizabeth Meyerand, Pat Turski, Soren Bentzen, Dana Tudorascu, and Rasmus Birn for agreeing to serve on my committee, and helping to guide me through this daunting process. I thank you for your time and effort. I know this can be very demanding, and I appreciate that each of you has taken the time to answer questions and serve as great sources of information. Your insights have guided me in directions I never anticipated. I am truly excited about this project, and I have you all to thank for that.

I owe a huge debt to Dr. John Hald and Pat Turski for their assistance in hand labeling the MRI data voxel-by-voxel. This would have been impossible without your expertise and willingness to share your time. Thank you.

Thank you to the entire Applied Neuro fMRI group. It has been a great pleasure getting to know you. I will truly miss working and playing with this great group of friends. I love that our group meetings could be productive yet crazy, and we always had a place to safely defend our ideas. We owe much of this to Dr. Cheng Guan Koay, whose relentless questioning, honest critiques, and caring nature make him a great asset to any group. Thank you to Dr. Rachel McKinsey for your loving support. You told me I could do it!

I would also like to thank the TomoTherapy group for giving me the great experience of working in the clinic. Drs. Ranjini Tolakanahalli, Dinesh Tewatia, and Mark Geurts for guiding me through this process, and helping mentor Athena and I with all things clinical.

I would like to thank my funding sources: Advanced Opportunity Fellowship, and NIH RO1 CA118365 for providing me with an education that I could not have otherwise obtained.

Thank you to Chad Moritz for helping me with clinical fMRI and making yourself available to answer questions. Thank you to Yin Huang for supplying the perfusion code. Thank you to Dr. Veena Nair for assisting me with AFNI codes.

Thank you to my undergraduate mentors, Drs. Nora Berrah, and Rene Bilodeau for providing me with the experience of hands-on experimental research. Without your influence, I would never have thought of a career in physics. You are both wonderful teachers, and your patience and support is much appreciated.

Finally, I want to thank my church for all the guidance and fellowship. You have made Madison my home. To my family, thank you. I am so blessed to have each of you in my life. Nothing I say could explain how weird and wonderful you all are. And to Athena, I can't believe I tricked you into marrying me! I love you always.

Abstract

Multiparametric MRI analysis of Glioblastoma Multiforme tissues using Support Vector Machines

Glioblastoma Multiforme (GBM) is the most prevalent and aggressive type of malignant brain tumor affecting humans. The median survival ranges from 4.5 months with no treatment, to 15 months with resection, radiation therapy, and chemotherapy. Recurrence occurs in nearly 95% of patients, and this is due in part to the highly infiltrative and heterogeneous nature of GBM. The ability to better characterize the heterogeneous makeup of these tumors could allow for more aggressive treatment. Our study consisted of collecting multiple MR images from patients with confirmed GBM, including: T2, post-contrast T1, perfusion based rCBV, diffusion (ADC), and a carbogen-based hypoxia map. These images were collected for each patient and coregistered to provide the data matrix. They were then analyzed by a radiologist and assigned tissue labels to complete the input matrix for the support vector machine classifier (SVM). This document will discuss the ability of a support vector machine classifier to effectively classify tumor, necrosis, edema, and non-enhancing (NCE) tumor tissues based on the multiparametric MRI data in patients with GBM. Specifically, we investigate and compare the methodology involved in the model selection and optimization process. Our goal is to provide a streamlined approach for dealing with large MRI datasets, and the implementation of the SVM framework on a voxel-by-voxel level.

Table of Contents

List of Figures

List of Tables

1. Introduction

Although glioblastomas are rare (incidence of 2-3 cases per 100,000 in Europe and North America), they still account for 52% of all functional tissue brain tumors (CBTRUS 2010). GBM is a Grade 4 tumor, characterized by the presence of necrotic tissue surrounded by anaplastic cells, and hyperplastic blood vessels. These characteristics differentiate GBM from Grade 3 astrocytomas (WHO 4th Edition). GBMs often form in the white matter, and can grow very large before symptoms are detected. They are usually quite aggressive, although about 10 percent are slower developing 'secondary GBMs', resulting from the degeneration of low-grade astrocytoma or anaplastic astrocytoma.

The current standard of care starts with surgical resection, which can be difficult due to the heterogeneous and infiltrative nature of the disease (Clarke 2010). Though difficult, complete resection of the enhancing tumor region was shown to increase the two year survival rate to 11.1 percent, vs. 2.6 percent for partial resection in a 2008 study (Pichlmeier).

After surgery, chemotherapy and radiation therapy are used. The combination of temozolomide with radiotherapy vs. radiotherapy alone has shown an increase in the two year survival rate from 10.4 percent to 26.5 percent with the combined treatment (Stupp 2005). The median overall survival also improved from 12 months (RT alone) to 14.6 months with combined treatment (Figure 1.1).

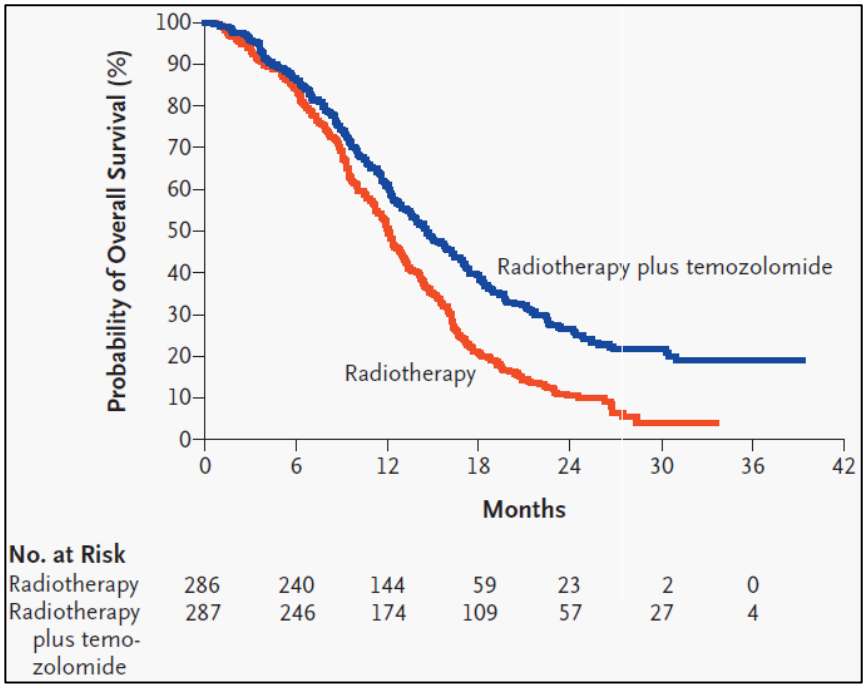


Figure 1.1: Kaplan-Meier estimates of overall survival by treatment group (Stupp 2005)

Although these results are significant, there is still much work to be done to improve the long-term survival outcomes. Other areas of ongoing research are in the search for molecularly targeted agents, immunotherapy, and antiangiogenic agents in the treatment of glioblastomas. Recent genomic analysis by Verhaak et al. (2010) has shown that there are four clinically relevant subtypes of glioblastomas: proneural, neural, classical, and mesenchymal (Figure 1.2).

The proneural subtype is associated with high rates of alteration in TP53 (the gene that encodes tumor protein 53 which functions as a tumor suppressor), and PDGFRA (the gene encoding a-type platelet-derived growth factor receptor). The neural subtype showed less TP53 and EGFR mutations than the other subtypes. The classical subtype often has higher than normal expression of epidermal growth factor receptor (EGFR), but rarely has a mutated TP53 gene.

The mesenchymal subtype is associated with alterations in NFI (the gene encoding neurofibromatosis type 1).

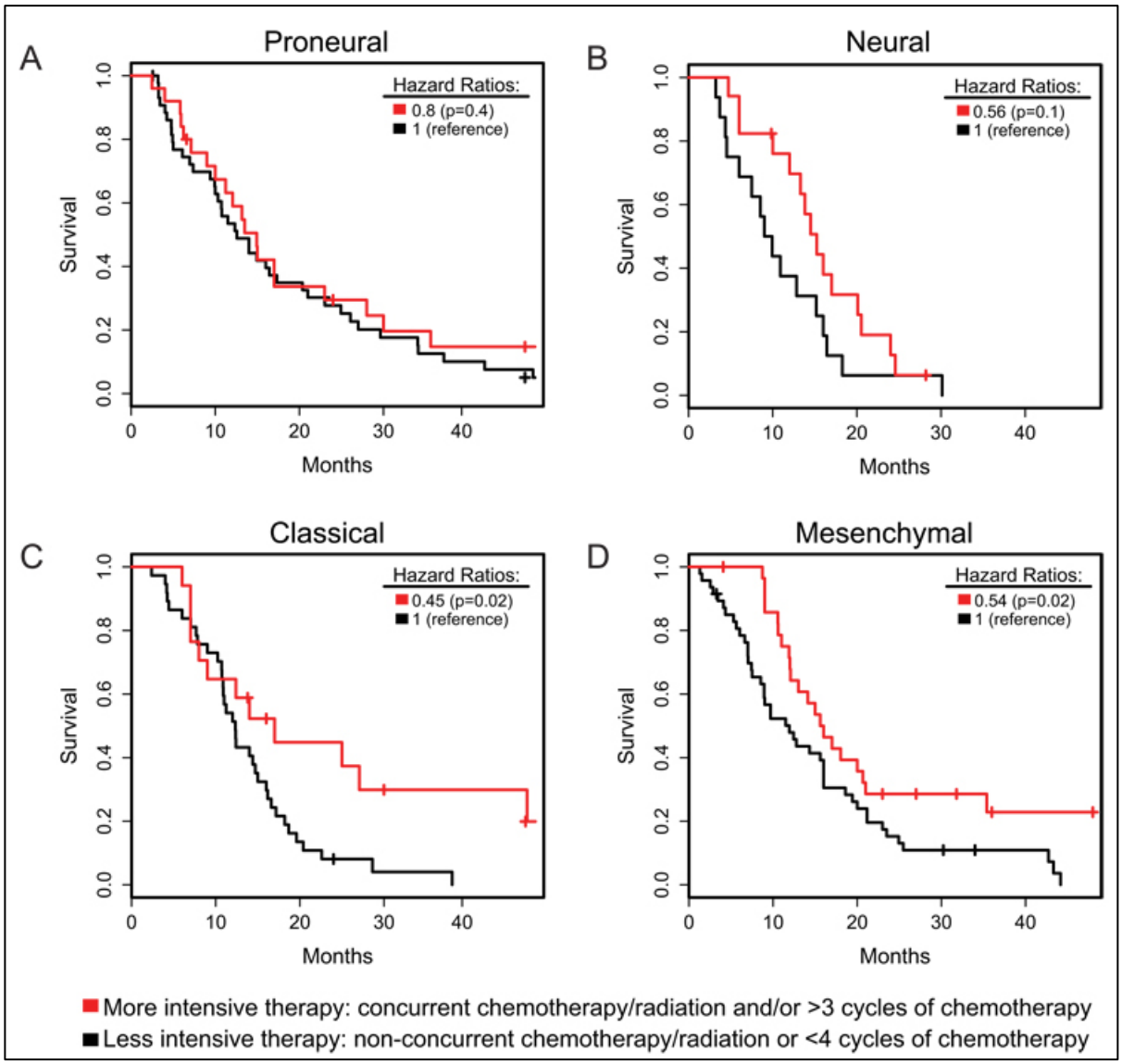


Figure 1.2: Survival by treatment type and tumor subtype (Verhaak 2010).

We can see from these studies, GBM is a very complex disease that requires more than one approach to treatment. This broad spectrum of approaches is encouraging the field to collaborate across disciplines. As we can see from the genetic subsets of GBM, and the differing responses to therapy, the field is moving towards a more individualized treatment approach.

Where some patients may improve, others may not respond as well. As these treatment options become more and more tailored to individual patients, it becomes increasingly important to develop our imaging capabilities in the hope of improving patient care.

My research focuses on the array of MR imaging techniques that help to provide anatomical and physiological information regarding the disease. Using a multiparametric dataset of post-contrast T1, T2, diffusion, perfusion, and hypoxia mapping allows us to gain insight into many different aspects of the disease process. However, it can be quite overwhelming, when dealing with such large datasets, to identify useful trends and relationships. This is a problem well suited for machine-learning algorithms.

However, this research will specifically focus on multiparametric MRI data for each patient. I will explain the implementation of SVM and how it is a useful tool for classifying different tissue types in GBM patients. I will also discuss common pitfalls and limitations of the proposed methods. Specifically, I will investigate the appropriate metrics for reporting and comparing SVM classifier results. Although SVMs have seen much attention in recent years, there is still disagreement as to the appropriate methodology.

Another aspect of interest is the model selection and construction methods. SVMs inherently require less user intervention than other machine learning techniques, such as Bayesian Networks. However, there still remain many preprocessing and model selection steps that affect the performance of the classifier. This work will focus on applying multiple model preparation methods to individual case studies, in an effort to compare how normalization choices, data balance, and inter vs. intra patient models affect the outcomes of a classifier using identical testing data.

We will also address the clinical feasibility of using SVMs for patient coding and analysis. Besides the question of performance, there is also a question of practicality and cost justification. Our goal is to assist radiologists as they identify and track the progress of various tissues. However, if our method requires too much time from the radiologist for labeling the input data, then it becomes less likely to be implemented clinically. We aim to address this by comparing simple models, which require very little radiologist labeling time, to complex models that require a great deal of effort from the radiologists.

2. Background

2.1 T1 and T2 mapping

GBM tumors can often be seen on MRI scans as contrast-enhancing lesions that are quite heterogeneous and extending into surrounding normal tissue. This occurs because of the breakdown of the blood-brain barrier associated with GBM. By using paramagnetic contrast agents (such as gadolinium), we can see T1-shortening in the tumor tissue, which results in an increase in signal on T1-weighted images. Figure 2.1 shows an example of a GBM tumor pre- and post-contrast.

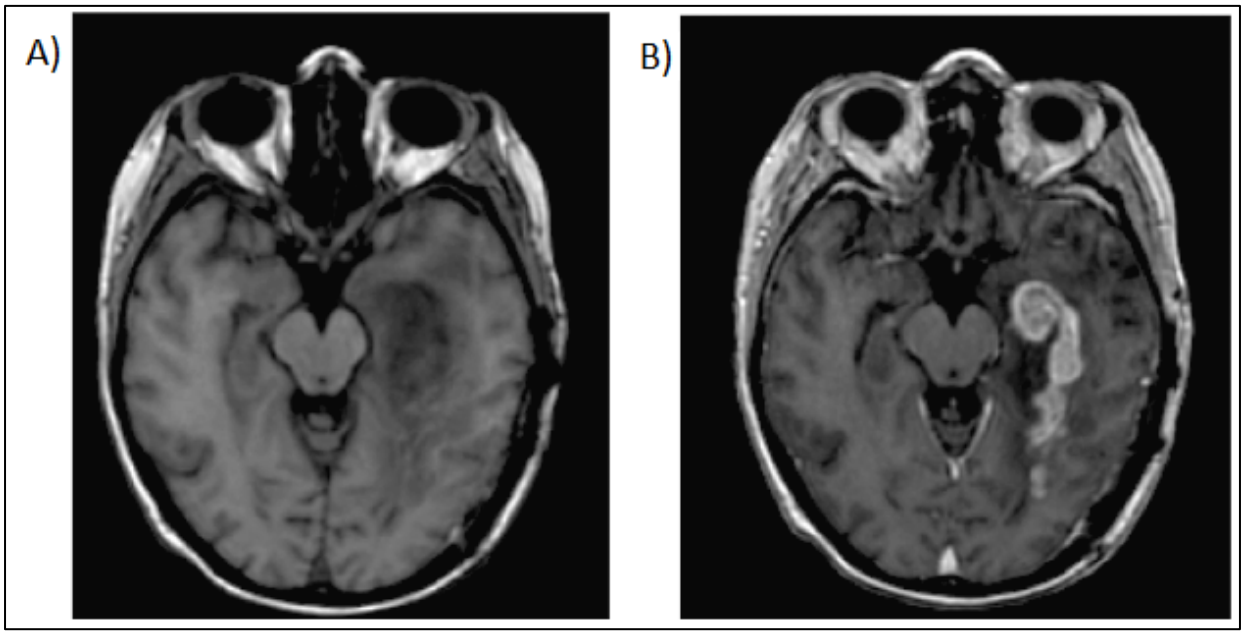


Figure 2.1: T1-weighted images of patient with GBM. A) Pre-contrast. B) Post-contrast. (Pirzkall 2009).

Although post-contrast T1 images provide excellent information about the tumor location, they are only sensitive to areas where there is a break down in the blood-brain barrier. For this reason, it is believed that post-contrast T1 scans tend to underestimate the tumor volume. The

blood-brain barrier can also be affected by therapy and can be a challenging factor when imaging post-treatment changes. In our study, we attempt to avoid these complications by only using scans starting at least 3 months post-treatment.

On the other hand, T2-weighted images tend to overestimate the tumor extent. This is because the high signal intensity from the surrounding edema as well as microscopic tumor extension can mask the tumor boundaries. A confounding factor with this modality is that radiation effects can also mimic microscopic tumor extension (Knopp 1999).

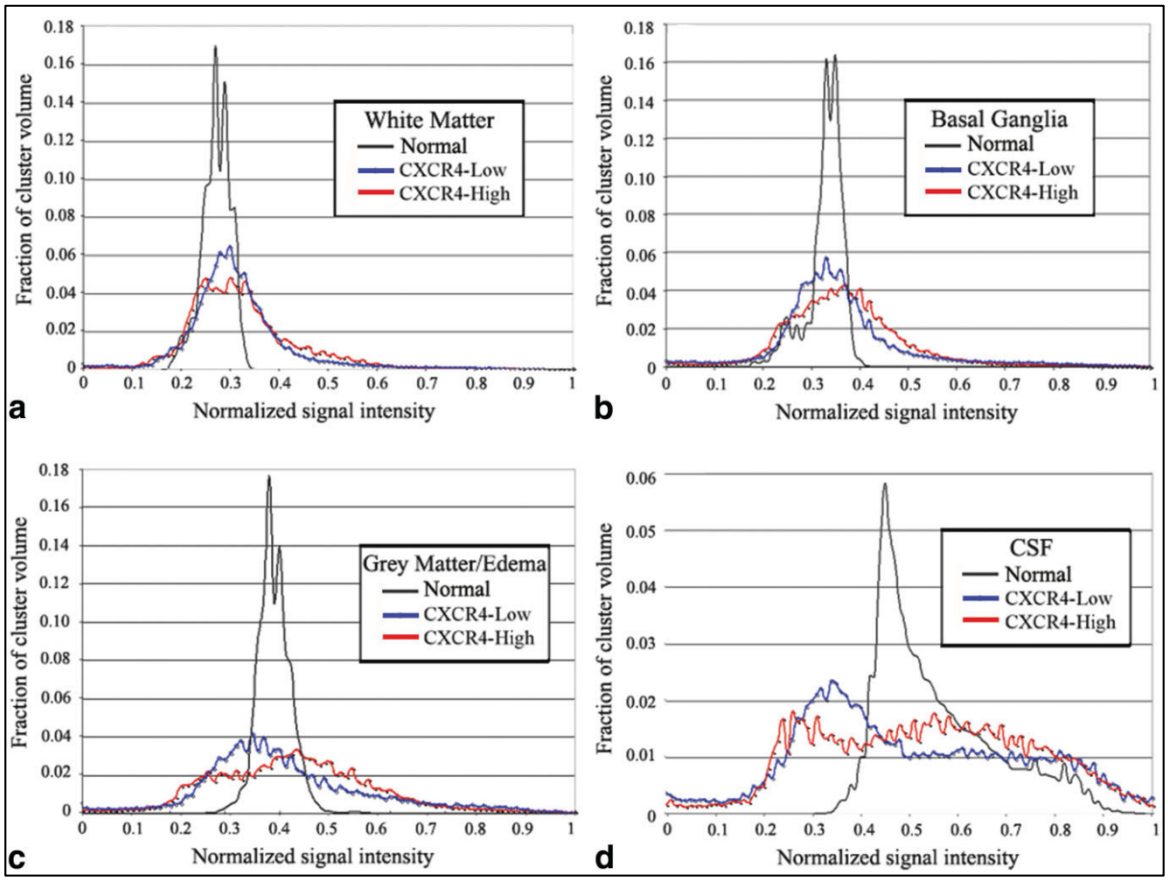


Figure 2.2: Average T2 histograms for each cluster volume comparing normal group to two patient groups. GBM patients were split into two groups based on their levels of chemokine receptor 4 (CXCR4 low and high). In every cluster the GBM groups display significantly higher standard deviations compared to normal subjects.

A 2009 study by McMillan et al. used T2 signal intensity histograms to investigate the differences in distributions between normal subjects and patients with GBM. They segmented their data to achieve four clusters of tissue types: white matter, basal ganglia, grey matter/edema, CSF. The results of these comparisons can be seen in figure 2.2. In each of the 4 clusters, it is clear that the GBM patients have a much broader range of signal intensities than the normal subjects. Although no comparison can be made in tumor tissue because normal subjects lack this by definition, it is particularly interesting that there are differences in some of the presumably healthy tissue. Although these modalities offer valuable information, more scans will help to address some of the confounding factors. The following modalities allow us to investigate some of the physiological aspects of the brain.

2.2 Diffusion imaging

Diffusion imaging has also shown positive results in assessing tumor regions (Castillo 2001). Since extracellular space decreases when cells are multiplying rapidly, this restricts the motion of water molecules outside of the cells as well as the more viscous intracellular space. The resulting low apparent diffusion coefficient (ADC) values are thought to indicate areas of high cellularity (Sugahara 1999).Studies by Kono et al. (2001), Bulakbasi et al. (2004), and Sugahara et al. (1999), all found that high-grade tumors tended to exhibit diffusion restriction with low ADC values.

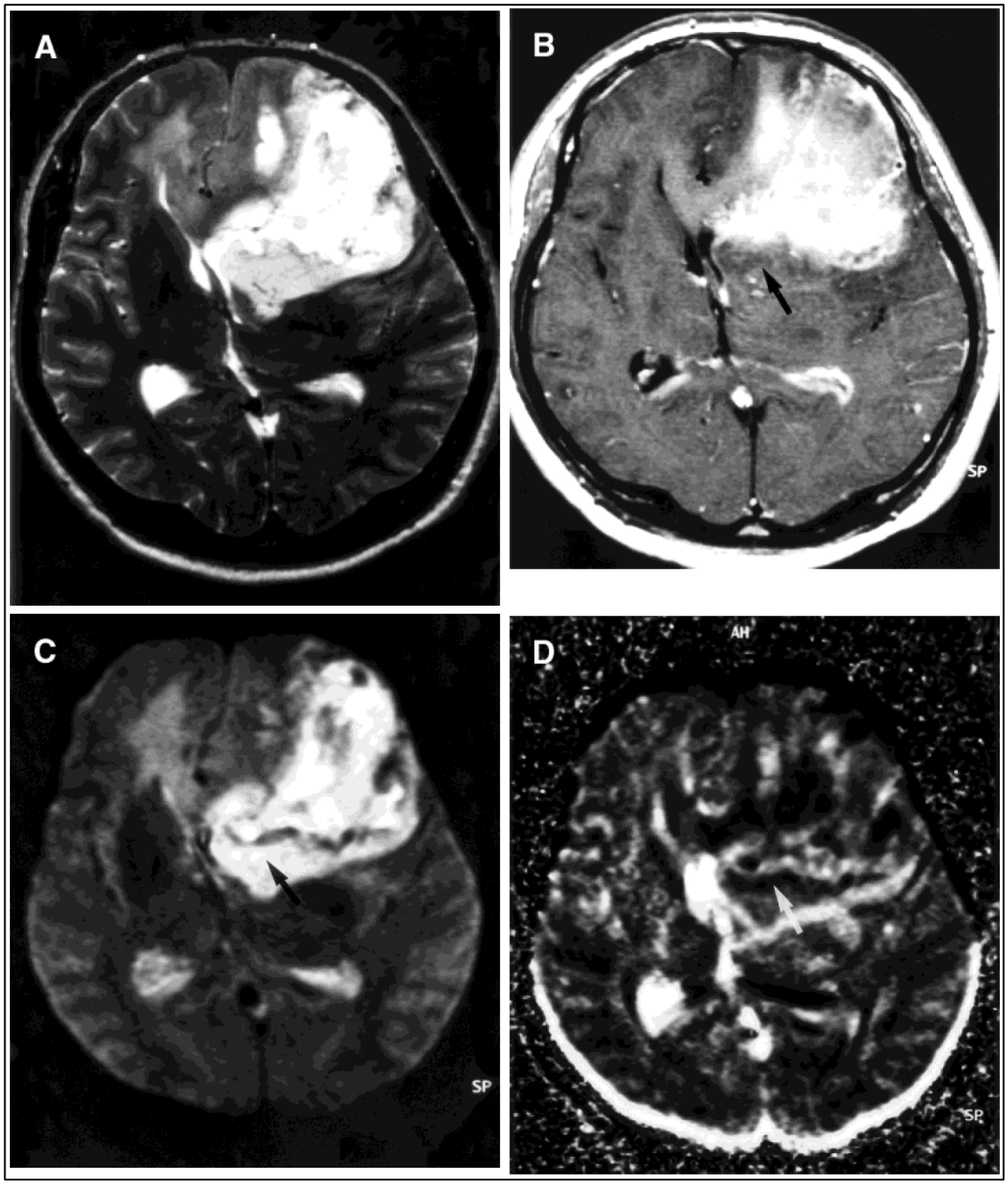


Figure 2.3: Man with GBM verified by surgical resection, and a histologic specimen showing marked hypercellularity of 24%. (A) T2-weighted image: shows an inhomogeneously hyperintense tumor in the left frontal lobe. (B) Contrast enhanced T1: the tumor is markedly enhanced, but some parts of the hyperintense T2 image are not enhanced (arrow). (C) Diffusion weighted image: the unenhanced areas are depicted as hyperintense (arrow), and the enhanced areas are depicted as inhomogeneously hypointense. (D) ADC map: hypointense areas are observed within the tumor, suggestive of the restriction of diffusion. The central part of the tumor is mostly hypointense (arrow), suggesting that the diffusion within the tumor is most restricted (Sugahara 1999).

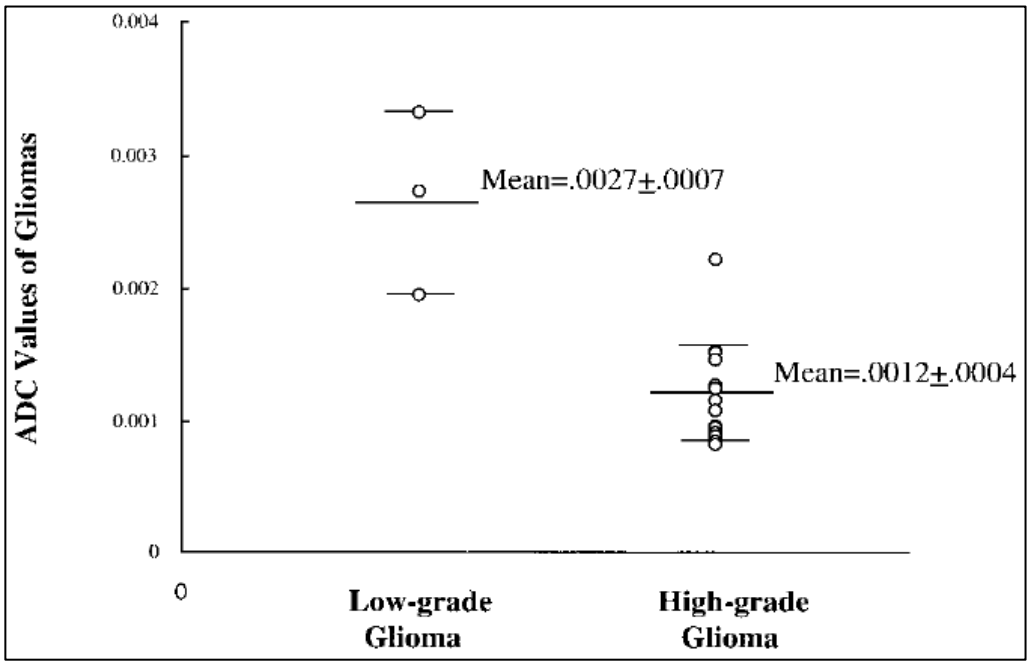


Figure 2.4: ADC values are lower for high-grade gliomas vs. low-grade gliomas. (Sugahara1999)

Although low ADC values are strongly associated with increased cellularity in the literature, there are other aspects of tumor development that can also impact the ADC values; such as cell density changes resulting from necrosis or apoptosis. Also, areas of vasogenic edema, where diffusion is less restricted, will increase ADC values. Edema can be complex when dealing with ADC values. The problem is that it can permeate tissue containing neoplastic cells, resulting in a mixture of edema and tumor cells. Since the added water content will increases the distance between cell membranes, the ADC values will increase, and potentially mask the hypercellular effects of the tumor tissue (Sugahara 1999). This can also occur with cystic changes or areas of necrosis. This makes it difficult to use ADC to demarcate between contiguous areas of tumor and edema, from areas of just edema (Muti 2002).

Castillo et al. (2001) found significant differences between ADC values in tumor tissue vs. normal tissue. They also found this for edema vs. normal tissue.

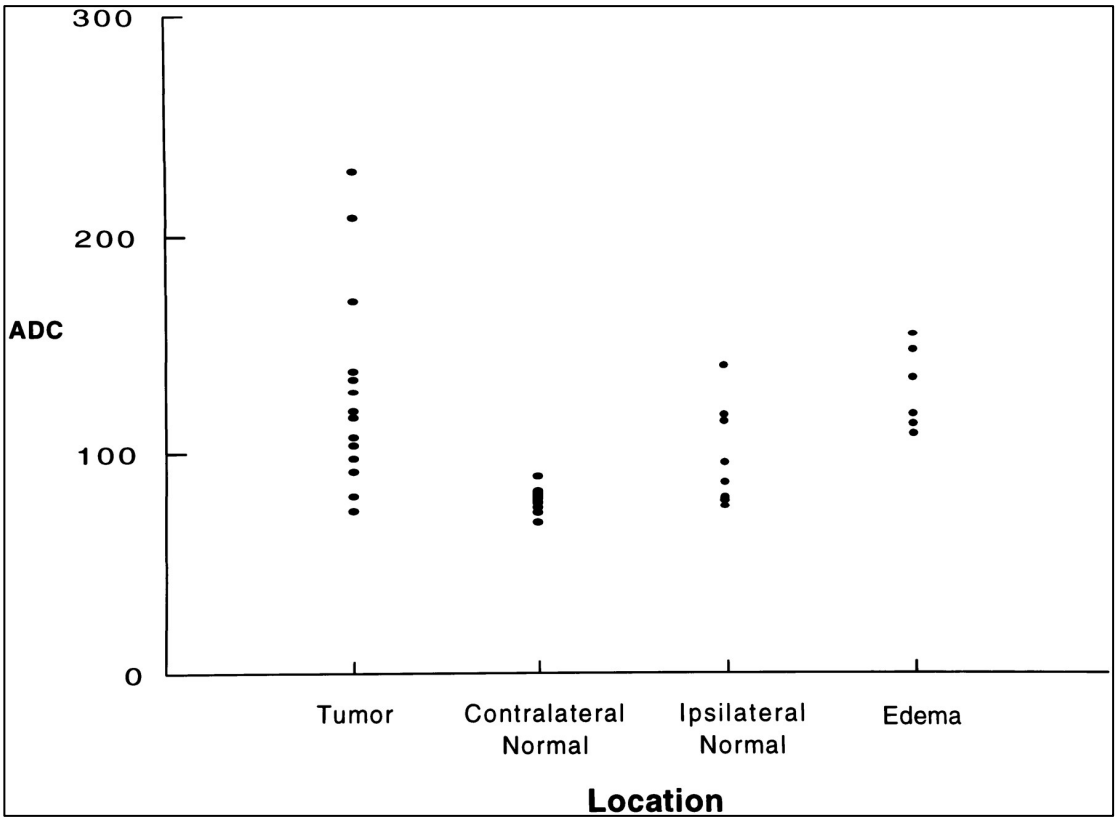


Figure 2.5: Plot of individual data points for tissue types. Note considerable overlap of single data points (Castillo 2001).

However, it must be noted that these differences were significant on the group analysis level, but we can see that the individual data points do not discriminate as clearly. While this is useful in group analysis, our study focuses primarily on tissue classification in the single subject setting. We are looking for differences that are robust enough to provide information on a patient-by-patient approach. Tien et al. (1994) found promising results in distinguishing between areas of enhancing tumor, non-enhancing tumor, cyst/necrosis, and edema using the ADC values of 10 patients with confirmed GBM (table 2.1). However, other authors have reported difficulties in achieving similar results (Kono 2001; Stadnik 2001). This could be due to the heterogeneous

makeup of GBM tumors. While some cases may show strong differences, others may not exhibit much variation across tissues.

Tissue	Diffusion Coefficient ($\times 10^{-3}$ mm^2/sec)		No. of Measurements
	Range	Mean ± SD	
CSF	2.1–4.6	3.2 ± 0.7	20
Gray matter	0.7–1.7	1.2 ± 0.2	34
White matter parallel to direction of diffusion gradient	1.2–1.8	1.5 ± 0.5	18
White matter orthogonal to direction of diffusion gradient	0.1–0.8	0.5 ± 0.2	56
Enhancing tumor	0.9–1.4	1.1 ± 0.2	20
Cystic/necrotic tumor	1.7–3.8	2.2 ± 0.9	14
Nonenhancing tumor in parallel white matter	1.1–1.8	1.6 ± 0.2	8
Edema in parallel white matter	1.3–2.1	1.8 ± 0.3	24

Table 2.1: ADC values in various brain tissues of 10 patients with confirmed GBM (Tien1994).

Since GBM often has a heterogeneous makeup, it makes the case for multiparametric analysis even stronger. ADC maps offer valuable information, and will become even more discriminatory when combined with multiple techniques. It is also important to note that these findings are all based on region of interest (ROI) based studies. While ROIs provide good information about mean values and variation across the region, we are interested in looking at the voxel level. Assuming our results are robust enough when combined with other types of scans, we will utilize the voxel level analysis to help delineate the boundaries between tissues.

2.3 Perfusion imaging

Physiological mapping of the tumor regions has been performed with perfusion weighted imaging. In dynamic contrast enhanced perfusion (DCE), a bolus injection of a paramagnetic contrast agent such as gadolinium diethylenetriaminepenta-acetic acid (Gd-DTPA) is used with perfusion weighted MRI acquisitions to obtain information about cerebral hemodynamics. The magnetic properties of tissue voxels change as the contrast passes through the blood vessels. This produces changes in T2* that allow for a signal representing cerebral blood flow (CBF) to be plotted vs. time. Using the central volume theorem (equation 1), we are also able to obtain measures of cerebral blood volume (CBV), and the mean transit time (MTT) (Rosen 1990; Ostergaard 1996).

1) $$MTT = \frac{CBV}{CBF_t}$$

Perfusion measurements have been used extensively to investigate brain tumors because of the valuable physiological information they represent. rCBV has been shown useful in determining areas of increased vascularity, or angiogenesis (Knopp 1999; Provenzale 2002). It is believed that the rapidly growing tumor needs more blood supply to survive, and induces neo-vascularization through protein signaling pathways such as vascular endothelial growth factors (VEGF). This leads to areas of increased CBV (see figure 2.6). These new vascular networks act to deliver nutrients and oxygen necessary for growth (Gillies 2000). Neovasculature is marked by lack of smooth muscle, large endothelial gaps (resulting in increased permeability to macromolecules), and chaotic, ineffective blood flow (Provenzale 2002). It is also can have both acutely and transiently collapsing vessels. These immature, fragile, and leaky vessels often

struggle to meet the needs of the developing tumor, and hypoxic or necrotic regions develop despite the elevated angiogenic activity (Gillies 2000).

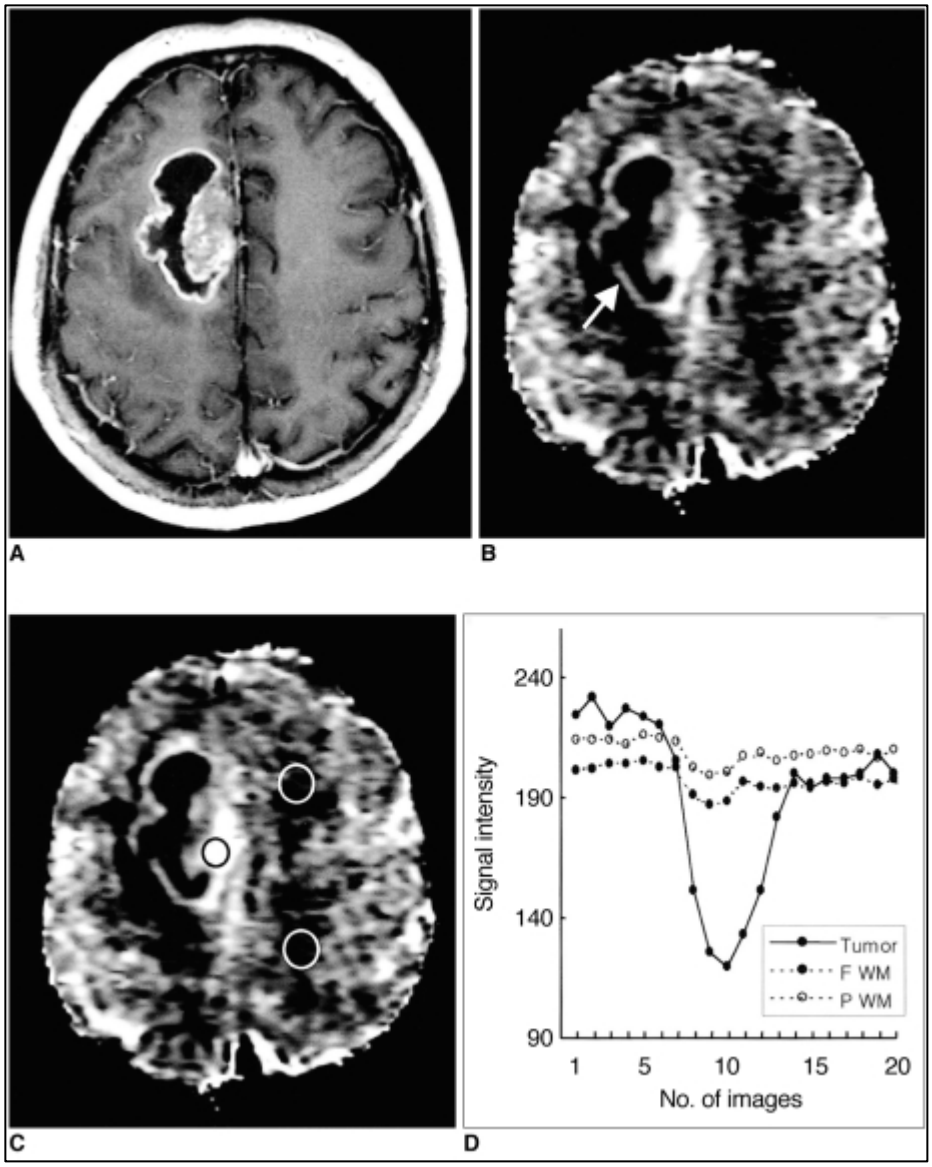


Figure 2.6: From Lee et al. (2001). GBM in 47-year-old man. A) Postcontrast T1-weighted image shows a ring-enhancing necrotic tumor in right frontal lobe. B) rCBV map shows high rCBV in solid portion of the tumor (arrow). C) rCBV map shows the placement of ROIs for measurement of rCBV time curves (black circle for tumor, white circles for normal white matter). D) Signal intensity time curves show different patterns of signal reduction between normal white matter and tumor.

Relative cerebral blood volume maps (rCBV) have also been the most widely used DSC parameter for investigating brain tumor grade (Knopp 1999; Shin 2002; Lee 2001).

