University of Alberta

PROGNOSIS OF GLIOBLASTOMA MULTIFORME USING TEXTURAL PROPERTIES ON MRI

by

Maysam Heydari

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Fall 2009
Edmonton, Alberta

Examining Committee

Russell Greiner, Computing Science

Osmar Zaïane, Computing Science

Matthew Brown, Psychiatry

Albert Murtha, Radiation Oncology

To my parents, Saeid and Shahnaz and my little sister, Mana,

For being there for me even when I was most bitter, frustrated and unpleasant,

And to my grandmother, Zahraa, and my uncle, Yahya, who passed away recently. May both rest in peace.

Abstract

This thesis addresses the challenge of prognosis, in terms of survival prediction, for patients with Glioblastoma Multiforme brain tumors. Glioblastoma is the most malignant brain tumor, which has a median survival time of no more than a year. Accurate assessment of prognostic factors is critical in deciding amongst different treatment options and in designing stratified clinical trials. This thesis is motivated by two observations. Firstly, clinicians often refer to properties of glioblastoma tumors based on magnetic resonance images when assessing prognosis. However, clinical data, along with histological and most recently, molecular and gene expression data, have been more widely and systematically studied and used in prognosis assessment than image based information. Secondly, patient survival times are often used along with clinical data to conduct population studies on brain tumor patients. Recursive Partitioning Analysis is typically used in these population studies. However, researchers validate and assess the predictive power of these models by measuring the statistical association between survival groups and survival times. In this thesis, we propose a learning approach that uses historical training data to produce a system that predicts patient survival. We introduce a classification model for predicting patient survival class, which uses texture based features extracted from magnetic resonance images as well as other patient properties. Our prognosis approach is novel as it is the first to use image-extracted textural characteristics of glioblastoma scans, in a classification model whose accuracy can be reliably validated by cross validation. We show that our approach is a promising new direction for prognosis in brain tumor patients.

Acknowledgements

I would like to thank my supervisor, **Dr. Russell Greiner**, for his seemingly unlimited patience, for his wisdom and helpful guidance throughout my time in graduate school. I would also like to thank all the members of the Brain Tumor Analysis Project, past and present, including **Dr. Jörg Sander** for his helpful and constructive criticism and advice, **Dr. Albert Murtha** for providing much medical insight, and for his suggestions on my project goals, **Dr. Matthew Brown** for his guidance and for informally acting as my second adviser, **Dr. Dana Cobzas** for her helpful ideas and discussions, **Bret Hoehn** for his technical assistance, running code, processing and providing MRI images and **Karteek Popuri** for his technical advice, encouragement and moral support.

Additionally, I would like to thank my undergraduate research adviser, **Dr. Guohui Lin** for his guidance and support for me in my first years of doing research and in applying to graduate school. Many thanks to all my committee members, including **Dr. Osmar Zaïane**, for agreeing to be on my defense committee and for their valuable advice on my dissertation. Last, but not least, I would like to thank all my friends, specially **Azad Shademan** and **Amir Massoud Farahmand**, for their encouragement as well as providing much appreciated moral support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis addresses the task of learning to predict the survival time of glioblastoma patients, using various features, including textural information extracted from their magnetic resonance (MR) images. We use MR images of patients diagnosed with glioblastoma, whose survival status is partially known. We develop a framework that extracts textural information from the MR images and uses them to predict the patient's survival outcome. Our goal is to show that textural properties about the glioblastoma tumor and the brain are predictive of patient survival outcome.

1.1 Motivation

Glioblastomas are the most difficult brain tumors to treat. Despite decades of research, advanced and aggressive treatments, scientist and clinicians still cannot treat glioblastomas very effectively. Hence, prognosis of glioblastomas is generally dismal. Currently, the most common methods used to predict prognosis are based on survival analysis of clinical or pathological features. However, the predictive accuracy of these methods is limited. It is well known that older patients who are diagnosed with glioblastomas have shorter survival times. But how much longer will old patient A live compared to old patient B? And why do some younger patients defy the trend and have short survival times? There are obviously more factors that influence survival of which we are unaware. Therefore, it is worth exploring other options and why not start by systematically analyzing the very appearances of the glioblastoma tumors on medical images? The intrinsic appearance of a glioblastoma on a magnetic resonance image should say something about how well-behaved or aggressive the tumor is or will be in the not so distant future. A very useful way to characterize the appearance of the tumors on MRI is to analyze their texture. When an oncologist describes a tumor from its MR image, he/she may talk about the irregularity of the tumor borders or heterogeneity of the tumor itself. He/she may describe the brain tissue surrounding the tumor as compressed with a smudged appearance. It is possible to measure these intuitive properties using texture analysis and see how they correlate with how a glioblastoma behaves over the course of time leading to the patient's inevitable death. We take up the challenge of prognosis for glioblastomas using textural

information.

Others have already applied textural information from medical images to help prognose other diseases. Yogesan et al. apply texture extraction methods to light microscopy images of nuclei to classify prostate cancer patients into two prognostic groups; good versus poor [66]. Geisler et al. apply texture analysis to optical images to determine predictive factors in prognosis of endometrial cancer [17]. Weyn et al. apply texture analysis on histological samples for survival prediction of malignant mesothelioma [63]. There have been no attempts, as of yet, to apply textural information towards prognosis of glioblastomas. Most texture analysis frameworks developed regarding brain tumors deal with tumor segmentation or tissue characterization. Therefore, our work is the beginning of a new direction in the battle with glioblastomas. The main challenge is that there is no benchmark to compare our prediction model with. The medical community mainly uses prediction models for designing stratified clinical trials and thus has a different understanding of survival prediction, which is not compatible with our definition of prediction. Therefore, while prediction models may not be meaningful at this time, our survival prediction model demonstrates that textural information from MRI are of prognostic value and can be used, along with other prognostic factors, for survival prediction and also to help in stratification of clinical trials.