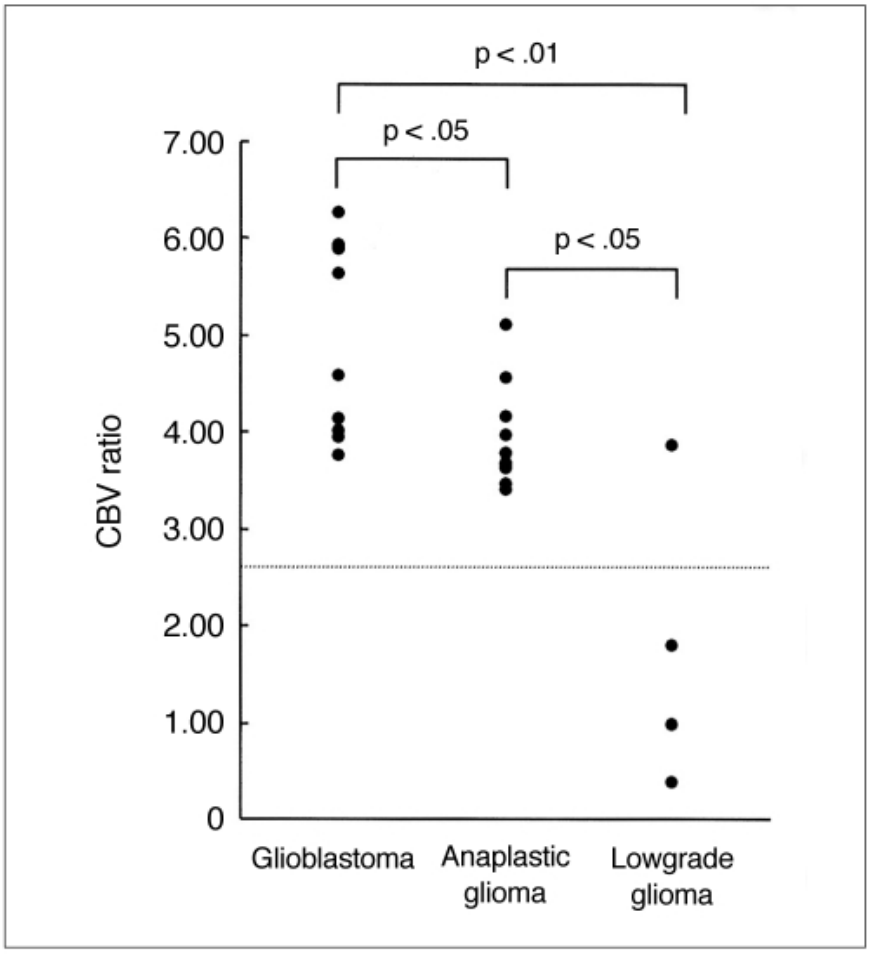


Figure 2.7: Plot of rCBV ratios in glioblastomas, anaplastic gliomas, and low-grade gliomas. The rCBV ratio is highest in GBMs and lowest in low-grade gliomas. A comparison of mean rCBV ratios in each tumor group shows statistically significant differences (Lee 2001).

It is widely accepted that microvascular blood volume is elevated in tumors, leading to an association between microvascular density and tumor energy metabolism (Aronen 2000). There

is also a clear correlation between malignancy and tumor neovascularity (irregular, enlarged, and distorted microvessels) (Boxerman 2006; Burger 1985; Papadimitriou 1975).

Despite these relationships, the literature has not readily agreed in terms of the usefulness of rCBV for tumor grading. The results from Lee et al. show significant differences between tumor grades (figure 2.7). These findings are consistent with other studies that also show rCBV to be a useful measure of tumor grade (Donahue 2000; Shin 2002; Jackson 2002; Aronen 2000). However, other studies have found no correlation or statistical significance when separating tumor grade by the use of rCBV maps (Hacklander 1995; Lam 2001).

Boxerman et al. (2006) proposed that these inconsistencies could be explained by the fact that tumor patients often have a breakdown in the blood-brain barrier, leading to contrast agent extravasation. This leakage effect contaminates the derived concentration-time data and can potentially mask useful differences in rCBV data. In patients with an intact blood brain barrier, the contrast agent is contained in the intravascular space. The resulting difference in susceptibility between the intravascular and extravascular space causes dephasing of the MR signal, and a loss on T2-weighted scans. However, in patients with blood-brain barrier breakdown, the contrast agent (Gd-DTPA) can leak into the interstitial space of the lesions. Since Gd-DTPA is also an effective T1 relaxation enhancer, it can produce T1 effect signal increases in areas where leakage is prominent. These T1 effects can then mask the susceptibility contrast loss we are attempting to measure in DSC perfusion imaging.

The method that Boxerman et al. (2006) developed for correcting the rCBV maps for leakage uses linear fitting to estimate the T1 contamination due to contrast agent extravasation. This technique is described in detail in the methods section (chapter 3.3), as it was also employed

in correcting our rCBV maps. The results from their leakage-corrected rCBV ($rCBV_{corrected}$) study showed that leakage correction has an effect on tumor grade prediction.

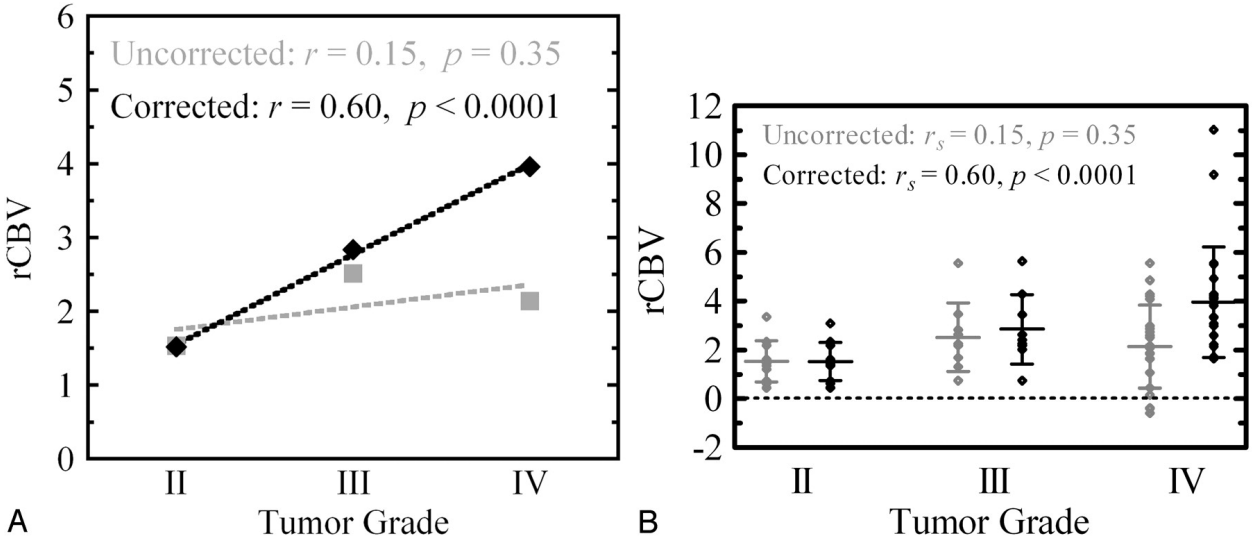


Figure 2.8: A) $rCBV_{corrected}$ estimates corrected for leakage correlate significantly with glioma tumor grade, whereas uncorrected rCBV does not. B) Despite significant correlation, there still exists moderate variability in $rCBV_{corrected}$ within each grade (Boxerman 2006).

rCBV has also been used to differentiate tumor recurrence from radiation necrosis (Sugahara 2000). They showed statistically significant results when comparing the two groups (figure 2.9). This is believed to occur because the vascularity of malignant tumor differs dramatically from that of irradiated brain tissue, specifically radiation necrosis.

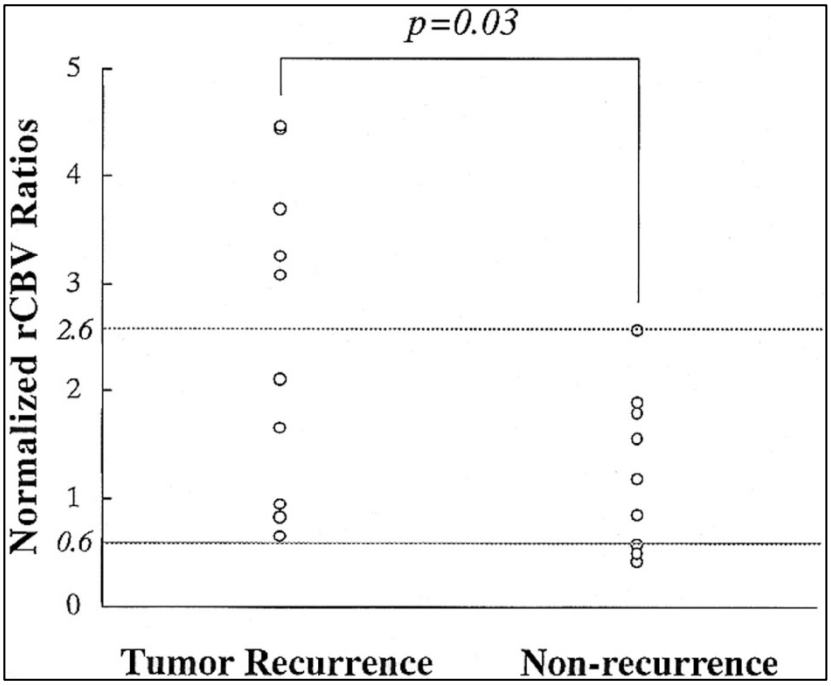


Figure 2.9: Relationship between normalized rCBV ratios in patients with tumor recurrence and those with radiation necrosis (Sugahara 2000).

In a study performed by Henry et al. (2000), they showed that rCBV values in tumor regions tended to be larger than normal appearing white matter, while surrounding areas of edema or gliosis tended to have less than normal values (figure 2.10).

Other studies have looked at using rCBV to perform tumor segmentations. Emblem et al. (2009) showed that tumor segmentation using knowledge-based fuzzy clustering conveyed similar values for diagnostic accuracy as manually selected volumes. Using a voxel by voxel method for analyzing the rCBV map, they found a sensitivity and positive predictive value of 69% and 73% respectively. This is compared to the manual segmentation values of 57% and 87% (sensitivity, PPV).

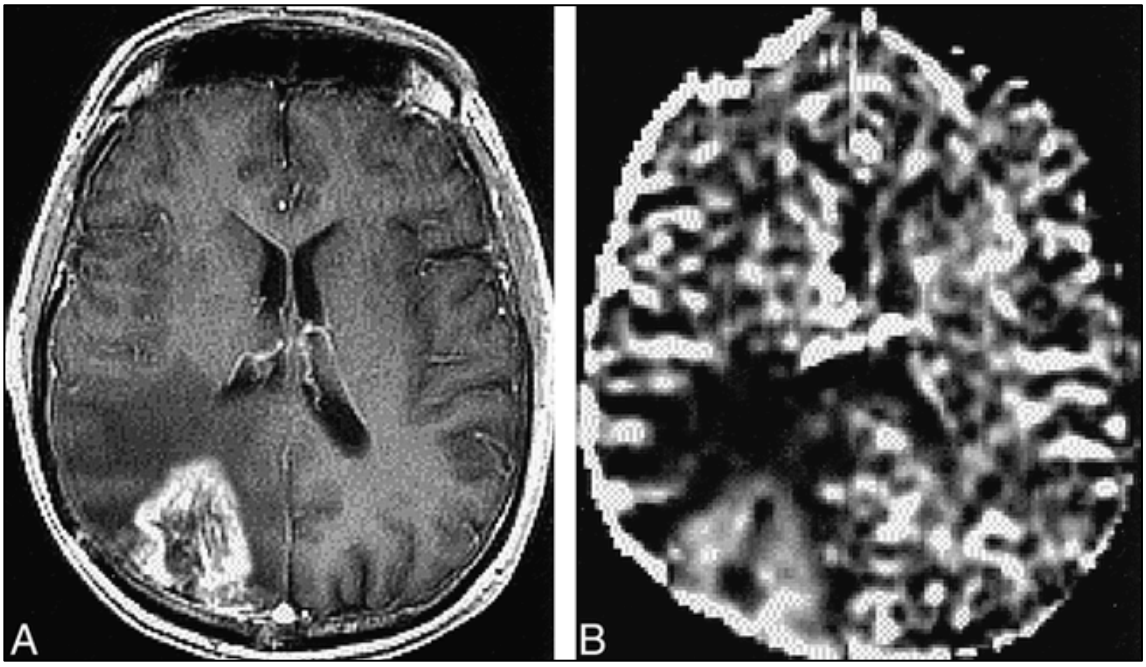


Figure 2.10: A) Post-contrast T1 image of patient with confirmed GBM. B) Elevated rCBV levels in enhancing region, and lower rCBV in surrounding edema/gliosis (Henry 2000).

DSC perfusion imaging has been shown to be useful when analyzing GBM tumor data. The literature has reported uses in tumor segmentation, identification of radiation necrosis, and mixed results with tumor grading. Although, improvements are made with leakage corrected methods, the heterogeneity of GBMs warrants the use of multiple scans and methods to achieve a more complete picture of the underlying physiology.

2.4 Carbogen-induced hypoxia mapping

As mentioned earlier, GBM tumors outgrow their blood supply, causing areas of hypoxia and necrosis. Hypoxic areas are of great importance to tumor physiology, because as noted in earlier sections, they help to promote new vasculature through signaling pathways such as VEGF. Another very important effect of hypoxic regions is that they are less sensitive to radiotherapy (Overgaard 1989; Teicher 1995; Harrison 2002).

Oxygen is a potent radiosensitizer, as it increases the effectiveness of radiation by forming DNA-damaging free radicals. Therefore, oxygen tension (pO_2) plays a large role in determining the success of radiation therapy. Radiation resistance begins at oxygen values less than 10 mm Hg, and is fully resistant below 0.5 mm Hg. The normal range for oxygen tension in healthy brain tissue is between 40 and 60 mm Hg, while the median value for tumor tissue is less than 10 mm Hg (Brown 1999).

The gold standard for measuring pO_2 is the use of an oxygen electrode (Horsman 1993). However, this method is also invasive, and can be subject to large intrinsic errors due to the limited access to tissues of interest (Al-Hallaq 1998). The heterogeneity of the tumor also limits these measurements because the electrode only detects a limited region without requiring a new placement.

However, blood oxygen level dependent (BOLD) imaging is non-invasive and has shown strong correlations to the electrode method (Al-Hallaq 1998; Dunn 2002). BOLD imaging is based on the T2* sensitivity to the paramagnetic effects of deoxygenated hemoglobin (Ogawa 1993). Since deoxyhemoglobin is paramagnetic it increases the differences in susceptibility between intravascular and extravascular tissue, resulting in dephasing and loss of T2* signal.

BOLD imaging is also an indirect measure of pO2, blood flow, and metabolic rate. This is because the amount of deoxyhemoglobin in the blood is a function of the oxygen supply and oxygen use. Although PET techniques may correlate slightly better with known pO_2 values, MRI is more widely available, has better resolution, and has been shown to be a reliable measure.

There has been a great deal of research investigating the effects of hyperoxygenation in improving radiation therapy. Carbogen breathing techniques (along with nicotinamide) have been shown to be useful in patients with head and neck cancer (Kaanders 2002), as well as in a recent randomized trial of patients with bladder carcinoma (Hoskin 2010). However, a 2003 study by Simon et al. found no significant improvement in a comparison of GBM patients undergoing radiotherapy and chemotherapy 'with' vs. 'without' carbogen and nicotinamide.

The carbogen gas mixture consists of a large amount of oxygen with a small amount of CO_2 mixed in to act as a vasodilator. Our study uses a 95% O_2 and 5% CO_2 mixture. Hyperoxygenation causes vasoconstriction in normal tissue. However, immature tumor vessels fail to exhibit vasoconstriction due to a lack of smooth muscle, which leads to a higher flow of oxygenated blood (Gilles 2000). This is due to the decreasing resistance of the tumor vessels relative to normal tissue.

A study by Kuperman et al. (1995) suggests that the largest changes in T2* during carbogen breathing are detected in hypoxic tumor regions which have low vascular density, but are not necrotic. In normal tissue, the rate of oxygen delivery is much higher than oxygen consumption, leading to small amounts of deoxyhemoglobin. However, in hypoxic tumor regions, the rate of oxygen consumption is limited by the reduced rate of oxygen delivery. This

leads to a large proportion of deoxyhemoglobin, as most available oxygen is used. Therefore, they proposed that by increasing the oxygen to the tissue through carbogen breathing, the relative amount of deoxyhemoglobin would decrease. In effect, if the rate of oxygen delivery is increased beyond the level of normal consumption, then less deoxyhemoglobin will be produced relative to the total blood flow. These decreases in deoxyhemoglobin due to carbogen breathing will result in a longer T2* and increased signal intensity. There are two reasons why this effect should favor hypoxic tumor tissue and not normal tissue. Firstly, as mentioned above, normal vasculature will undergo vasoconstriction, attenuating the increase in available oxygen. Secondly, since the ratio of oxygen delivery to oxygen consumption is so much higher in normal tissue, any increases in the rate of oxygen delivery will have a much greater relative effect on the lower ratio associated with oxygen delivery and consumption in the hypoxic tumor regions.

Kuperman et al. (1995) studied these changes by using mammary adenocarcinomas implanted in the hind limbs of rats. Their results showed that significant increases in image intensity occurred in tumor centers and rims, as compared to much smaller responses in the surrounding muscle. The largest responses were found in the rims of the tumors, where the relatively low density of blood vessels lead to chronically hypoxic tissue, limited by the range of oxygen diffusivity throughout the tissue (figure 2.11, 2.12).

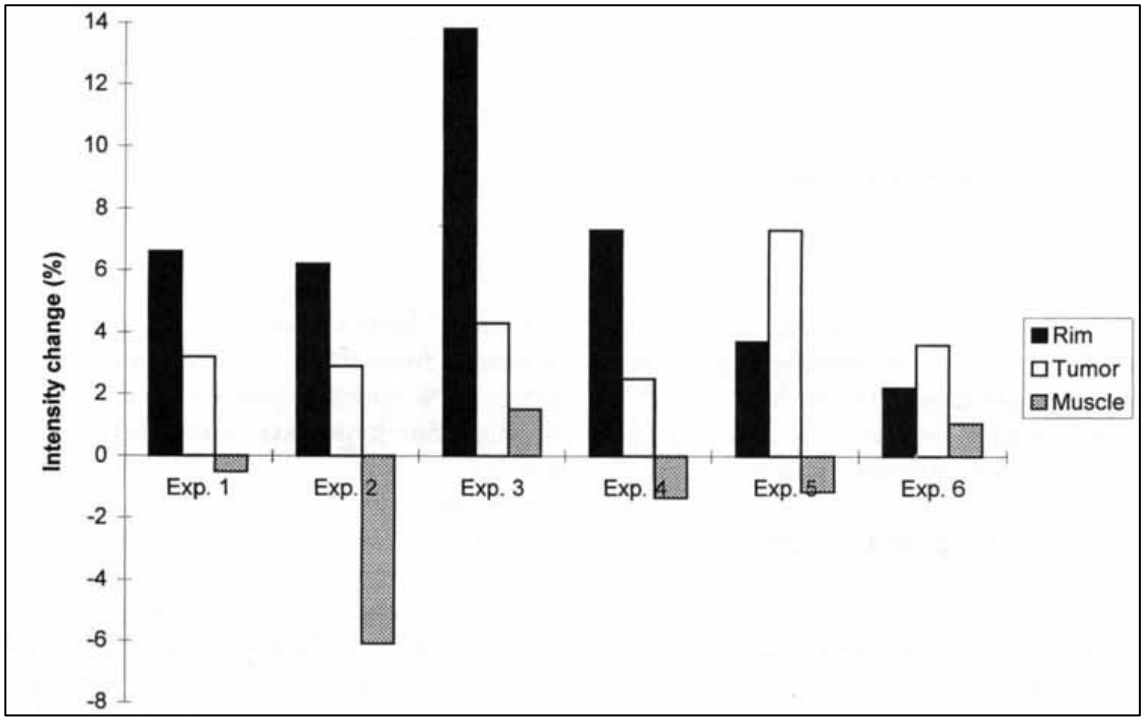


Figure 2.11: Maximum change in intensity during carbogen breathing (using rat model) for three regions of interest (Kuperman 1995).

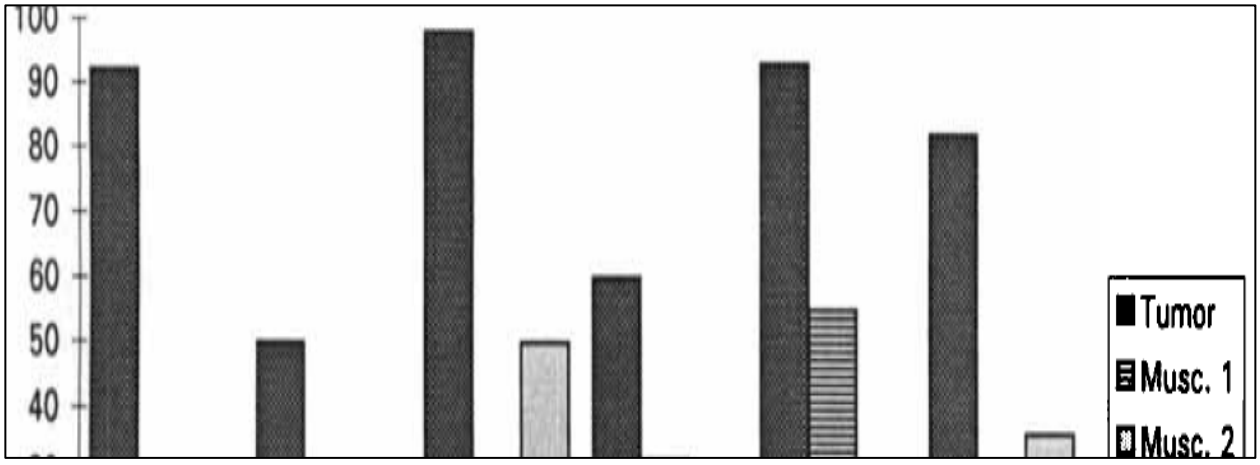


Figure 2.12: The percentage of pixels in which an increase in intensity greater than 1% was detected during carbogen breathing in rats. Data collected for tumor tissue as well as two muscle tissues (Kuperman 1995).

A later study by Dunn et al. (2002) validated these results, and found that carbogen breathing reduces the average R2* in intracranial rat tumors. This study also benefited from the use of electron paramagnetic resonance (EPR) oximetry and BOLD imaging. Coupling these two techniques allowed for further confirmation that BOLD imaging is capable of measuring the changes associated with carbogen breathing. An example of the R2* effects can be seen in figure 2.13.

Figure 2.13: R_2* images of CNS tumor in rat. A) air-breathing B) Carbogen-breathing C) signal intensity difference; bright=increase in SI, indicating a large reduction in R_2* relaxation rates (Dunn 2002).

These findings support the use of carbogen breathing in our study of GBM. The aggressive fast growing nature of GBM makes it a good candidate for investigation with carbogen breathing techniques. We expect that these tumors will often outgrow their blood supply, and as the immature vasculature struggles to keep up, there will be regions of the tumor that are chronically hypoxic. Although some tumor tissue will be well-perfused and not experience as many changes, we expect that hypoxic regions of the tumor will show a signal change consistent with previous findings.

2.5 Multiparametric Approaches

Although each of the imaging techniques mentioned earlier provides useful information, the complexity and heterogeneous nature of GBM makes it difficult to characterize the tumor using single modalities. Many of the promising results have shown that including multiple imaging techniques allows for a more robust model. By using multiparametric approaches we are hoping to identify trends that withstand the particular limitations and variations associated with each single technique in isolation.

Henry et al. (2000) found a strong correlation between MR spectroscopy and rCBV in identifying recurrent tumor in glioma patients. Di Costanzo et al. (2006) used spectroscopy, diffusion and perfusion imaging in the delineation of GBM. They saw trends that showed elevated rCBV in tumor regions when compared to edema and normal tissue, as well as differences in metabolite ratios. However, the use of an ROI method with low number of patients limited their results. This is a potential limiting factor of ROI methods. On one hand, ROI methods may be more generalizable, in that they see more data, and look for broader trends. However, they require more patients and cannot be used for segmentation with the flexibility that voxel-wise methods can.

A study by McMillan et al. in 2007 showed the use of CSI, diffusion, perfusion, and hypoxia imaging for analyzing GBM tumors. They presented a method based on overlapping regions of thresholded parameter maps. They refer to this as the percent overlap method (POM). They started by thresholding each of the distinct parameter based on a cutoff value associated with the contralateral healthy hemisphere. They then investigated the voxels that survived this thresholding, and looked for overlapping voxels between the parameters. The assumption is that

recurrent tumor voxels will have extreme values for multiple parameters. The results of their method can be seen in figure 2.14. They compare their method (POM) to the previously established ISODATA method, which is based on an iteratively self-organizing variation of the K-means algorithm (McMillan 2007). Their results show similar spatial location, but some mismatch between tumor extent. The advantage of POM over ISODATA is that it is insensitive to the addition of parameters that do not hold useful information. However, the addition of low information non-discriminatory parameters to ISODATA can negatively affect the algorithm.

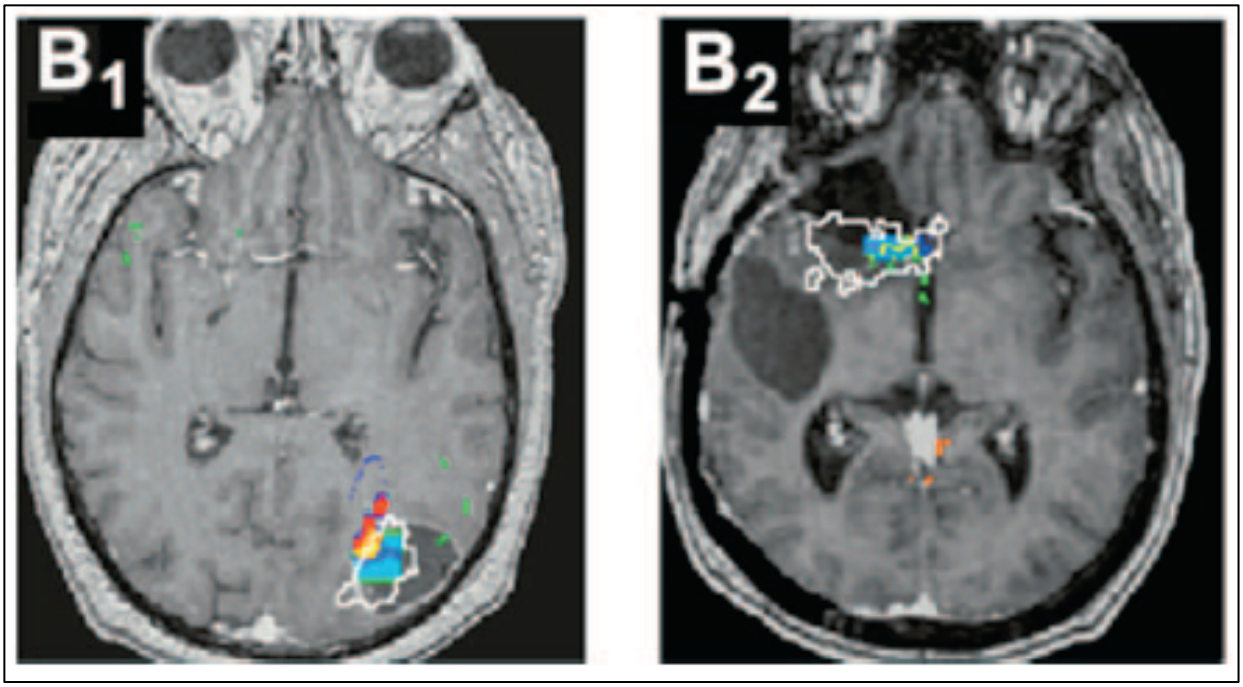


Figure 2.14: Two patients with recurrent GBM. Percent Overlap Method (POM) in color. Compared to ISODATA results in white (McMillan 2007).

A 1998 study by Clark et al. used T1, T2, and Proton Density (PD), images as input to a knowledge based (KB) classification system. They used this system to segment tumor volumes from patients with GBM. Their algorithm consisted of a multiple step process, whereby healthy voxels are stripped away on whole-volume datasets. The whole-volume approach is a convenient feature in order to reduce human intervention time. However, the results were mixed in this study. Table 2.2 shows the results of the KB system when compared against the radiologist's margins (treated as ground truth).

Patient	Scan	True Positive	False Positive	False Negative	Tumor Size	Percent Match	Corr. Ratio	"True" False Positive
1	Base	6921	2700	234	7155	0.97	0.78	80
1	R1	7038	3879	196	7234	0.97	0.70	467
1	R2	7285	4869	176	7461	0.98	0.65	496
1	R3	6206	3261	166	6372	0.97	0.72	227
1	R4	5930	3130	48	5978	0.99	0.63	47
2	Base	7892	5976	408	8300	0.95	0.54	18
2	R1	10092	3481	1059	11151	0.91	0.75	66
2	R2	14822	4961	1012	15834	0.94	0.78	219
3	Base	8917	1635	581	9498	0.94	0.85	47
3	R1	5003	2619	169	5172	0.97	0.71	89
4	Base	3054	1536	75	3129	0.98	0.73	124
4	R1	3627	2082	659	4286	0.85	0.43	1092
4	R2	2506	1020	1103	3609	0.69	0.46	495
5	Base	829	573	173	1002	0.83	0.54	161
6	Base	1425	624	0	1425	0.96	0.78	53
7	Base	177	175	0	177	1.00	0.51	54

Table 2.2: Comparison of Knowledge Based tumor segmentation versus hand-labeled segmentation per volume of patients with GBM (Clark 1998).

Table 2.2 shows two columns for *false positive* and *"true" false positive*. This may be a slightly confusing way to label the voxels, because the *false positive* column includes all of the false positive voxels. However, the *"true" false positive* column only reports a reduced subset of the

false positives; it only includes voxels which have no spatial contact with the main tumor volume.

Although their percent matching values are quite high (mostly above 90 percent), they include the true negatives to calculate this value. This is a common problem when dealing with unbalanced data sets. Due to the fact that the negative examples far outweigh the positive examples (roughly 10 to 1 ratio), the percent matching can actually be quite high, while still having a relatively low positive predictive value (PPV). These issues will be addressed in further detail in the methods section (Chapter 3.12). Although the results may not be as robust as they initially appear, their method is admittedly more automated than my work. Future implementations of my work could hopefully modify some of these multi-step approaches to assist in the reduction of processing time.

A recent study by Su et al. (2012) used SVM for segmenting GBM tissues. They used T1, T2, post-contrast T1, and FLAIR images for their analysis. They attempted to segment tissues on a voxel-wise basis into enhancing tumor, necrosis, and edema. This study is closely related to our particular project because they are focusing on a voxel by voxel classification of multiple tissues using SVM. Our study differs, as we also include ADC maps, rCBV maps, and carbogen breathing $\Delta T2^*$ maps. Although their results show an interesting improvement from using an active learning approach, like the Clark study, there is potential for improvement in the selection of classifier metrics. Specifically, the reported true positive rate and false negative rate both depend on the same two elements of the confusion matrix – true positives and false negatives. These measures seem to be somewhat redundant. It would be beneficial to include a metric such as positive predictive value, because PPV also depends on false positives which have not been previously reported.

(a) SVM active learning		
GBM tissue	TPR (average)	FNR (average)
Enhanced Tumor	88.4%	2.8%
Necrosis	86.3%	4.5%
Edema	82.5%	5.1%
(b) Knowledge-based fuzzy clustering method		
GBM tissue	TPR (average)	FNR (average)
Enhanced Tumor	77.7%	4.4%
Necrosis	78.4%	9.3%
Edema	74.5%	8.1%
(c) SVM with randomly selected samples		
GBM tissue	TPR (average)	FNR (average)
Enhanced Tumor	81.6%	5.3%
Necrosis	79.2%	8.1%
Edema	76.4%	6.8%

Table 2.3: Evaluation of GBM segmentation results using 3 different classifiers (Su 2012).

A 2011 study by Hu et al. looked at using SVM to classify tumor recurrence from radiation necrosis. They utilized post-contrast T1, T2, FLAIR, PD, rCBV, and ADC as features in a voxel-wise analysis. Distinguishing between tumor recurrence and radiation necrosis is a very useful project, because they are frequently indistinguishable using conventional MRI. This is because both the large endothelial gaps associated with GBM neovasculature and the gaps produced by radiation injury to native brain capillaries, result in compromise of the blood-brain barrier (New 2001). This leads to similar enhancement of post-contrast T1 images. The results of their SVM were compared against radiologist confirmed labels and can be seen in figure 2.15.

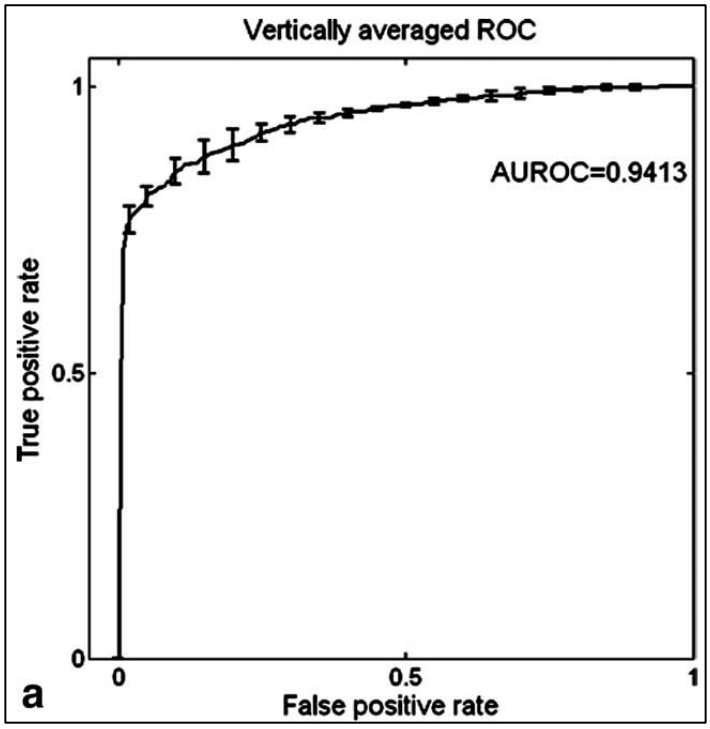


Figure 2.15: Averaged ROC curve for five test datasets using SVM to classify tumor recurrence vs. radiation necrosis (Hu 2011).

These results show promising performance of the SVM to classify recurrence vs. radiation. However, there is no information in the paper about the balance between the two classes. Although they used balanced training data, it is unlikely that they used balanced testing data due to the large amount of healthy tissue when compared to pathological tissue. This problem of unbalanced testing again presents a problem. The ROC curve is not as informative as it could be in the balanced case. Again this occurs because of the large ratio of negative examples. This makes it very easy to obtain a small false positive rate. In effect, the error introduced by the false positives can be masked by the overpowering class of true negatives. This is not to say that this was bad work by any means, although it would be useful to include a measure of positive predictive value, or the actual numbers from the confusion matrix. The concept is however, quite important and intriguing. Admittedly, I would have liked to include

radiation necrosis as one of our tissue types. However, our patients lacked enough radiation necrosis examples to be included in our modeling.

The benefits of multiparametric approaches are evident from many of these studies. We see that the literature supports the premise of added value from each of our scans: post-contrast T1, T2, ADC, rCBV, ΔT2*. The methods to be discussed will show how we built a classifier using support vector machine that is capable of handling multiple binary decisions; specifically, distinguishing between tumor, edema, necrosis, healthy, and non-enhancing tumor voxels. Although the performance of our classifiers ranges from poor to outstanding, I will outline the methodology used to establish a useful and reliable classifier. Specifically, the issues of classifier performance metrics will be addressed, as there is much confusion in the literature.

References

1. Al-Hallaq H, River J, Zamora M, Oikawa H, Karczmar G. Correlation of Magnetic Resonance and Oxygen Microelectrode Measurements of Carbogen-Induced Changes in Tumor Oxygenation 1. 1998;41(1):151–159.

2. Aronen HJ, Pardo FS, Kennedy DN, et al. High microvascular blood volume is associated with high glucose uptake and tumor angiogenesis in human gliomas. Clin Cancer Res 2000;6(6):2189-2200.

3. Boxerman JL, Schmainda KM, Weisskoff RM. Relative cerebral blood volume maps corrected for contrast agent extravasation. AJNR Am J Neuroradiol 2006;27(4):859-867.

4. Brown JM. The hypoxic cell: a target for selective cancer therapy--eighteenth Bruce F. Cain Memorial Award lecture. Cancer Res 1999;59(23):5863-5870.

5. Bulakbasi N, Guvenc I, Onguru O, Erdogan E, Tayfun C, Ucoz T. The added value of the apparent diffusion coefficient calculation to magnetic. J Comput Assist Tomogr 2004;28(6):735-746.

6. Burger PC, Vogel FS, Green SB, Strike TA. Glioblastoma multiforme and anaplastic astrocytoma.Pathologic criteria and prognostic implications. Cancer 1985;56(5):1106-1111.

7. Castillo M, Smith J, Kwock L, Wilber K. Apparent Diffusion Coefficients in the Evaluation of High-grade Cerebral Gliomas. 2001.

8. Clark MC, Hall LO, Goldgof DB, Velthuizen R, Murtagh FR, Silbiger MS. Automatic tumor segmentation using knowledge-based techniques. IEEE Trans Med Imaging 1998;17(2):187-201.

9. Di Costanzo A, Trojsi F, Giannatempo GM, et al. Spectroscopic, diffusion and perfusion magnetic resonance imaging at 3.0 Tesla in. J ExpClin Cancer Res 2006;25(3):383-390.

10. Donahue KM, Krouwer HG, Rand SD, et al. Utility of simultaneously acquired gradient-echo and spin-echo cerebral blood volume and morphology maps in brain tumor patients. MagnReson Med 2000;43(6):845-853.

11. Dunn JF, O'Hara JA, Zaim-Wadghiri Y, et al. Changes in oxygenation of intracranial tumors with carbogen: a BOLD MRI and EPR. J MagnReson Imaging 2002;16(5):511-521.

12. Emblem KE, Nedregaard B, Hald JK, Nome T, Due-Tonnessen P, Bjornerud A. Automatic glioma characterization from dynamic susceptibility contrast imaging: brain tumor segmentation using knowledge-based fuzzy clustering. J MagnReson Imaging 2009;30(1):1-10.

13. Gillies RJ, Bhujwalla ZM, Evelhoch J, et al. Applications of magnetic resonance in model systems: tumor biology and physiology. Neoplasia 2000;2(1-2):139-151.

14. Hacklander T, Hofer M, Reichenbach J, et al. [Possibilities of the use of MR tomography-based cerebral blood volume maps in. Rofo 1995;163(6):484-489.

15. Harrison L, Chadha M, Hill R, Hu K, Shasha D. Impact of Tumor Hypoxia and Anemia on Radiation Therapy Outcomes. 2002.

16. Henry RG, Vigneron DB, Fischbein NJ, et al. Comparison of relative cerebral blood volume and proton spectroscopy in patients with treated gliomas. AJNR Am J Neuroradiol 2000;21(2):357-366.

17. Horsman MR, Khalil AA, Nordsmark M, Grau C, Overgaard J. Relationship between radiobiological hypoxia and direct estimates of tumour oxygenation in a mouse tumour model. RadiotherOncol 1993;28(1):69-71.

18. Hoskin PJ, Rojas AM, Bentzen SM, Saunders MI. Radiotherapy with concurrent carbogen and nicotinamide in bladder carcinoma. J ClinOncol 2010;28(33):4912-4918.

19. Hu X, Wong KK, Young GS, Guo L, Wong ST. Support vector machine multiparametric MRI identification of pseudoprogression. J MagnReson Imaging 2011;33(2):296-305.

20. Jackson A, Kassner A, Annesley-Williams D, Reid H, Zhu XP, Li KL. Abnormalities in the recirculation phase of contrast agent bolus passage in cerebral gliomas: comparison with relative blood volume and tumor grade. AJNR Am J Neuroradiol 2002;23(1):7-14.

21. Kaanders JH, Pop LA, Marres HA, et al. ARCON: experience in 215 patients with advanced head-and-neck cancer. Int J RadiatOncolBiolPhys 2002;52(3):769-778.

22. Knopp EA, Cha S, Johnson G, et al. Glial neoplasms: dynamic contrast-enhanced T2*-weighted MR imaging. Radiology 1999;211(3):791-798.

23. Kono K, Inoue Y, Nakayama K, et al. The role of diffusion-weighted imaging in patients with brain tumors. AJNR Am J Neuroradiol 2001;22(6):1081-1088.

24. Kuperman V, River JN, Lewis MZ, Lubich LM, Karczmar GS. Changes in T2*-weighted images during hyperoxia differentiate tumors from normal tissue. MagnReson Med 1995;33(3):318-325.

25. Lam WW, Chan KW, Wong WL, Poon WS, Metreweli C. Pre-operative grading of intracranial glioma. ActaRadiol 2001;42(6):548-554.

26. Lee SJ, Kim JH, Kim YM, et al. Perfusion MR imaging in gliomas: comparison with histologic tumor grade. Korean J Radiol 2001;2(1):1-7.

27. McMillan KM, Ehtesham M, Stevenson CB, Edgeworth ML, Thompson RC, Price RR. T2 detection of tumor invasion within segmented components of glioblastoma. J MagnReson Imaging 2009;29(2):251-257.

28. McMillan KM, Rogers BP, Koay CG, Laird AR, Price RR, Meyerand ME. An objective method for combining multi-parametric MRI datasets to characterize. Med Phys 2007;34(3):1053-1061.

29. Muti M, Aprile I, Principi M, et al. Study on the variations of the apparent diffusion coefficient in areas of solid. MagnReson Imaging 2002;20(9):635-641.

30. New P. Radiation injury to the nervous system. CurrOpinNeurol 2001;14(6):725-734.

31. Ogawa S, Menon RS, Tank DW, et al. Functional brain mapping by blood oxygenation level-dependent contrast magnetic. Biophys J 1993;64(3):803-812.

32. Ostergaard L, Weisskoff RM, Chesler DA, Gyldensted C, Rosen BR. High resolution measurement of cerebral blood flow using intravascular tracer bolus passages. Part I: Mathematical approach and statistical analysis. MagnReson Med 1996;36(5):715-725.

33. Overgaard J. Sensitization of hypoxic tumour cells--clinical experience. Int J RadiatBiol 1989;56(5):801-811.

34. Papadimitrou JM, Woods AE. Structural and functional characteristics of the microcirculation in neoplasms. J Pathol 1975;116(2):65-72.

35. Pirzkall A, McGue C, Saraswathy S, et al. Tumor regrowth between surgery and initiation of adjuvant therapy in patients. NeuroOncol 2009;11(6):842-852.

36. Provenzale JM, Wang GR, Brenner T, Petrella JR, Sorensen AG. Comparison of permeability in high-grade and low-grade brain tumors using dynamic susceptibility contrast MR imaging. AJR Am J Roentgenol 2002;178(3):711-716.

37. Rosen BR, Belliveau JW, Vevea JM, Brady TJ. Perfusion imaging with NMR contrast agents. MagnReson Med 1990;14(2):249-265.

38. Shin JH, Lee HK, Kwun BD, et al. Using relative cerebral blood flow and volume to evaluate the histopathologic. AJR Am J Roentgenol 2002;179(3):783-789.

39. Simon JM, Noel G, Chiras J, et al. Radiotherapy and chemotherapy with or without carbogen and nicotinamide in inoperable biopsy-proven glioblastoma multiforme. RadiotherOncol 2003;67(1):45-51.

40. Stadnik TW, Chaskis C, Michotte A, et al. Diffusion-weighted MR imaging of intracerebral masses: comparison with conventional MR imaging and histologic findings. AJNR Am J Neuroradiol 2001;22(5):969-976.

41. Su P, Zhong X, Chi L, Jianhua Y, Wong S. Support vector machine (SVM) active learning for automated Glioblastoma segmentation. Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on; 2012. p. 598-601.

42. Sugahara T, Korogi Y, Kochi M, et al. Usefulness of diffusion-weighted MRI with echo-planar technique in the evaluation of cellularity in gliomas. J MagnReson Imaging 1999;9(1):53-60.

43. Sugahara T, Korogi Y, Tomiguchi S, et al. Posttherapeuticintraaxial brain tumor: the value of perfusion-sensitive. AJNR Am J Neuroradiol 2000;21(5):901-909.

44. Teicher BA. Physiologic mechanisms of therapeutic resistance. Blood flow and hypoxia.HematolOncolClin North Am 1995;9(2):475-506.

45. Tien RD, Felsberg GJ, Friedman H, Brown M, MacFall J. MR imaging of high-grade cerebral gliomas: value of diffusion-weighted echoplanar pulse sequences. AJR Am J Roentgenol 1994;162(3):671-677.

3. Methods

The 5 patients presented for case studies are as follows:

Patient A Male, 50 years, Post-craniotomy, Post radiotherapy

Patient B – Male, 61 years, Post XRT, Chemo, Surgical Resection

Patient C – Female, 47years, Post-Resection, Post-Radiotherapy

Patient D – Female, 59 years, Post XRT, Chemo, Radiosurgery

Patient E – Male, 33years, Post-Resection, Post-Craniectomy

Post-contrast T1 3d SPGR	TR = 20ms, TE = 6 ms, ϕ = 30º, FOV = 22cm, slice = 1.5mm, gap = 0 mm, matrix = 256 x 256
T2 Fast Spin Echo	TR = 3.4s, TE = 89 ms, ϕ = 90º, FOV = 22cm, slice = 3mm, gap = 1 mm, matrix = 256 x 256
Diffusion-Weighted EPI	TR = 5.0s, TE = 100ms, ϕ = 90º, FOV = 22cm, slice= 5mm, gap = 0 mm, matrix = 128 x 128, b=0 and b=1000 s/mm^2
Perfusion Imaging	TR = 1.5s, TE = 60 ms, ϕ = 60º, FOV = 22cm, slice = 5mm, gap = 1 mm, matrix = 128 x 128
Δ T2* Hypoxia Imaging	TR = 2.0s, TE = 35-80 ms, ϕ = 90º, FOV = 22cm, slice = 5mm, gap = 1 mm, matrix = 128 x 128

Table 3.1: MRI scan parameters

3.2 Diffusion Imaging

Diffusion-weighted imaging (DWI) is a type of MR imaging that examines the degree of random Brownian motion of water molecules in tissues (Carr and Purcell 1954; Woessner 1961; Stejskal and Tanner 1965). The differences in the amount of restricted motion can tell us information about normal vs. abnormal structures. For our study we used a typical Spin-Echo EPI sequence.

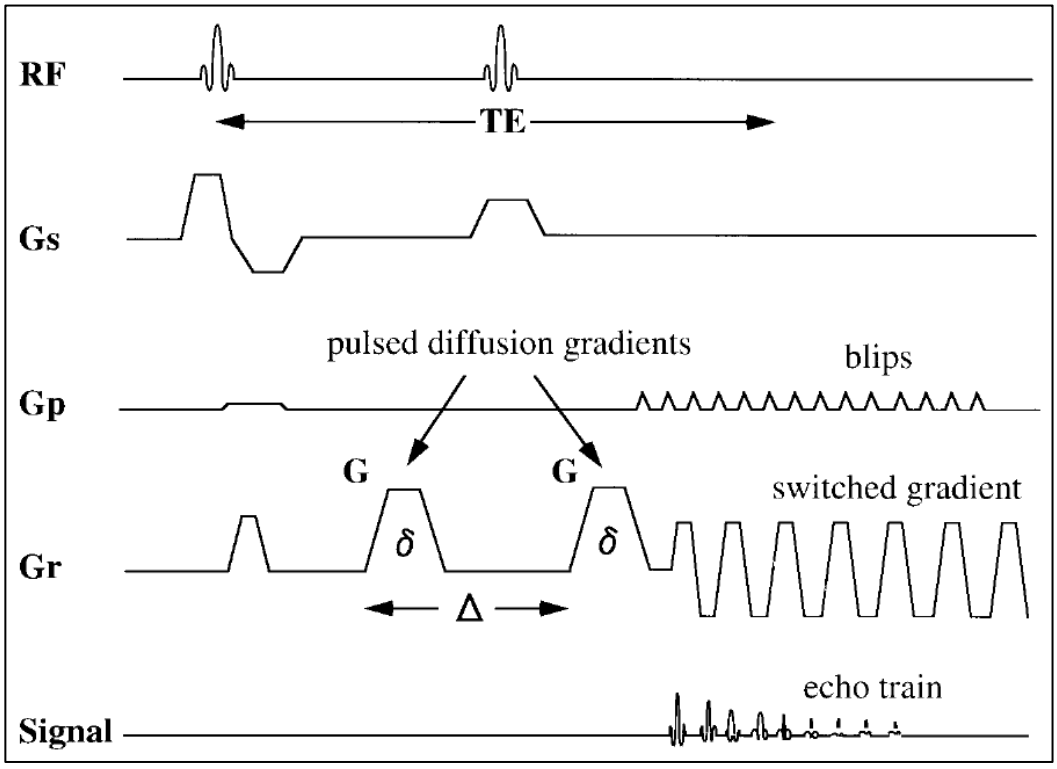


Figure 3.1: Typical Spin-Echo EPI Diffusion Sequence (Sugahara 1999)

Gs = slice selection gradient waveforms, **Gp** = phase-encoding waveforms, **Gr** = frequency-encoding waveforms. Diffusion sensitivity is due mainly to the diffusion gradients, **G**. The duration of **Δ** is determined by the time between the leading edges of the diffusion gradients.

An early description of DWI by Stejskal and Tanner (1965) uses a spin-echo T2-weighted pulse sequence with two extra gradient pulses. This allows us to measure the rate of movement along one direction (take x for example). These gradient pulses are equal and opposite for all points at the same x location; but the strength of the two balanced gradients increases as we move along the x direction. Therefore if a voxel of tissue contains water that has no net movement between the balanced gradient pulses, then the diffusion signal will cancel out, and the resultant signal intensity of that voxel will be equal to the signal intensity of an image using the same sequence without the diffusion gradients.

However, if the water molecules have a net movement in the x direction, then the two gradients are no longer equal in magnitude and they do not cancel. This is because a moving water molecule would be subject to the first gradient pulse at one location and magnitude, and then upon arriving at a new x location, it experiences the second gradient pulse with a different magnitude corresponding to that new location. The difference in gradient magnitude is proportional to the displacement in the x direction. Therefore, faster moving water protons undergo a larger net dephasing.

This means that the resulting signal intensity for a voxel with moving water protons is equal to its 'normal' T2-weighted intensity minus the diffusion term. We then calculated the signal intensity for a given voxel using the equation 1:

$$1) \qquad SI = SI_0 e^{-bD}$$

SI_0 = the signal intensity on the T2-weighted (b = 0 sec/mm^2) image

$$b = \gamma^2 G^2 \delta^2 (\Delta - \frac{\delta}{3})$$

γ = gyromagnetic ratio

G = magnitude of the two balanced DW gradient pulses

δ = width of the two balanced DW gradient pulses

Δ = time between the two balanced DW gradient pulses

According to Fick's law, true diffusion is the net movement of molecules due to a concentration gradient. However, with MRI, we cannot differentiate between concentration gradients, pressure gradients, thermal gradients, or ionic interactions. For this reason, the term apparent diffusion coefficient is substituted for true diffusion (Schaefer 2000):

2) $$SI = SI_0 e^{-b\ ADC}$$

3) $$ADC = \frac{1}{b}\ln(\frac{SI_0}{SI})$$

Therefore, if a molecule moves faster and experiences a greater difference in diffusion pulses, then the net dephasing is higher, and results in a lower intensity on a diffusion weighted scan. And since the diffusion weighted intensity (SI) is inversely proportional to the ADC value, it results in a low ADC value. Lower ADC values reflect less water motion, and higher ADC values reflect more motion.

The method above is described for one direction. However, it is easily repeated for additional directions (equation 4), and we are able to gain a better picture about the diffusivity of water molecules in the brain, and measure which directions are more dominant in their diffusivity:

4) $$ADC_0 = \frac{[ADC_x + ADC_y + ADC_z]}{3}$$

For instance, white matter is especially anisotropic, and it is believed that the predominant diffusion direction often coincides with the direction of fiber tracts (Moseley 1990). Although in our study, we are also concerned with the areas of reduced diffusivity due to the hypercellularity of the tumor.

For our study, diffusion gradients were applied in three orthogonal directions, and the acquisition parameters for the diffusion-weighted EPI images were TR = 5.0s, TE = 100ms, ϕ = 90º, field of view (FOV) = 22cm, slice = 5mm, gap = 0 mm, matrix = 128 x 128, b=0 and b=1000 s/mm2 (see table 3.1 for summary), resulting in whole brain coverage. After obtaining the B_0 and B_{1000} images we applied the eddy current correction feature from the FMRIbrain Software Library (FSL). FSL is a comprehensive library of analysis tools for functional, structural and diffusion MRI brain imaging written mainly be members of the Analysis Group, FMRIB, and Oxford (Jenkinson 2012). Eddy currents in the gradient coils induce (approximate) stretches and shears in the diffusion weighted images. These distortions are different for different gradient directions. FSL's Eddy Current Correction corrects for these distortions, and for simple head motion, using affine registration to a reference volume (Jenkinson 2001; Jenkinson 2002; Greve 2009). Affine registration uses 12 degrees of freedom to coregister the images (3 rotations, 3 translations, 3 scalings, and 3 skews/shears). After the eddy current correction is completed, we then calculated the ADC values, and produced a voxel-by-voxel whole brain ADC map for each subject and scan date.

3.3 Perfusion Imaging

The perfusion imaging was performed using a gradient echo EPI pulse sequence with the following acquisition parameters: TR = 1.5s, TE = 60 ms, $\phi = 60^{o}$, FOV = 22cm, slice = 5mm, gap = 1 mm, matrix = 128 x 128 (Table 3.1)

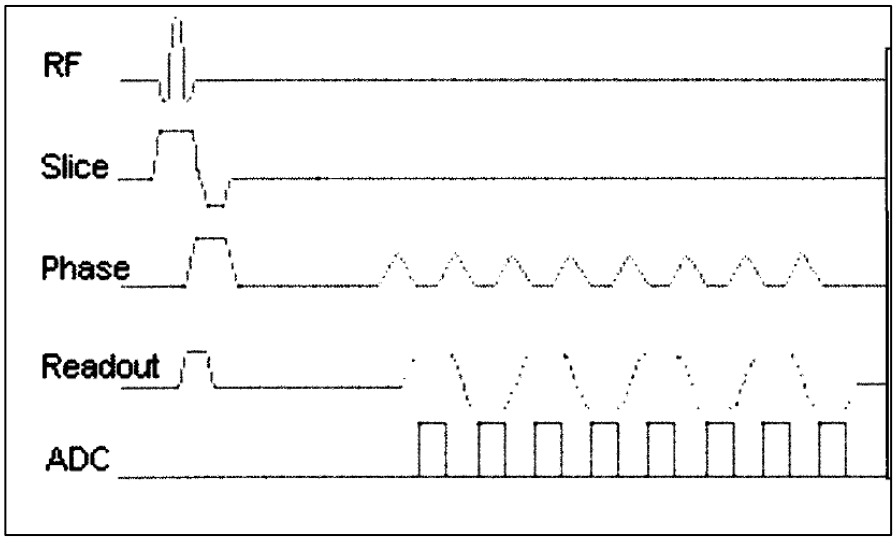


Figure 3.2: Gradient Echo EPI pulse Sequence (adapted from mr-tip.com)

A gadodiamide contrast agent (Omniscan, Nycomed-Amersham, Princeton, NJ) was injected at 4.0 ml/s followed by a saline flush using a power injector (Spectris, MEDRAD, Indianola, PA) 13s after scan initiation. The amount of contrast administered was 0.1 mmol/kg of body weight. The images were acquired in the axial plane and whole brain coverage was achieved.

When paramagnetic contrast agents (such as gadodiamide) are confined to the intravascular space, they produce signal intensity losses on T2-weighted scans. This is due to the differences in local magnetic susceptibility between the vessels and their surrounding tissue. Therefore, both intravascular and extravascular spins undergo a reduction in T2* that leads to a

large transient signal loss. In normal white matter with a standard dose of contrast agent (0.1mmol/kg of body weight), the signal loss is approximately 25 percent (Rosen 1990; Chan 2002).

In a typical (no leakage correction) method, the relative cerebral blood volume (rCBV) is estimated by integrating the relaxivity-time curve. However, tumor patients often do not have an intact blood-brain barrier (BBB), and a correction must be made to account for the leakage of contrast agent into the extravascular space. Our perfusion maps were generated using a technique developed be Boxerman et al. (2006) to correct for leakage effects. The dynamic signal intensity curve ***S(t)*** was converted to a relaxivity-time curve, which is a parameter related to the concentration of gadolinium in each voxel:

1) $$\Delta\widetilde{R2}^*(t) = -\frac{1}{TE}\ln\frac{S(t)}{S_0}$$

$\Delta\widetilde{R2}^*(t)$ refers to the contaminated voxel level time curve

$\Delta R2^*(t)$ refers to the uncontaminated (true) voxel level time curve

Assuming that the true $\Delta R2^*(t)$ for each voxel is a scaled version of the whole-brain $\overline{\Delta R2^*}(t)$:

2) $$\Delta R2^*(t) = K_1\overline{\Delta R2^*}(t)$$

Then, the relaxivity change for each voxel is approximated as a linear combination of the whole-brain average contaminated non-enhancing voxels $\overline{\Delta\widetilde{R2}^*}(t)$, the uncontaminated whole-brain average $\overline{\Delta R2^*}(t)$, and its time integral:

$$3) \qquad \Delta\widetilde{R2}^{*}(t) \approx K_1\overline{\Delta R2^{*}}(t) - K_2 \int_0^t \overline{\Delta R2^{*}}(t')dt'$$

$K_1\overline{\Delta R2^{*}}(t)$ reflects the uncontaminated $\Delta R2^{*}(t)$ for each voxel (T2* relaxivity term), and the K_2 term reflects the effects of leakage (T1 relaxivity term). Equation (3) has 2 unknowns, K_1 and K_2. Multiplying the measured brain-averaged log-signal change and its time integral, K_1 and K_2 can be determined by a simple linear least-squares fitting. Then a corrected $\Delta\widetilde{R2}^{*}(t)$ can be computed:

$$4) \qquad \Delta R2^{*}(t)_{corrected} = \Delta\widetilde{R2}^{*}(t) + K_2 \int_0^t \overline{\Delta R2^{*}}(t')dt'$$

The final step is to integrate $\Delta R2^{*}(t)_{corrected}$ to obtain a corrected estimate of the blood volume.

$$5) \qquad rCBV_{corrected} = rCBV + K_2 \int_0^T dt'' \int_0^{t''} \overline{\Delta R2^{*}}(t')dt'$$

The double integral term is the same for all voxels, and T is the end time point for numerical integration of the bolus. Because the time integral is the same for all voxels, this gives us a simple correction proportional to K_2 which varies for each voxel (Boxerman 2006).

These calculations were performed using MATLAB (Mathworks Corp., Natick, MA), with perfusion codes written by Yin Huang (2009).

3.4 Carbogen Protocol

The carbogen breathing ΔT2* protocol was performed using a multi-echo EPI sequence with ten different echo times (TE) in order to quantify T2*. The parameters were: TR = 2.0s, TE = 35-80 ms, $\phi = 90^{o}$, FOV = 22cm, slice = 5mm, gap = 1 mm, matrix = 128 x 128 for full brain coverage. A breathing mask was attached to the patient to allow for the delivery of a carbogen gas mixture. During the first multi-echo scan, the patient wore the mask, but the gas tube was not connected. This allowed for normal, unaltered oxygen levels. We referred to this as the 'air' breathing condition. A full set of multi-echo EPI images were obtained under this condition, with full volumes acquired for each varying TE.

For the second breathing condition, the gas tube was connected to the breathing mask and the patient breathed a carbogen gas mixture (95% O_2, 5% CO_2). The gas flow was adjusted to a rate of 20 l/min. After breathing for a 15 minute equilibrium period, a second multi-echo scan is obtained under the carbogen breathing conditions (the patient continued to breathe the carbogen during this 'carbogen' acquisition). A second set of multi-echo EPI images were then obtained under this carbogen condition using the same parameters as the air condition. Patients were instructed to only breathe through the mask (later patients utilized nose clips) (McMillan 2006).

Each condition (air vs. carbogen) was post-processed separately at first. A full volume was collected at each of 10 echo times. These volumes were first realigned to each other using SPM realign tool. This tool utilizes a 6 parameter (rigid body) spatial translation (Friston 1995, Thevenaz 2000). This results in 10 realigned volumes for each breathing condition. Therefore, we are then able to analyze how the signal intensity changes voxel-by-voxel for each of the 10 echo times.

We then calculated T2* for each breathing condition separately by fitting the signal intensity (SI) vs. echo time (TE) to a single exponential function using MATLAB (Mathworks Corp., Natick, MA) according to equation 1:

$$1) \qquad \ln(S) = -\frac{1}{T2^*}(TE) + \ln(S_0)$$

Where $\boldsymbol{S}$ and $\boldsymbol{S_0}$ are the signal intensities at TE and TE=0 respectively (Dunn 2002; McMillan 2006).

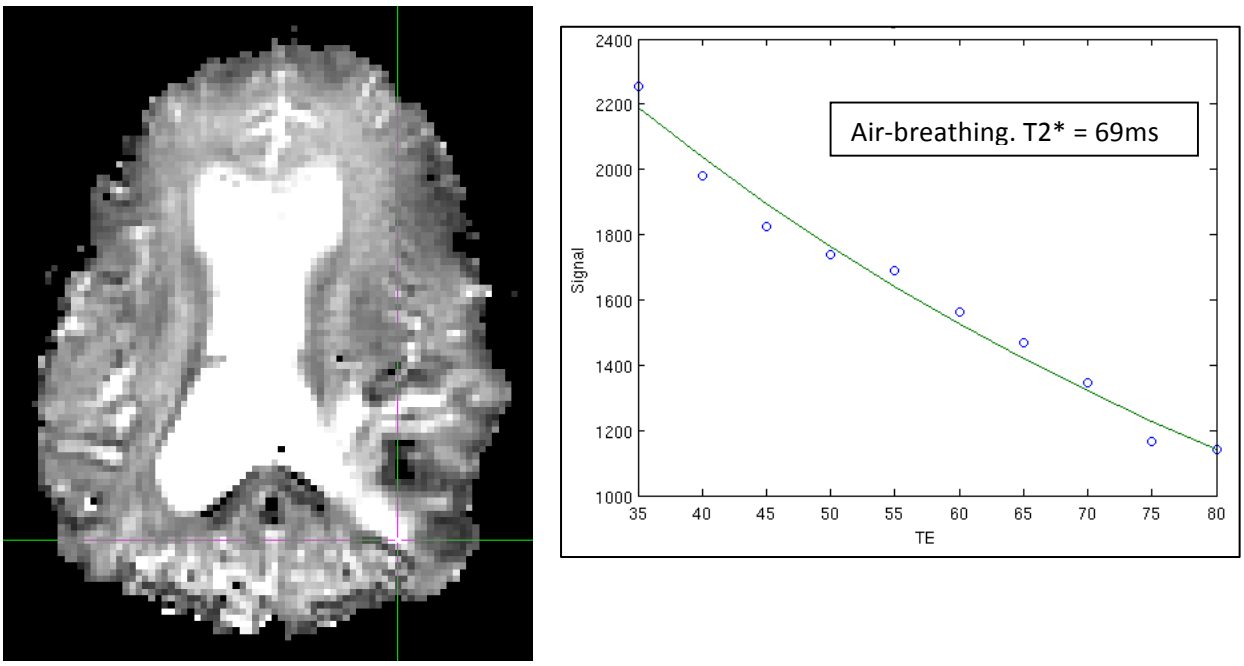


Figure 3.3: A) Multi-echo T2*$_{air}$ map. B) Single voxel multi-echo fit (T2*$_{air}$ = 69ms)

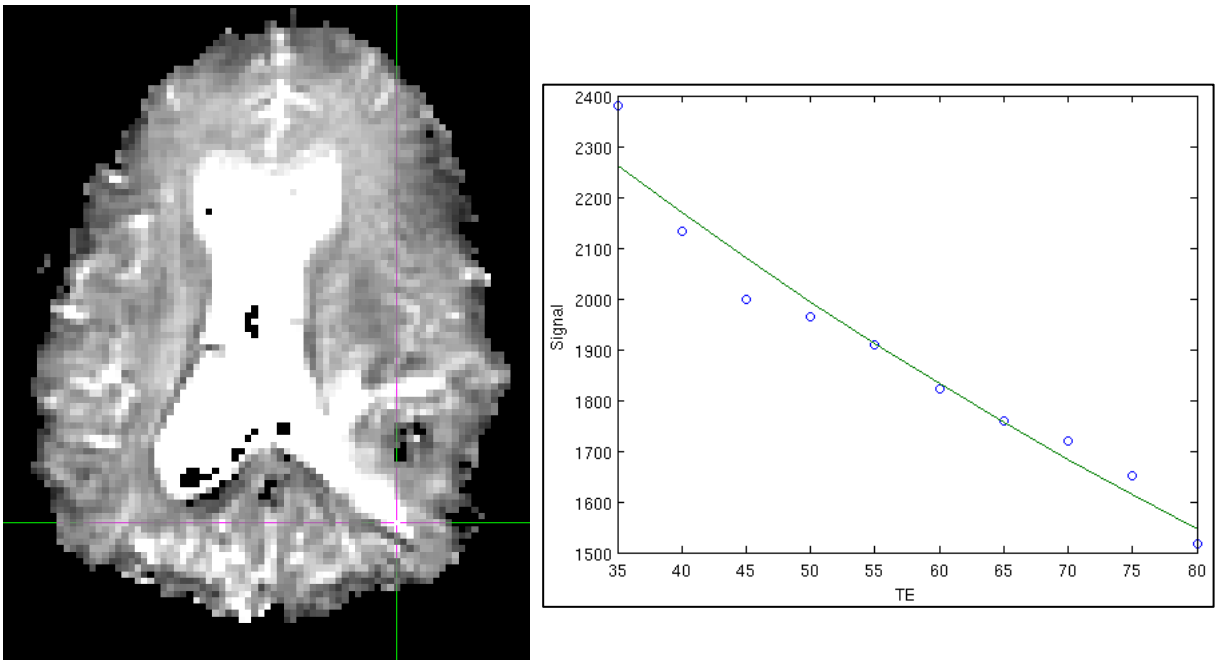


Figure 3.4: A) Multi-echo T2*$_{carbogen}$ map. B) Single voxel fit (T2*$_{carbogen}$ = 118ms)

After obtaining a separate T2* map for both the air condition and the carbogen condition, we then used SPM's 12 parameter affine coregistration tool (Collignon 2007) to register the air and carbogen images to each other. Upon registration and reslicing (see chapter 3.4 for more details), we then subtracted the T2*$_{air}$ from the T2*$_{carbogen}$ image to obtain a carbogen-air T2* map; or ΔT2* map for each subject.

Figure 3.5: Δ T2* map

3.5 Registration Techniques

The first step after acquiring our scans is to coregister all of them to a common grid, so that each voxel matches from scan to scan. This is a two-step process. First, each volume within a single scan needs to be properly aligned. In the case of post-contrast T1 and T2, there is only one volume acquired, so there is no need for this step. However, with diffusion, perfusion, and ΔT2*, we proceeded by aligning the within-scan volumes first.

For the perfusion scan, we collect between 35 and 45 volumes per scan, to establish the time-course curve. We then used SPM's re-align tool, which uses a 6-parameter rigid body spatial translation (Friston 1995, Thevenaz 2000). All of the volumes are acquired using the same prescription (positioning parameters of the scanner),and there is no major source of distortion between volumes. Therefore, since we are scanning the same patient under extremely similar conditions, we can expect no major distortions. In this case 3 rotations and 3 translations is an adequate tool for re-aligning within patient data.

However, with our ΔT2* scan, we needed a multi-step approach. As stated earlier, we calculate two multi-echo T2* maps for each patient. Each map (one for air-breathing, one for carbogen-breathing) is calculated from ten volumes that are aligned using SPM's 6-parameter re-alignment tool. We then have a $T2^*_{air}$ and $T2^*_{carbogen}$ map for each patient. Since each map is generated from the same patient using the same scan parameters, we would ideally be able to use another rigid body transformation to align the $T2^*_{air}$ and $T2^*_{carbogen}$ maps to each other. However, a complication arose because of the necessary 12 minute gap between scans needed for carbogen saturation to reach equilibrium. There were often times major shifts in alignment over this time period, and in most cases, this resulted in mismatched slice locations between the

two breathing conditions. Due to these greater differences, a 6-parameter model failed to re-align the images properly. We found more success using SPM's 12-parameter affine transformation, which also uses a more complex interpolation method (tri-linear) for reslicing the data. All scans were individually visually inspected for confirmation. Once aligned, the difference map was computed to produce the ΔT2*map.

The diffusion scan produced two volumes that were first aligned to each other using FSL's eddy correct method, which uses a 12-parameter affine registration (Jenkinson 2001; Jenkinson 2002; Greve 2009). Since the diffusion gradients introduce distortions a simple rigid body method is not preferred. Once aligned, an ADC map was computed.

Upon completion of the individual re-alignment steps, we then needed to re-align each individual scan type to a common volume. This process is done using SPM's 12-parameter affine coregistration tool. All maps were realigned and resliced using tri-linear interpolation to match the high-resolution T2 anatomical image. All images were stored as 3d NIFTI files. A flow chart of this process can be seen in figure 3.6.

Although many of these details may seem of lesser import, I believe them to be extremely relevant to the sources of error in our study. It is my opinion that, in some cases, a significant amount of information was lost to the interpolation process needed for the coregistration and reslicing step. Future studies could benefit from using more similar scan prescriptions in order to avoid slice mismatch as much as possible.

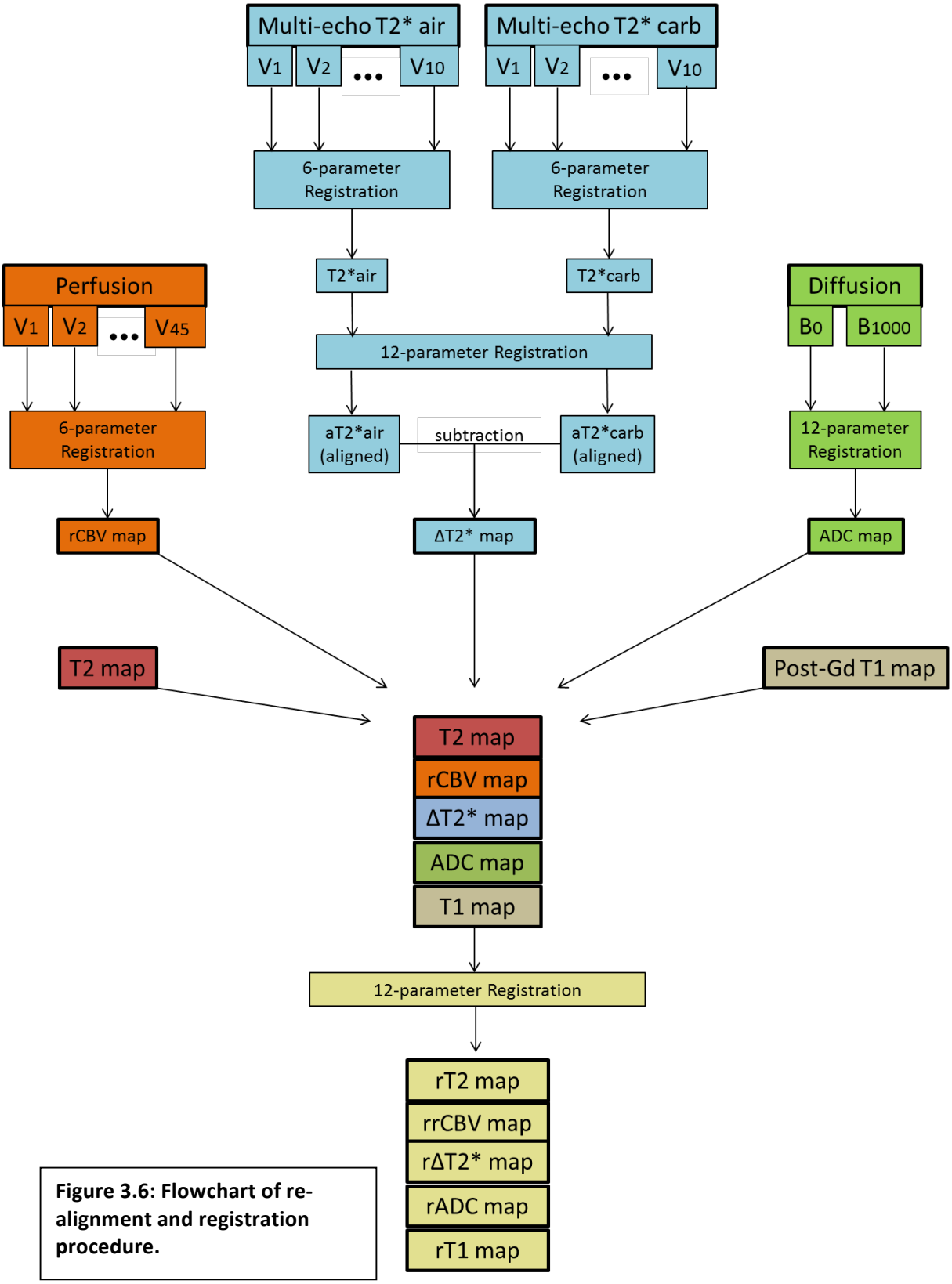


Figure 3.6: Flowchart of re-alignment and registration procedure.

3.6 Tissue Labels

Upon realignment of all 5 scans, each voxel possessed 5 unique feature values. The next step in preparing the data for a machine learning analysis, was to provide the corresponding tissue labels for each voxel. This was performed by expert consensus analysis, consisting of two neuroradiologists: Dr. John Hald (Department of Neuroradiology, Oslo University Hospital, Oslo, Norway) and Dr. Patrick Turski (Department of Radiology, University of Wisconsin, Madison, WI). The labeling was performed in FSL using multiple MR scans and overlays to accurately define the tissue boundaries. Serial scans were also used to investigate later developments, helping to disambiguate areas of concern: such as, eliminating the possibility of radiation necrosis as the enhancing region continued to grow over time. Patient history and physicians notes were also referenced at times. This process was done for each patient on a slice-by-slice, voxel-by-voxel basis.

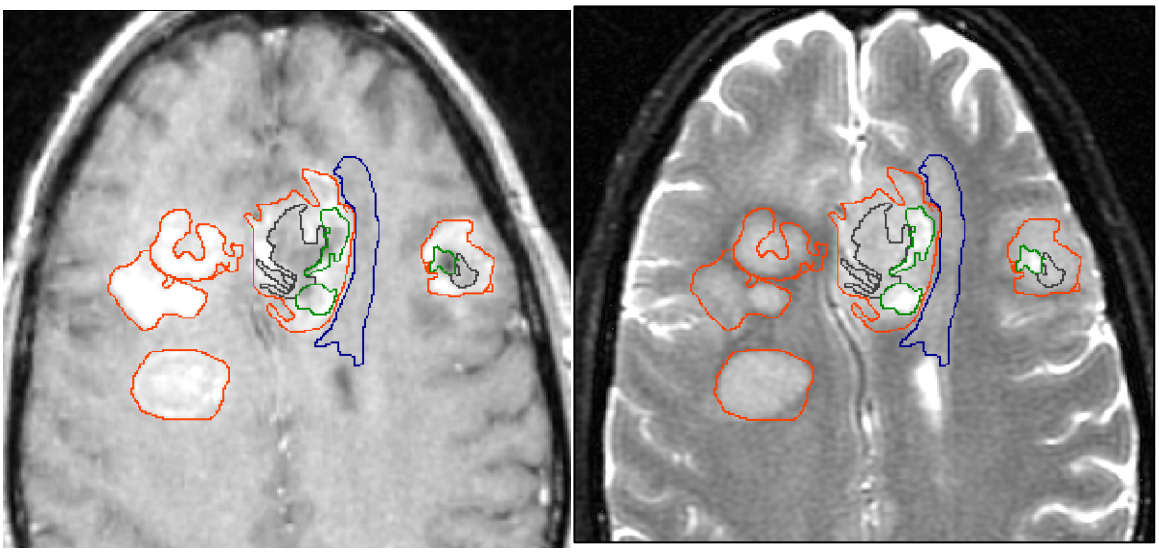


Figure 3.7: Neuroradiologist drawn full tissue labels. Red=tumor, Blue=edema, Green=necrosis, Grey=non-enhancing tumor.

The voxels were assigned to 1 of 5 classes: enhancing tumor, necrosis, edema, non-enhancing tumor, and healthy tissue. Although various other types of classifications exist, we felt like certain tissue types would be too similar to distinguish, and were therefore pooled together. For instance, we included gliosis in the edema category. We also included cyst in the necrosis category. Although these instances have differing properties, they might be more suited for ROI analysis. For example edema can cause effacement of the sulci due to mass effect; while gliosis often leads to an enlargement of the sulci due to scarring and loss of tissue, creating more space. Although these two tissues may have similar voxel-wise properties, if entered into a ROI level machine learning algorithm, we could include information about the shape and texture of the ROI. This is one of the limitations of this study, but there is also much to gain from a voxel-wise analysis. For this reason we are able to create boundaries in future cases, which would not be possible with ROI methods.

The process of individually labeling each voxel is quite time consuming. Therefore, we used each expertly assigned label as our ground truth for the remainder of the study. However, we also prepared a smaller subset of labels for training data that attempted to speed up the workflow. The idea for this proposed method, was that due to the large numbers of voxels in the total dataset, we could train the classifier on a much smaller subset of voxels and still achieve reasonable results. To test this, we used quickly drawn labels from a single slice rather than the entire brain. Not only does this reduce the slice number, but we also selected voxels well within the boundaries of their respective tissue margins; thereby eliminating the time intensive task of outlining the borders. An example of these labels can be seen in figure 3.8.

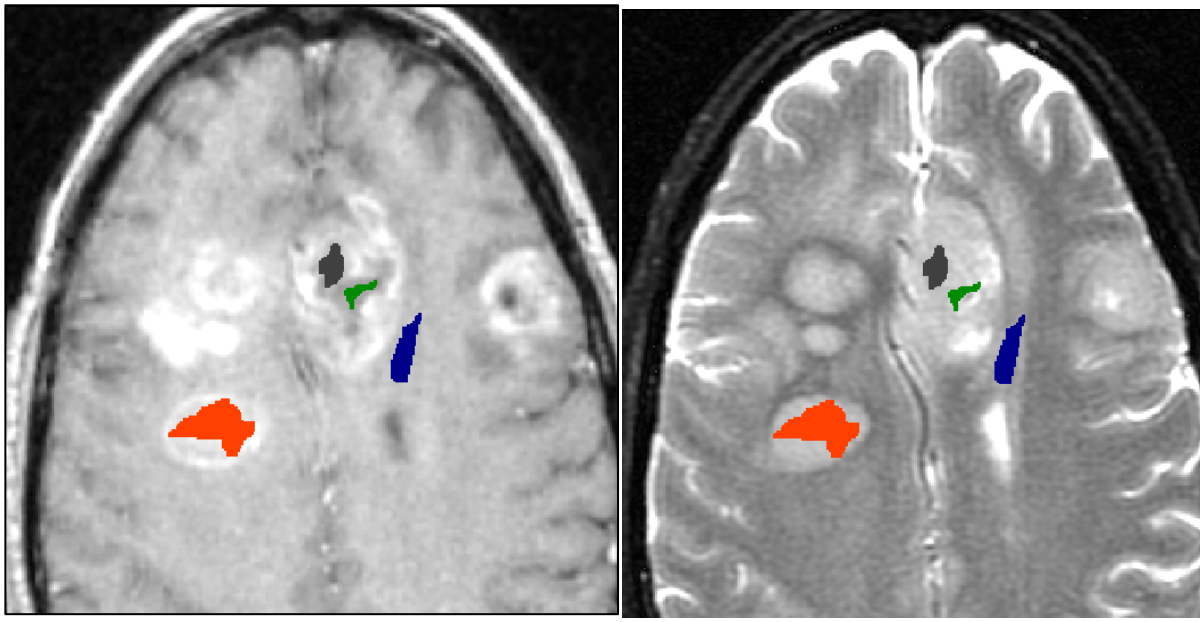


Figure 3.8: Fast subset tissue labels used only for training. Red=tumor, Blue=edema, Green=necrosis, Grey=non-enhancing tumor.

3.7 SVM Theoretical Description

The SVM is a type of learning algorithm that attempts to discriminate between data points (X_i) belonging to two classes, defined by class labels ($Y_i \in \{+1, -1\}$). If we have n data points with D dimensions, then X_i has D features represented by $x_1, x_2 \ldots x_D$. Our data then takes the form:

$$\{X_i, Y_i\} \text{ where } i = 1 \ldots n, \quad Y_i \in \{+1, -1\}, \quad X \in R^D$$

Let us first consider the simplest form, where we assume the data is linearly separable. If our data has two features (D=2), then we can plot those features x_1 vs. x_2 on a graph and draw a line between the two groupings of points X_i. On one side of the line the points X_i have labels $Y_i = +1$, and on the other side the points have labels $Y_i = -1$. In many instances the data can have more

than two features and the dimension is therefore D > 2. In these cases the points $\boldsymbol{X_i}$ must be separated by a hyperplane which is described by the equation $\prod_{w,b} = \boldsymbol{w} \cdot \boldsymbol{X_i} + b = 0$ where:

$\boldsymbol{w}$ is the weight vector and is normal to the hyperplane.

b is referred to as the bias

$b / \|\boldsymbol{w}\|$ is the perpendicular distance from the hyperplane to the origin.

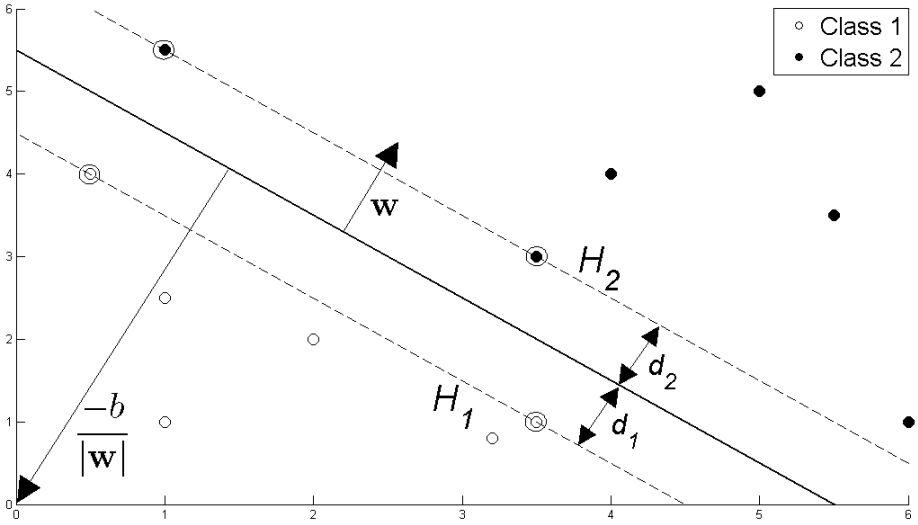


Figure 3.9: Hyperplane through two linearly separable classes (Fletcher 2009).