1.2 Thesis Outline

In Chapter 2, we first introduce glioblstoma multiformes and describe their biological and appearance-based properties (Section 2.1). Then we discuss the common approaches to the prognosis of glioblastomas in the medical community (Section 2.1.1 and Section 2.2). Finally, we discuss what constitutes texture, common approaches to texture analysis and their application on magnetic resonance images (Section 2.3).

In Chapter 3, we motivate our survival prediction framework (Section 3.1). Then, before we describe it in detail, we will give an overview of the full framework (Section 3.2) and introduce the main ideas behind it (Section 3.2.1, Section 3.2.2 and Section 3.2.3). Then we will describe our MRI data that we will be using to test the survival predictability of our framework (Section 3.3).

In Chapter 4, we describe our framework in detail, introducing the methods that we use in texture extraction (Section 4.1), and how we combine the texture features (Section 4.2).

In Chapter 5, we explain our experimental setup (Section 5.1 and Section 5.2) and discuss the results (Section 5.3, Section 5.4 and Section 5.5). Finally, we conclude the thesis in Chapter 6, with discussing challenges we encountered and possible future work (Section 6.1), and then summarizing our contribution and ending with some concluding remarks(Section 6.2).

Chapter 2

Background and Related Work

Prognosis is the prospect of survival and recovery from a disease as anticipated from the usual course of that disease or indicated by special features of the case [1]. In clinical trials for new treatments, assessment of prognosis is used to help stratify patients into homogeneous risk groups based on prognostic factors. Therefore, prognosis assessment is essential in understanding the effects of new treatments. There are different ways to conduct prognostic analysis. Statistical surveys of populations help clinicians estimate a likely outcome for a patient based on groups of other patients in the population with common clinical and diagnostic characteristics. Some studies categorize patients based on *survival rates*. In other words, patients in a population are grouped based on the number of years they survived after being diagnosed with the disease. Another way to characterize groups of patients is to refer to *progression-free survival*, which is time, after treatment, one lives without the disease recurring or progressing any further. Such population studies are designed to help in stratification of clinical trials and not for survival prediction. Also, since every patient is unique with different responses to the disease and the consequent treatments, clinicians cannot accurately predict individual outcomes based on population studies, as such populations based on a few diagnostic and prognostic characteristics are still quite heterogeneous. A patient's prognosis may even change over time based on their responses to treatment. Moreover, statistical surveys based on populations may also change depending on what populations and in what span of time the surveys were conducted and which clinical parameters were included.

Other than population surveys, there are also specific *markers*, discovered in the past decade or so that can help distinguish subclasses of a certain disease based on their correlation with the disease outcome. Genetic mutations in tumors are one such example. The prognostic value of these markers becomes established when their presence is associated with better or worse prognosis. Some genetic mutations, or combinations of them, present in a tumor can lead to better responses to certain treatments, and thus lead to better prognosis. In this thesis we address the problem of prognosis in terms of survival in a certain class of high grade brain tumors called *glioblastoma multiforme*. In the following sections, we present some background information on these tumors on their biological characteristics and the challenges they present to clinicians and researchers. Then we discuss prog-

nostic factors used by clinicians and in research studies. Such factors include populations surveys and molecular and genetic markers, which are described in more detail.

2.1 Glioblastoma Multiforme

Glioblastoma Multiforme (or just glioblastoma) is a malignant type of a glioma brain tumor, which occurs in people of all age groups, but is mostly prevalent in ages 65 to 75 [49]. *Gliomas* grow from *glial* cells, which support nerve cells in the central nervous system by providing nutrition and protection. Glioblastomas are the most common gliomas and the most aggressive brain tumors in general. These tumors typically rise in the deep white matter but soon infiltrate the gray matter and other structures as well [49] . They are highly invasive to the neighboring tissues. A glioblastoma often infiltrates the adjacent hemisphere of the brain, invading through the corpus callosum, which connects the two cerebral hemispheres. When this happens, it produces a symmetric appearance, which resembles a butterfly and so is commonly referred to as a *butterfly glioma* [49]. The growth rate of glioblastomas is so high that they tend to deplete their blood supply resulting in central *necrotic* regions, which contain of only dead cells. In many cases, there is also a significant amount of fluid accumulation, called *vasogenic edema*, around the tumor. A high growth rate together with the surrounding edema can cause *mass effect*, compressing the brain tissue against the *cranium* (i.e. skull) leading to a condition called *intracranial pressure*. If not treated through surgical removal of the tumor mass, the patient will soon die because the increased mass effect limits the blood supply to the neighboring brain tissue.

The four-tiered grading system of the World Health Organization (WHO) categorizes glioblastomas as grade IV gliomas, the most malignant, based on pathological, clinical and prognostic characteristics[33]. These tumors are pathologically composed of poorly differentiated (i.e. *anaplastic*) and heterogeneous cancerous cells. Well-differentiated cells are ones that grow and specialize normally in healthy tissues. But in glioblastomas, the cells have lost any resemblance to their originating normal and specialized cells. This loss of differentiation is called *anaplasia* and is indicative of biological aggressiveness [39]. According to the WHO grading system, the most common histopathological characteristics of glioblastomas are the following [49, 34]:

Cellular pleomorphism is high variability in size and shape of tumor cells. **Dense cellularity** is high density of cells in tumor tissue. Variable **mitotic activity**, which is the degree of population cell growth, paired with high cellularity and cellular pleomorphism are indicative of abnormally high and unregulated rate of cell growth and division in glioblastomas. **Neovascularity**, which is the presence of blood vessels in tissues that do not normally contain them, is also prevalent in glioblastomas. But, the most distinctive characteristic of glioblastomas is the microscopic presence of regions with **pseudopalisading necrosis**, which are regions of dead cells surrounded by a dense ring of cancerous cells.