The support vectors are the points (or vectors in high-dimensional space) that are closest to the hyperplane. In the figure above, the support vectors are $\boldsymbol{H_1}$ and $\boldsymbol{H_2}$. The aim of a Support Vector Machine (SVM) is to orientate the hyperplane in a way that maximizes the margin between the support vectors and the hyperplane (Cortes 1995). The underlying idea is that choosing a hyperplane farthest from the observed data points, minimizes the risk of misclassifying future data points.

We select the variables $\boldsymbol{w}$ and b so that our training data can be described by:

$$\boldsymbol{w} \cdot \boldsymbol{X_i} + b \geq +1 \quad for \quad \boldsymbol{Y_i} = +1$$

$$\boldsymbol{w} \cdot \boldsymbol{X_i} + b \leq -1 \quad for \quad \boldsymbol{Y_i} = -1$$

We then solve

$$\boldsymbol{max_{(w,b)}\ (min_i\ d(\prod_{w,b}, X_i))}$$

where $\boldsymbol{d(\prod_{w,b}, X_i) = |w \cdot X_i + b| / \|w\|}$ is the distance between data point $\boldsymbol{X_i}$ and the plane $\prod_{w,b}$, subject to the constraint that this plane still separates the classes (Ben-Hur 2010). This can be reduced to the Primal Optimization Problem:

$$\boldsymbol{min_{(w,b)}\ (\tfrac{1}{2})\|w\|^2}$$

$$\text{subject to} \quad \boldsymbol{Y_i\ (w^T\ X_i + b)} \geq 1$$

Generally, there are only a small number of data points which will attain equality in the constraint above. These small subsets of data points are termed the support vectors because they support (constrain) the hyperplane. This is also a type of data compression because these few support vectors contain all the information necessary to solve the primal problem and derive the decision rule.

Up to this point we have considered a linear boundary, but in many instances, a non-linear boundary is better suited to separate the data. The goal with SVMs is to achieve the flexibility of a non-linear classifier while taking advantage of the simple training algorithms and computational efficiency of the linear classifier. This is achieved by mapping our data from the input space $\mathcal{X}$ to feature space $\mathcal{F}$ using a non-linear function $\phi : \mathcal{X} \rightarrow \mathcal{F}$. In the space $\mathcal{F}$ the discriminant function is:

$$f(\boldsymbol{X}) = \boldsymbol{w^T \phi(X)} + b$$

In order to solve the problem of increasing complexity, a kernel method is used to avoid mapping the data to a high dimensional feature space. The weight vector can be expressed as a linear combination of the training examples:

$$\boldsymbol{w} = \sum_{i=1..n} \alpha_i \boldsymbol{X}_i$$

$$f(\boldsymbol{X}) = \sum_{i=1..n} \alpha_i \boldsymbol{X}_i^T \boldsymbol{X} + b$$

In the feature space F this expression takes the form:

$$f(\boldsymbol{X}) = \sum_{i=1..n} \alpha_i \phi(\boldsymbol{X}_i)^T \phi(\boldsymbol{X}) + b$$

The kernel function is defined as:

$$k(\boldsymbol{X}_i, \boldsymbol{X}) = \phi(\boldsymbol{X}_i)^T \phi(\boldsymbol{X})$$

In terms of the kernel function, the discriminant function is:

$$f(\boldsymbol{X}) = \sum_{i=1..n} \alpha_i k(\boldsymbol{X}_i, \boldsymbol{X}) + b$$

Because the kernels depend only on the dot products of the two vectors, the linear decision boundary can be 'kernelized' and cast into a higher dimensional space without explicitly computing the mapping ϕ.

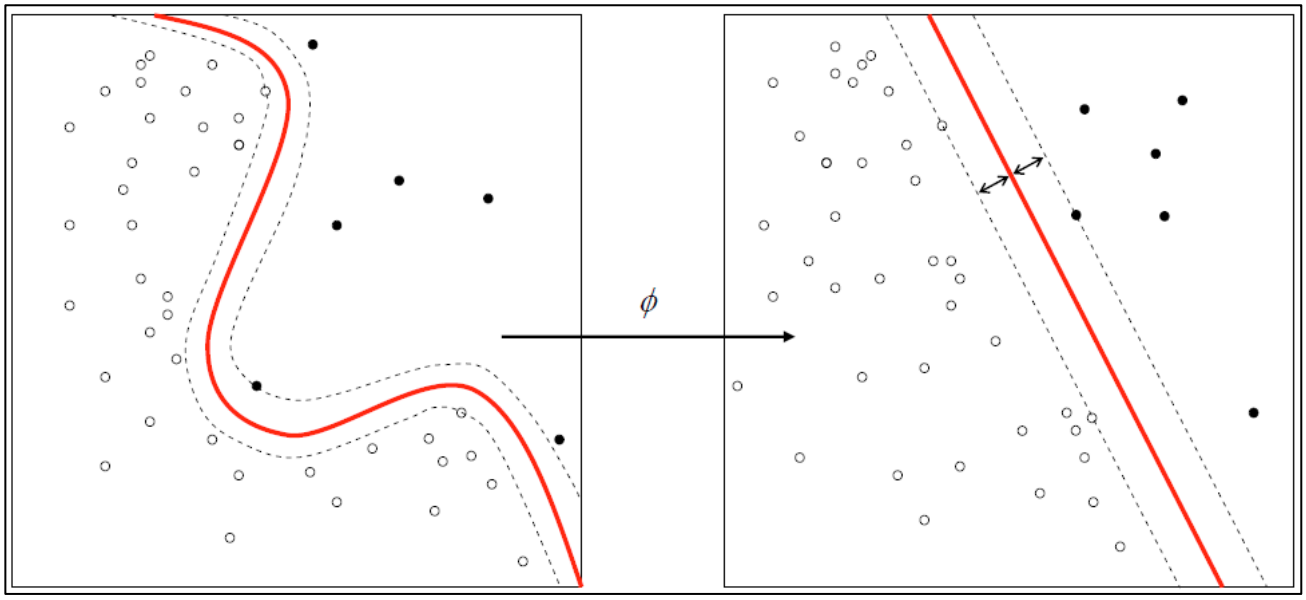


Figure 3.10: Nonlinear boundary becomes linear when kernel is applied.

Some examples of commonly used kernels in SVMs are:

Linear Kernel: $k(\boldsymbol{X}_i, \boldsymbol{X}) = \boldsymbol{X}_i^T \boldsymbol{X}$

Radial Basis Kernel: $k(\boldsymbol{X}_i, \boldsymbol{X}) = \exp-[(\|\boldsymbol{X}_i - \boldsymbol{X}\|^2)/ 2\sigma^2]$

Polynomial Kernel: $k(\boldsymbol{X}_i, \boldsymbol{X}) = (\boldsymbol{X}_i \cdot \boldsymbol{X} + a)^b$

Gaussian Kernel: $k(\boldsymbol{X}_i, \boldsymbol{X}) = \exp[-\gamma(\|\boldsymbol{X}_i - \boldsymbol{X}\|^2]$

The figure below shows how data can be remapped into higher dimensional space. On the left side of figure 3.11, we can see that the data is not linearly separable in the 2-D space. However, after applying the Radial Basis Function Kernel this data becomes linearly separable in the higher dimensional space.

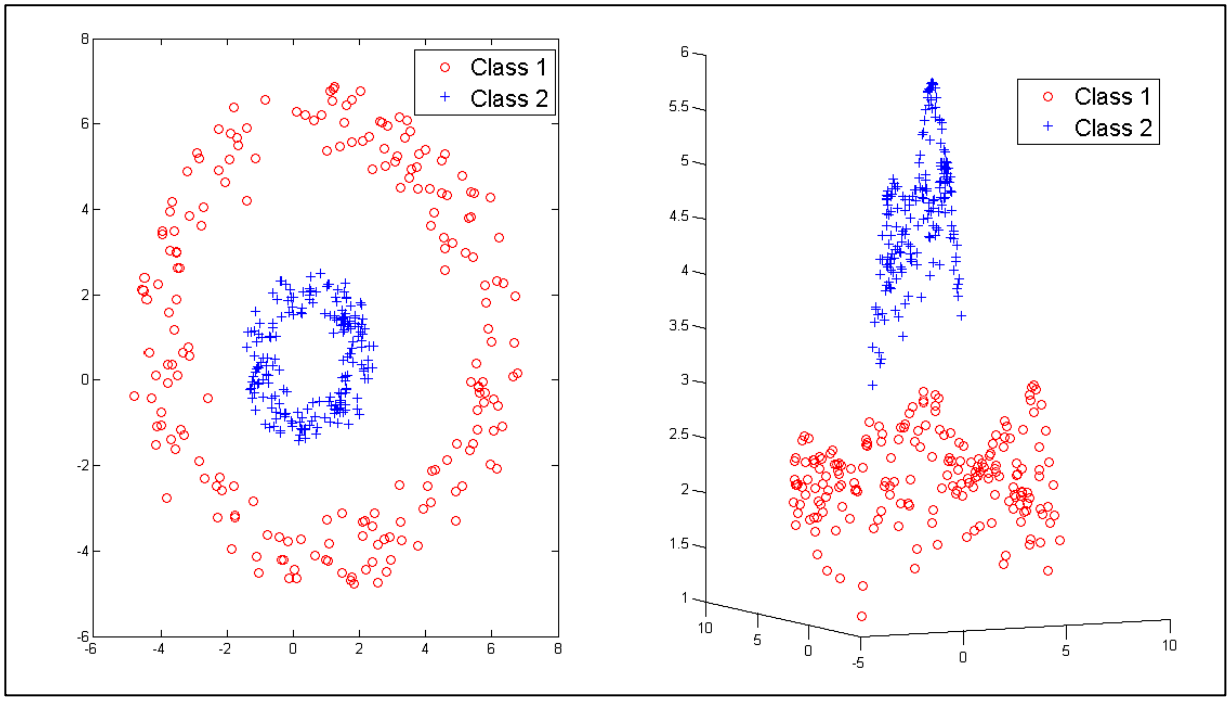


Figure 3.11: Data remapped using Radial Basis Kernel (Fletcher 2009).

In many cases we cannot completely separate the training data without running the risk of overfitting the data, and sacrificing the algorithm's generalizability to new testing data. In order to avoid overfitting, a soft-margin constant *C* is introduced. This allows the algorithm to ignore points close to the boundary when they could be masking more general trends. In figure 3.12 below, we can see two distinct groupings of data with a few outliers closer to the boundary. If we used a high value for *C* (as seen below left), then a large penalty is assigned for margin errors, and the result is very similar to a hard margin (as derived above) where no mistakes are permitted. We would have to use the closest points to the margin as the support vectors and this would mask the more general trend observed by the two rather large groupings. Essentially we are allowing the outliers to determine the margin. However, if we use a lower value for *C* (below right), then a few errors are permitted, and the classifier is not controlled by the outliers.

In the figure below, it is evident that the more general trend has been captured by the added flexibility of a lower soft-margin constant C .

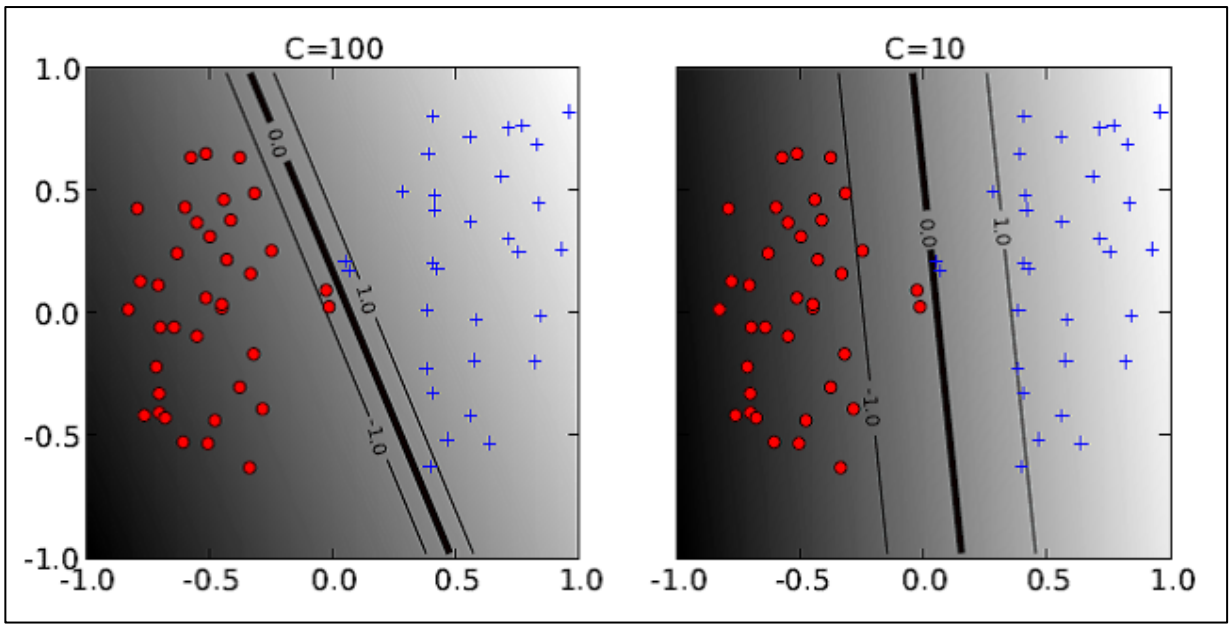


Figure 3.12: Overfitting is avoided by adjusting the soft-margin constant C (Ben-Hur 2010).

The optimization now becomes:

$$\boldsymbol{min}_{(w,b)} \, (\tfrac{1}{2})\|\boldsymbol{w}\|^2 + \boldsymbol{C} \sum_{i=1..n} \boldsymbol{\varepsilon_i}$$

$$\text{subject to} \quad \boldsymbol{Y_i} \, (\boldsymbol{w}^T \, \boldsymbol{X_i} + b) \geq 1 - \boldsymbol{\varepsilon_i}, \quad \boldsymbol{\varepsilon_i} \geq 0$$

Where $\boldsymbol{\varepsilon_i} \geq 0$ are the slack variables that allow an example to be a margin error ($0 \leq \boldsymbol{\varepsilon_i} \leq 1$), or to be misclassified ($\boldsymbol{\varepsilon_i} > 1$).

The final step is to specify the kernel parameters, if any exist in the kernel you have selected. Kernel parameters can adjust the flexibility of the classifier (Nilsson 2006). The linear

classifier has no kernel parameters and thus, no added flexibility. However, if we examine the Gaussian kernel as an example, we can see how the kernel parameter γ affects the decision boundary.

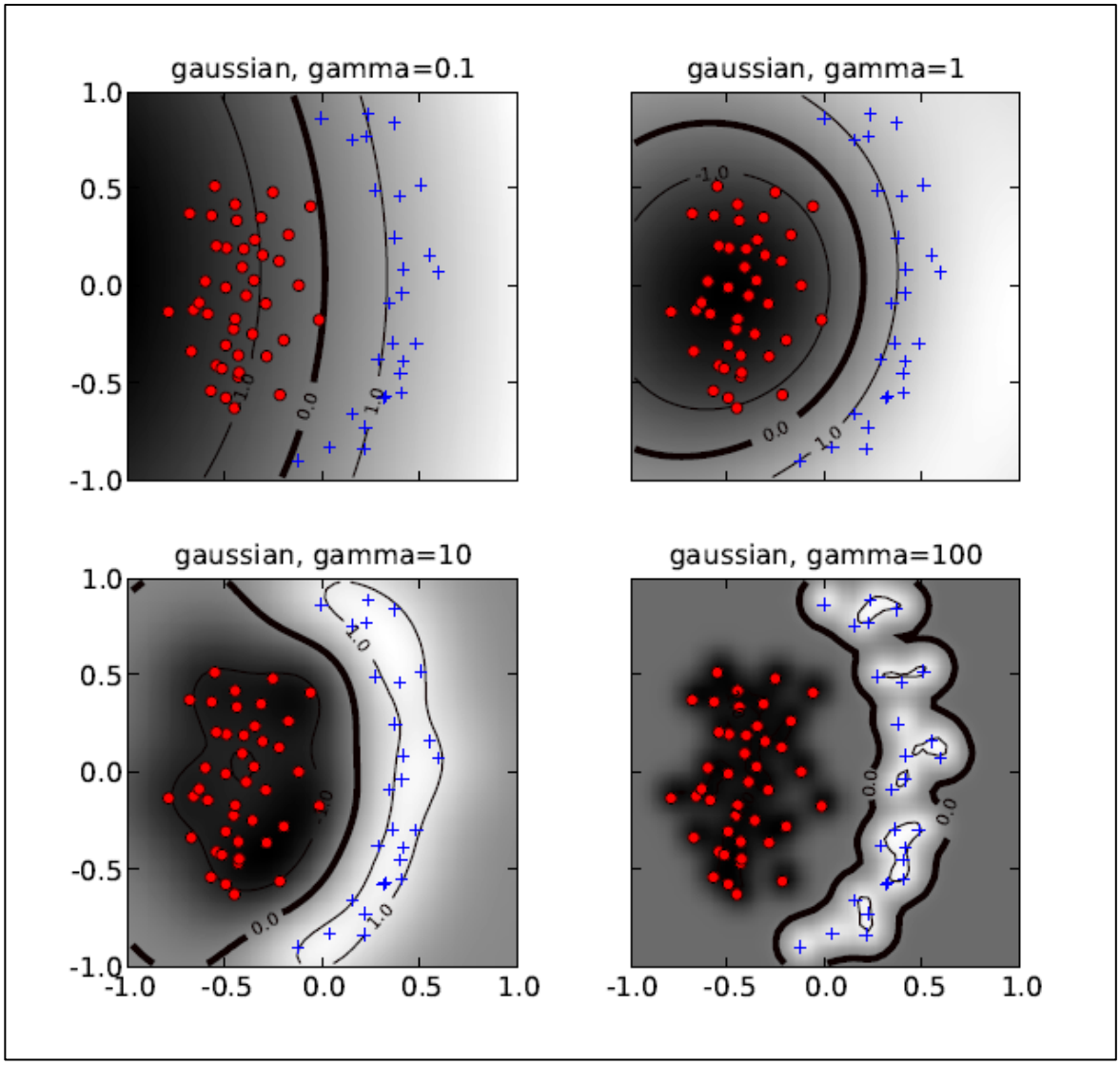


Figure 3.13: The Gaussian kernel parameter (γ) affects the flexibility of the decision boundary (Ben-Hur 2010).

Small values of γ (upper left) are less flexible and lead to nearly linear decision boundaries, while large values of γ lead to overfitting (bottom right). By selecting the soft-margin constant C and the hyperparameter γ, we can adjust the flexibility and generalizability of the classifier. Once

the margin has been determined, then new data can be evaluated and classified according to which side of the margin it falls on.

3.8 Data Preparation

The implementation of the SVM was done using Spider for MATLAB. Spider is a machine learning toolbox developed by Jason Weston, and it is highly customizable to perform many types of algorithms. The first step in using Spider was to prepare the data into an *n* x *D* matrix '*A*'. The matrix needs to be arranged with the rows being the examples and the columns being the features. Our examples were *n* voxels, and each voxel had a set of *D* MR values taken from the coregistered MR modalities which represented the features. The labels were arranged in another matrix '*B*' with *n* rows and one column assigning the class label (table 3.2).

Features [D]

Examples [n]	**A=nx5**	T1	T2	ADC	rCBV	BOLD	**B=nx1**	Label
	Voxel 1						Voxel 1	
	Voxel 2						Voxel 2	
	Voxel 3						Voxel 3	
	Voxel 4						Voxel 4	
	Voxel …						Voxel …	

Table 3.2: SVM input matrices. Matrix A is arranged as *voxels* by *features*. Matrix B is *voxels* by *labels.*

This results in a matrix with whole brain coverage. At a voxel dimension of roughly $15mm^3$(the resampling resolution), the whole brain volume consists of about 80,000 voxels. This can cause very slow processing times when training the SVM, and is unnecessary due to the overwhelming amount of healthy voxels. In most cases, the pathological tissue is constrained to a smaller portion of the brain, and can be selected as a 'quadrant' of interest. By reducing the dataset to roughly 20,000 voxels, we can include all of the diseased tissue, as well as a large amount of healthy tissue. This process is beneficial to us in terms of computational cost, but more importantly it actually improves the performance of the classifier.

In the case of large class imbalances, the SVM can show bias towards the majority class. In our case, the whole brain matrix can sometimes have on the magnitude of 100:1 imbalances between healthy tissue and disease tissue. A paper by Chawla et al. (2002) shows how under-sampling the majority class can act to alleviate the class imbalance dilemma. In essence, we are artificially balancing our data set by ignoring a large portion of healthy tissue. Of course, as can be seen in figure 3.14, we still need to include enough healthy voxels to adequately train the decision boundary. It is important to note that this is only necessary for the training portion of the SVM. Once the algorithm is trained, it can easily be applied to imbalanced or larger data sets for the testing phase. Most of our cases end up near 10:1 imbalance due to the inclusion of other disease classes. This is a result of using a multi-class approach, as we are interested in classifying many different tissue types.

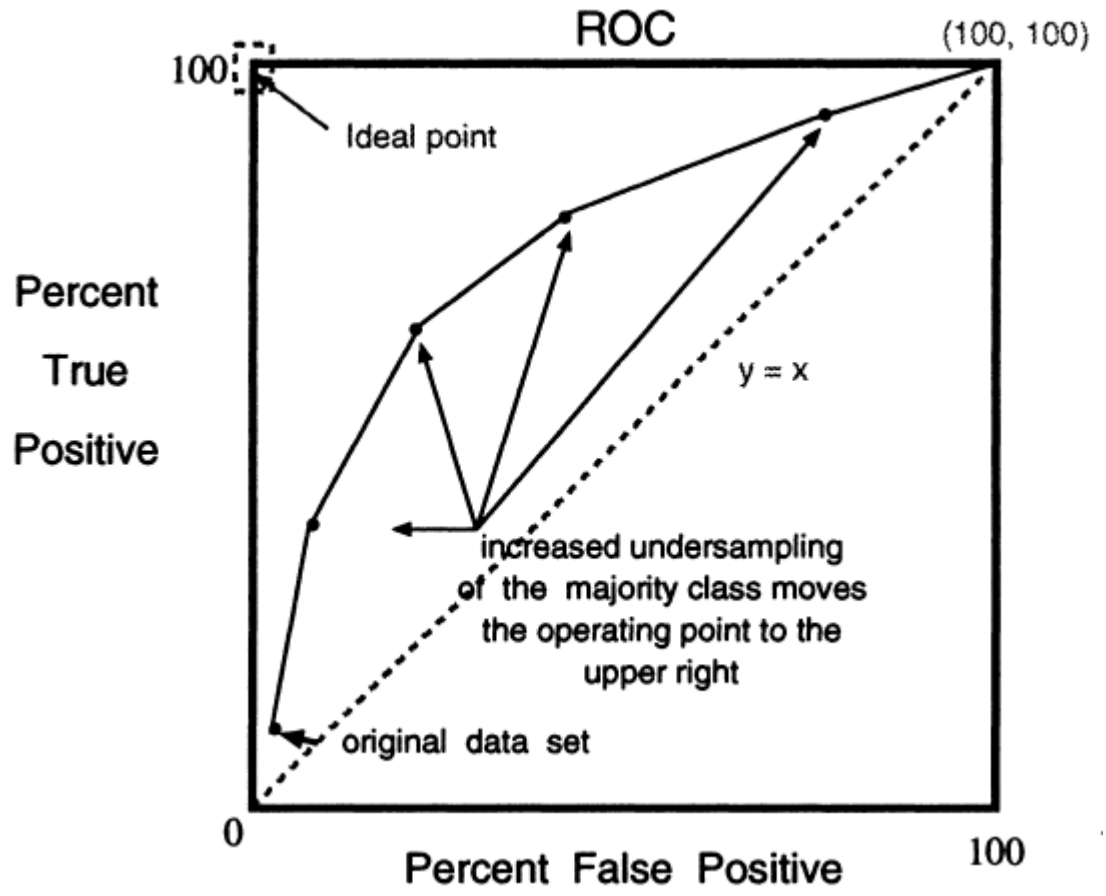


Figure 3.14: Increased under-sampling of majority class will move the performance from the lower left point to the upper right (Chawla 2002).

3.9 Normalization

SVM performance is also strongly affected by scaling differences between individual features. In order to ensure that no single feature is inappropriately favored, we first normalize the data. This is performed by transforming the data such that the mean of each feature column is equal to zero, with standard deviation 1. Normalization as a whole, is well accepted in the literature. However, it is unclear what methods of normalization perform best for MRI data. We present two normalization techniques in the results section. The first method is the standard

approach of normalizing after data reduction (post-reduction ROI). However, we found that this was possibly skewing the data due to the limited healthy voxels contributing to the distribution. We then implemented a second technique; where we first normalized the entire dataset (Whole-Brain method), and then selected our reduced data from the pre-normalized matrices.

3.10 Data Reduction using Positions Matrix

This two-step normalization is actually more complicated than it appears. Because we are dealing with imaging data, we not only wanted our testing results to be reported numerically, but we also wanted to visually asses the classifier performance. When dealing with ROI methods, this is a trivial step; because the entire ROI is either classified as negative or positive and no added spatial information is needed. However, by using voxel-wise analysis, we wanted to see the output for each individual voxel translated back to native space. This means that the output from the SVM needs spatial information as well as class information.

Therefore, the original voxel location (x,y,z) from the 3d coregistered NIFTI files must be preserved throughout the entire workflow. However, the input matrices for SPIDER require us to build each individual feature column, by arranging the voxels to be in a single vector rather than a 3d matrix. If we were selecting the entire brain, we could simply apply a transform to the native 3d matrix that converted it to a vector, and then apply the inverse transform to the output vector to preserve the spatial coordinates. However, since we select reduced datasets within the full volume, we no longer preserve the original order within the input vector.

To work around this problem, we wrote a MATLAB code that allowed us to preserve the original coordinates. This was accomplished by creating a ‘Position Matrix’ which contained the reduced voxel set as rows, and their corresponding spatial coordinates as columns.

		Position		
	P= nx3	X	Y	Z
Voxels [n]	Voxel 1			
	Voxel 2			
	Voxel 3			
	Voxel 4			
	Voxel …			

Table 3.3: Position Matrix. Matrix X is arranged as *voxels* by *position* (x,y,z).

The position matrix row numbers represent the exact same voxel and voxel position as the equivalent reduced data feature matrix. For instance, if we selected the 1st voxel in the reduced feature matrix, it would not correlate to the first voxel in the native brain feature matrix. However, it does match the position matrix precisely; and since the position matrix retains the spatial coordinates throughout the processing steps, we can then use it to transform back to the native whole-brain space. In this way, the SVM predictions can be represented in a 3d NIFTI mask, which can be overlaid on anatomical images or other label images. This is an important step because we do not want to lose the benefit of visual inspection of the classifier. As will be discussed in coming sections, the choice of evaluation metrics can lead us to interpret results in a variety of ways. However, a visual inspection is quite direct, and should serve to complement

our numerical reporting. An example of our positional matrix re-mapping can be seen in figure 3.15, which shows the results of a tumor classifier compared to the radiologist labels.

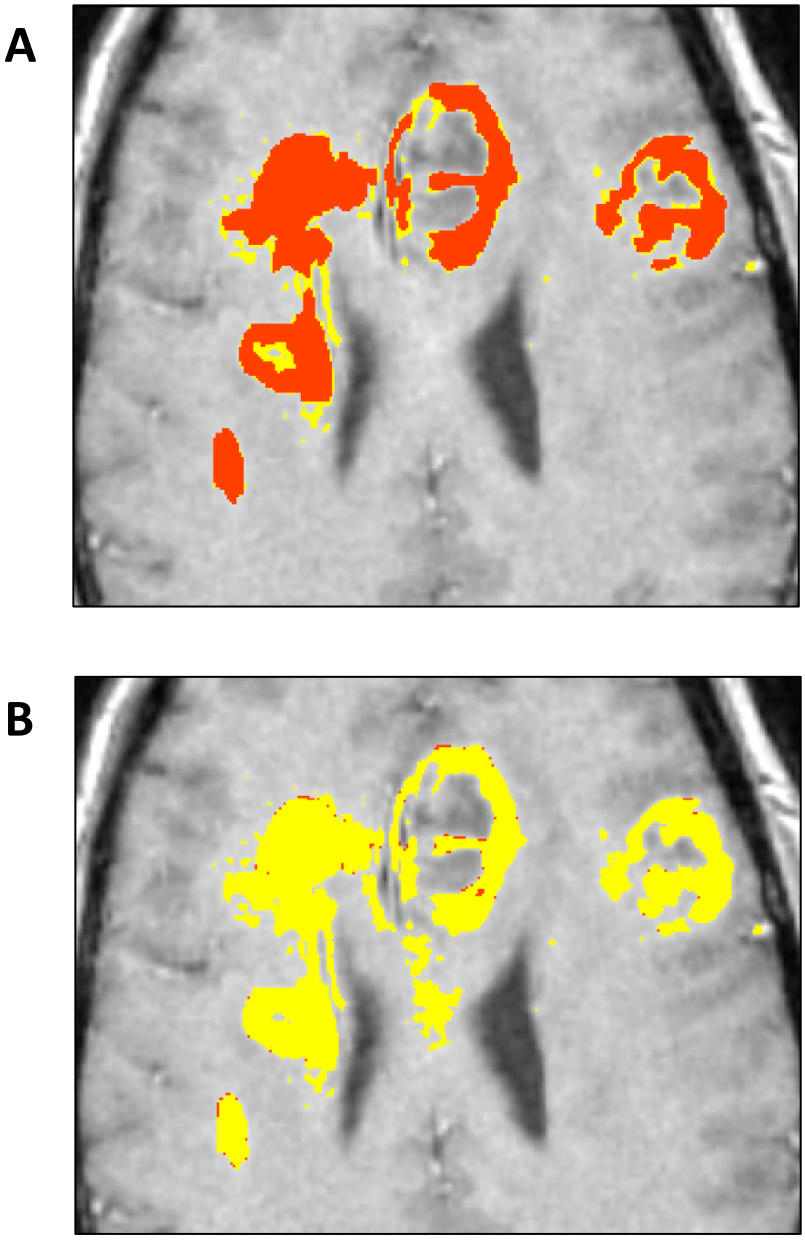


Figure 3.15: SVM prediction in yellow mapped back to native space using positional matrix technique. Red voxels represent original radiologist labels for tumor. A) Labels overlaid on SVM predictions. B) SVM predictions overlaid on Labels.

3.11 SVM Training using Cross Validation

After completion of the normalization and data reduction steps, the SVM data is ready for training. The first step was to separate the data into training and testing sets. Depending on the size of the tumor, this usually consisted of two slices for training and two slices for testing in order to investigate SVM performance on the same scan date. For between patients or between scan date SVMs, we were able to use larger testing sets, due to the fact that we were not required to use any available data for training.

The training step was performed on each patient and each tissue type separately, in order to provide a patient specific and tissue specific model. This means that for each patient, we constructed at least 5 two-class models: tumor, necrosis, edema, non-enhancing tumor, and healthy tissue. These models were constructed to perform a binary classification. For example, the tumor model strictly classifies voxels as either tumor or not tumor. We also trained a multi-class model which takes all tissue types into account. This is performed by using a one-vs.-rest SVM, which iteratively finds the best tissue class fit by comparing all five model weights for each voxel.

We also chose a few representative patients to compare alternative methods, which resulted in additional models. For these patients we compared different labeling techniques using a quickly drawn sub-set for time efficiency. These models are referred to as our 'quick' models. We trained all 6 quick models to correspond to their normal counterparts, giving us 12 models for these patients. We then performed direct comparisons between 'tumor' and 'quick tumor' models for the same patient; and repeated the process for necrosis, edema, non-enhancing tumor, healthy, and multi-class models.

We also compared normalization techniques between models. This consisted of comparing the 'post-reduction' normalization to the 'pre-reduction' normalization models. This was done by directly comparing 6 post-reduction models to their corresponding 6 pre-reduction models for the same patient.

In order to train each of the individual models mentioned above, we started by using a cross-validation technique. This is done to select our optimal hyperparameters as mentioned in the theoretical outline (C and kernel parameters). We selected the radial basis function kernel, which allows us to adjust the width of the kernel using σ.

$$\textit{Radial Basis Kernel:} \qquad k(\boldsymbol{X}_i, \boldsymbol{X}) = \exp-[(\|\boldsymbol{X}_i - \boldsymbol{X}\|^2)/\, 2\sigma^2]$$

This resulted in two hyperparameters, C and σ, that required cross-validation optimization. By adjusting the hyperparameters, we can change the flexibility and response of the model. In most cases we are able to find a good fit for the testing data by choosing the appropriate values. However, in order for the testing data to remain independent, we are not able to use it to tune our hyperparameters. Any choices about the construction of the model must be made using only the training data.

The cross-validation (CV) process involves separating the training data into multiple folds, and training the CV model on all but one fold. The left-out fold is then used to test the CV model; and the process is repeated, cycling through each fold. This is referred to as leave-one-out cross validation. By separating the training data into temporary training and testing folds, we are able to assess the initial performance of our models. As long as the training set is large enough to contain a representative distribution of feature values, then the model is able to fit the

general trends present in the independent testing data. However, problems of overfitting can arise when using the wrong metrics to tune the hyperparameters, or if the training data is too small to capture the more general trends.

Using cross-validation allows us to construct many different models, with varying hyperparameters, to select the most appropriate fit. This is referred to as a grid-search. Our grid-search consisted of constructing 36 CV models (with C values ranging from 1 to 20, and Gamma values ranging from 0.5 to 10) for each single 'final' model trained. As mentioned above, some of our patients had up to 18 final models constructed. This resulted in 648 CV models trained for each of these cases. This is very computationally demanding, and is one of the reasons we chose to investigate the reduced data models. We also had to choose a method for selecting which CV model performed the best. Due to the fact that we preferred a single-value metric to sort through our CV results, we chose the Matthew's Correlation Coefficient (MCC). Further details concerning performance metrics will be discussed in the following sections. After choosing the CV model with the highest MCC value, we then selected those C and Gamma values as our optimal hyperparameters. The full training model is then trained on the optimized hyperparameters and includes all of the training data. Once the model is complete, we are ready for testing.

3.12 Testing Methods and Performance Metrics

The testing phase of the SVM is not nearly as computationally demanding. The model is already trained and can easily be applied to various testing data. In order to asses within session performance, we tested the SVM first on single scan sessions of the same patient. We then

extended this to within patient testing, using the same patient but different scan dates. Our final test condition was to look at between patient results, using a model constructed on one patient to test data in a different patient.

One of the areas of difficulty in this study was deciding what performance metrics to use. How do we assess the quality of our SVM? In the field of machine learning, a confusion matrix is a table that describes the performance of the classifier. An example can be seen in table 3.4, showing a confusion matrix for a tumor classifier. Each tissue type and model produces a different confusion matrix. However, for the sake of consistency, we treat the tissue of interest (tumor in the case below) as the positive class, and all others voxels as the negative class.

		Predicted Class: Tumor	Predicted Class: Not Tumor
Actual Class	Tumor	True Positive (TP)	False Negative (FN)
	Not Tumor	False Positive (FP)	True Negative (TN)

Table 3.4: Confusion matrix for binary tumor classifier.

The confusion matrix provides the complete raw results for classifier performance. However, many other measures have been used to summarize specific aspects of the confusion matrix. For instance:

$$Sensitivity = \frac{TP}{(TP+FN)} (recall \text{ or } true\ positive\ rate)$$

$$Specificity = \frac{TN}{(TN+FP)} (true\ negative\ rate)$$

$$PPV = \frac{TP}{(TP+FP)} (positive\ predictive\ value \text{ or } precision)$$

$$Accuracy = \frac{TN+TN}{(P+N)} (1 - class\ loss)$$

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(P\times N\times P'\times N')}} (Matthew's\ Correlation\ Coefficient)$$

Where: $P = actual\ positives = TP + FN$

$$P' = predicted\ positives = TP + FP$$

$$N = actual\ negatives = TN + FP$$

$$N' = predicted\ negatives = TN + FN$$

Figure 3.16: Terminology and derivations from a confusion matrix

Although early machine learning research focused on accuracy, the limitation of this metric were quickly discovered (Chawla 2005). Specifically, the accuracy does nothing to inform us about how the loss is distributed across classes. The receiver operator characteristic (ROC) soon emerged as a popular choice for a performance measure (Ferri 2004). The ROC graph is a method of representing the trade-offs between benefits (true positives) and costs (false negatives). One of the common methods for comparing models is using the area under the curve (AUC) for the receiver operator characteristic (ROC).

However, ROC does not always give us a complete picture. The ROC is a function of sensitivity and 1-specificity (false positive rate). Although, these metrics are generally useful, in the presence of large class imbalances, they can be misleading. For example, we can compare model A and B below, which are taken from two different necrosis models for the same patient and same scan date (table 3.5). Which model appears superior? Judging from the ROC metrics alone, we would choose model A. However, if we visually inspect the output of these two classifiers, we see a different story (figure 3.17).

Model	**Sensitivity**	**Specificity**	**ROC**
A	0.9600	0.9281	0.9278
B	0.6800	0.9997	0.8353

Table 3.5: Comparison of two necrosis models from the same patient using only ROC metrics.

Upon visual inspection of figure 3.17 it is quite clear that model B is a more desirable classifier. Although it misses a few actual necrosis voxels leading to false negatives, it gives us a much

better representation of where the necrosis actual lies. This is due to the fact that model A has so many false positives, making it difficult to determine any consistent area of focus. So what leads to this misleading ROC result?

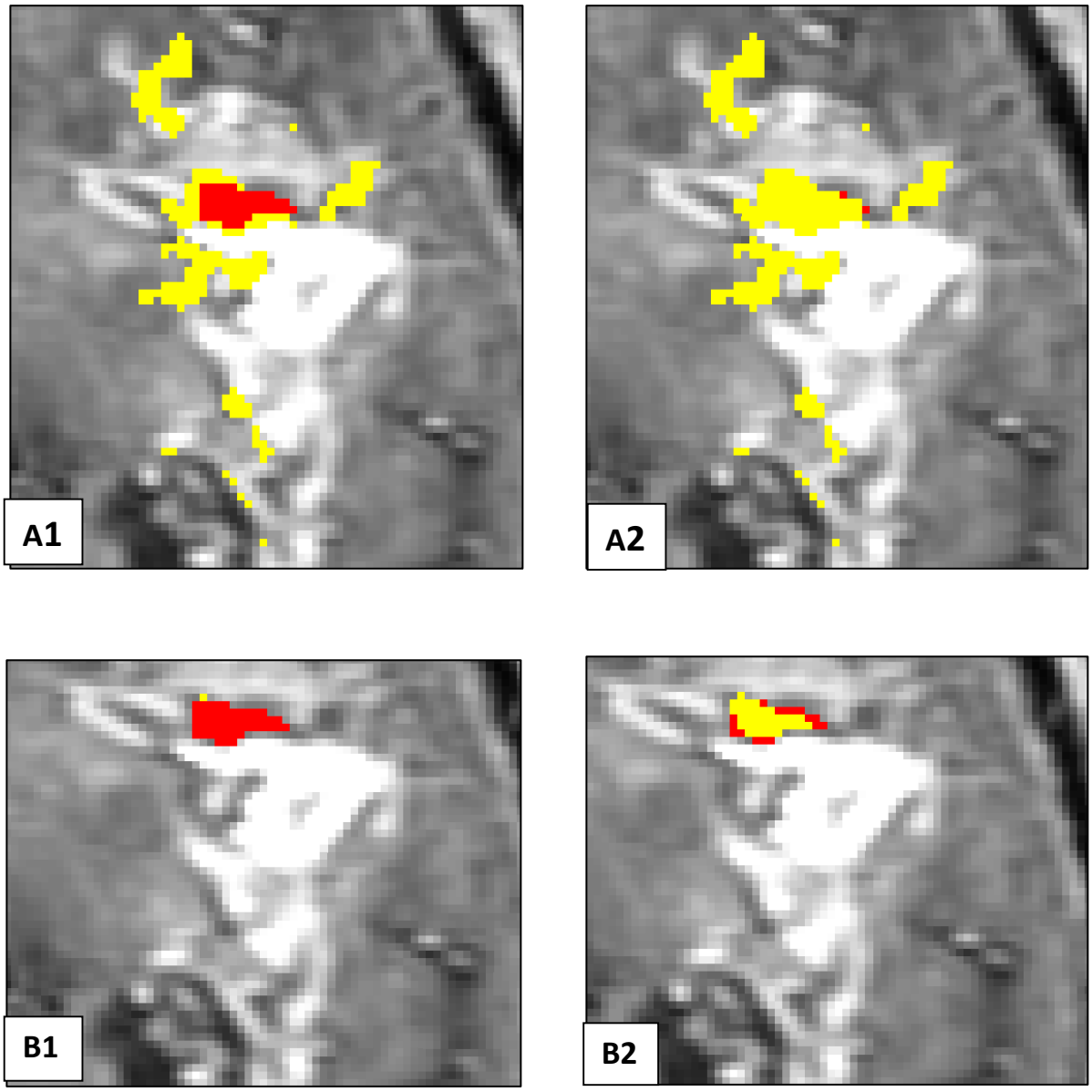


Figure 3.17: Comparison of necrosis model A and necrosis model B for same patient. Red = Radiologist Labels, yellow = SVM predictions. A1 shows Labels over SVM. A2 shows SVM over Labels. B1 shows Labels over SVM. B2 shows SVM over Labels.

In this case, the reason that the ROC fails to choose the more appropriate model, is because of the highly imbalanced classes. The actual numbers from the confusion matrix can be seen in table 3.6.

Model	TP	FN	FP	TN
A	48	2	230	2968
B	34	16	1	3197

Sensitivity	Specificity	ROC
0.9600	0.9281	0.9278
0.6800	0.9997	0.8353

Table 3.6: Confusion matrices for model A and B. Two necrosis models from the same patient.

It is clear that model A is superior to B when only considering sensitivity. However, when it comes to specificity, model B is superior. The problem arises because the ROC statistic weighs the sensitivity and specificity equally. However, in this case, specificity is not as responsive due to the highly imbalanced classes.

$$Specificity = \frac{TN}{(TN + FP)} \qquad When\ TN \gg FP, \quad specificity \approx 1$$

Due to the overwhelming number of true negatives, the false positives are not counted as heavily. In essence, there is a large difference in the false positive errors between the two models that is not appropriately represented in the ROC statistic. So the question becomes, how do we account for false positives?

As noted in figure 3.17, positive predictive value (PPV) accounts for false positives. Not only does it account for false positives, but it also relates them to true positives rather than true

negatives. In this way, the highly imbalanced negative class (TN) does not enter the equation. PPV can be thought of as: out of all the predicted positives, what percentage were actual positives. As we can see from figure 3.17, model A predicted a lot of necrosis voxels that were not actual necrosis, leading to a low PPV.

PPV and sensitivity have also been used in other areas of machine learning research that contain large class imbalances. In the information retrieval domain a classifier can take a document and identify occurrences of useful words. Due to the large amount of non-relevant words usually present, these classifiers also suffer from large class imbalances (Dumais 1998). In the case of information retrieval, PPV and sensitivity are the accepted metrics (although they are referred to as precision and recall). Multiparametric MRI datasets also follow tend to have large class imbalances. It is most often the case that the tissue of interest (necrosis in this case), is only a small portion of the total voxels. It is for this reason that we chose to focus on PPV and sensitivity as our performance metrics of choice. Table 3.7 clearly shows the inferiority of PPV for model A. The relatively large difference in false positives between model A (230 FP) and B (1 FP) is clearly accounted for in the PPV metric.

Model	PPV	Sensitivity
A	0.173	0.9600
B	0.971	0.6800

ROC
0.9278
0.8353

Table 3.7: Comparing Positive Predictive Value (PPV) and Sensitivity vs. ROC for two necrosis models from the same patient.

So why is ROC used so often? This is because ROC measures are independent of the prevalence (relative amount of actual positives to actual negatives) of the disease within a single model. This means that for a single model A, we can expect the ROC measure to be close to 0.93 even if we test it on a different proportion of actual necrosis vs. healthy tissue voxels in the same patient. The ROC is a characteristic of the rate of TP and TN independent of the prevalence within the chosen sample. However, this is assuming we have a single model that we are testing on different samples from the same population. This is not the case in our study. We are comparing multiple models that change their performance as they are applied to different populations of voxels (new patients). And since the training of the models is dependent on prevalence, we see this reflected in the ROC differences between models.

PPV is however, influenced by the prevalence of the disease in the population being tested (Mausner 1985). Therefore, if we test in a setting where necrosis voxels are a high percentage of the total voxels, then it is more likely that a voxel that tests positive actually reflects a true positive necrosis voxel. The opposite is also true. In a low prevalence setting, the PPV will decline for the same exact test (see figure 3.18). However, if the prevalence is not changing, then this is not a concern. For this reason, we use PPV as a metric to tune our grid search parameters. Since we are comparing different models as applied to the same datasets, the prevalence is not changing (only the models are changing). This is the same reason why we do not prefer, ROC during grid search; because the models themselves are changing. To summarize, if we want to compare a single model applied to different patients, then ROC could be used (although we prefer MCC). However, if we want to compare multiple models applied to the same patient, then PPV is preferred.

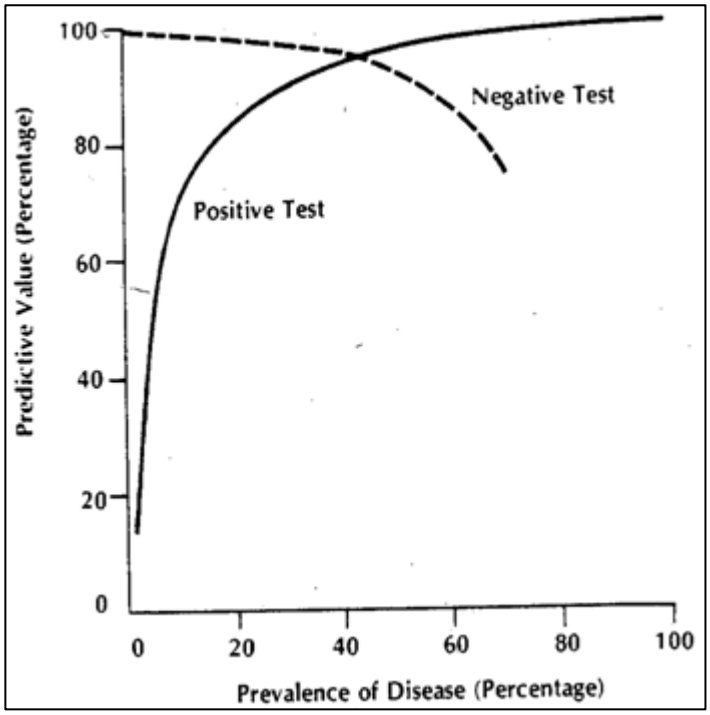


Figure 3.18: Relationship between disease prevalence and predictive value in a test with 95% sensitivity and 85% specificity (Mausner 1985).

Although PPV and sensitivity are quite beneficial in dealing with our class imbalances, we needed a single metric to use for grid search optimization. When choosing between 36 grid search models, it is necessary to rank them in order to select the optimal hyperparameters. The F-score was chosen because it is a weighted average of the PPV and sensitivity.

$$F_\beta = (1 + \beta^2) \times \frac{PPV \times Sensitivity}{(\beta^2 \times PPV) + Sensitivity}$$

Where β^2 can be adjusted to change the relative weighting of PPV and Sensitivity. The F_1-score weights both equally. The F_2-score weights sensitivity higher than PPV. And the $F_{0.5}$-score

weights PPV higher than sensitivity. By using the F-score we are able to adjust the weighting of the selection process if needed.

Although the F-score is a beneficial metric for our grid selection process, there has been a move towards using the Matthews Correlation Coefficient (MCC) as a way to assess the performance of a classifier. It takes into account TP, FP, TN, FN and is regarded as a useful metric for imbalanced data. It is also not as sensitive to changes in prevalence as PPV. Therefore it is a very robust measure that is useful for comparing multiple models. It is used as a reference measure for fields such as bioinformatics and network topologies (Jurman 2012). It was recently chosen as the accuracy index in the FDA-led initiative MAQC-II for comparing 13,000 microarrays and genotyping models. MCC also has the added flexibility of being a useful metric for multi-class classifiers. One of the benefits is that it includes all four values from the confusion matrix, yet still remains relatively unaffected by class imbalances. Due to the fact that it is gaining general acceptance as a classifier measure, and because it survives the class imbalances present in our data, we also include the MCC values as a means of comparison. Although they are not used for the grid search selection, because this is an isolated process with static prevalence values, allowing for our own preferred F-score to be tune appropriately. However, when reporting the final testing results, we include the MCC as a way to translate our results into commonly used metrics among various disciplines.

References

1. Ben-Hur A, Weston J. A user's guide to support vector machines. Methods MolBiol 2010;609:223-239.

2. Boxerman JL, Schmainda KM, Weisskoff RM. Relative cerebral blood volume maps corrected for contrast agent extravasation. AJNR Am J Neuroradiol 2006;27(4):859-867.

3. Carr H. Effects of Diffusion on Free Precession in Nuclear Magnetic Resonance Experiments. Physical Review 1954;94(3):630-638.

4. Chan JH, Tsui EY, Chau LF, et al. Discrimination of an infected brain tumor from a cerebral abscess by combined MR perfusion and diffusion imaging.Comput Med Imaging Graph 2002;26(1):19-23.

5. Chawla N. Data mining for imbalanced datasets: An overview. US: Springer: 2005.

6. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 2002;16(1):321-357.

7. Collignon A, Maes F, Delaere, Vandermeulen, Suetens P, Marchal G. Automated multi-modality image registration based on information theory. Information Processing in Medical Imaging (1995), pp 263-274 2007.

8. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20(3):273-297.

9. Dumais S, Platt J, Sahami M, Heckerman D. Inductive Learning Algorithms and Representations for Text Categorization. Proceedings of the Seventh International Conference on Information and Knowledge Management; 1998. p. 148-155.

10. Dunn JF, O'Hara JA, Zaim-Wadghiri Y, et al. Changes in oxygenation of intracranial tumors with carbogen: a BOLD MRI and EPR. J MagnReson Imaging 2002;16(5):511-521.

11. Ferri C, Flach P, Orallo J, Lachice N. First Workshop on ROC Analysis in AI.ECAI; 2004.

12. Fletcher T. Support Vector Machines Explained. 2009.

13. Friston KJ, Frith CD, Frackowiak RS, Turner R. Characterizing dynamic brain responses with fMRI: a multivariate approach. Neuroimage 1995;2(2):166-172.

14. Greve DN, Fischl B. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 2009;48(1):63-72.

15. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion. Neuroimage 2002;17(2):825-841.

16. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. Neuroimage 2012;62(2):782-790.

17. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal 2001;5(2):143-156.

18. Jurman G, Riccadonna S, Furlanello C. A comparison of MCC and CEN error measures in multi-class prediction. PLoS One 2012;7(8):e41882.

19. Mausner J, Kramer S. Mausner and BahnEpidemiolgy: An Introductory Text. Philadelphia: WB Saunders: 1985.

20. McMillan KM, Rogers BP, Field AS, Laird AR, Fine JP, Meyerand ME. Physiologic characterisation of glioblastoma multiforme using MRI-based hypoxia. J ClinNeurosci 2006;13(8):811-817.

21. Moseley M, Cohen Y, Kucharczyk J, et al. Diffusion-weighted MR imaging of anisotropic water diffusion in cat central nervous system. 1990.

22. Nilsson R, Bjorkegren J, Tegner J. A Flexible Implementation for Support Vector Machines. The Mathematica Journal 2006;10(1):14.

23. Rosen BR, Belliveau JW, Vevea JM, Brady TJ. Perfusion imaging with NMR contrast agents. MagnReson Med 1990;14(2):249-265.

24. Schaefer PW, Grant PE, Gonzalez RG. Diffusion-weighted MR imaging of the brain. Radiology 2000;217(2):331-345.

25. Stejskal EO, Tanner JE. Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient. Journal of Chemical Physics 1965;42(1):5.

26. Sugahara T, Korogi Y, Kochi M, et al. Usefulness of diffusion-weighted MRI with echo-planar technique in the evaluation of cellularity in gliomas. J MagnReson Imaging 1999;9(1):53-60.

27. Thevenaz P, Blu T, Unser M. Interpolation Revisited. IEEE Transactions on Medical Imaging 2000;19(7):20.

28. D.E. W. Effects of Diffusion in Nuclear Magnetic Resonance Spin-Echo Experiments. The Journal of Chemical Physics 1961;34(6):5.

4. Results

4.1 Training

The first step was to choose the optimal hyperparameters for model training. Grid search results were sorted according to their F2-score. The highest F2-score parameters were chosen for full model training. F2 was used due to the fact that it weights sensitivity slightly higher than PPV. In the overwhelming majority of cases, whether we used F1 or F2 did not change which grid search model was selected. The few cases where F1 and F2 disagreed were mostly limited to times when the SVM struggled to perform. In these instances, it was often the case that sensitivity struggled more than PPV. For this reason, when F1 and F2 disagreed, F2 was chosen as the optimizing metric. An example from a non-contrast enhancing (NCE) tumor model is shown below (table 4.1). The red section shows model A and B training results, and the green section shows their corresponding testing results on new data. Model A produces better results, based on F2 selection. F2 is used for future optimizations.

	Gamma	C	MCC	Sensitivity	PPV	Specificity	F1	F2
A	1	10	0.7703	0.8167	0.8305	0.9491	0.8235	0.8194
B	0.5	20	0.7757	0.8133	0.8414	0.9532	0.8271	0.8188

	Gamma	C	MCC	Sensitivity	PPV	Specificity	TP	FN	FP	TN
A	1	10	0.5297	0.6929	0.517	0.9048	273	121	255	2423
B	0.5	20	0.4303	0.599	0.4419	0.8887	236	158	298	2380

Table 4.1: Red) Best grid search model results comparing F1 and F2. If using F1, model B is selected. If using F2, model A is selected. Green) Results from testing data, using model A vs. B.

An example of the cross-validation grid search results are shown in table 4.2. After selecting the optimal hyperparameters from the grid search, the full model is trained.

Patient A.1 Tumor

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
0.5	3	0.8683	0.9384	0.8883	0.9685	0.9127	0.8979
0.5	2	0.8709	0.9417	0.8883	0.9703	0.9142	0.8985
10	3	0.8921	0.9708	0.8867	0.9856	0.9268	0.9023
1	1	0.8839	0.9489	0.8983	0.9739	0.9229	0.9080
1	2	0.9021	0.9471	0.925	0.9721	0.9359	0.9293
1	3	0.9009	0.9425	0.9283	0.9694	0.9353	0.9311
1	6	0.9112	0.9523	0.9317	0.9748	0.9419	0.9357
6	2	0.9227	0.9671	0.9317	0.9829	0.9491	0.9386
3	1	0.9253	0.9705	0.9317	0.9847	0.9507	0.9392
1	20	0.9202	0.9606	0.935	0.9793	0.9476	0.9400
2	1	0.9266	0.9689	0.935	0.9838	0.9516	0.9416
1	10	0.924	0.964	0.9367	0.9811	0.9502	0.9420
10	6	0.9266	0.9625	0.9417	0.9802	0.9520	0.9458
2	2	0.9344	0.966	0.9483	0.982	0.9571	0.9518
2	3	0.9319	0.9612	0.95	0.9793	0.9556	0.9522
6	3	0.9408	0.9727	0.95	0.9856	0.9612	0.9545
3	2	0.946	0.9745	0.955	0.9865	0.9647	0.9588
2	10	0.9537	0.9845	0.955	0.9919	0.9695	0.9608
2	6	0.9511	0.9779	0.9583	0.9883	0.9680	0.9622
3	10	0.9576	0.9863	0.9583	0.9928	0.9721	0.9638
10	10	0.955	0.9813	0.96	0.9901	0.9705	0.9642
2	20	0.9589	0.9846	0.9617	0.9919	0.9730	0.9662
3	20	0.9641	0.9914	0.9617	0.9955	0.9763	0.9675
3	6	0.9589	0.9765	0.97	0.9874	0.9732	0.9713
3	3	0.9577	0.9717	0.9733	0.9847	0.9725	0.9730
6	6	0.9705	0.98	0.9817	0.9892	0.9808	0.9814
6	10	0.9782	0.9883	0.9833	0.9937	0.9858	0.9843
10	20	0.9807	0.9899	0.985	0.9946	0.9874	0.9860
6	20	0.9808	0.9835	0.9917	0.991	0.9876	0.9900

Table 4.2: Grid search sorting by F2-score for training the tumor model using cross-validation. Optimal hyperparameters in red (Gamma=6, and C=20) were then used to train full tumor model.

After training the full model using Gamma=6 and C=20, the full model is tested on separate testing data from the same patient and same session.

4.2 Testing

Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
tumor	6	20	0.973	0.990	0.973	0.995

Table 4.3: Testing results for full model – same patient, same session.

TP	FN	FP	TN
949	26	10	2087

Table 4.4: Confusion matrix for full model – same patient, same session.

ADC	T1	T2	hypoxia	rCBV
0.14	1.00	0.15	0.00	0.14

Table 4.5: Relative weighting of each feature for full model – same patient, same session.

The 3 previous tables summarize the results for a single tissue, single patient. Table 4.3 highlights the MCC value as our preferred statistic. However, we also include ROC measures

(sensitivity and specificity) for use in comparing with outside work. Table 4.4 shows the confusion matrix, which enables us to retroactively calculate many other metrics of interest if so desired. It is important to remember that a MCC value of 0.5 is not comparable to a sensitivity or specificity of 0.5 (AUROC=0.5 is equivalent to chance guessing). MCC on the other hand has a range from -1 to +1. 0 represents the level equivalent to chance guessing for a classifier.

Table 4.5 shows the weighted contribution of each individual feature in constructing the full model. In this case the, the T1 scan provides the most information, while the hypoxia scan contributes no useful information. A summary of these tables will be presented in each of the following case studies.

The following four cases were selected for presentation in order to investigate the differences between model construction, and how these affect testing performance in our study. The first case, Patient A, will investigate the effect of the data normalization procedure. The two normalization techniques presented in Chapter 4 consist of normalizing after data selection (ROI method), or pre-normalizing the entire brain before constructing the reduced data input matrices (Whole-Brain method).

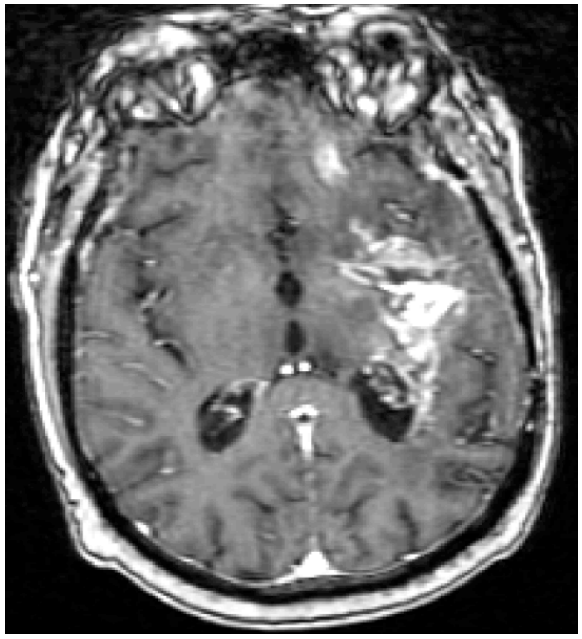
The second case (Patient B), compares the differences between training model size. The first model is the Full-brain model, where the majority of the slices are used for training. However, a few slices are kept separate for testing. The next method is the Single-slice method. This consists of using a single-slice of accurately labeled data for the training set. The final model is constructed using the Quick method. This is done by selecting a small subset of representative voxels for each tissue type. Whereas the first two methods consist of a voxel-by-voxel approach to include every edge and detail possible, the third method only selects a few

voxels from the centers of tissue ROIs. The vast majority of labeling time is spent carefully defining the tissue boundaries. Therefore, we were interested in seeing if the radiologist's labeling time could be reduced significantly by selecting rough approximations of the ROIs, rather than detailed voxel-painting.

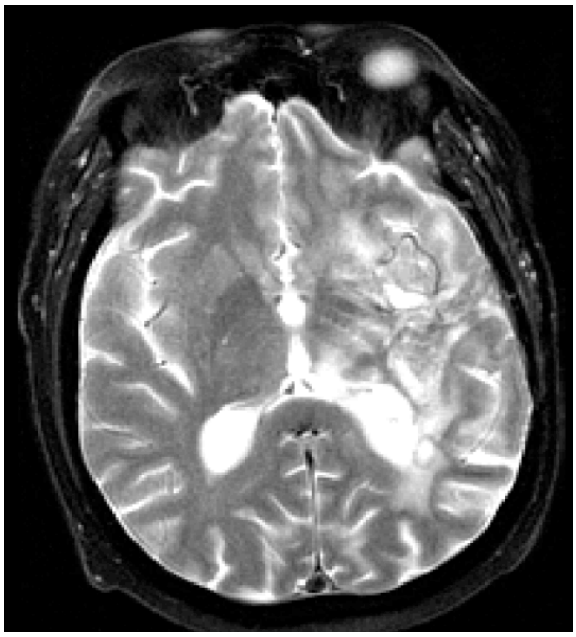
The third case compares the use of a two-class SVM vs. a multi-class SVM. The two-class model is the standard binary classification model described in Chapter 4. For this process, we address only one tissue type at a time. For example, a tumor-based SVM model classifies each voxel as either 'tumor' or 'not tumor'. This process is then repeated for each tissue type. In this case, a voxel could be classified as 'tumor' by the tumor model, 'edema' by the edema model, and 'necrosis' by the necrosis model. Although only one classification would agree with the radiologist labels. The multi-class SVM differs, in that it considers each tissue type before assigning the best fit classification. In this way, it will not assign multiple predictions to a single voxel. The multi-class model cycles through each tissue type, and weighs each single tissue voxel prediction against the rest of the tissues. This approach is called a one-versus-rest multi-class SVM.

The fourth case serves to investigate the differences within patients and between patients in model performance. The first method used is Intrasession, which is the standard method used for our basic analysis. This consists of training the model on a portion of the data, and then testing the model on a separate portion of the data from within the same session. The second method, Intersession, consisted of training on one session and testing on a new session, while still using the same patient. The final method was Interpatient, where we trained on one patient, and tested on a new patient.

Patient A3 Male, 50 years, Post-craniotomy, Post radiotherapy



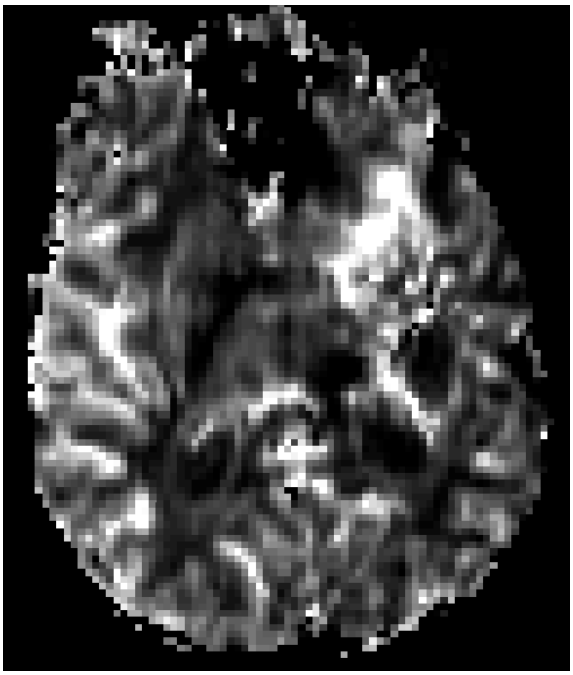
GD-T1



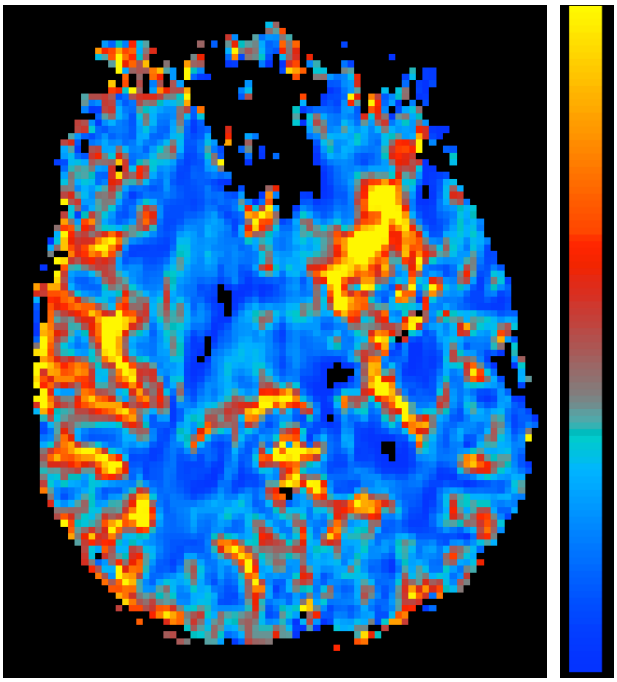
T2

Δ T2*

ADC



rCBV-grayscale



rCBV-color

Figure A3.1: Multiparametric feature maps for patient A, session 3 (A3).

Session	Date relative to First Session
A1	0 months
A2	3 months
A3	6.5 months

Table A3.1: Session list for patient A.

Tumor Training Grid-Search A3

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
30	3	0.519	0.996	0.328	1.000	0.494	0.379
10	0.5	0.666	0.998	0.513	1.000	0.678	0.568
30	6	0.753	0.994	0.638	0.999	0.777	0.687
10	1	0.831	0.993	0.751	0.998	0.855	0.790
30	10	0.861	0.990	0.800	0.998	0.885	0.832
0.5	0.5	0.880	0.963	0.853	0.990	0.905	0.873
6	0.5	0.883	0.989	0.833	0.997	0.905	0.860
0.5	30	0.898	0.950	0.893	0.985	0.921	0.904
0.5	10	0.898	0.950	0.895	0.985	0.922	0.905
0.5	6	0.899	0.950	0.896	0.985	0.922	0.906
0.5	1	0.904	0.954	0.898	0.986	0.926	0.909
0.5	3	0.904	0.953	0.900	0.986	0.926	0.910
6	1	0.926	0.984	0.902	0.996	0.942	0.918
10	3	0.934	0.986	0.914	0.996	0.949	0.928
30	30	0.938	0.983	0.922	0.995	0.952	0.934
1	0.5	0.939	0.958	0.949	0.987	0.954	0.951
1	30	0.943	0.964	0.949	0.989	0.957	0.952
1	10	0.947	0.969	0.949	0.991	0.959	0.953
1	6	0.947	0.971	0.949	0.991	0.960	0.953
3	0.5	0.948	0.981	0.940	0.994	0.960	0.948
10	6	0.948	0.981	0.940	0.994	0.960	0.948
3	1	0.948	0.970	0.952	0.991	0.961	0.955
6	3	0.951	0.972	0.953	0.991	0.963	0.957
1	3	0.951	0.967	0.958	0.990	0.963	0.960
3	3	0.951	0.956	0.970	0.986	0.963	0.967
1	1	0.952	0.961	0.966	0.988	0.964	0.965
10	10	0.954	0.967	0.962	0.990	0.965	0.963
3	10	0.955	0.965	0.966	0.989	0.965	0.966
3	6	0.956	0.964	0.969	0.989	0.966	0.968
6	6	0.957	0.963	0.973	0.988	0.968	0.971
3	30	0.958	0.970	0.966	0.991	0.968	0.967
10	30	0.959	0.963	0.975	0.988	0.969	0.973
6	10	0.961	0.962	0.979	0.988	0.970	0.976

Table A3.2: Grid search sorting by F2-score for patient A3-tumor model. Optimal hyperparameters in red (Gamma=6, and C=10) were then used to train the full tumor model.

Tumor Results Whole-brain vs. ROI Normalization A3

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
ROI-Norm	tumor	6	10	0.721	0.997	0.619	0.999
Brain-Norm	tumor	6	10	0.965	0.960	0.992	0.980

TP	FN	FP	TN
650	401	2	2195
1043	8	43	2154

ADC	T1	T2	Hypoxia	rCBV
0.03	1.00	0.02	0.00	0.06
0.22	1.00	0.25	0.21	0.00

Table A3.3: Tumor Results - Performance metrics, confusion matrix, model weighting.

Both reduced ROI and Whole-Brain normalization methods performed quite well, with MCCs above 0.70 (Table A3.3). However, the sensitivity of the ROI method is substantially lower than the Whole-Brain approach. This is evident from the large amount of false negatives missed in the ROI method. The feature weighting is more widely dispersed in the Whole-brain method. While the ROI method only focuses on T1 values.

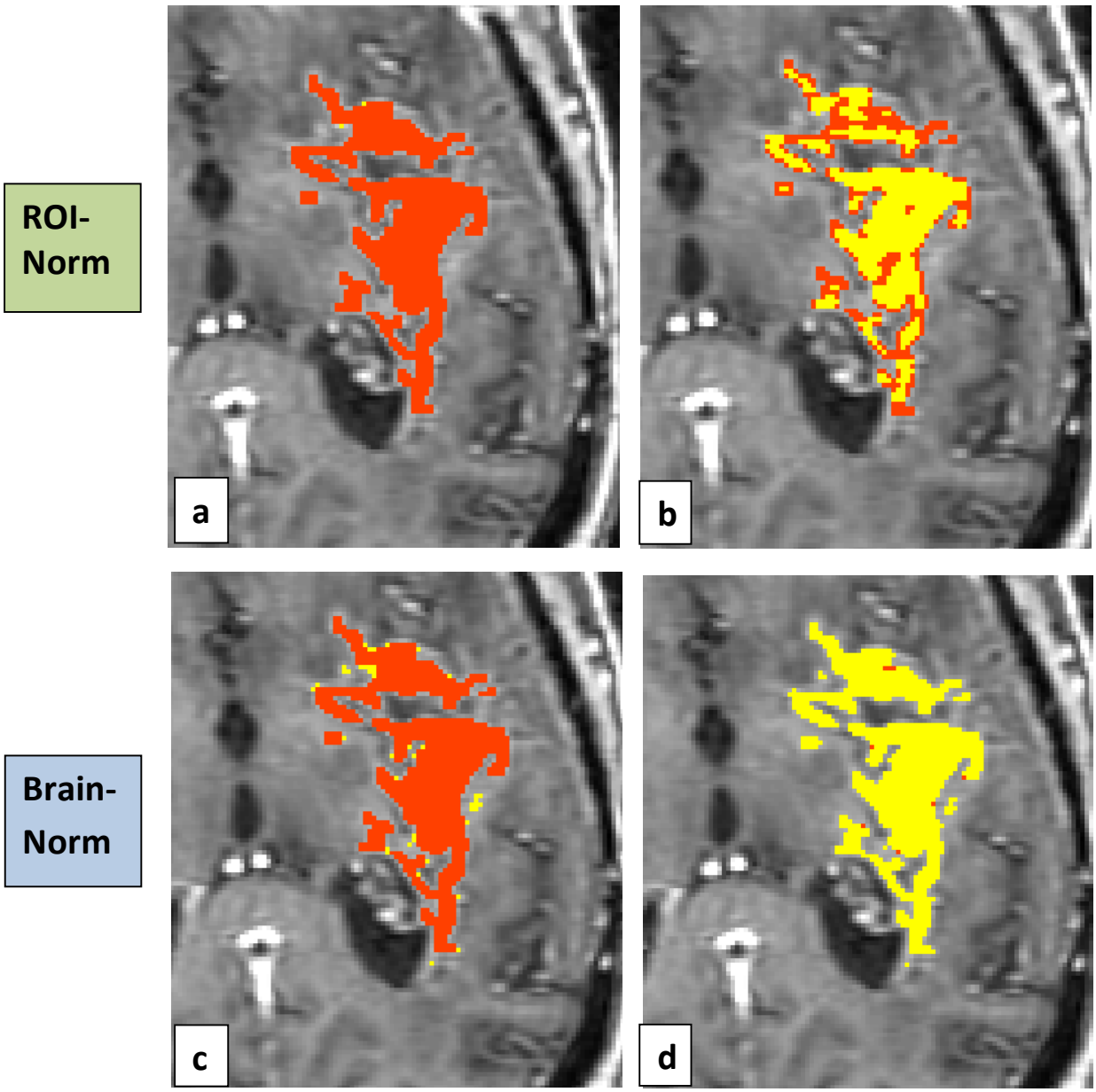


Figure A3.2: Comparing results of Whole-brain vs. ROI Normalization tumor models on patient A3. Red = Labels, yellow = SVM. a) ROI-Norm - Labels overlaid on SVM. b) ROI-Norm - SVM overlaid on Labels. c) Brain-Norm – Labels over SVM. d) Brain-Norm - SVM over Labels.

Necrosis Training Grid-Search A3

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
30	6	0.113	1.000	0.013	1.000	0.026	0.017
10	1	0.394	1.000	0.160	1.000	0.276	0.192
30	10	0.497	1.000	0.253	1.000	0.404	0.298
6	0.5	0.559	1.000	0.320	1.000	0.485	0.370
0.5	0.5	0.571	1.000	0.333	1.000	0.500	0.385
0.5	1	0.647	0.944	0.453	0.999	0.613	0.506
6	1	0.696	0.974	0.507	1.000	0.667	0.561
10	3	0.696	0.974	0.507	1.000	0.667	0.561
3	0.5	0.715	0.976	0.533	1.000	0.690	0.586
30	30	0.715	0.976	0.533	1.000	0.690	0.586
10	6	0.742	0.977	0.573	1.000	0.723	0.625
0.5	3	0.711	0.880	0.587	0.997	0.704	0.629
0.5	10	0.698	0.833	0.600	0.996	0.698	0.636
0.5	30	0.698	0.833	0.600	0.996	0.698	0.636
0.5	6	0.713	0.865	0.600	0.997	0.709	0.639
6	3	0.752	0.939	0.613	0.999	0.742	0.659
3	1	0.770	0.941	0.640	0.999	0.762	0.684
10	10	0.770	0.941	0.640	0.999	0.762	0.684
6	6	0.771	0.925	0.653	0.998	0.766	0.694
1	0.5	0.779	0.942	0.653	0.999	0.772	0.696
10	30	0.746	0.836	0.680	0.995	0.750	0.706
6	10	0.742	0.813	0.693	0.994	0.748	0.714
6	30	0.762	0.809	0.733	0.994	0.769	0.747
3	3	0.781	0.846	0.733	0.995	0.786	0.753
3	6	0.785	0.826	0.760	0.994	0.792	0.772
3	30	0.771	0.784	0.773	0.992	0.779	0.775
3	10	0.788	0.817	0.773	0.994	0.795	0.782
1	30	0.812	0.866	0.773	0.996	0.817	0.790
1	3	0.814	0.855	0.787	0.995	0.819	0.799
1	1	0.833	0.894	0.787	0.997	0.837	0.806
1	6	0.824	0.847	0.813	0.995	0.830	0.820
1	10	0.830	0.859	0.813	0.995	0.836	0.822

Table A3.4: Grid search sorting by F2-score for patient A3-necrosis model. Optimal hyperparameters in red (Gamma=1, and C=10) were then used to train the full necrosis model.

Necrosis Results Whole-brain vs. ROI Normalization A3

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
ROI-Norm	necrosis	3	10	0.733	0.696	0.780	0.995
Brain-Norm	necrosis	1	10	0.837	0.927	0.760	0.999

TP	FN	FP	TN
39	11	17	3181
38	12	3	3195

ADC	T1	T2	Hypoxia	rCBV
1.00	0.27	0.08	0.79	0.00
0.92	0.89	1.00	0.01	0.00

Table A3.5: Necrosis Results - Performance metrics, confusion matrix, model weighting.

Again, both methods perform reasonably well classifying necrosis, but the Whole-brain method was superior. The major discrepancy is in the PPV. The ROI method produced more false positives, which led to a lower PPV. The outliers can be seen in yellow in figure A3.3a. Although, there are a few outliers, the general trend and location of the necrosis voxels is consistent.

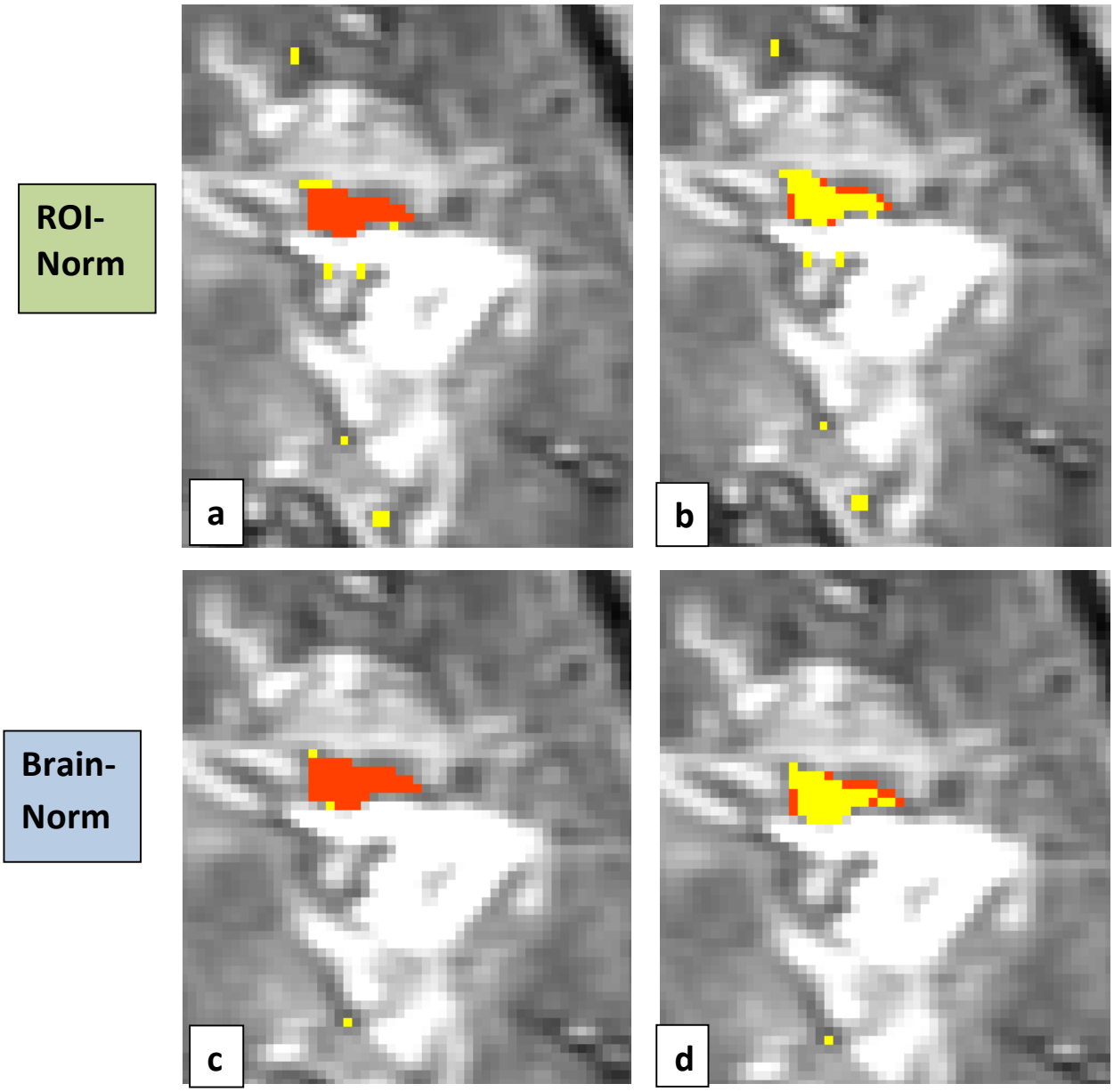


Figure A3.3: Comparing results of Whole-brain vs. ROI Normalization necrosis models on patient A3. Red = Labels, yellow = SVM. a) ROI-Norm - Labels overlaid on SVM. b) ROI-Norm - SVM overlaid on Labels. c) Brain-Norm – Labels over SVM. d) Brain-Norm - SVM over Labels.

Edema Training Grid-Search A3

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
10	1	0.367	0.707	0.379	0.920	0.494	0.418
10	2	0.428	0.689	0.510	0.882	0.586	0.538
10	3	0.463	0.699	0.556	0.878	0.619	0.579
6	1	0.504	0.721	0.596	0.882	0.652	0.617
10	6	0.502	0.715	0.600	0.878	0.653	0.620
10	10	0.518	0.719	0.622	0.876	0.667	0.639
6	2	0.534	0.719	0.650	0.870	0.683	0.663
10	20	0.538	0.720	0.656	0.870	0.686	0.667
6	3	0.544	0.725	0.658	0.872	0.690	0.671
6	6	0.572	0.740	0.683	0.877	0.710	0.694
6	10	0.581	0.743	0.693	0.877	0.717	0.702
3	1	0.593	0.759	0.690	0.888	0.723	0.703
6	20	0.597	0.758	0.699	0.886	0.727	0.710
3	2	0.596	0.756	0.700	0.884	0.727	0.710
3	3	0.602	0.754	0.711	0.882	0.732	0.719
2	1	0.612	0.762	0.717	0.886	0.739	0.725
3	6	0.619	0.770	0.717	0.891	0.742	0.727
2	2	0.638	0.778	0.736	0.893	0.757	0.744
3	10	0.639	0.776	0.740	0.891	0.758	0.747
2	3	0.640	0.777	0.740	0.892	0.758	0.747
3	20	0.648	0.774	0.757	0.887	0.765	0.760
2	6	0.661	0.783	0.765	0.892	0.774	0.769
1	1	0.678	0.802	0.767	0.904	0.784	0.774
2	10	0.669	0.781	0.781	0.888	0.781	0.781
2	20	0.679	0.779	0.799	0.884	0.789	0.795
1	2	0.696	0.801	0.796	0.899	0.799	0.797
1	3	0.697	0.795	0.806	0.894	0.800	0.803
0.5	6	0.720	0.822	0.806	0.911	0.813	0.809
0.5	1	0.725	0.830	0.804	0.916	0.817	0.809
0.5	10	0.723	0.825	0.806	0.913	0.815	0.809
0.5	20	0.722	0.821	0.810	0.910	0.815	0.812
0.5	3	0.737	0.839	0.810	0.921	0.824	0.815
1	6	0.708	0.795	0.822	0.892	0.808	0.817
1	10	0.717	0.801	0.828	0.895	0.814	0.822
0.5	2	0.737	0.831	0.819	0.915	0.825	0.822

Table A3.6: Grid search sorting by F2-score for patient A3-edema model. Optimal hyperparameters in red (Gamma=0.5, and C=2) were then used to train the full edema model.