Glioblastomas are classified into two subtypes based on their histopathological grading when

they first occur. Primary glioblastomas, seen mostly in older patients, are malignant and aggressive from the point they are diagnosed without any previous history of lower grade gliomas. Secondary glioblastomas are initially diagnosed as lower grade and less malignant gliomas. But they return as malignant and highly invasive tumors after 5 to 10 years of the original diagnosis and subsequent treatments [34]. The genetic profile, in terms of mutations, of these two subtypes is somewhat different and there has been debate on whether they should be treated as separate diseases [34]. But once the patient is diagnosed with a glioblastoma, regardless of the clinical history, the degree of aggressiveness and malignancy is the same for both subtypes.

The mutation profile in the genetic makeup of glioblastomas is very complex. There are always a variety of genetic abnormalities present at the same time [49]. *Oncogenes* are mutated genes that can turn a normal cell to a cancerous one. *Tumor-suppressor* genes are ones that control the normal growth of cells and prevent abnormal growth. Both genetic mutations caused by oncogenes and malfunctioning of tumor-suppressor genes such as p53 can be present in glioblastomas. Malfunctioning of tumor-suppressor genes on chromosome 10 is prevalent in cases of glioblastomas and not very common in other forms of cancer [49]. Abnormally high expression or mutation of epidermal growth factor receptor, which is normally involved in promoting healthy cell growth and division, is also present in many cases of primary glioblastomas [49, 62].

Magnetic Resonance Imaging (MRI) is a common medical imaging technology used to visualize a patient's brain. The most common MRI modalities used to assess glioblastomas are FLuid Attenuated Inversion Recovery (FLAIR), T1 and T2-weighted modalities. T1-weighted modalities highlight fat tissue in the brain and FLAIR and T2-weighted modalities highlight tissue with higher concentration of water. T1-weighted scans are usually followed by post-contrast T1-weighted scans taken after the patient is injected with a contrast agent, which is usually gadolinium. The contrast agent enables clinicians to locate areas of contrast enhancement on the brain scans, which is an increased T1 signal intensity in post-contrast T1-weighted scans. Contrast enhancement is indicative of disruption in the blood-brain barrier. This abnormality is typically seen in brain tumors and other brain diseases. The most common appearance characteristic of a glioblastoma on MRI is a heterogeneous mass with central regions of necrosis or hemorrhage and non-uniform borders surrounded by extensive vasogenic edema [49] (Figure 2.1). Necrosis is displayed on MRI as a region of low T1 signal intensity (dark) and high T2 and FLAIR signal (bright), located within the tumor, and is typically surrounded by a contrast enhancing ring. This enhancing ring, as seen on post-contrast T1 scans, is typically thick and has a wavy and irregular appearance with a shaggy inner margin [57]. The enhancing rim is usually thicker on the side of the tumor closer to the cortical surface of the brain, rather than the deeper white matter [57]. The necrotic region itself is not contrast enhancing. The enhancing ring around the necrotic region has the highest concentration of neovascularity, which is an indication of blood-brain barrier disruption [57], which is the reason it enhances, as mentioned earlier. Edema is displayed on MRI as a region of high T2 and FLAIR signal intensity