Edema Results Whole-brain vs. ROI Normalization A3

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
ROI-Norm	Edema	0.5	1	0.536	0.633	0.705	0.848
Brain-Norm	Edema	0.5	2	0.527	0.672	0.631	0.886

TP	FN	FP	TN
620	260	359	2009
555	325	271	2097

ADC	T1	T2	Hypoxia	rCBV
0.55	1.00	0.66	0.00	0.58
0.76	1.00	0.88	0.11	0.00

Table A3.7: Edema Results - Performance metrics, confusion matrix, model weighting.

In the edema model, the ROI method outperformed the Whole-brain method by a narrow margin. Although the PPV is higher in the Whole-brain, the sensitivity is lower, leading to a lower overall MCC value. While the ROI method gained information from the rCBV feature, the Whole-brain method did not. This is our first example of an MCC value near 0.5. Upon inspection of figure A3.4, it is quite clear that this classifier is performing well above chance. Although there is noise in the predictions, the overall form is visually recognizable.

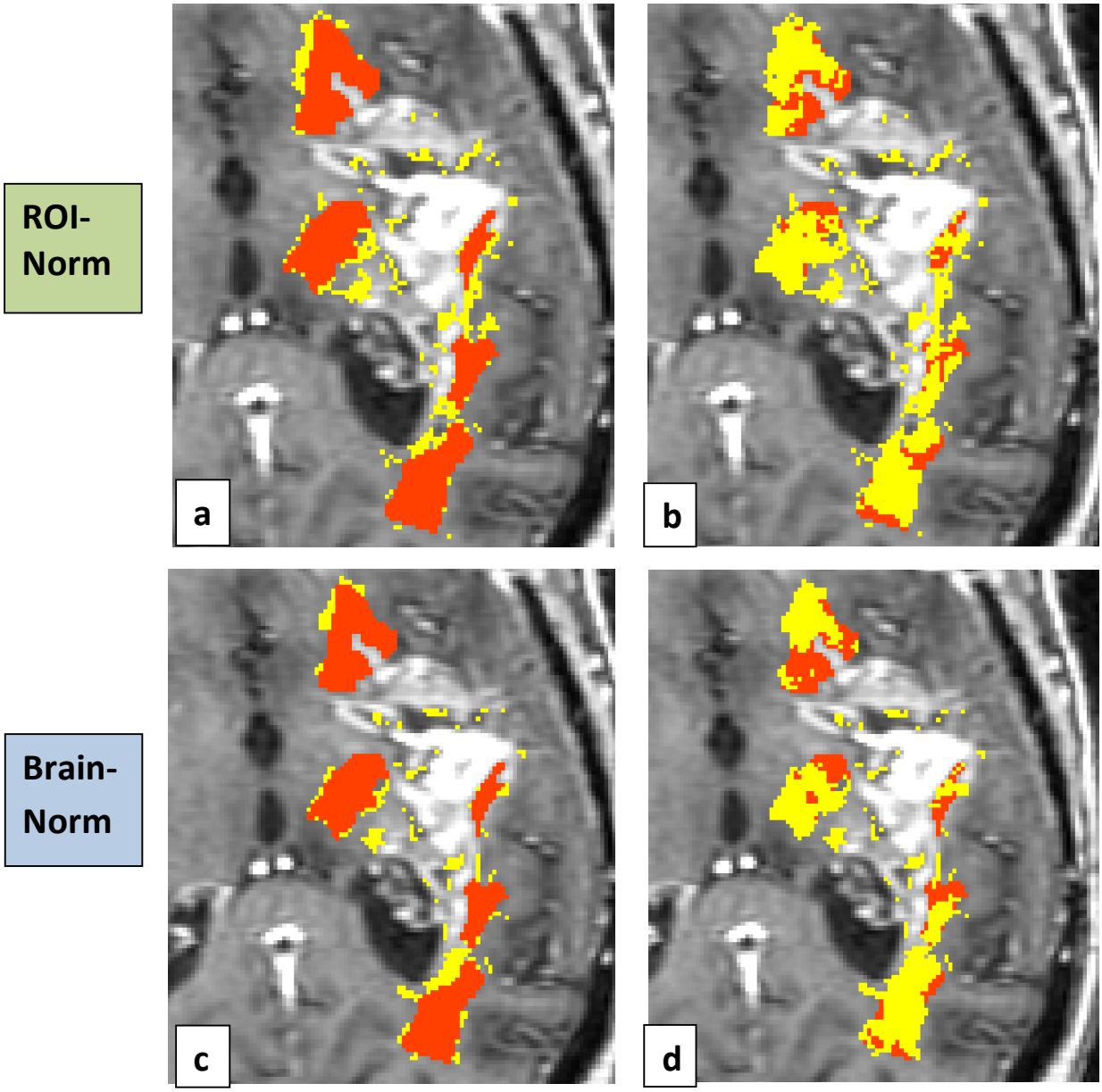


Figure A3.4: Comparing results of Whole-brain vs. ROI Normalization edema models on patient A3. Red = Labels, yellow = SVM. a) ROI-Norm - Labels overlaid on SVM. b) ROI-Norm - SVM overlaid on Labels. c) Brain-Norm – Labels over SVM. d) Brain-Norm - SVM over Labels.

NCE Training Grid-Search A3

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
6	2	0.028	0.500	0.003	0.999	0.006	0.004
6	3	0.194	0.864	0.055	0.998	0.103	0.067
10	20	0.206	0.875	0.061	0.998	0.113	0.074
6	6	0.299	0.776	0.150	0.992	0.251	0.179
6	10	0.368	0.772	0.225	0.987	0.348	0.262
3	1	0.384	0.789	0.236	0.988	0.364	0.275
6	20	0.398	0.781	0.257	0.986	0.386	0.296
3	2	0.438	0.764	0.317	0.981	0.448	0.359
3	3	0.451	0.745	0.346	0.977	0.472	0.387
2	1	0.475	0.749	0.378	0.975	0.502	0.419
3	6	0.482	0.743	0.392	0.974	0.513	0.433
2	2	0.489	0.738	0.406	0.972	0.524	0.446
2	3	0.489	0.727	0.415	0.970	0.528	0.454
3	10	0.495	0.739	0.415	0.971	0.531	0.455
3	20	0.516	0.736	0.450	0.969	0.558	0.488
2	6	0.571	0.752	0.525	0.966	0.618	0.558
1	1	0.604	0.755	0.576	0.964	0.654	0.605
2	10	0.600	0.744	0.579	0.961	0.652	0.606
2	20	0.609	0.734	0.605	0.957	0.664	0.627
1	2	0.618	0.740	0.614	0.958	0.671	0.635
1	3	0.655	0.754	0.663	0.958	0.706	0.679
0.5	1	0.700	0.805	0.689	0.968	0.742	0.709
1	6	0.678	0.752	0.706	0.955	0.728	0.715
1	10	0.687	0.752	0.723	0.954	0.737	0.729
0.5	10	0.704	0.774	0.729	0.959	0.751	0.738
0.5	6	0.705	0.772	0.732	0.958	0.751	0.740
1	20	0.701	0.764	0.735	0.956	0.749	0.740
0.5	3	0.719	0.792	0.735	0.962	0.762	0.746
0.5	20	0.710	0.772	0.741	0.957	0.756	0.747
0.5	2	0.723	0.799	0.735	0.964	0.766	0.747

Table A3.8: Grid search sorting by F2-score for patient A3-NCE model. Optimal hyperparameters in red (Gamma=0.5, and C=2) were then used to train the full NCE model.

NCE Results **Whole-brain vs. ROI Normalization** **A3**

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
ROI-Norm	NCE	0.5	2	0.239	0.334	0.348	0.895
Brain-Norm	NCE	0.5	2	0.423	0.548	0.439	0.945

TP	FN	FP	TN
149	279	297	2523
188	240	155	2665

ADC	T1	T2	Hypoxia	rCBV
0.01	1.00	0.20	0.00	0.14
0.65	1.00	0.66	0.55	0.00

Table A3.9: NCE Results - Performance metrics, confusion matrix, model weighting.

Both models performed poorly on the NCE tissue classification. The Whole-brain model performed significantly better at distinguishing true negatives, leading to a higher PPV and MCC.

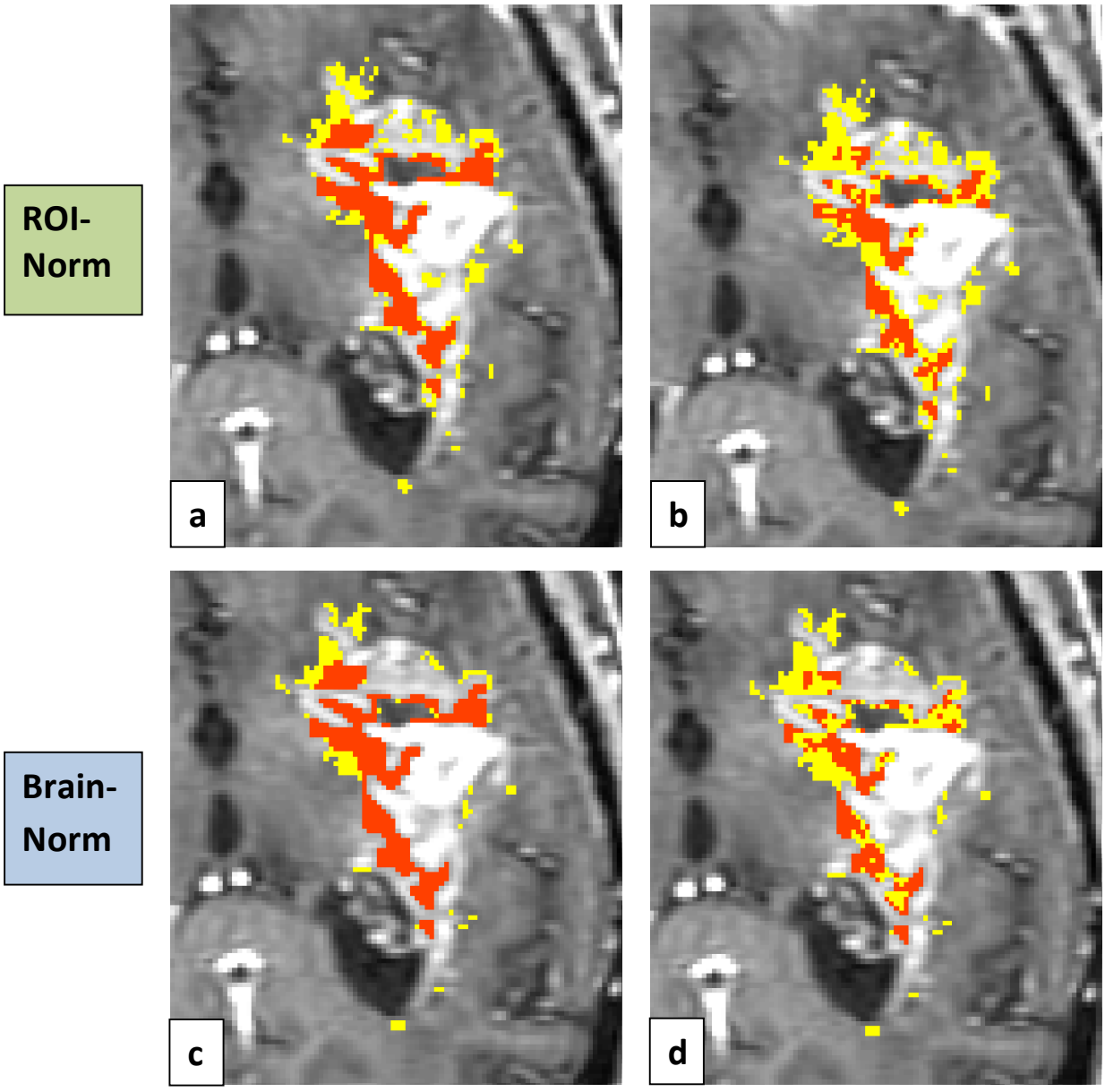


Figure A3.5: Comparing results of Whole-brain vs. ROI Normalization NCE models on patient A3. Red = Labels, yellow = SVM. a) ROI-Norm - Labels overlaid on SVM. b) ROI-Norm - SVM overlaid on Labels. c) Brain-Norm – Labels over SVM. d) Brain-Norm - SVM over Labels.

Summary of MCCs

	ROI-Norm	Brain-Norm
tumor	0.721	0.965
necrosis	0.733	0.837
edema	0.536	0.527
NCE	0.239	0.423

Table A3.10: Comparing MCCs of tissue classes using Reduced ROI Normalization vs. Whole-brain normalization.

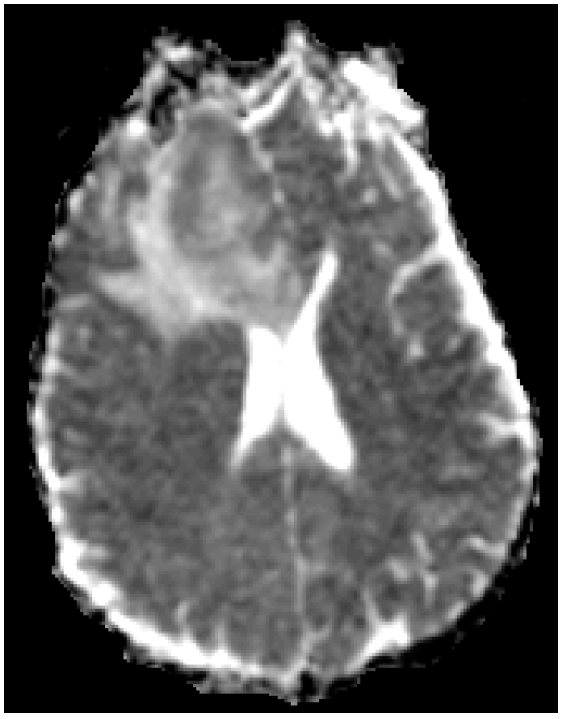
The overall impression of the two methods is clear from the data. Although in the case of edema, the ROI method performed slightly better than the Whole-brain approach, the majority of the tissues saw a recognizable improvement in the Whole-brain normalization method. This method is applied to the remaining datasets.

Patient B1 – Male, 61 years, Post XRT, Chemo, Surgical Resection

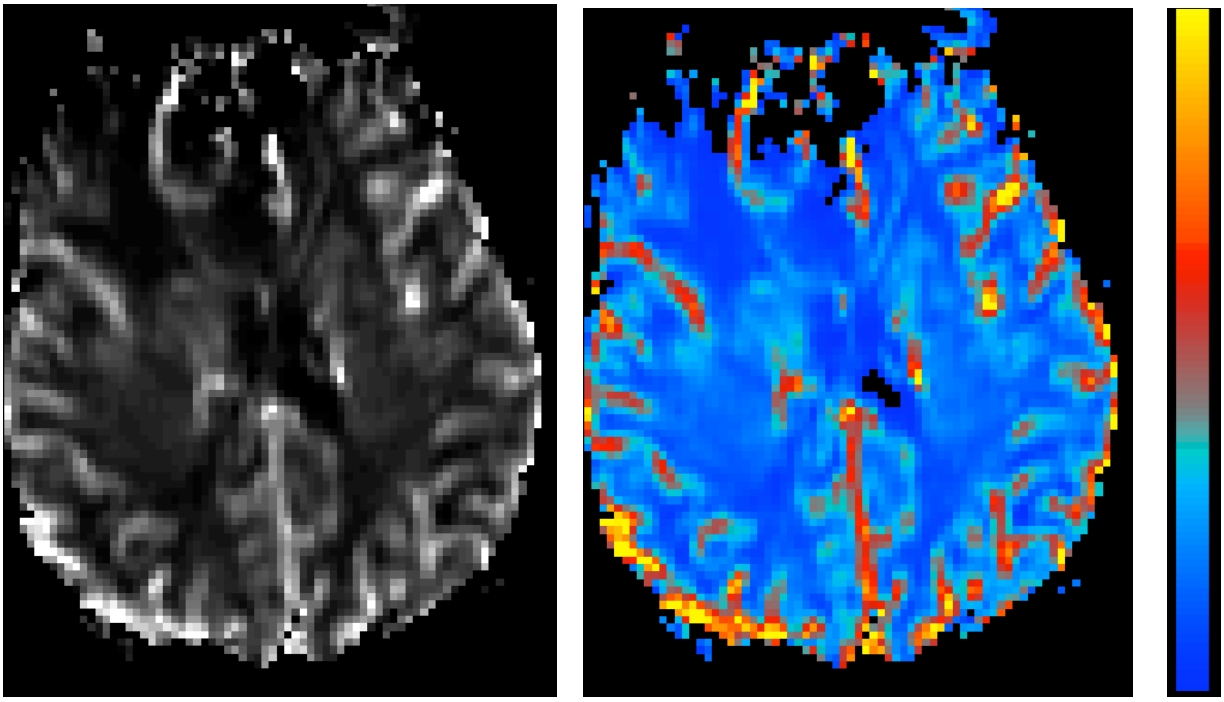
GD-T1

T2

Δ T2*



ADC



rCBV-greyscale rCBV-color

Figure B1.1: Multiparametric feature maps for patient B, session 1 (B1).

Session	Date relative to First Session
B1	0 months
Single session only	

Table B1.1: Session list for patient B.

Tumor Training Grid-Search B1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
0.5	3	0.962	0.981	0.962	0.994	0.9712	0.9656
0.5	6	0.964	0.981	0.965	0.994	0.9728	0.9682
0.5	10	0.964	0.981	0.965	0.994	0.9728	0.9682
0.5	20	0.964	0.981	0.965	0.994	0.9728	0.9682
8	1	0.975	0.984	0.978	0.995	0.9809	0.9790
1	1	0.971	0.975	0.981	0.992	0.9779	0.9798
1	10	0.975	0.981	0.981	0.994	0.9810	0.9810
1	20	0.975	0.981	0.981	0.994	0.9810	0.9810
1	6	0.979	0.987	0.981	0.996	0.9841	0.9822
6	1	0.971	0.969	0.987	0.990	0.9780	0.9835
8	2	0.971	0.969	0.987	0.990	0.9780	0.9835
1	2	0.979	0.984	0.984	0.995	0.9841	0.9841
8	3	0.967	0.960	0.991	0.987	0.9750	0.9842
1	3	0.981	0.987	0.984	0.996	0.9857	0.9847
3	1	0.969	0.963	0.991	0.988	0.9766	0.9849
6	2	0.969	0.963	0.991	0.988	0.9766	0.9849
2	1	0.971	0.966	0.991	0.989	0.9780	0.9855
3	2	0.973	0.969	0.991	0.990	0.9796	0.9861
6	3	0.971	0.963	0.994	0.988	0.9782	0.9874
8	6	0.971	0.963	0.994	0.988	0.9782	0.9874
8	10	0.971	0.963	0.994	0.988	0.9782	0.9874
2	2	0.979	0.978	0.991	0.993	0.9843	0.9880
2	3	0.979	0.978	0.991	0.993	0.9843	0.9880
3	3	0.979	0.978	0.991	0.993	0.9843	0.9880
3	10	0.979	0.978	0.991	0.993	0.9843	0.9880
6	6	0.973	0.966	0.994	0.989	0.9797	0.9880
2	6	0.981	0.981	0.991	0.994	0.9858	0.9886
3	6	0.981	0.981	0.991	0.994	0.9858	0.9886
2	10	0.983	0.984	0.991	0.995	0.9873	0.9892
2	20	0.983	0.984	0.991	0.995	0.9873	0.9892
6	10	0.979	0.975	0.994	0.992	0.9843	0.9899
8	20	0.979	0.975	0.994	0.992	0.9843	0.9899
3	20	0.985	0.984	0.994	0.995	0.9890	0.9918
6	20	0.985	0.984	0.994	0.995	0.9890	0.9918

Table B1.2: Grid search sorting by F2-score for patient B1-tumor model. Optimal hyperparameters in red (Gamma=6, and C=20) were then used to train the tumor models.

Tumor Results Quick, Single-Slice, and Full Models B1

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Quick	tumor	6	20	0.913	0.948	0.920	0.984
Slice	tumor	6	20	0.982	0.978	0.994	0.993
Full	tumor	6	20	0.992	0.997	0.991	0.999

TP	FN	FP	TN
619	54	34	2112
669	4	15	2131
667	6	2	2144

ADC	T1	T2	Hypoxia	rCBV
0.00	1.00	0.01	0.01	0.06
0.14	1.00	0.10	0.12	0.00
0.12	1.00	0.12	0.10	0.00

Table B1.3: Tumor Results - Performance metrics, confusion matrix, model weighting.

The tumor classifier performed quite well regardless of which level of detail was used for the training data. The Quick method lagged behind the Single-slice and Full-brain models only slightly. This is perhaps due to its sole focus on T1 in the feature weighting.

Tumor Results **Quick, Single-Slice, and Full Models** **B1**

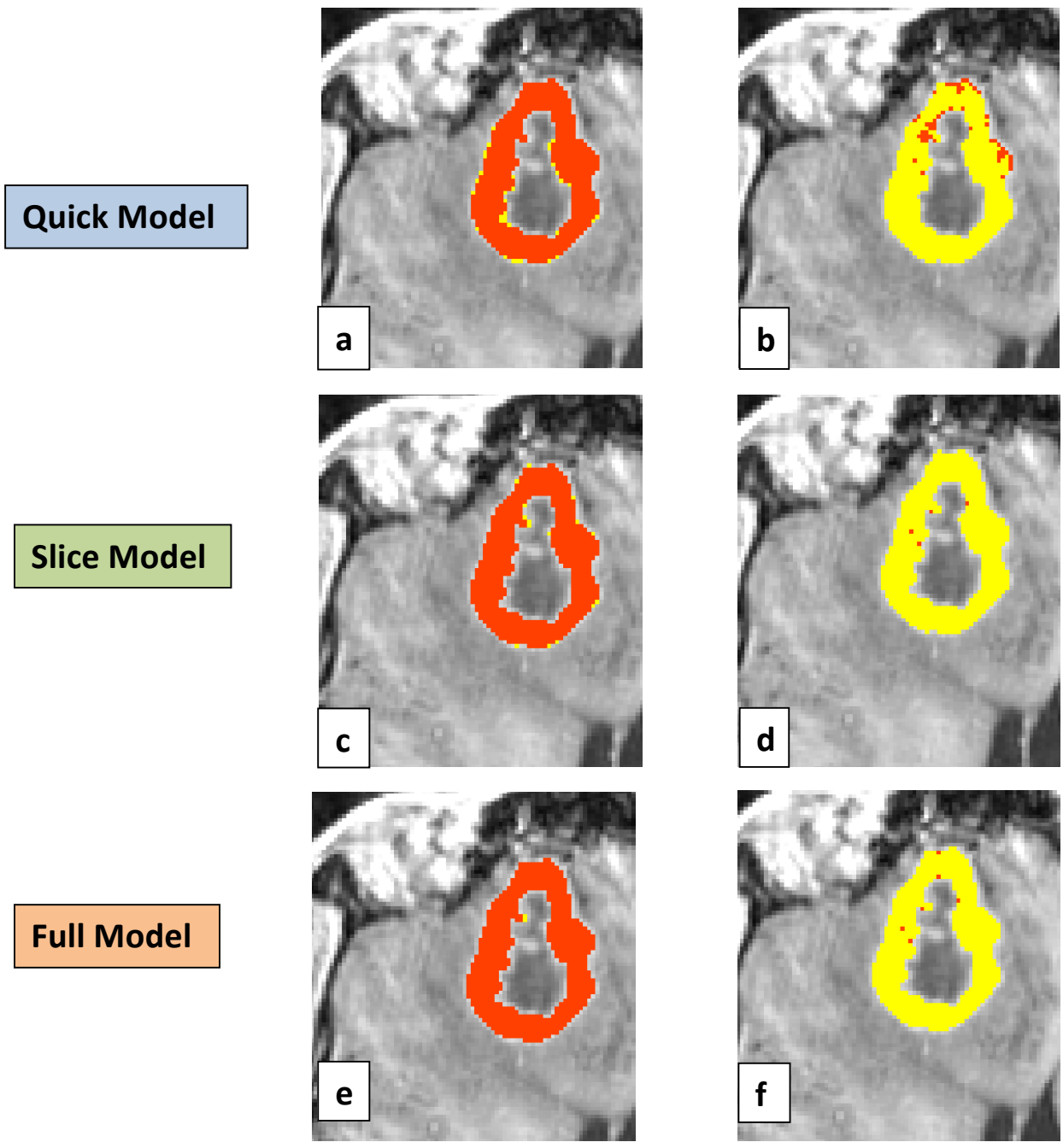


Figure B1.2: Comparing results of Quick, Single-Slice, and Full volume tumor model on patient B1. Red = Labels, yellow = SVM. a) Quick - Labels overlaid on SVM. b) Quick - SVM overlaid on Labels. c) Single-slice – Labels over SVM. d) Single-slice - SVM over Labels. e) Full – Labels overlaid on SVM. f) Full – SVM overlaid on Labels.

Necrosis Training Grid-Search B1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
6	1	0.214	1.000	0.052	1.000	0.0981	0.0637
8	2	0.313	1.000	0.110	1.000	0.1977	0.1335
8	3	0.633	1.000	0.432	1.000	0.6036	0.4877
6	2	0.700	1.000	0.523	1.000	0.6865	0.5778
6	3	0.776	1.000	0.632	1.000	0.7747	0.6825
8	6	0.776	1.000	0.632	1.000	0.7747	0.6825
3	1	0.814	1.000	0.690	1.000	0.8168	0.7359
8	10	0.830	1.000	0.716	1.000	0.8346	0.7592
6	6	0.834	1.000	0.723	1.000	0.8390	0.7650
6	10	0.850	0.992	0.755	0.999	0.8571	0.7926
8	20	0.854	0.992	0.761	0.999	0.8613	0.7984
3	2	0.863	1.000	0.768	1.000	0.8686	0.8051
3	3	0.854	0.976	0.774	0.997	0.8633	0.8075
2	1	0.862	0.992	0.774	0.999	0.8696	0.8097
6	20	0.862	0.976	0.787	0.997	0.8714	0.8188
3	6	0.874	0.977	0.807	0.997	0.8834	0.8356
2	2	0.878	0.984	0.807	0.998	0.8866	0.8367
2	6	0.870	0.962	0.813	0.996	0.8811	0.8389
3	20	0.870	0.962	0.813	0.996	0.8811	0.8389
3	10	0.874	0.969	0.813	0.996	0.8842	0.8400
2	3	0.878	0.977	0.813	0.997	0.8873	0.8411
1	2	0.867	0.941	0.826	0.993	0.8797	0.8466
1	1	0.874	0.955	0.826	0.995	0.8858	0.8488
2	20	0.874	0.955	0.826	0.995	0.8858	0.8488
2	10	0.878	0.962	0.826	0.996	0.8889	0.8499
1	20	0.872	0.923	0.852	0.990	0.8859	0.8650
1	3	0.883	0.943	0.852	0.993	0.8949	0.8684
1	10	0.876	0.924	0.858	0.990	0.8896	0.8704
1	6	0.891	0.938	0.871	0.992	0.9030	0.8835
0.5	1	0.899	0.944	0.877	0.993	0.9097	0.8900
0.5	3	0.902	0.951	0.877	0.994	0.9127	0.8912
0.5	6	0.899	0.938	0.884	0.992	0.9103	0.8943
0.5	2	0.903	0.945	0.884	0.993	0.9133	0.8954
0.5	20	0.895	0.899	0.916	0.986	0.9073	0.9126
0.5	10	0.911	0.934	0.910	0.991	0.9216	0.9144

Table B1.4: Grid search sorting by F2-score for patient B1-Necrosis model. Optimal hyperparameters in red (Gamma=0.5, and C=10) were then used to train the necrosis models.

Necrosis Results Quick, Single-Slice, and Full Models B1

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Quick	Necrosis	0.5	10	0.513	0.541	0.574	0.953
Slice	Necrosis	0.5	10	0.629	0.777	0.554	0.984
Full	Necrosis	0.5	10	0.637	0.794	0.554	0.986

TP	FN	FP	TN
144	107	122	3446
139	112	40	2528
139	112	36	2532

ADC	T1	T2	Hypoxia	rCBV
0.18	1.00	0.08	0.11	0.00
0.54	1.00	0.29	0.39	0.00
0.53	1.00	0.27	0.32	0.00

Table B1.5: Necrosis Results - Performance metrics, confusion matrix, model weighting.

While the Single-slice and Full-brain methods performed reasonably well, the Quick model was lacking in its PPV. This is an interesting example of how sensitivity and specificity may not always be completely informative. Judging from sensitivity and specificity alone, we would assume the Quick model to perform equally well, if not better. However, upon visual inspection (figure B1.3), it can be confirmed that the Single-slice and Full-brain methods show improvement. This is evidenced in their higher MCC values, further validating the use of this performance metric.

Necrosis Results Quick, Single-Slice, and Full Models B1

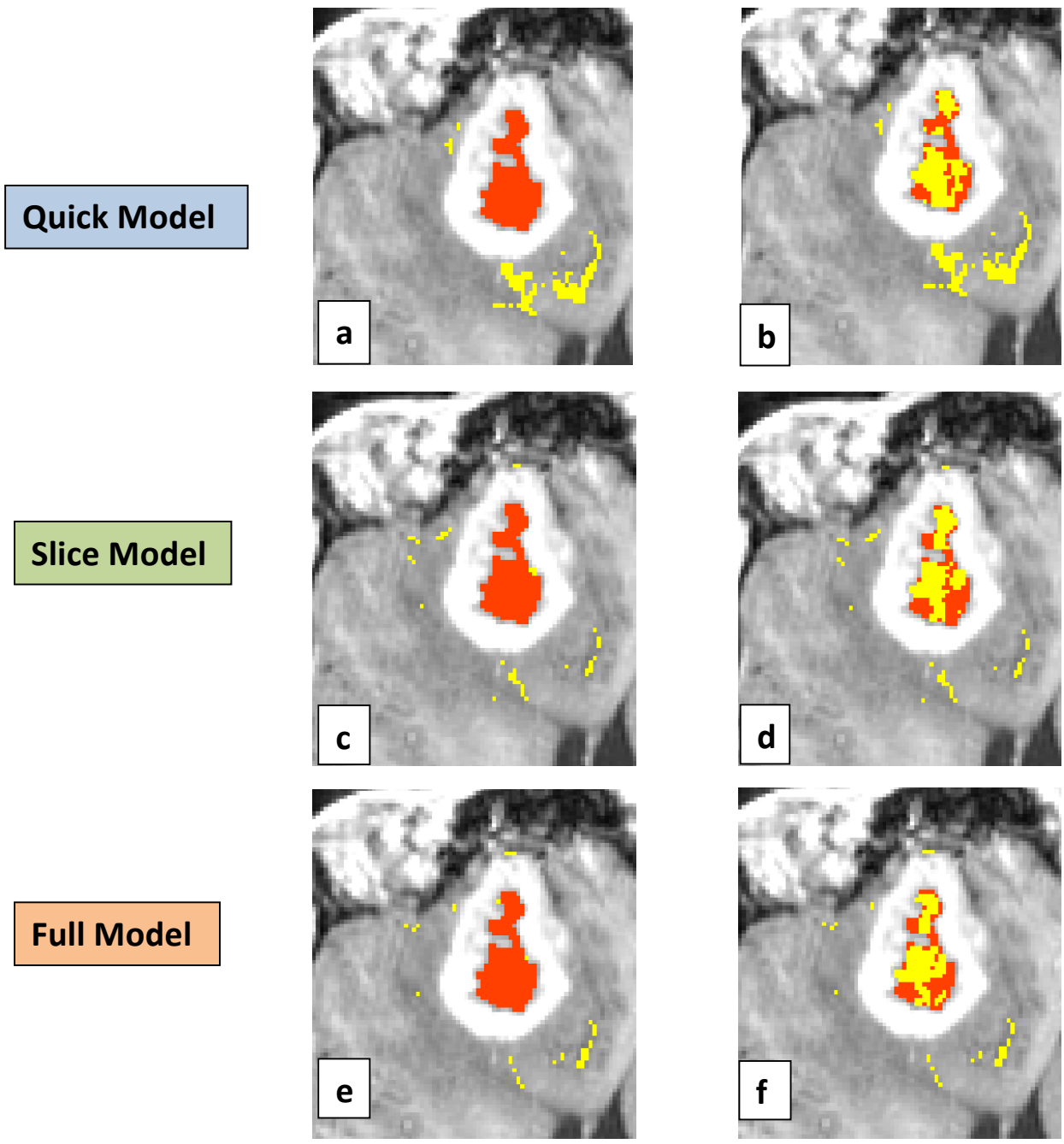


Figure B1.3: Comparing results of Quick, Single-Slice, and Full volume necrosis models on patient B1. Red = Labels, yellow = SVM. a) Quick - Labels overlaid on SVM. b) Quick - SVM overlaid on Labels. c) Single-slice – Labels over SVM. d) Single-slice - SVM over Labels. e) Full – Labels overlaid on SVM. f) Full – SVM overlaid on Labels.

Edema Training Grid-Search B1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
8	1	0.693	0.753	0.923	0.778	0.8295	0.8830
6	1	0.704	0.762	0.925	0.787	0.8353	0.8866
6	2	0.712	0.771	0.923	0.798	0.8399	0.8877
8	2	0.707	0.763	0.926	0.789	0.8369	0.8884
8	6	0.724	0.783	0.921	0.812	0.8460	0.8894
8	3	0.712	0.768	0.926	0.794	0.8397	0.8896
8	10	0.731	0.788	0.923	0.817	0.8499	0.8921
6	3	0.729	0.784	0.926	0.812	0.8489	0.8937
8	20	0.741	0.792	0.930	0.820	0.8552	0.8986
6	6	0.742	0.793	0.930	0.821	0.8559	0.8989
6	10	0.754	0.802	0.932	0.831	0.8620	0.9026
6	20	0.771	0.816	0.934	0.846	0.8711	0.9076
3	1	0.774	0.817	0.937	0.846	0.8731	0.9106
3	2	0.798	0.839	0.939	0.867	0.8862	0.9172
2	1	0.818	0.859	0.939	0.886	0.8971	0.9219
3	3	0.810	0.847	0.945	0.874	0.8930	0.9234
2	2	0.830	0.867	0.945	0.893	0.9040	0.9281
3	6	0.833	0.870	0.945	0.896	0.9056	0.9287
3	20	0.836	0.872	0.945	0.898	0.9072	0.9294
2	6	0.849	0.886	0.945	0.911	0.9145	0.9324
3	10	0.837	0.869	0.950	0.894	0.9077	0.9328
2	3	0.841	0.873	0.950	0.898	0.9101	0.9338
1	1	0.864	0.903	0.943	0.926	0.9225	0.9346
2	10	0.860	0.896	0.947	0.919	0.9203	0.9359
2	20	0.866	0.899	0.950	0.921	0.9239	0.9396
1	2	0.893	0.917	0.961	0.936	0.9388	0.9522
1	20	0.906	0.935	0.958	0.951	0.9463	0.9531
1	6	0.900	0.926	0.961	0.943	0.9431	0.9539
1	3	0.899	0.922	0.963	0.940	0.9424	0.9548
1	10	0.904	0.927	0.963	0.944	0.9449	0.9558
0.5	1	0.914	0.942	0.960	0.957	0.9507	0.9560
0.5	2	0.916	0.939	0.965	0.954	0.9519	0.9597
0.5	10	0.918	0.939	0.967	0.954	0.9528	0.9612
0.5	20	0.916	0.936	0.969	0.951	0.9520	0.9620
0.5	6	0.919	0.939	0.969	0.954	0.9538	0.9627
0.5	3	0.919	0.938	0.971	0.953	0.9538	0.9638

Table B1.6: Grid search sorting by F2-score for patient B1-Edema model. Optimal hyperparameters in red (Gamma=0.5, and C=3) were then used to train the edema models.

Edema Results Quick, Single-Slice, and Full Models B1

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Quick	Edema	0.5	3	0.543	0.875	0.504	0.957
Slice	Edema	0.5	3	0.765	0.862	0.841	0.920
Full	Edema	0.5	3	0.731	0.820	0.847	0.889

TP	FN	FP	TN
530	521	76	1692
884	167	142	1626
890	161	196	1572

ADC	T1	T2	Hypoxia	rCBV
0.86	1.00	0.75	0.00	0.65
0.80	1.00	1.00	0.00	0.80
0.74	0.94	1.00	0.00	0.58

Table B1.7: Edema Results - Performance metrics, confusion matrix, model weighting.

The Quick model failed to perform as well as the other two models. This was due to its lower sensitivity, which is a result of the high number of false negatives. This is also evident in figure B1.4. The Single-slice method showed the highest MCC value, outperforming the Full-brain method due to the elevated specificity.

Edema Results Quick, Single-Slice, and Full Models B1

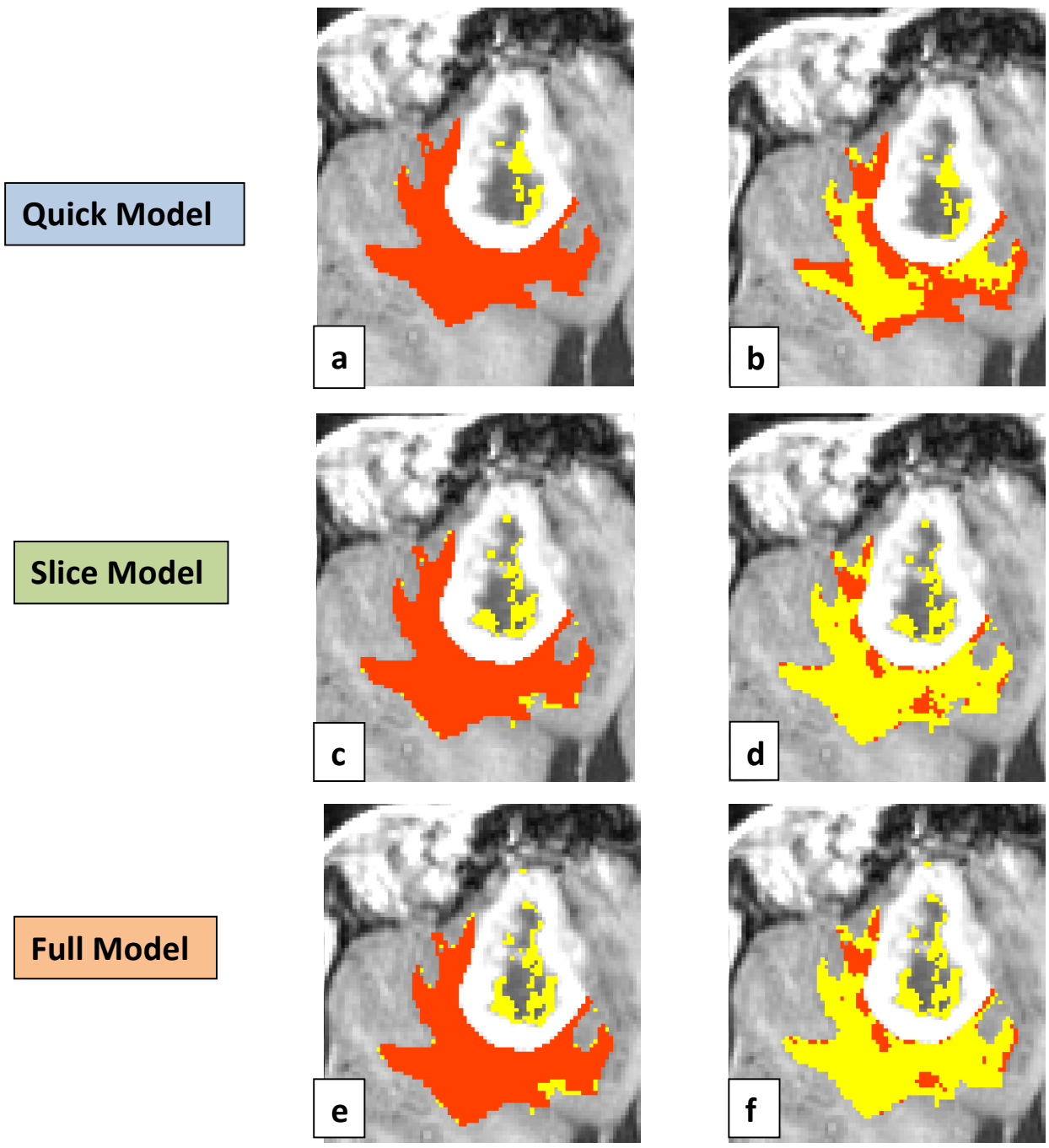


Figure B1.4: Comparing results of Quick, Single-Slice, and Full volume edema models on patient B1. Red = Labels, yellow = SVM. a) Quick - Labels overlaid on SVM. b) Quick - SVM overlaid on Labels. c) Single-slice – Labels over SVM. d) Single-slice - SVM over Labels. e) Full – Labels overlaid on SVM. f) Full – SVM overlaid on Labels.

NCE Training Grid-Search B1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
2	10	0.125	1.000	0.016	1.000	0.0323	0.0204
2	20	0.253	0.833	0.082	0.999	0.1493	0.1000
1	2	0.332	0.889	0.131	0.999	0.2285	0.1580
0.5	1	0.529	0.815	0.361	0.996	0.5000	0.4059
1	3	0.577	0.957	0.361	0.999	0.5239	0.4120
1	6	0.579	0.794	0.443	0.994	0.5684	0.4856
1	10	0.723	0.864	0.623	0.995	0.7238	0.6598
0.5	2	0.690	0.755	0.656	0.989	0.7017	0.6734
0.5	3	0.741	0.786	0.721	0.990	0.7521	0.7333
1	20	0.752	0.790	0.738	0.990	0.7627	0.7475
0.5	10	0.767	0.754	0.803	0.987	0.7778	0.7929
0.5	20	0.767	0.754	0.803	0.987	0.7778	0.7929
0.5	6	0.793	0.803	0.803	0.990	0.8033	0.8033

Table B1.8: Grid search sorting by F2-score for patient B1-NCE model. Optimal hyperparameters in red (Gamma=0.5, and C=6) were then used to train the NCE models.

NCE Results Quick, Single-Slice, and Full Models B1

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Quick	NCE	0.5	6	0.320	0.315	0.385	0.968
Slice	NCE	0.5	6	0.538	0.552	0.558	0.983
Full	NCE	0.5	6	0.589	0.663	0.548	0.989

TP	FN	FP	TN
40	64	87	2628
58	43	47	2668
57	47	29	2686

ADC	T1	T2	Hypoxia	rCBV
0.69	1.00	0.12	0.00	0.13
0.62	1.00	0.42	0.38	0.00
0.62	1.00	0.52	0.42	0.00

Table B1.9: NCE tumor Results - Performance metrics, confusion matrix, model weighting.

The Quick model performed poorly on the NCE data, while the Single-slice and Full-brain methods performed only moderately well. The Full-brain model had a higher PPV.

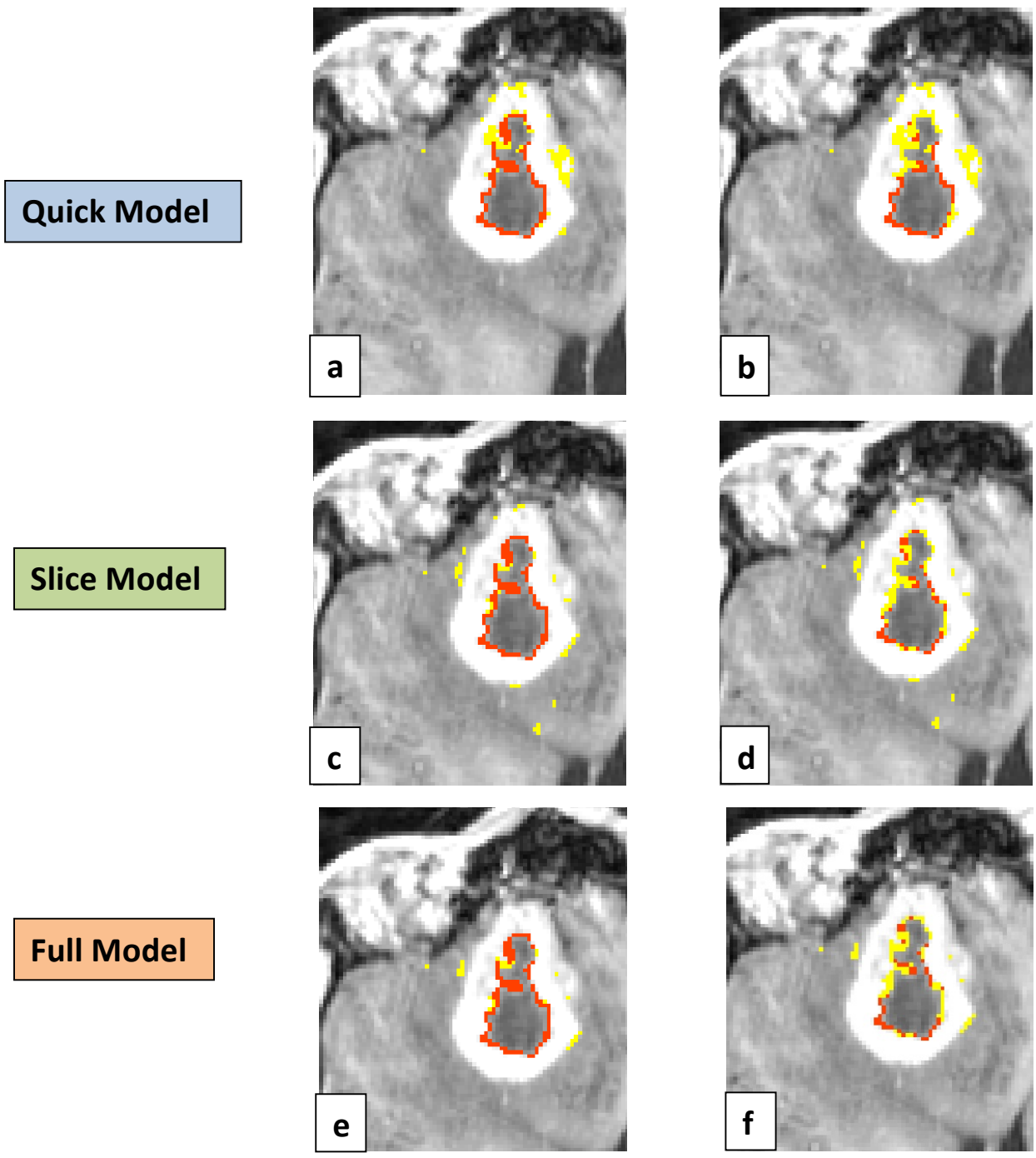


Figure B1.5: Comparing results of Quick, Single-Slice, and Full volume NCE models on patient B1. Red = Labels, yellow = SVM. a) Quick - Labels overlaid on SVM. b) Quick - SVM overlaid on Labels. c) Single-slice – Labels over SVM. d) Single-slice - SVM over Labels. e) Full – Labels overlaid on SVM. f) Full – SVM overlaid on Labels.

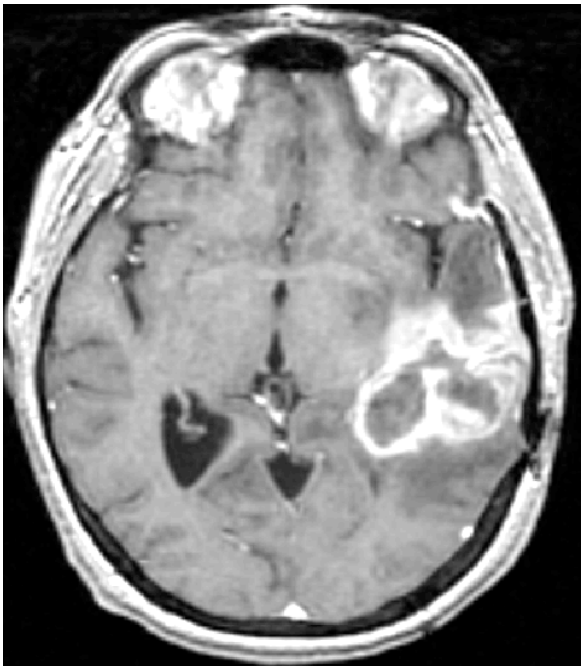
Summary of MCCs for Comparison of 3 models

	quick	slice	full
tumor	0.913	0.982	0.992
necrosis	0.513	0.629	0.637
edema	0.543	0.765	0.731
nce	0.320	0.538	0.589

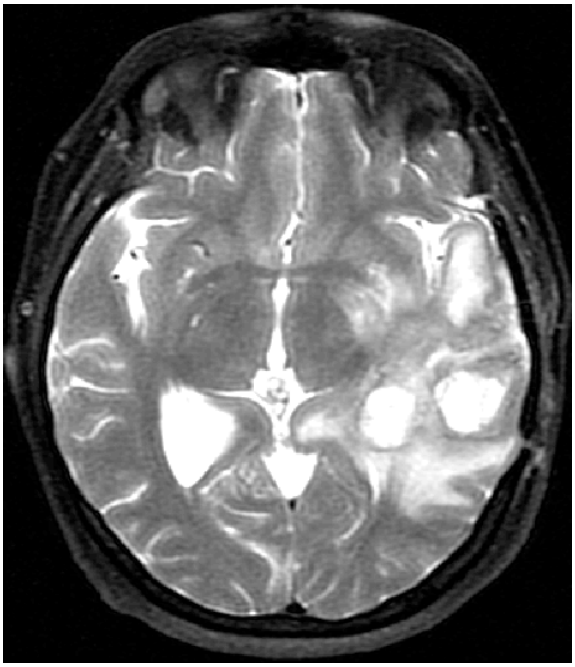
Table B1.10: Comparing MCCs of tissue classes across 3 different sized training models: quick, single-slice, and full volume.

Although the all three method performed quite well on tumor data, the Quick method failed to perform as well on the remaining tissues. This is not surprising, as it includes the least amount of training data, and most likely does not adequately capture the full spectrum of tissue characteristics. However, on a positive note, the Single-slice method compared quite well to the Full-brain model, and even outperformed it on the edema data. This is of great interest due to the time-saving benefits of requiring a single radiologist labeled slice.

Patient C1 – Female, 47years, Post-Resection, Post-Radiotherapy

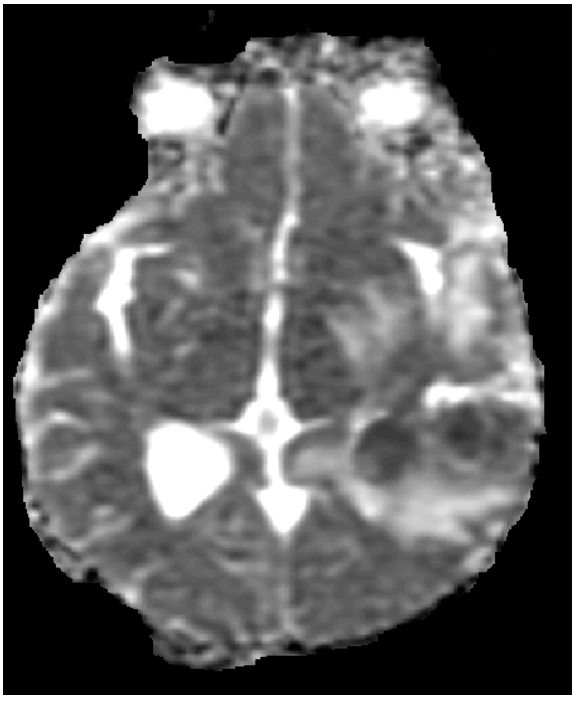


GD-T1

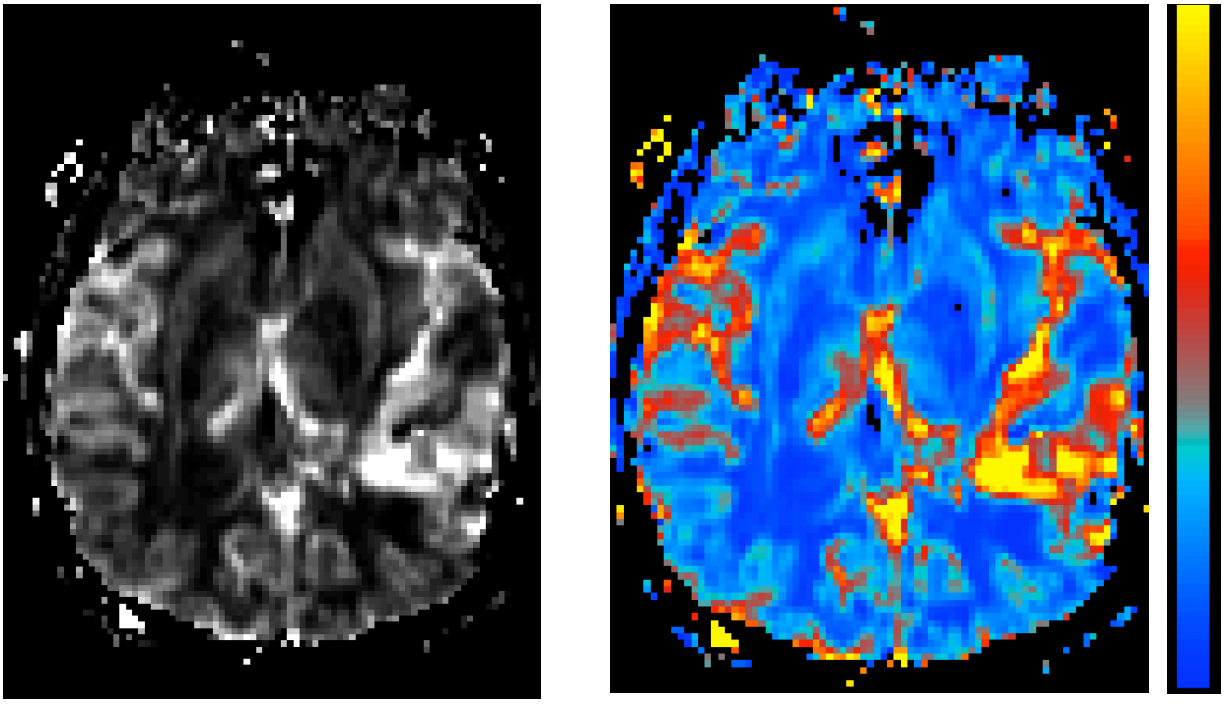


T2

Δ T2*



ADC



rCBV-grayscale rCBV-color

Figure C1.1: Multiparametric feature maps for patient C, session 1 (C1).

Session	Date relative to First Session
C1	0 months
Single session only	

Table C1.1: Session list for patient C.

Tumor Training Grid-Search C1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
1	1	0.984	0.988	0.990	0.995	0.9890	0.9896
1	3	0.986	0.989	0.992	0.995	0.9903	0.9913
4	6	0.987	0.990	0.992	0.996	0.9911	0.9918
7	1	0.988	0.992	0.991	0.996	0.9913	0.9912
6	1	0.989	0.991	0.993	0.996	0.9921	0.9929
1	6	0.989	0.995	0.990	0.998	0.9922	0.9909
4	3	0.991	0.992	0.995	0.997	0.9936	0.9947
7	6	0.991	0.993	0.995	0.997	0.9936	0.9942
4	1	0.991	0.992	0.995	0.997	0.9938	0.9947
6	3	0.992	0.993	0.995	0.997	0.9942	0.9949
6	6	0.992	0.993	0.995	0.997	0.9942	0.9949

Table C1.2: Grid search sorting by F2-score for patient C1-tumor model. Optimal hyperparameters in red (Gamma=6, and C=6) were then used to train the full tumor model.

Tumor Results **Two-class vs. Multi-class SVM** **C1**

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Two-class	tumor	6	6	**0.991**	0.996	0.991	0.998
Multi-class	tumor	5	10	**0.950**	0.984	0.947	0.993

TP	FN	FP	TN
11440	101	46	25540
10930	613	175	25420

ADC	T1	T2	Hypoxia	rCBV
0.04	1	0.01	0.12	0
0.42	1	0.82	0.10	0.44

Table C1.3: Tumor Results - Performance metrics, confusion matrix, model weighting.

Both SVM models performed very well on the tumor data, however Multi-class had slightly worse sensitivity. The MCC values are both good, but the Two-class method is preferred. This is confirmed in figure C1.2. The feature weighting is of interest since Multi-class is much more dispersed. This is perhaps because the Multi-class model chooses one set of weights to use for all tissue types, while the Two-class model is tissue specific.

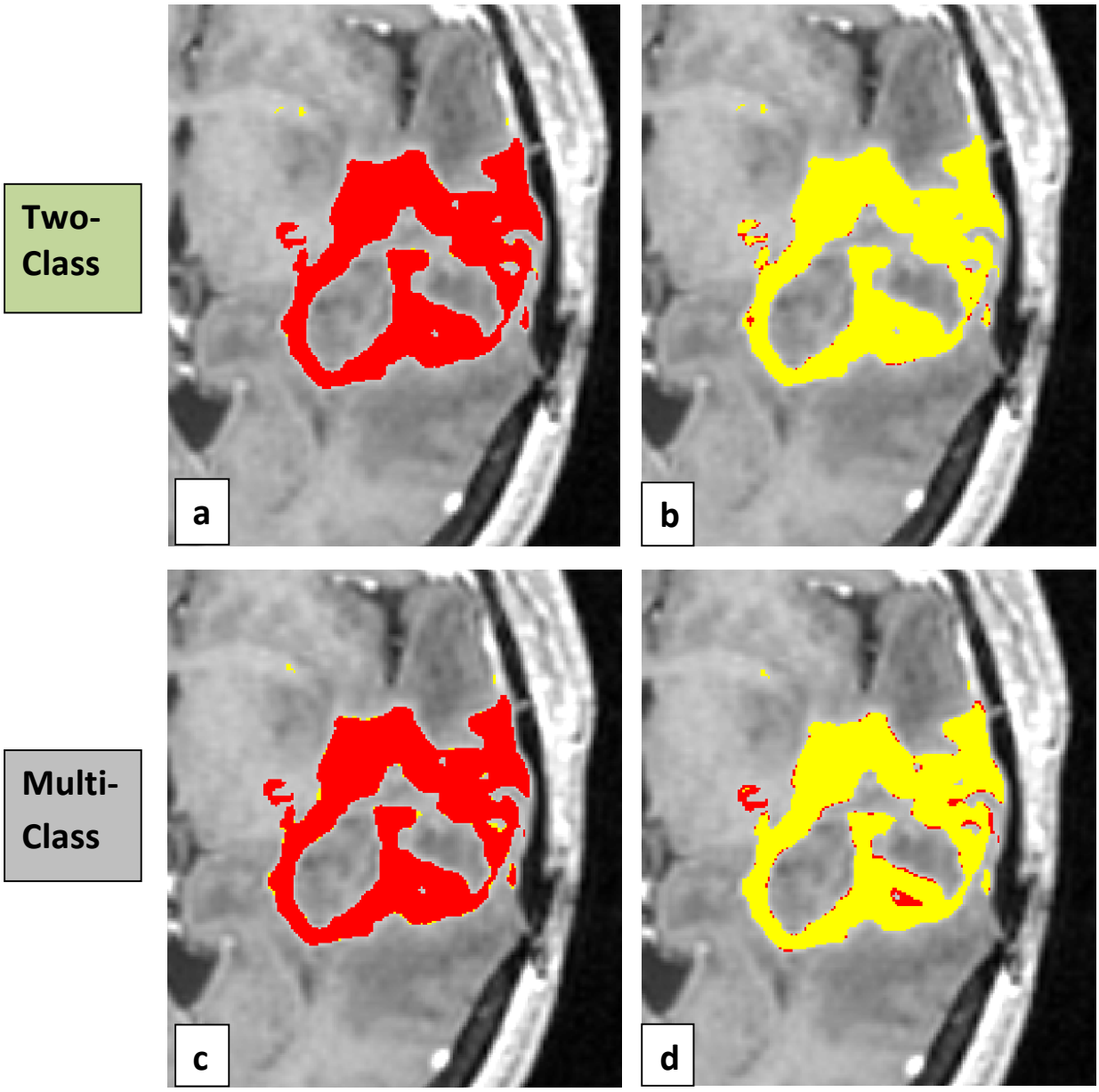


Figure C1.2: Comparing results of Two-Class vs. Multi-Class tumor model on patient C1. Red = Labels, yellow = SVM. a) Two-Class - Labels overlaid on SVM. b) Two-Class - SVM overlaid on Labels. c) Multi-Class – Labels over SVM. d) Multi-Class - SVM over Labels.

Necrosis Training Grid-Search C1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
3	1	0.881	0.928	0.850	0.995	0.8875	0.8647
3	5	0.912	0.961	0.875	0.997	0.9163	0.8913
3	10	0.916	0.963	0.882	0.998	0.9209	0.8973
1.5	1	0.920	0.965	0.887	0.998	0.9246	0.9018
1	1	0.935	0.980	0.901	0.999	0.9386	0.9155
0.7	0.5	0.935	0.978	0.902	0.999	0.9387	0.9166
0.8	0.5	0.935	0.978	0.902	0.999	0.9387	0.9166
0.6	0.5	0.940	0.982	0.907	0.999	0.9431	0.9214
0.8	1.5	0.945	0.978	0.919	0.999	0.9479	0.9305
1.5	5	0.946	0.979	0.921	0.999	0.9488	0.9319
1.5	10	0.950	0.975	0.933	0.998	0.9535	0.9409
0.7	1.5	0.953	0.982	0.931	0.999	0.9559	0.9408
1	5	0.953	0.981	0.933	0.999	0.9560	0.9419
0.8	5	0.959	0.982	0.941	0.999	0.9613	0.9491
1	10	0.960	0.984	0.941	0.999	0.9621	0.9494
0.6	1.5	0.960	0.984	0.941	0.999	0.9621	0.9494

Table C1.4: Grid search sorting by F2-score for patient C1 - Necrosis model. Optimal hyperparameters in red (Gamma=0.6, and C=1.5) were then used to train the full necrosis model.

Necrosis Results Two-class vs. Multi-class SVM C1

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Two-class	Necrosis	0.6	1.5	**0.775**	0.677	0.908	0.984
Multi-class	Necrosis	5	10	**0.630**	0.666	0.620	0.988

TP	FN	FP	TN
1214	123	579	35220
829	508	415	35380

ADC	T1	T2	Hypoxia	rCBV
0.58	1	0.87	0	0.49
0.42	1	0.82	0.10	0.44

Table C1.5: Necrosis Results - Performance metrics, confusion matrix, model weighting.

Both models perform reasonably well. The Multi-class model performs worse due to the lower sensitivity. Figure C1.3 shows some interesting distinctions. The Two-class model is preferred due to the elevated accuracy in the center of the necrotic area. However, the Multi-class model is less noisy outside of the area of interest. This is due to the one-vs-rest method. The Multi-class model may have had a reasonably high 'necrosis weight' assigned to those same noisy voxels seen in the Two-class model. However, if the one-vs-rest algorithm returned a higher weighting for classifying those noisy voxels as edema, then they will not be called necrosis.

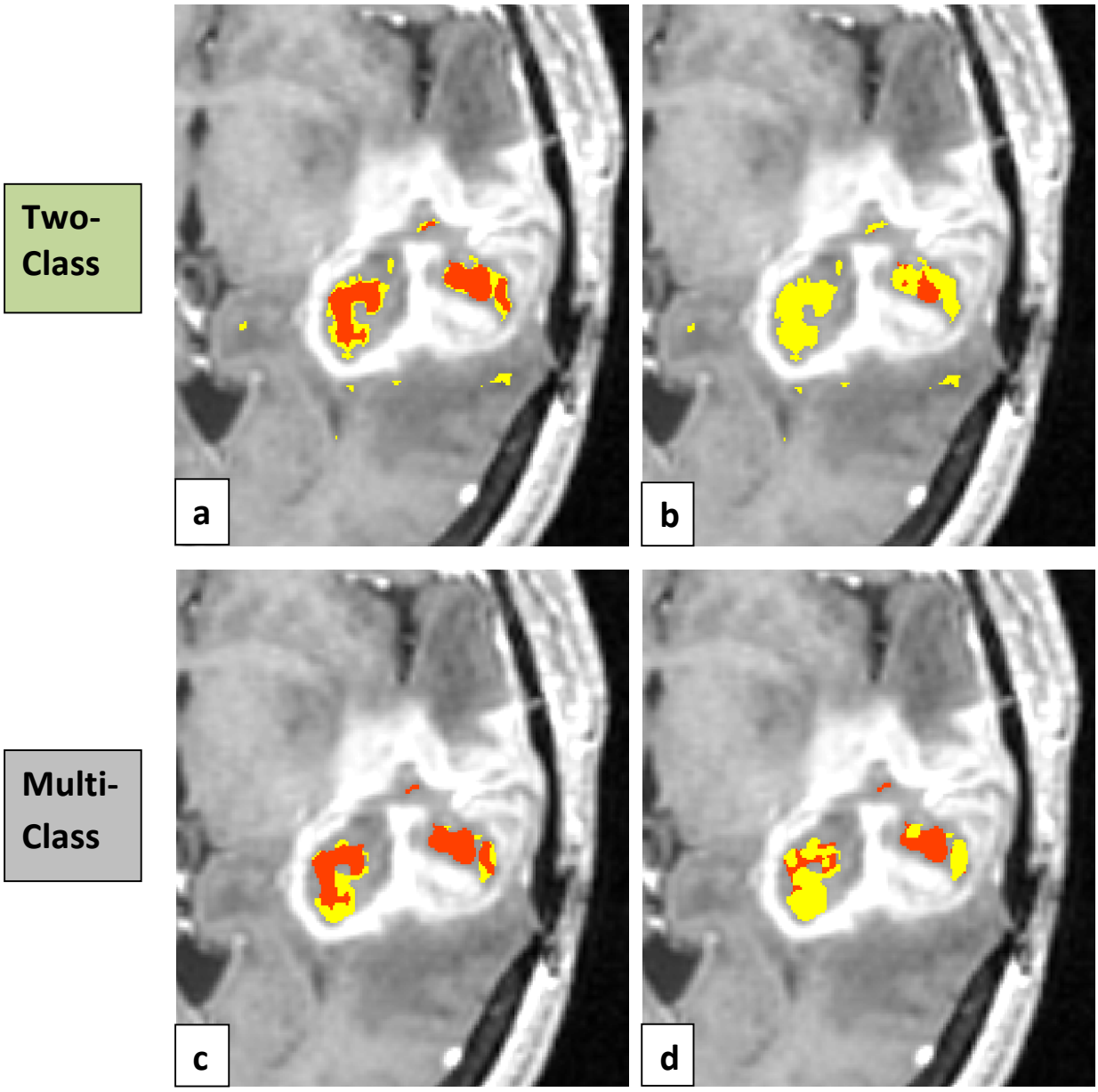


Figure C1.3: Comparing results of Two-Class vs. Multi-Class Necrosis model on patient C1. Red = Labels, yellow = SVM. a) Two-Class - Labels overlaid on SVM. b) Two-Class - SVM overlaid on Labels. c) Multi-Class – Labels over SVM. d) Multi-Class - SVM over Labels.

Edema Training Grid-Search C1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
3	0.5	0.786	0.895	0.783	0.968	0.8350	0.8028
3	1	0.807	0.894	0.816	0.967	0.8532	0.8303
3	1.5	0.822	0.897	0.836	0.967	0.8653	0.8476
1.5	0.5	0.847	0.915	0.855	0.973	0.8839	0.8660
1.5	1	0.864	0.919	0.877	0.973	0.8974	0.8852
1	0.5	0.868	0.921	0.882	0.974	0.9009	0.8892
1.5	1.5	0.867	0.916	0.885	0.972	0.9003	0.8912
1	1	0.879	0.924	0.896	0.975	0.9097	0.9013
1	1.5	0.882	0.922	0.902	0.974	0.9117	0.9056
0.8	1.5	0.886	0.925	0.906	0.975	0.9152	0.9095
1	5	0.888	0.922	0.911	0.973	0.9162	0.9130
0.8	5	0.891	0.927	0.910	0.975	0.9184	0.9135
0.5	1.5	0.894	0.932	0.910	0.977	0.9207	0.9143
1	10	0.891	0.924	0.915	0.974	0.9191	0.9164
0.8	10	0.896	0.927	0.918	0.975	0.9225	0.9195

Table C1.6: Grid search sorting by F2-score for patient C1 - Edema model. Optimal hyperparameters in red (Gamma=0.8, and C=10) were then used to train the full edema model.

Edema Results **Two-class vs. Multi-class SVM** **C1**

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Two-class	Edema	0.8	10	**0.776**	0.867	0.803	0.954
Multi-class	Edema	5	10	**0.742**	0.769	0.865	0.903

TP	FN	FP	TN
8097	1983	1245	25810
8719	1361	2614	24440

ADC	T1	T2	Hypoxia	rCBV
0.86	0.85	1.00	0.00	0.74
0.42	1.00	0.82	0.10	0.44

Table C1.7: Edema Results - Performance metrics, confusion matrix, model weighting.

Both models performed well as confirmed in figure C1.4. The Two-class model shows higher PPV and Specificity, while the Multi-class shows higher Sensitivity.

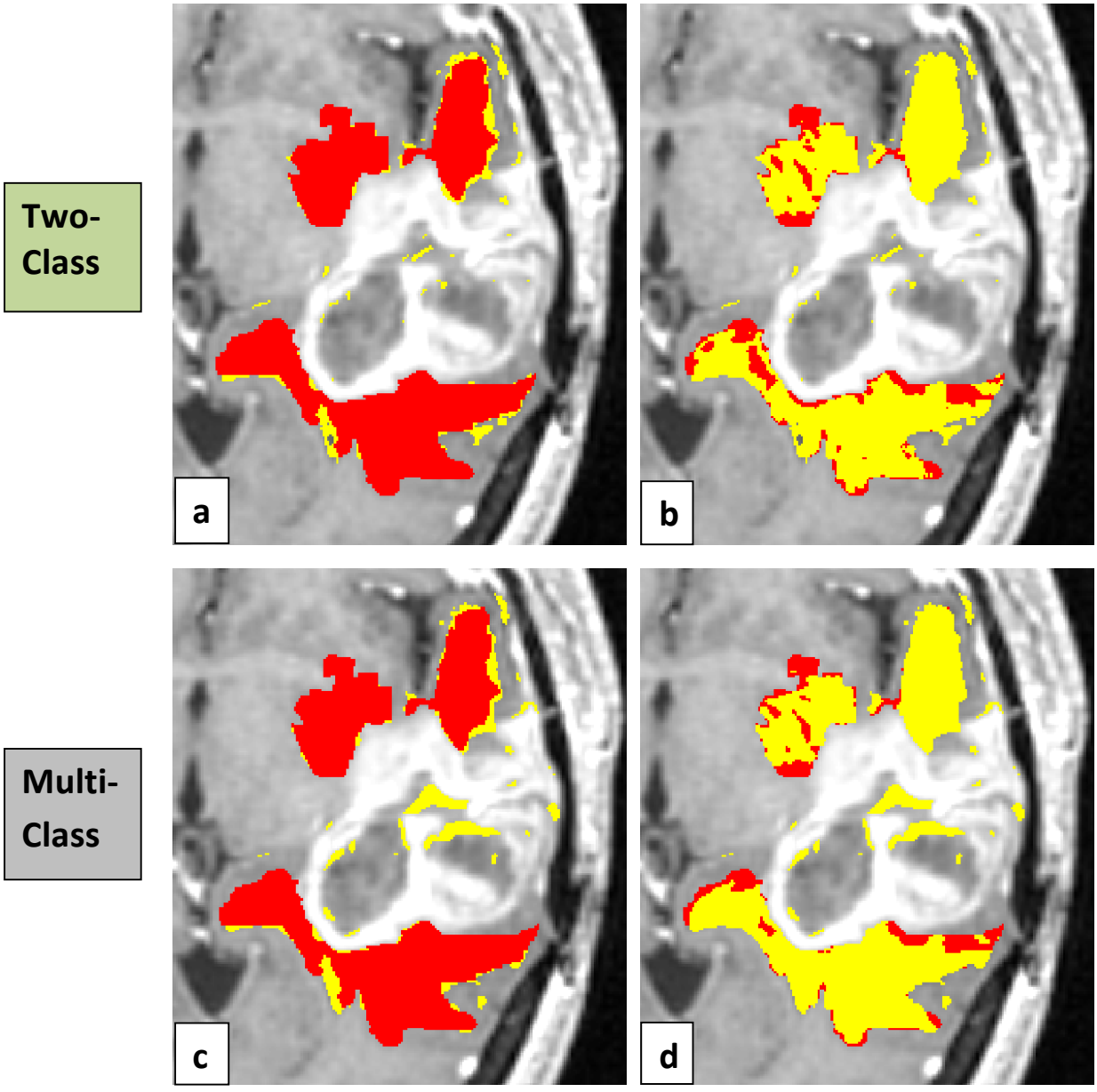


Figure C1.4: Comparing results of Two-Class vs. Multi-Class Edema model on patient C1. Red = Labels, yellow = SVM. a) Two-Class - Labels overlaid on SVM. b) Two-Class - SVM overlaid on Labels. c) Multi-Class – Labels over SVM. d) Multi-Class - SVM over Labels.

NCE Training Grid-Search C1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
6	1	0.214	0.977	0.057	1.000	0.1074	0.0700
6	3	0.569	0.933	0.400	0.994	0.5603	0.4520
6	6	0.579	0.871	0.451	0.986	0.5942	0.4990
3	1	0.609	0.908	0.469	0.990	0.6183	0.5191
3	3	0.631	0.914	0.496	0.990	0.6435	0.5463
3	6	0.676	0.906	0.567	0.988	0.6977	0.6131
1	1	0.771	0.941	0.686	0.991	0.7934	0.7251
1	3	0.809	0.938	0.747	0.990	0.8318	0.7788
1	6	0.820	0.943	0.762	0.990	0.8424	0.7919
1	6	0.820	0.943	0.762	0.990	0.8424	0.7919
1	10	0.825	0.936	0.775	0.989	0.8478	0.8026

Table C1.8: Grid search sorting by F2-score for patient C1- NCE model. Optimal hyperparameters in red (Gamma=1, and C=10) were then used to train the full NCE model.

NCE Results **Two-class vs. Multi-class SVM** **C1**

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Two-class	NCE	1	10	**0.540**	0.656	0.567	0.942
Multi-class	NCE	5	10	**0.482**	0.670	0.453	0.956

TP	FN	FP	TN
3456	2641	1810	29230
2759	3338	1359	29680

ADC	T1	T2	Hypoxia	rCBV
0.03	1.00	0.35	0.34	0.00
0.42	1.00	0.82	0.10	0.44

Table C1.9: NCE Results - Performance metrics, confusion matrix, model weighting.

Both models perform moderately well at classifying the interior NCE voxels in figure C1.5. However, they both struggled to identify voxels in outside of the enhancing region. The Two-class model is preferred despite moderate performance.

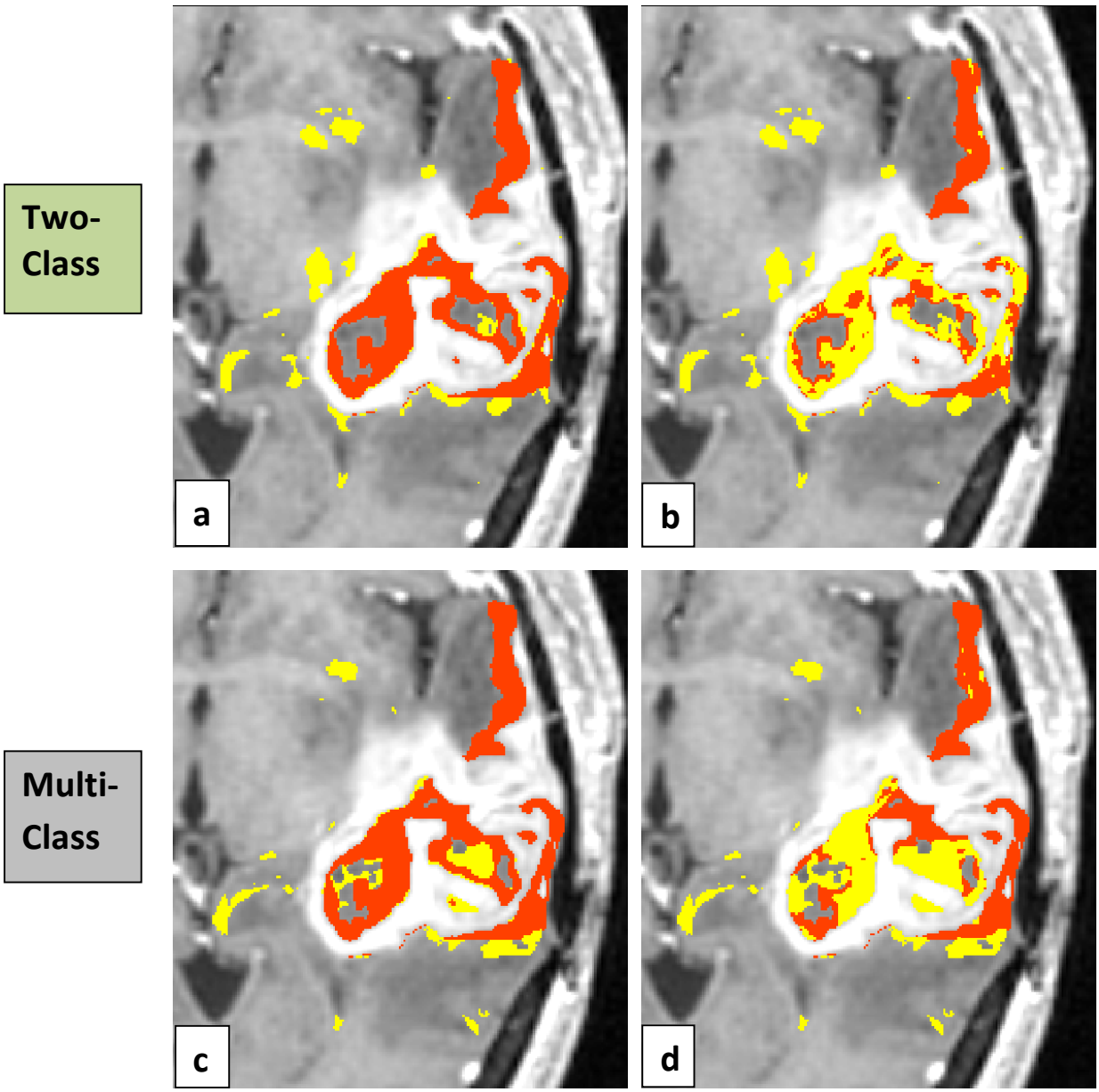


Figure C1.5: Comparing results of Two-Class vs. Multi-Class NCE model on patient C1. Red = Labels, yellow = SVM. a) Two-Class - Labels overlaid on SVM. b) Two-Class - SVM overlaid on Labels. c) Multi-Class – Labels over SVM. d) Multi-Class - SVM over Labels.

Summary of MCCs

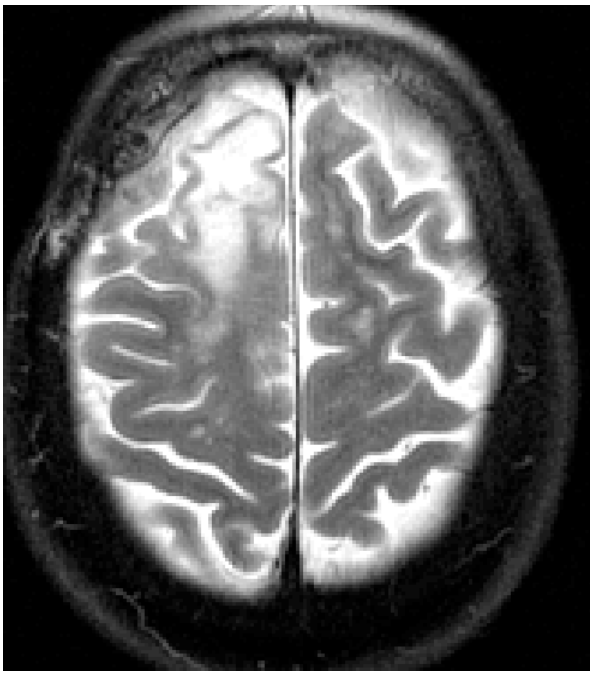
	Two-Class	Multi-Class
tumor	0.991	0.950
necrosis	0.775	0.630
edema	0.776	0.742
NCE	0.540	0.482

Table C1.10: Comparing MCCs of tissue classes using Two-Class vs. Multi-Class SVM.

A side-by-side comparison of the Two-class and Multi-class models shows a slight favor towards the Two-class approach. The largest gains were in the necrosis and NCE models. Overall they are moderately similar.