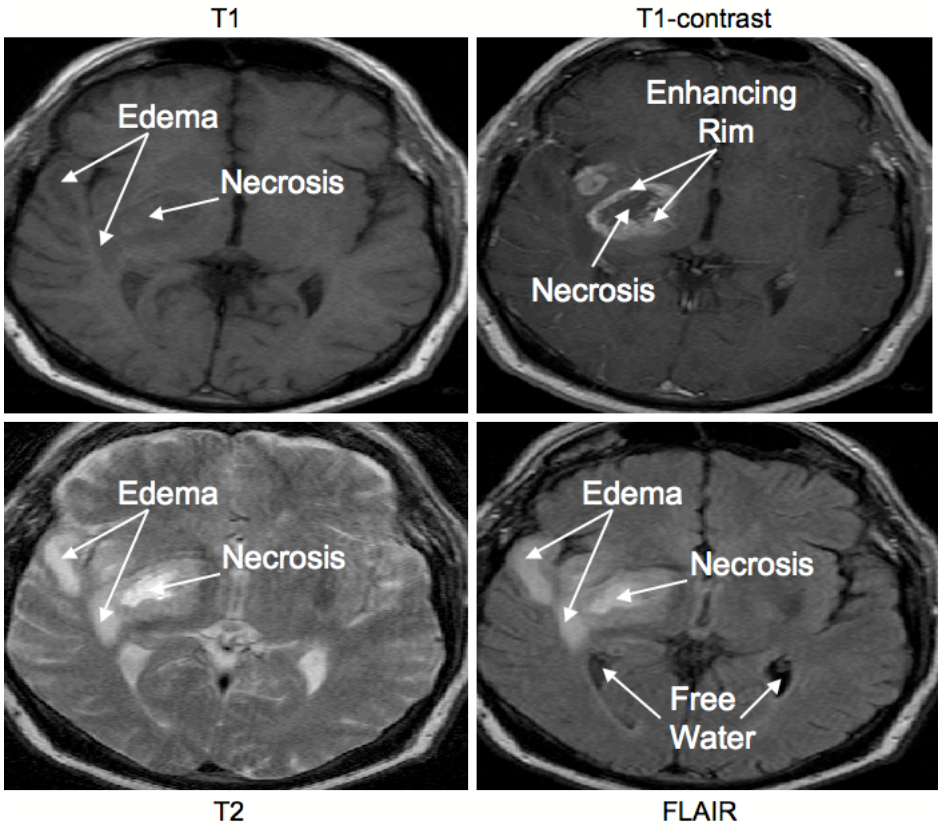


Figure 2.1: MRI scans of a glioblastoma tumor. A T1-weighted image, the T1-weighted image after the injection of contrast agent, a T2-weighted image and the corresponding FLAIR image. Enhancing rim can be seen in T1-contrast. Edema appears dark in T1 and T1-contrast, and bright in FLAIR and T2. Edema appears as bright as the ventricles in T2 but in FLAIR, the ventricles appear dark because they contain free water. Necrosis appears less bright in T1 and T1-contrast, and bright in FLAIR and T2.

(bright) and low T1 signal (dark) (Figure 2.1). The high T2 signal of vasogenic edema is at the same level of intensity as the T2 signal of the fluid filling the ventricles. However, the ventricles appear dark in FLAIR images because they contain free moving water, which has a low signal on FLAIR images. Mass effect can be seen on FLAIR, T1 and T2-weighted scans as areas where the tissue adjacent to the tumor has been deformed, smudged or pushed to one side as a result of pressure from the edema. In general, edema, border definition and tumor heterogeneity are best observed on FLAIR and T2-weighted images.

2.1.1 Prognosis

Due to their high degrees of malignancy, glioblastomas cannot be cured. Instead, they are treated to prolong survival of the patients but with very limited success. Treatments typically include combinations of surgery, radiotherapy and chemotherapy [62]. The high invasive and aggressive nature of glioblastomas makes treatment very difficult. Additionally, because these heterogeneous tumors are

composed of a mixture of different tumor cell types, treatments may destroy certain types of tumor cells but allow other types to survive and result in recurrence of the tumor. As a result, glioblastomas are highly recurrent tumors. As these tumors return, they become more devastating to the patient, as they infiltrate and destroy more brain tissue, eventually leading to death. Researchers have discovered many prognostic factors pertaining to survival in glioblastoma patients. These factors range from clinical data to molecular, genetic, histopathological and image-based properties. However, some of these findings appear to be conflicting and others require more confirmation from other sources. For instance, the presence of necrosis in glioblastoma with an *oligodendroglioma* component (a lower grade glioma) is associated with poor prognosis. But it is debatable whether these tumors, in general, have a more favorable prognosis than standard glioblastomas [33]. Pope et al. , on the other hand, report that the existence of a large oligodendroglioma component is associated with favorable prognosis [46].

The most widely confirmed predictive factors for prognosis are the age at which the patient is diagnosed, and the Karnofsky Performance Status (KPS) test [62]. The purpose of the KPS test is to assess the patient's general health and is a score that is a multiple of 10 from 0 (death) to 100 (full health). Older glioblastoma patients typically have very poor prognosis. Additionally, they are less tolerant to certain critical treatments than younger patients. Younger age, moreover, is typically associated with more favorable prognosis. According to statistical reports from the Neuro-Oncology program at the University of California at Los Angeles, the median survival times for young glioblastoma patients in the age ranges of 20-35 and 35-50 are 971 and 714 days respectively [60]. This is in contrast with the much shorter median survival times of older patients in the age ranges 50-70 and 70-100, which are 461 and 362 days respectively [60].

Regardless of age, however, the prognosis for glioblastoma is poor despite two decades of research and clinical trials and the availability of aggressive treatments. The reported median survivals for glioblastoma patients of all ages range from 9, 12 to 15 months at most [34, 39, 42, 62, 28]. Despite low rates of survival, there is a small percentage of glioblastoma patients with mysteriously long survival times. Krex et al. analyzed the largest group of long-term survival patients to uncover the prognosis factors that lead to such long survivals [28]. Their study confirms the general belief that younger age at diagnosis and high KPS scores are the most important factors indicative of a favorable prognosis. They also found that a histological subtype of glioblastoma called *giant cell glioblastoma* is over-represented in their group of long-term survival patients. Their study also confirms a major finding that MGMT methylation is strongly associated with long term survival. Martinez et al. also conducted a long-term survival study and confirmed the prognostic power of MGMT methylation in glioblastomas [36]. MGMT is a DNA-repair gene whose high level of activity in cancer cells is believed to cause resistance to chemotherapy. Deactivation of MGMT through methylation, which is a type of chemical DNA modification, reduces DNA repair in cancer cells. Methylation of MGMT is more prevalent among patients with long-term survivals of 3 or more

years [36]. Hegi et al. have in fact used statistical studies to demonstrate that, regardless of type of treatment received, glioblastoma patients with methylated MGMT survive significantly longer than patients with unmethylated MGMT [22]. Additionally, patients with methylated MGMT, who receive both radiotherapy and chemotherapy with agent temozolomide, survive even longer than patients with this genetic advantage but without receiving both treatments simultaneously.