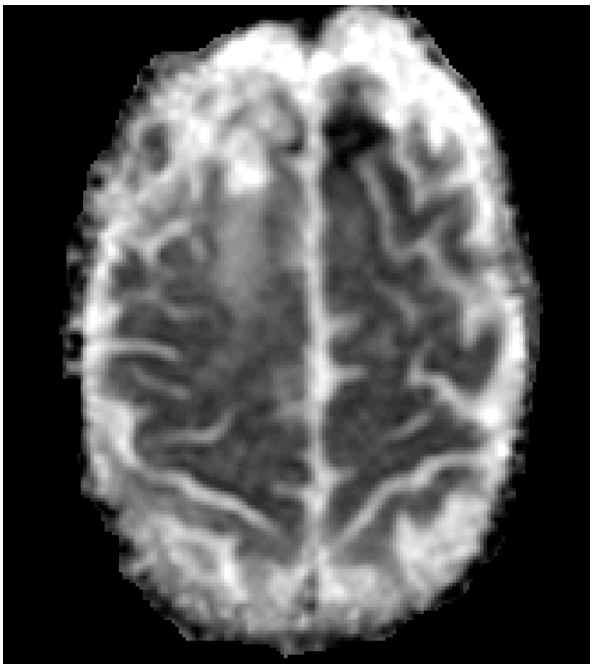
Patient D1 – Female, 59 years, Post XRT, Chemo, Radiosurgery

GD-T1

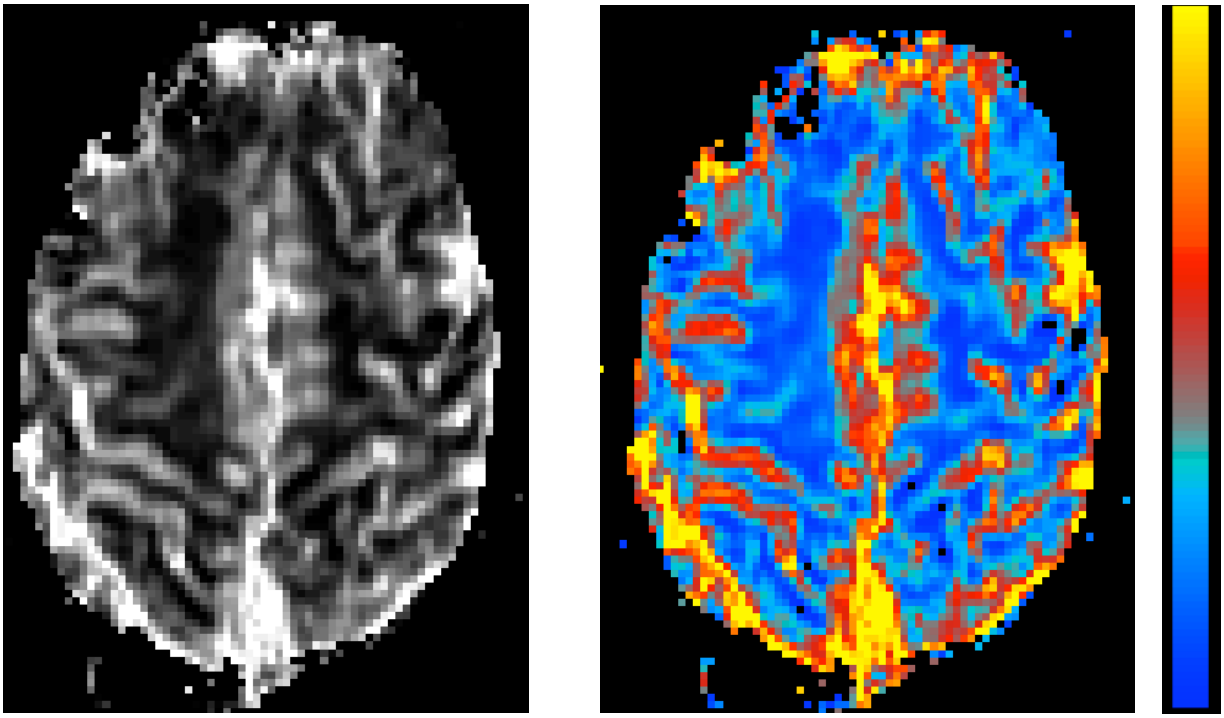


T2

Δ T2*



ADC



rCBV-grayscale rCBV-color

Figure D1.1: Multiparametric feature maps for patient D, session 1 (D1).

Session	Date relative to First Session
D1	0 months
D2	2.5 months

Table D.1: Session list for patient D.

Tumor Training Grid-Search Session D1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
10	1	0.710	1.000	0.537	1.000	0.6986	0.5916
10	2	0.768	1.000	0.621	1.000	0.7663	0.6720
6	1	0.803	1.000	0.674	1.000	0.8050	0.7207
10	3	0.830	1.000	0.716	1.000	0.8344	0.7589
6	2	0.882	1.000	0.800	1.000	0.8889	0.8333
10	6	0.882	1.000	0.800	1.000	0.8889	0.8333
6	3	0.895	0.988	0.832	0.999	0.9029	0.8587
10	10	0.895	0.988	0.832	0.999	0.9029	0.8587
3	1	0.907	0.988	0.853	0.999	0.9152	0.8766
2	1	0.914	0.988	0.863	0.999	0.9214	0.8856
3	2	0.914	0.988	0.863	0.999	0.9214	0.8856
6	6	0.920	0.988	0.874	0.999	0.9274	0.8944
10	20	0.926	0.988	0.884	0.999	0.9333	0.9032
6	10	0.939	1.000	0.895	1.000	0.9444	0.9139
1	1	0.939	0.989	0.905	0.999	0.9451	0.9208
3	3	0.939	0.989	0.905	0.999	0.9451	0.9208
3	6	0.939	0.989	0.905	0.999	0.9451	0.9208
6	20	0.945	1.000	0.905	1.000	0.9503	0.9228
0.5	1	0.951	0.989	0.926	0.999	0.9565	0.9382
1	20	0.939	0.957	0.937	0.994	0.9468	0.9408
0.5	20	0.934	0.938	0.947	0.991	0.9424	0.9454
2	2	0.957	0.989	0.937	0.999	0.9621	0.9468
2	3	0.957	0.989	0.937	0.999	0.9621	0.9468
0.5	10	0.952	0.968	0.947	0.996	0.9574	0.9514
0.5	6	0.958	0.978	0.947	0.997	0.9626	0.9534
1	10	0.958	0.978	0.947	0.997	0.9626	0.9534
0.5	2	0.964	0.989	0.947	0.999	0.9678	0.9554
0.5	3	0.964	0.989	0.947	0.999	0.9678	0.9554
2	20	0.964	0.989	0.947	0.999	0.9678	0.9554
3	10	0.964	0.989	0.947	0.999	0.9678	0.9554
3	20	0.964	0.989	0.947	0.999	0.9678	0.9554
1	2	0.970	0.989	0.958	0.999	0.9733	0.9640
2	6	0.970	0.989	0.958	0.999	0.9733	0.9640
2	10	0.976	1.000	0.958	1.000	0.9785	0.9660
1	3	0.976	0.989	0.968	0.999	0.9787	0.9725
1	6	0.976	0.989	0.968	0.999	0.9787	0.9725

Table D1.2: Grid search sorting by F2-score for patient D1-tumor model. Optimal hyperparameters in red (Gamma=1, and C=6) were then used to train the full tumor model.

Tumor Results Intrasession, Intersession, Interpatient Models D1

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Intrasession	tumor	1	6	**0.963**	0.961	0.983	0.987
Intersession	tumor	6	6	**0.965**	1.000	0.947	1.000
Interpatient	tumor	3	10	**0.977**	0.965	1.000	0.988

TP	FN	FP	TN
297	5	12	931
286	16	0	943
302	0	11	932

ADC	T1	T2	Hypoxia	rCBV
0.00	1.00	0.09	0.12	0.23
0.34	1.00	0.35	0.00	0.29
0.10	1.00	0.11	0.03	0.00

Table D1.3: Tumor Results - Performance metrics, confusion matrix, model weighting.

All methods showed outstanding performance on the tumor data. Visually (Figure D1.2), Intersession and Interpatient would be preferred.

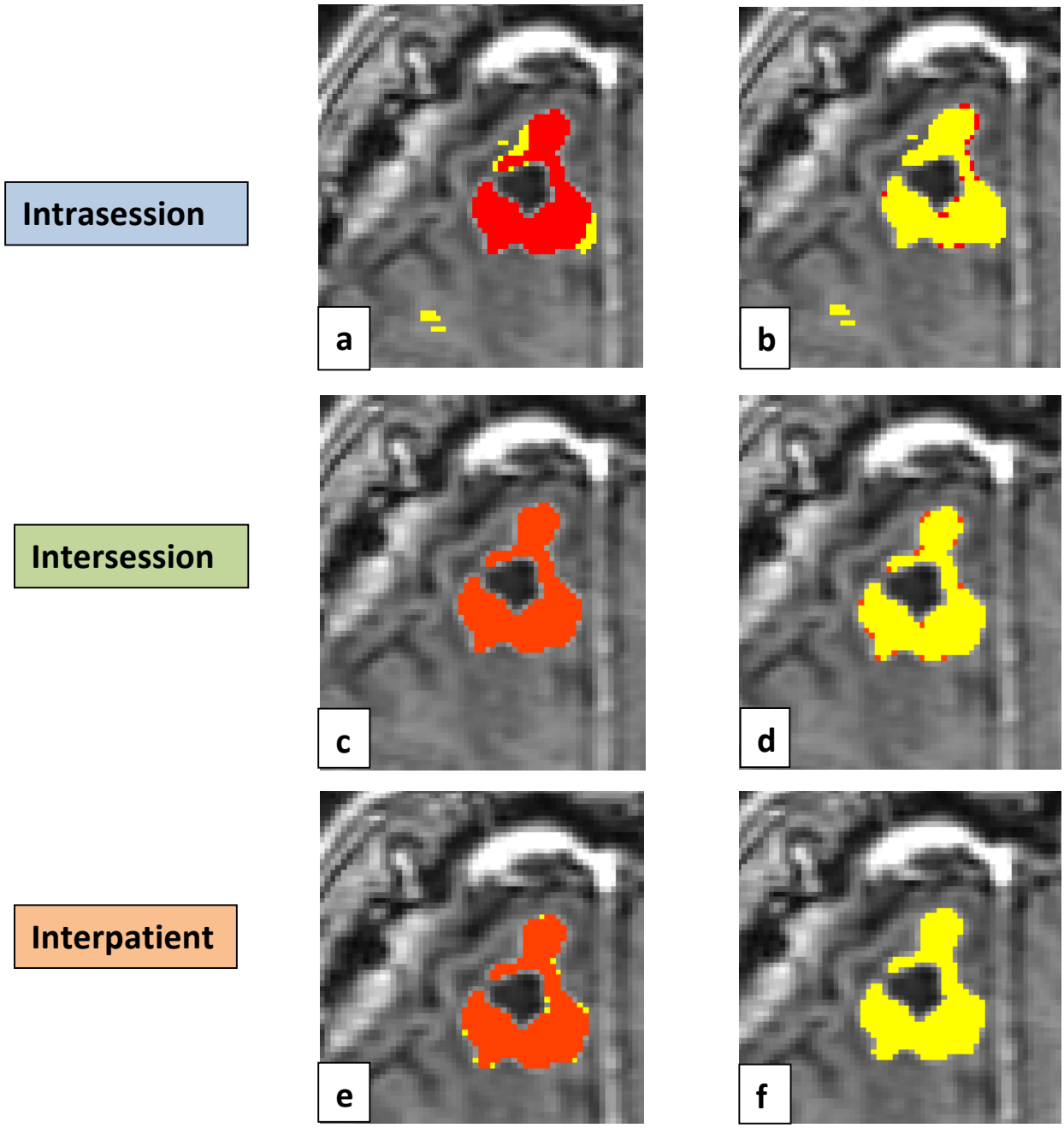


Figure D1.2: Comparing results of Intrasession, Intersession, and Interpatient tumor models on patient D1. Red = Labels, yellow = SVM. a) Intrasession - Labels overlaid on SVM. b) Intrasession - SVM overlaid on Labels. c) Intersession – Labels over SVM. d) Intersession - SVM over Labels. e) Interpatient – Labels overlaid on SVM. f) Interpatient – SVM overlaid on Labels.

Necrosis Training Grid-Search D1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
10	1	0.800	0.943	0.765	0.982	0.8448	0.7950
10	2	0.823	0.950	0.793	0.984	0.8643	0.8198
6	2	0.834	0.961	0.797	0.987	0.8715	0.8254
10	6	0.834	0.961	0.797	0.987	0.8715	0.8254
10	10	0.834	0.961	0.797	0.987	0.8715	0.8254
6	1	0.830	0.951	0.802	0.984	0.8700	0.8277
10	3	0.830	0.951	0.802	0.984	0.8700	0.8277
6	3	0.834	0.956	0.802	0.986	0.8721	0.8285
10	20	0.823	0.936	0.807	0.978	0.8664	0.8294
6	6	0.827	0.941	0.807	0.980	0.8685	0.8302
6	10	0.833	0.937	0.820	0.978	0.8747	0.8412
3	1	0.837	0.942	0.820	0.980	0.8769	0.8420
6	20	0.833	0.932	0.825	0.976	0.8753	0.8444
3	2	0.840	0.942	0.825	0.980	0.8796	0.8459
2	1	0.847	0.943	0.834	0.980	0.8851	0.8538
3	3	0.847	0.943	0.834	0.980	0.8851	0.8538
3	6	0.837	0.924	0.839	0.973	0.8792	0.8545
2	20	0.850	0.939	0.843	0.978	0.8884	0.8608
2	2	0.853	0.943	0.843	0.980	0.8905	0.8616
3	20	0.840	0.916	0.853	0.969	0.8830	0.8645
3	10	0.847	0.925	0.853	0.973	0.8873	0.8661
2	6	0.853	0.934	0.853	0.976	0.8915	0.8677
2	10	0.856	0.939	0.853	0.978	0.8937	0.8685
2	3	0.850	0.925	0.857	0.973	0.8899	0.8699
1	2	0.870	0.949	0.862	0.982	0.9034	0.8780
1	1	0.870	0.945	0.866	0.980	0.9039	0.8810
1	3	0.873	0.945	0.871	0.980	0.9065	0.8849
1	20	0.890	0.955	0.885	0.984	0.9187	0.8980
1	6	0.890	0.951	0.889	0.982	0.9190	0.9010
1	10	0.893	0.955	0.889	0.984	0.9212	0.9019
0.5	1	0.899	0.965	0.889	0.987	0.9257	0.9036
0.5	10	0.903	0.952	0.908	0.982	0.9292	0.9163
0.5	20	0.906	0.956	0.908	0.984	0.9314	0.9171
0.5	6	0.913	0.961	0.912	0.986	0.9362	0.9218
0.5	2	0.922	0.971	0.917	0.989	0.9431	0.9273
0.5	3	0.922	0.971	0.917	0.989	0.9431	0.9273

Table D1.4: Grid search sorting by F2-score for patient D1-Necrosis model. Optimal hyperparameters in red (Gamma=0.5, and C=3) were then used to train the full necrosis model.

Necrosis Results Intrasession, Intersession, Interpatient Models D1

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Intrasession	Necrosis	0.5	3	0.726	0.726	0.757	0.983
Intersession	Necrosis	1	10	0.708	0.648	0.814	0.974
Interpatient	Necrosis	1	10	0.753	0.870	0.671	0.994

TP	FN	FP	TN
53	17	20	1155
57	13	31	1144
47	23	7	1168

ADC	T1	T2	Hypoxia	rCBV
0.00	1.00	0.45	0.26	0.34
0.37	1.00	0.85	0.00	0.07
0.75	1.00	0.82	0.00	0.63

Table D1.5: Necrosis Results - Performance metrics, confusion matrix, model weighting.

All three models performed quite well. However, the Intersession method did slightly worse, specifically in the area of PPV. This is an interesting example because if we only looked at sensitivity and specificity values (which make up the ROC curve), we would be led to choose the Intersession model as the top choice. However, our MCC values rank Intersession the lowest of the 3. A visual inspection of the data in figure D1.3 confirms that Intersession would not be the best choice.

Necrosis Results Intrasession, Intersession, Interpatient Models D1

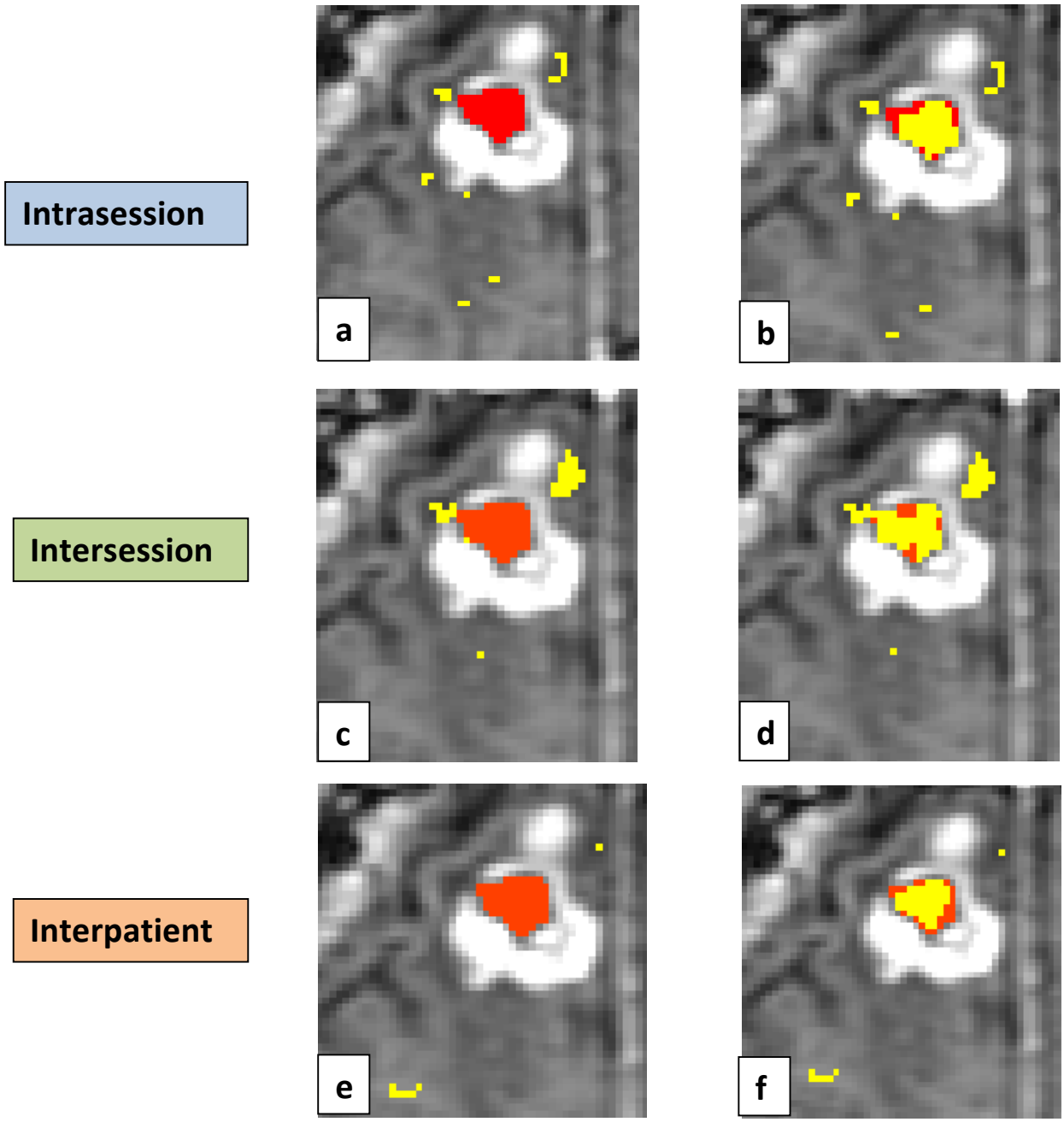


Figure D1.3: Comparing results of Intrasession, Intersession, and Interpatient necrosis models on patient D1. Red = Labels, yellow = SVM. a)Intrasession - Labels overlaid on SVM. b) Intrasession - SVM overlaid on Labels. c) Intersession – Labels over SVM. d) Intersession - SVM over Labels. e) Interpatient – Labels overlaid on SVM. f) Interpatient – SVM overlaid on Labels.

Edema Training Grid-Search D1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
3	2	0.235	1.000	0.075	1.000	0.1397	0.0921
3	3	0.608	0.878	0.540	0.971	0.6686	0.5849
2	1	0.714	0.829	0.751	0.940	0.7882	0.7656
3	6	0.735	0.817	0.798	0.931	0.8076	0.8019
2	2	0.747	0.833	0.798	0.939	0.8153	0.8049
3	10	0.764	0.844	0.812	0.942	0.8277	0.8183
1	1	0.774	0.850	0.822	0.944	0.8353	0.8270
2	3	0.783	0.859	0.826	0.948	0.8421	0.8325
1	2	0.781	0.831	0.855	0.933	0.8426	0.8497
3	20	0.788	0.833	0.864	0.933	0.8479	0.8574
1	3	0.792	0.833	0.869	0.933	0.8505	0.8612
0.5	1	0.806	0.856	0.864	0.944	0.8598	0.8622
2	6	0.795	0.834	0.873	0.933	0.8532	0.8651
2	20	0.795	0.834	0.873	0.933	0.8532	0.8651
1	6	0.803	0.833	0.887	0.931	0.8591	0.8758
1	10	0.806	0.836	0.887	0.933	0.8610	0.8766
2	10	0.801	0.826	0.892	0.928	0.8578	0.8780
0.5	20	0.815	0.848	0.887	0.939	0.8669	0.8790
1	20	0.808	0.830	0.897	0.930	0.8623	0.8826
0.5	10	0.827	0.857	0.897	0.942	0.8761	0.8884
0.5	6	0.825	0.850	0.901	0.939	0.8747	0.8905
0.5	2	0.826	0.847	0.906	0.937	0.8753	0.8935
0.5	3	0.827	0.844	0.911	0.935	0.8759	0.8965

Table D1.6: Grid search sorting by F2-score for patient D1-Edema model. Optimal hyperparameters in red (Gamma=0.5, and C=3) were then used to train the full edema model.

Edema Results Intrasession, Intersession, Interpatient Models D1

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Intrasession	edema	0.5	3	0.743	0.874	0.768	0.948
Intersession	edema	1	20	0.730	0.786	0.854	0.892
Interpatient	edema	1	5	0.000	0.000	0.000	1.000

TP	FN	FP	TN
304	92	44	805
338	58	92	757
0	396	0	849

ADC	T1	T2	Hypoxia	rCBV
0.00	0.78	1.00	0.20	0.03
0.20	0.53	1.00	0.00	0.25
0.65	1.00	0.97	0.19	0.00

Table D1.7: Edema Results - Performance metrics, confusion matrix, model weighting.

The Intrasession and Intersession models performed quite well on the edema data. MCC values above 0.7 are confirmed in the figures below. However, the between patient model performed very poorly. It classified everything as 'not edema' despite a large amount of actual edema voxels. The Interpatient feature weighting varies from the other models in that it weighs the ADC and T1 input higher.

Edema Results Intrasession, Intersession, Interpatient Models D1

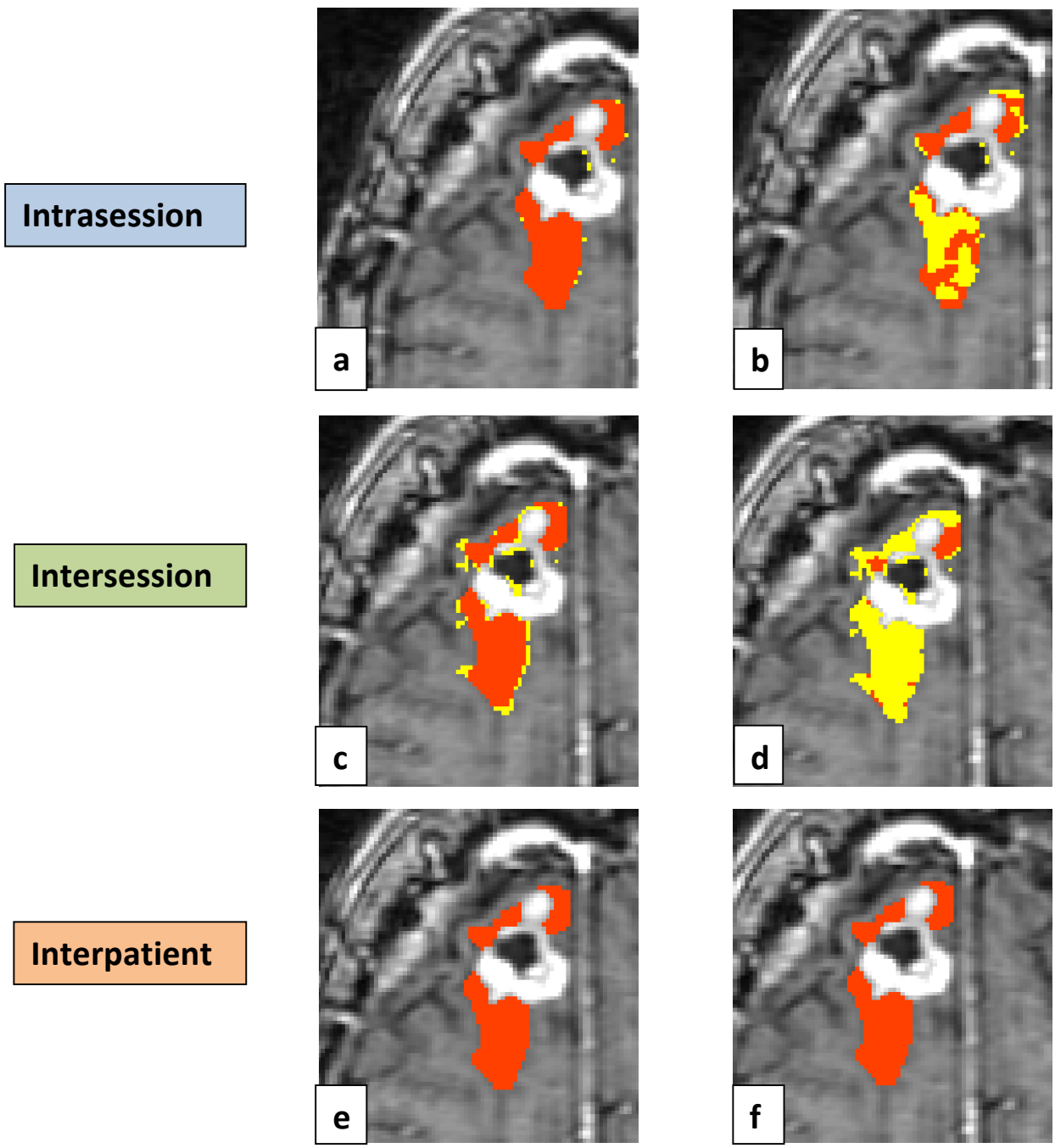


Figure D1.4: Comparing results of Intrasession, Intersession, and Interpatient edema models on patient D1. Red = Labels, yellow = SVM. a)Intrasession - Labels overlaid on SVM. b) Intrasession - SVM overlaid on Labels. c) Intersession – Labels over SVM. d) Intersession - SVM over Labels. e) Interpatient – Labels overlaid on SVM. f) Interpatient – SVM overlaid on Labels.

NCE Training Grid-Search D1

rbf	C	MCC	PPV	Sensitivity	Specificity	F1	F2
2	20	0.305	0.778	0.132	0.997	0.2258	0.1584
1	2	0.377	1.000	0.151	1.000	0.2622	0.1818
1	3	0.428	0.813	0.245	0.996	0.3768	0.2851
0.5	1	0.535	0.944	0.321	0.999	0.4789	0.3696
1	6	0.558	0.840	0.396	0.994	0.5384	0.4430
1	10	0.602	0.885	0.434	0.996	0.5823	0.4832
0.5	2	0.673	0.963	0.491	0.999	0.6500	0.5440
0.5	3	0.637	0.844	0.509	0.993	0.6353	0.5533
1	20	0.637	0.844	0.509	0.993	0.6353	0.5533
0.5	10	0.702	0.909	0.566	0.996	0.6976	0.6122
0.5	6	0.714	0.938	0.566	0.997	0.7059	0.6147

Table D1.8: Grid search sorting by F2-score for patient D1-NCE model. Optimal hyperparameters in red (Gamma=0.5, and C=6) were then used to train the full NCE model..

NCE Results Intrasession, Intersession, Interpatient Models D1

Model	Tissue	rbf	C	MCC	PPV	Sensitivity	Specificity
Intrasession	NCE	0.5	6	0.357	0.273	0.500	0.980
Intersession	NCE	1	20	0.307	0.444	0.222	0.996
Interpatient	NCE	1	6	-0.057	0.009	0.389	0.379

TP	FN	FP	TN
9	9	24	1203
4	14	5	1222
7	11	762	765

ADC	T1	T2	Hypoxia	rCBV
0.00	1.00	0.18	0.14	0.32
0.26	1.00	0.50	0.00	0.17
0.61	1.00	0.65	0.00	0.39

Table D1.9: NCE tumor Results - Performance metrics, confusion matrix, model weighting.

Although all three methods perform poorly on NCE data, the Interpatient model is by far the worst. It has almost no PPV as it fails to correctly classify negative voxels. This is confirmed in figure D1.5. It is clear that the Interpatient model is wrongly attempting to classify edema voxels as NCE.

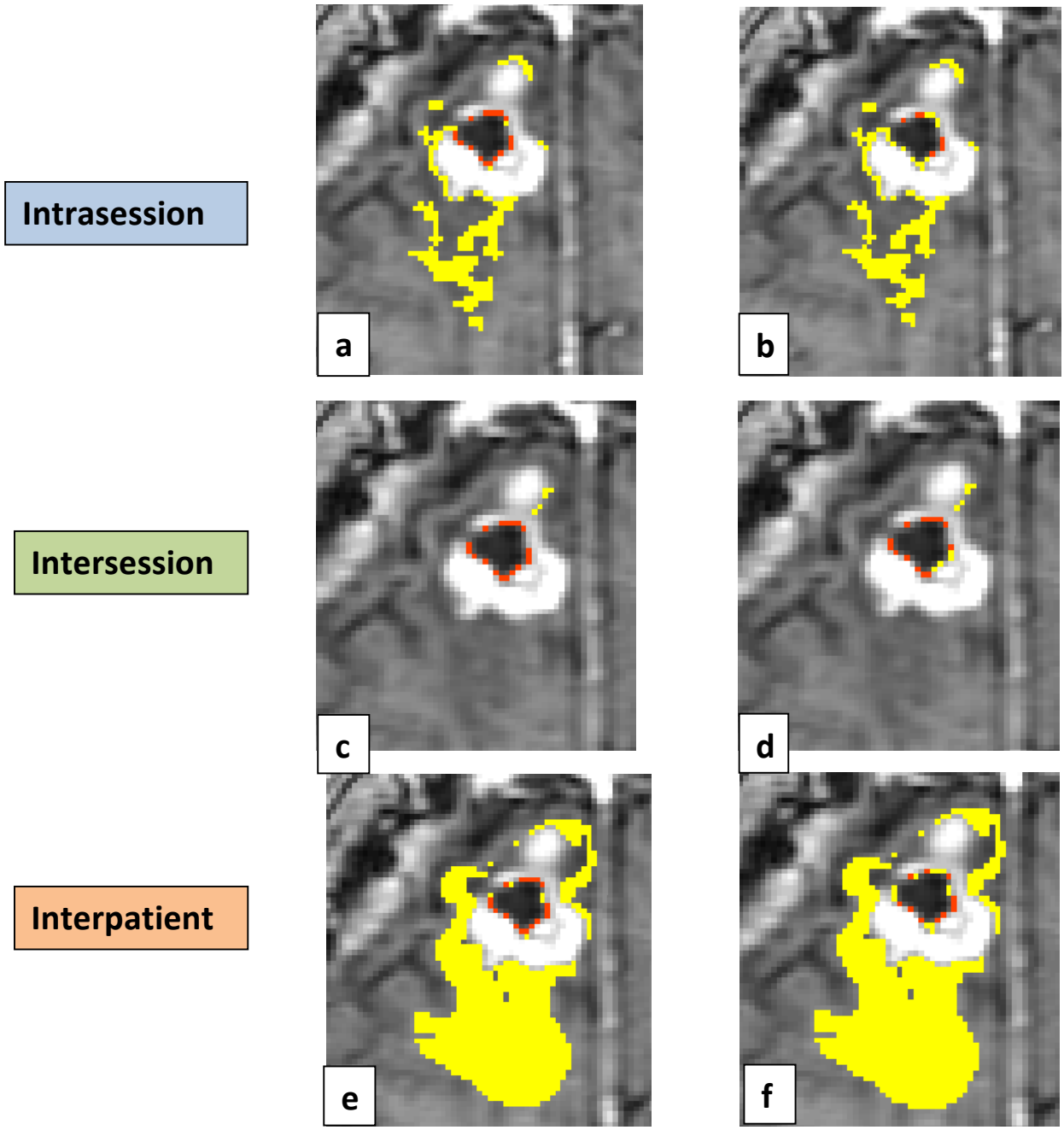


Figure D1.5: Comparing results of Intrasession, Intersession, and Interpatient NCE models on patient D1. Red = Labels, yellow = SVM. a)Intrasession - Labels overlaid on SVM. b) Intrasession - SVM overlaid on Labels. c) Intersession – Labels over SVM. d) Intersession - SVM over Labels. e) Interpatient – Labels overlaid on SVM. f) Interpatient – SVM overlaid on Labels.

Summary Intrasession, Intersession, Interpatient Models D1

Summary of MCCs for Comparison of 3 models

	Intrasession	Intersession	Interpatient
tumor	0.963	0.965	0.977
necrosis	0.726	0.708	0.753
edema	0.743	0.730	0.000
nce	0.357	0.307	-0.057

Table D1.10: Comparing MCCs of tissue classes across 3 different models: Intrasession, Intersession, and Interpatient.

When comparing all three methods, it is clear that the Interpatient model does not perform as well on edema and NCE data. Although it did perform well on tumor and necrosis voxels, the poor response to the remaining tissue types may be due to patient specific differences.

5. Discussion

The results from our four case studies have helped us to construct a streamlined approach to dealing with multiparametric MRI tumor data. The performance of each individual tissue classifier varies from patient to patient, and ranges from poor to excellent. However, there are also general trends that can be seen in our selected cases.

As can be evidenced from the case study of patient A, the Whole-brain normalization approach outperforms the Reduced ROI normalization. This result seems intuitive due to the fact that a reduced ROI is of arbitrary size, thereby affecting the distribution and ultimately the normalization of the voxel values. However, it is unclear in the literature how most researchers apply normalization to MRI data; specifically in the case of voxel-wise analysis where data set reduction is almost always necessary.

Compared to ROI based analysis methods, voxel-by voxel approaches require much more processing power, and as can be seen from some of our Single-slice vs. Full-brain model comparisons, there is often little advantage gained once a given training data quantity is reached. It is thereby useful to reduce the data, to overcome processing limitations. Therefore, we want to normalize the data to the entire brain for all five feature maps, at the same time. This is easily performed by Spider once the features are implemented into the input matrices. However, it is difficult to retain the position information for each voxel if we select our reduced data from the input matrices.

We successfully solved this problem using our Position Matrix code that allows us to extract the whole-brain normalized input matrices from Spider, then reduce them to a reasonably

sized ROI, and return them to Spider while saving the Position Matrix information to be reunited with the output predictions. This is an important step, because as we can see from various examples (Figure D1.3 and D1.5), there is much to be gained from visually inspecting our output data. In figure D1.3, it is obvious that the Intersession model performs worse. However, if we only looked at ROC metrics, this would not be clear (another instance where our MCC values achieve clarity). On the other hand, figure D1.5 shows an example where MCC, nor any derivation of the confusion matrix, would capture the trend, which is immediately obvious upon visual inspection of the data. It is clear that the Interpatient model is classifying all edema voxels as NCE voxels. The value of visual inspection must be preserved when possible.

The second case study serves to illustrate the difference between Quick, Single-slice and Full-brain models. The purpose of this test was to try and find the minimum amount of information needed to successfully characterize the tissue classes. Due to the extremely time-intensive requirements of tissue labeling, it is not very convenient for a radiologist to sit down and label an entire brain volume. To this note, we owe much appreciation to Dr. Turski and Dr. Hald for their generous contributions to the project. However, this is not a practical model for implementation in the clinic. As we can see from table B1.10, the Quick model does not match up to the performance of the Single-slice and Full-brain models. However, the Single-slice method does perform as well as the Full-brain alternative, and seems to include enough training data to fully characterize the remaining slices for testing.

This is an extremely promising result because this would require only a single representative slice to be labeled by the radiologist. This would be a 5 to 10 fold improvement in time savings over the Full-brain method. Table 5.1 shows confirmation of these findings in a separate case study using patient D.

Summary of MCCs for Comparison of 3 models Using Patient D

	quick	slice	full
tumor	0.865	0.963	0.989
necrosis	0.820	0.726	0.507
edema	0.596	0.743	0.692
nce	0.135	0.357	0.357

Table 5.1 Comparing MCCs of tissue classes across 3 different sized training models in Patient D: quick, single-slice, and full volume.

The third case study was applied to patient C in order to investigate the difference between a Two-class SVM and a Multi-class SVM. Table C1.10 shows the results of these comparisons favoring the Two-class model. Although the Multi-class model performs equally, albeit poorly with the NCE data, it does not respond as well to the necrosis and edema data. The Multi-class model does, however, perform slightly better on the tumor data (MCC_{Multi}=0.989 vs. MCC_{2class}=0.963). However, this is not enough to justify the use of the Multi-model. Although the Multi-model could ideally improve workflow efficiency by reducing the number of models, it is not a relevant improvement if it cannot perform as well as the standard Two-class model.

The failure of the Multi-model to perform as well is most likely due to the selection of the hyperparameters. While the Two-class model allows for individual tuning of the parameters for each separate tissue class model, the Multi-model requires us to apply a single set of hyperparameters to the entire model, resulting in the same hyperparameters for each tissue. This lack of flexibility is a trade-off of this particular Multi-model and gives us adequate motivation to look towards a different implementation of the Multi-model SVM in future studies. Although

the one-vs-rest method was selected because of available software choices and ease of comparison, it has not proved to be equal. There are however, more complicated implementations that allow for individual tuning of the tissue hyperparameters which could be investigated in future work. Another example of Two-class outperforming Multi-class can be seen in table 5.2 showing this comparison for patient E.

Summary of MCCs on Patient E

	Two-Class	Multi-Class
tumor	0.952	0.907
necrosis	0.977	0.973
edema	0.695	0.665
NCE	0.692	0.471

Table 5.2 Comparing MCCs of tissue classes using Two-Class vs. Multi-Class SVM on patient E.

The investigation of Intrasession, Intersession, and Interpatient models applied to patient D showed some interesting results. It is not surprising that the Interpatient model performed the worst out of the three. This can be seen in the summary table D1.10. The Interpatient model did perform well on tumor and necrosis data, but performed very poorly on the edema and nce data. This is perhaps to be expected, due to the heterogeneous nature of GBM. It seems likely that the precise mixture of complicated tissues such as edema, tumor, and non-enhancing tumor could be quite patient specific. However the ability to truly measure the between patient reliability would

require many more patients. For the purpose of this study, it is clear that the individual makeup of a single patient can often be characterized using the SVM approach, whether it be Intrasession or Intersession modeling.

This study also suffers from a few limitations worth noting. Firstly, although we have five different feature maps, they suffer from different starting origins, slice locations and resolutions. This leads to a significant loss of information in the coregistration and reslicing procedure. The interpolation methods used for this, although advanced, cannot fully recreate the original data as it would exist in a natively equivalent space. This is perhaps one of the reasons why the T1 and T2 features are often so strongly weighted. They have the highest resolution and therefore suffer less from interpolation error (especially T2 since all other scans are coregistered to native T2 space). Although the original design of this study was not intended for machine learning implementation, the data has served to show the promise of this technique under somewhat strained conditions.

Another challenge with this study is that the scan time for each patient was close to 90 minutes long. Coupled with the fact that these patients are often suffering from disease-related complications, this can lead to an increased amount of patient motion. Many of these scans can suffer greatly from motion artifacts, especially the $\Delta T2^*$ hypoxia mapping using a multi-echo EPI acquisition. Due to the distortion introduced by EPI, and the large time gap required for Carbogen equilibrium between the two subtraction images (12 minutes), there was often too much movement to achieve a reliable measure.

It is also important to note the difficulty of assigning accurate tissue labels to each voxel. Although our labels were agreed upon by two expert Neuroradiologists, the nature of GBM

tumors makes it sometimes difficult to discern certain tissues. Due to the infiltrative and heterogeneous nature of the disease, there is sometimes a mix of tissues in any given voxel. This is one of the reasons that the NCE and edema tissues were more difficult to classify. That being said, this also speaks to the strength of SVMs. While some machine-learning algorithms would be lead astray be a couple instances of mislabeled voxels, SVM is able to dismiss these outliers using the soft-margin concept. This is, of course, within limits, but the flexibility of SVM allows us to gain valuable information from sometimes noisy data.

Looking forward, an area of interest would be in comparing SVM methods to Bayesian networks. This could be interesting because of the fundamental differences between their construction methods. While the SVM can be considered more of a black-box, in that we do not directly affect the weightings of the classifier, it also boasts a reduced amount of user bias. The Bayesian networks, on the other hand, are inherently open to user involvement. This would offer the radiologist a chance to apply custom weightings to the network relationships. Although the downside of this method is that it is inherently less objective.

In conclusion, it has been shown that the SVM framework is well-suited for dealing with multi-parametric MRI data. We have shown that the SVM is a flexible machine-learning implementation that can respond well, despite limitations in the data. That being said, our future investigations could benefit greatly from common resolution and starting prescriptions between MRI scans. The major benefit of this study is in the streamlined methodology and handling of the SVM processing steps. Although SVM is often touted as a highly objective classifier, it is clear that user-biased can be introduced through the selection of the pre-processing steps and modeling decisions.

Our approach provides a case-by-case comparison of these methods. The results suggest that a flexible approach to dealing with large amounts of MRI data should include the proposed Whole-brain normalization, and can benefit from the added flexibility of multiple Two-class SVMs, rather than a more rigid Multi-class model. Also, we have shown that the selection of performance metrics has a significant effect on the training optimization. Although ROC has value in that it is widely used and generally understood, it has been shown to be inappropriate for model training optimization. It is more suited for final reporting of the testing performance, although this is also not completely informative. We have shown that the MCC metric serves as a valuable tool in simplifying the confusion matrix. Although it should be noted, that whenever possible, we suggest including the confusion matrix in the results, as well as a visual representation of the predictions. In doing this, we are able to retroactively calculate any number of performance metrics using the confusion matrix, thereby making our results more flexible to compare with other studies.

Perhaps the most important finding is that our Single-slice models performed equally to their Full-brain counterparts. This alone helps to make this work more clinically relevant. The time needed for radiologist involvement could now be limited to a single-slice. Considering that the purpose of this work is to ultimately assist the radiologist in analyzing the images, it is of great importance to optimize their time investment.