Tumors are genetically complex and highly individualized per patient. Population and clinical studies are incapable of bringing these highly individual genetic characteristics to light [39]. Also, understanding the complex gene expression patterns in tumors can provide more information on the underlying biology than do individual genes in identifying molecular markers for prognosis [40]. Recently, there have been many studies based on microarray analysis to uncover complex genetic prognostic factors in glioblastoma [41, 16, 7, 50, 65, 5, 30]. Microarray analysis studies thousands of gene expressions simultaneously, which can provide new prognostic information by helping researchers distinguish between morphologically and pathologically similar glioblastomas. Pattern recognition techniques applied to microarray analysis help better understand the complex interplay of genes in glioblastoma patients. We describe below how this has enabled researchers to find new predictive markers associated with prognosis.

Nutt et al. use gene expression features to classify high grade gliomas (anaplastic oigodendroglioma versus glioblastomas) [41]. They found that survival correlation to gene-expression based classification is stronger than to pathology-based classification. They identify 20 genes out of thousands, whose expression strongly correlates with survival. Freije et al. use gene expression features to classify high grade glioma patients into separate groups based on patient survivals[16]. They use hierarchical clustering to group tumors based on over-expressed genes. They distinguish 44 genes out of thousands, based on which 4 subtypes of high grade gliomas are determined. These subtypes highly correlate with patient survival times. Among the groups with the poorest survival times, the genetic profiles of the tumors were defined by an over-expression of a set of extra-cellular matrix related genes. This, the authors reason, could be a cause of extensive local invasion, which leads to poorer prognosis for patients with these tumors. The most favorable prognosis was associated with glioblastomas with over-expressed genes that are involved in neurogenesis (neuronal development). The authors use the name, ProNeural, for this group. The authors of this study later make an intriguing discovery in a more recent study, that this subtype of glioblastomas is most commonly diagnosed in younger patients [30]. In fact, they demonstrate that the genetic profile of ProNeural glioblastomas is a stronger predictor of outcome than age. They moreover demonstrate that among patients with ProNeural glioblastomas, age is not a strong predictor of survival and that among the non-ProNeural patients, even the young patients had short survival times. They conclude that the survival advantage of younger patients is due to the more prevalent ProNeural types of glioblastomas among them.

Rich et al. use gene-expression profiling to identify three genes that are responsible for cellu-

lar motility and thus tumor migration, leading to poor patient survivals [50]. Yamanaka et al. use multivariate analysis on microarrays to identify 21 genes whose expression can predict patient survival [65]. Carlson et al. discover that when there is little to no amount of edema, the expression of *Vascular Endothelial Growth Factor* (VEFG) is related to patient survival, with high expression levels associated with low survival rates [5]. VEGF refers to a group of signaling proteins involved in *angiogenesis*, the growth of blood vessels from existing ones. On the other hand, with extensive amount of edema present, the overexpression of the NPTX2 gene was associated with low survival rates [5]. According to Chakravarti et al. , the activation of the phosphatidylinositol 3-kinase (PI3K) pathway in glioblastomas is believed to lead to poor response to radiation and thus to a shorter survival time in patients [7]. The PI3K is involved in the regulation of cell survival among other cellular functions. Members of this pathway are known to suppress *apoptosis*, which is a form of cell "suicide" that triggers when a cell becomes abnormal. This lack of programmed cell death in cancerous cells makes this type of genetic mutation challenging to treat.

MRI is routinely used in diagnosing brain tumors. We believe that effectively incorporating image-based features along with other prognostic features can help in prognosis. There have been studies that link appearance characteristics of glioblastomas to patient survivals. In these studies, experts visually inspect MRI scans and manually assign scores to pre-defined features. For example, the severity of vasogenic edema may receive a score of 0 for no edema and a score of 2 for an extensive amount of edema causing significant mass effect. Hammoud et al. found that the amount of tumor necrosis on pre-surgery images is the strongest prognostic factor in a group of patients with median age of 50 [20]. Hence, they found that smaller necrotic region is associated with favorable prognosis. They also found that minimal contrast enhancement is associated with favorable prognosis. The prognostic value of necrosis was confirmed by Pierallini et al. [44], who found that low necrosis to tumor mass ratio is associated with favorable prognosis. This finding is in accordance with the general understanding that the presence of extensive necrosis is indicative of high aggression. Pope et al. discovered that presence of non-enhancing tumor and extent of edema are of high prognostic value [46]. A non-enhancing tumor is one that has a region of solid cancerous tissue that does not enhance on a post-contrast T1 scan. These non-enhancing regions show bright on T2-weighted scans, but with a lower T2 signal than edema. According to the authors, the presence of a non-enhancing tumor is associated with favorable prognosis especially in older patients. They also confirm the belief that the presence of an increased amount of edema is indicative of poor prognosis. Multifocality and the presence of satellites were found to be of some prognostic value when coupled with other factors (multifocality refers to the existence of lesions that are not connected to the main lesion or its surrounding edema and satellites are lesions that are not connected to the main tumor lesion but within the surrounding abnormality or edema). They show through multivariate analysis that glioblastoma patients with non-enhancing tumors but without edema, satellites or multifocality had significantly more favorable prognosis than patients without non-enhancing tumors but with

edema and either of satellites or multifocality. They also find that the existence of a large oligodendroglioma (a lower grade glioma) component within the glioblastoma is associated with favorable prognosis [46].

2.2 Population Studies

What we have discussed in the previous section is by no means an exhaustive survey of all the recent findings in gene-expression profiling pertaining to prognosis of glioblastomas. But it demonstrates the immense interest developed recently in the use of gene-expression analysis to guide prognosis and treatment in glioblastomas and cancer in general. However, despite numerous studies based on gene-expression profiling, clinicians still rely on statistical studies based on populations as the main practical tool in prognosis of glioblastomas. Many studies based on gene-expression profiling are complex and not quite practical for use in prognosis. The concept of profiling gene-expressions is relatively new in prognosis of glioblastomas and thus many of the findings are not widely confirmed in the literature. Moreover, the studies are mostly based on smaller datasets of patients, a factor which reduces their level of reliability. Population studies, on the other hand, are based on much simpler prognostic factors and are conducted on much larger datasets.

The study of patient survival in populations using statistical tools gives rise to *survival analysis* in biomedical research. In this context, survival analysis is concerned with survival data, which contains information about patients on time to a certain event. In the case of survival for glioblastoma patients, the survival data contains, for each patient, the time from diagnosis to death or to the last follow-up. What makes survival data distinct is the fact that the information on survival (whether patient died and if so, the time of death) may not be complete for all patients in the data. Often, at the time when data collection is complete, some patients are still alive due to longer survival times. Also, patients in a clinical study do not all enter the study at the same time. Some patients may enter the study closer to the end of the study and once data collection is complete, these patients may still be alive. Data on patients whose time of death is not known (but the time of their last follow-up is known) are called *censored observations* [12]. Due to the presence of censored observations in survival analysis, specialized statistical tools have been developed, which have been consequently used in the study of prognosis in glioblastomas.

Curran et al. originally used Recursive Partitioning Analysis (RPA) on a dataset of 1578 glioma patients in the Radiation Therapy Oncology Group (RTOG) clinical trials to group patients into prognostic classes, which are significantly distinct from each other in terms of survival times [11]. RPA[1] is a nonparametric partitioning method that is used to study the full interaction between prognostic factors and patient survival. The RPA method has been developed by Ciampi et al. for use on censored survival data [9]. RPA searches for the best possible value for each prognostic variable that splits the data in a way that is most statistically significant with respect to survival times. Ini-

[1] Otherwise known as RECursive Partition and AMalgamation (RECPAM) [9]

tially, the entire data is considered. After determining a significant split, which separates the patients into statistically different survival classes, the process is repeated recursively on the data associated with each split. RPA repeats this process, producing subsequent splits, until no further significant splits can be made. This produces a binary decision tree, whose primary decision node is a split of a prognostic variable, which splits the entire data into two survival groups. The subsequent decision nodes are cut points of other prognostic variables, which, in turn, split the "'current" sets into smaller sets. This process terminates either when each resulting node's associated subset of patients contains a minimum pre-set number of patients, or when no more statistically significant splits can be made. Finally, when the tree formation is complete, any two leaf nodes, whose associated subsets of patients are statistically similar with respect to survival, are *amalgamated* or combined to form one survival class (hence, the tree becomes a directed acyclic graph). This final step is useful: even though the RPA method determines splits in a way that left and right terminal nodes represent significantly distinct subsets of the data, it is still possible that nodes from different parents are statistically similar.

Figure 2.2 shows the RPA decision tree developed by Curran et al. . This tree specifies six RTOG-RPA survival classes for glioma patients, where class I, with median survival time of 58.6 months, is the group of patients with the most favorable prognosis and class VI, with median survival time of 4.6 months, is the group with the poorest prognosis. Also, age is found to be the most significant pre-treatment prognostic factor in glioma patients, where patients older than 50 have significantly shorter survival times. In glioblastoma patients, after age, KPS and extent of surgery are found to be the most significant prognostic factors. The authors note that despite many advances in treatment of glioblastomas, at the time of their study (which was conduced in 1993), pre-treatment prognostic factors such as age and KPS affect survival more than treatment variations [11]. The full list of prognostic factors used by Curran et al. is as follows:

- Patient-specific pre-treatment: age, sex, race, duration of symptoms, neurologic functional class and KPS.
- Treatment-based: extent of surgery, total radiotherapy dose and fraction size, interfraction interval and type of agents used for chemotherapy.
- Tumor-specific: location, size and histology of tumor. Tumor size is the only image-based factor used in this study.

An important objective of the study by Curran et al. and many studies that follow below, aside from determining the significance of various prognostic factors, is to determine significantly different patient risk groups to help in the design and stratification of clinical trials. The RPA method used by Curran et al. allows them to fully utilize the interaction between the different prognostic factors to partition the patients into statistically distinct risk groups [11]. Being able to determine

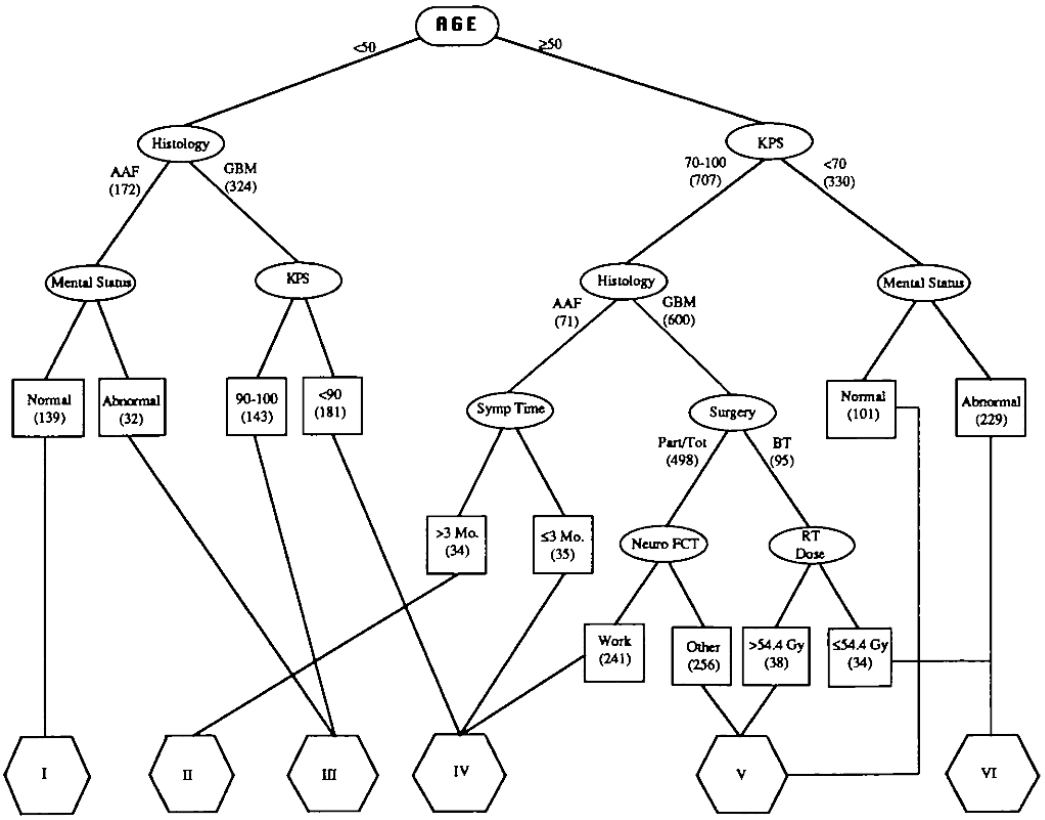


Figure 2.2: Recursive partitioning decision tree by Curran et al. [11]. Age is the primary prognostic factor, splitting the data at 50. In total, 12 terminal nodes are obtained. Terminal nodes with similar survival profiles are *amalgamated* producing a total of six survival classes. Classes III, IV, V and VI result from amalgamation. KPS = Karnofsky Performance Status, BT = biopsy, Neuro FCT = neurological function, RT = Radiotherapy, Symp Time = Duration of symptom signs, GBM = glioblastoma. Figure re-produced with the permission of Oxford University Press.

risk groups, which in turn helps in stratification of clinical trials, is the reason the RPA method has become the preferred statistical analysis method over another prominent method, the Cox proportional hazards model, which is used to determine prognostic factors in survival data by estimating hazard ratios [58].

To determine whether two given splits are statistically significant with respect to survival, many authors including Curran et al. , use the Kaplan-Meier Product-Limit method to estimate survival functions for each survival class [12], and then use a statistical significance test such as the Wilcoxon rank sum test or the logrank test to confirm that each survival class is significantly distinct from the others [12]. Kaplan-Meier estimators [25] are widely used in the study of prognosis in populations to estimate survival functions in survival data containing censored observations. For example, for each survival class in Figure 2.2, the Kaplan-Meier estimator produces a function plot, which can be seen in Figure 2.3. The function plots for all survival classes start at full patient participation at the beginning of the study. As months pass, each class loses patients, due to either death or censored

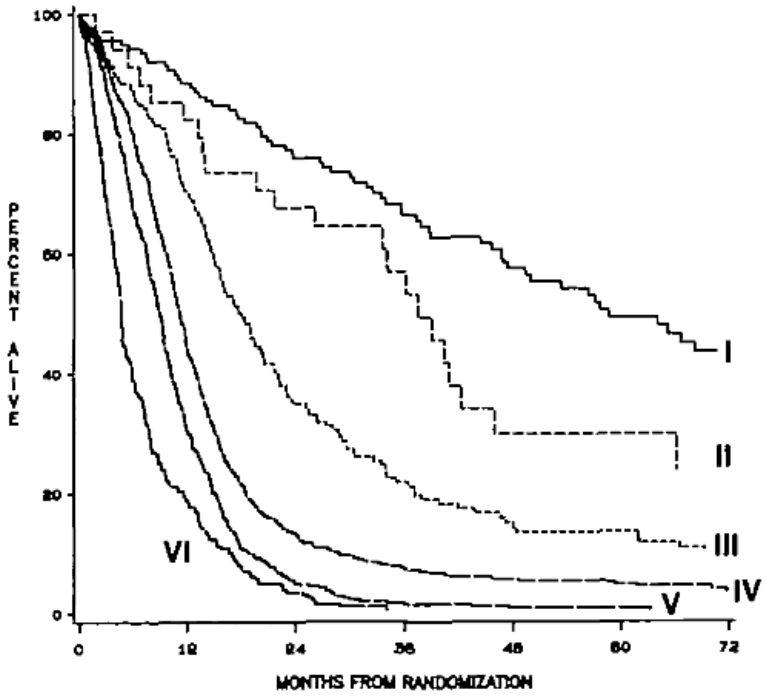


Figure 2.3: Kaplan-Meier plots for RTOG-RPA survival classes obtained by Curran et al. [11]. The Kaplan-Meier estimator for each survival class is the probability that a patient in that class survives at a given time. Figure re-produced with the permission of Oxford University Press.

observations. This loss of patients, due to both causes, is reflected in the plot of a survival class that changes *only* at times when a death has occurred. In other words, every time a death occurs, the total number of patient losses (due to death and censored observations) since the *last* time a death occurred, are explicitly shown in the Kaplan-Meier survival plot.

The Kaplan-Meier estimate of a survival function is expressed as:

$$\text{KM}(t) = \prod_{x \leq t} (1 - \frac{d(x)}{n(x)}) \tag{2.1}$$

where $d(x)$ is the the number of deaths at time x and $n(x)$ is the number of patients still in the study prior to time x. Note that in the function, as stated earlier, only events at time x, where a death occurs (i.e. $d(x) > 0$), contribute to the final product [2].

Due to the presence of censored observations, the logrank test is a preferred test of statistical significance between the Kaplan-Meier estimates for survival classes. One purpose behind using Kaplan-Meier estimates along with statistical significance tests is to help determine splits in recursive partitioning methods such as the RPA. However, many authors also use Kaplan-Meier estimators to demonstrate and confirm the predictive power of the obtained survival classes. To confirm that survival classes obtained from a partitioning method have strong predictive power for use in prognosis, authors typically try to demonstrate that the survival classes are significantly different at high levels of significance when tested on new datasets. In fact, Scott et al. validate the predictive

power of Curran et al. 's survival classes on a different RTOG dataset of 615 patients [54]. Scott et al. use the prognostic criteria in the decision tree obtained by Curran et al. to partition their new set of patients into survival classes. Then they use Kaplan-Meier estimators to demonstrate that the RTOG-RPA survival classes have strong predictive power by performing significance tests with significance levels of 0.0001 (of survival class being different) between most of their survival classes [54]. Many of the studies, mentioned in the previous section, use Kaplan-Meier estimators and plots with statistical significance testing such as the logrank test in order to demonstrate their model's predictive power [41, 16, 5, 30].

Shaw et al. apply the RPA method to a new dataset of 1672 patients diagnosed with only glioblastomas [55], as opposed to the RTOG-RPA classes, which were based on glioma patients [11]. This study simplifies the RTOG-RPA classes from the previous papers and combines certain survival classes from the RTOG-RPA classes to obtain only three survival classes: RPA class III with median survival time of 17.1 months, class IV with median survival time of 11.2 months, and classes V and VI combined with median survival time of 7.5 months. They also reduce the number of significant prognostic factors to only age, KPS, extent of surgery and neurological function (a measure of patient's ability to work). Another study, Mirimanoff et al. , uses 573 glioblastoma patients from EORTC[2] and NCIC[3] clinical trials to verify the predictive power of three of the RTOG-RPA classes (III to V) [38]. They also demonstrate that the combination of radiotherapy and temozolomide agent may have a more positive effect on patient survival than the use of radiotherapy alone. Note that as mentioned in the previous section, glioblastoma patients with methylated MGMT highly benefit from the use of both treatments [22].

Shakima et al. use a smaller dateset of 86 glioblastoma patients that fall in the RTOG-RPA classes V and VI and determine small pre-surgical tumor sizes as the predictor for longer survival times [56] in their dataset. Pichlmeier et al. use a dataset of 243 glioblastoma patients to investigate the influence of the extent of tumor resection on survival times in addition to verifying the predictive power of the RTOG-RPA survival classes [43] using Kaplan-Meier estimates with significance testing. They discover that in patients who fall in RTOG-RPA classes IV and V, complete resection in which all contrast-enhancing tumor region is surgically removed, improves survival times significantly more than incomplete resection.

In conclusion, despite many advances in cancer treatment and numerous clinical trials, pre-treatment prognostic factors are still significant predictors of glioblastoma patient survival. Many of the studies mentioned earlier use both pre-treatment factors and treatment variations. However, pre-treatment factors such as age and KPS have been widely confirmed to be the most important prognostic factors. A treatment factor, extent of surgery (resection), is also found to be significant in affecting patient survival. However, despite the frequent use of MRI in diagnosis and prognosis, image-based factors are rarely used in many of the studies mentioned in this section. For example,

[2]European Organization for Research and Treatment of Cancer
[3]National Cancer Institute of Canada

the only image-based factor used by Curran et al. is tumor size (maximum diameter < 5.0 cm or ≥ 5.0 cm) [11]. In our work, we only take into account pre-treatment factors. These factors include image-based textural properties in pre-surgical scans and patient-related information such as age and gender. Our study is the first to use image-based textural properties from MRI in prognosis of glioblastoma.

In biomedical research, the predictive power of a survival prediction method is determined by measuring the level of statistical significance at which the system can discriminate between survival classes. Kaplan-Meier estimators are used widely with statistical testing methods such as the logrank test to measure the predictive power of survival prediction methods using censored survival data. In our work, we adapt a different understanding of prediction and measures of prediction. We define prediction accuracy as the total number of patients whose survival class has been correctly determined by the prediction method. We will describe more of our approach to survival prediction in the later chapters.

Finally, unlike population studies, gene-expression profiling can personalize prognosis. Every patient has a unique genetic profile and therefore, personalized approaches are required for more reliable prognoses. Population studies capture features in the patient population that most commonly affect patient survival times. For example, age is detrimental in prognosis with patients older than 50 commonly having much shorter survival times. However, population studies do not explain why some young patients have similar survival times as much older patients. But as we mentioned earlier, researchers have discovered a gene-expression profile in certain patients, which is a stronger predictor of survival than age. Therefore, certain younger patients, who lack this gene-expression profile, do not benefit from its favorable prognostic effects. In our approach, we attempt to capture personalized features based on textural properties of a patient's brain scan images and then combine these features with other more general factors such as age, therefore, taking advantage of both unique personal factors and information learned from populations in the prognosis of glioblastoma patients.

2.3 Texture Analysis

Defining what constitutes texture has always been a topic of interest in the fields of image processing, computer graphics and computer vision. The main challenge has been to describe the properties of texture in an image numerically for meaningful quantitative analysis. Quantitative analysis in textures is essential in many tasks such as classification of images based on their textures, segmentation of an image into homogeneous regions, synthesizing texture for computer graphics and image retrieval based on texture [26, 37]. However, it is very difficult to describe in precise terms what we visually perceive as texture, even though being able to visually distinguish one texture from another comes to us naturally. As a result, there is no unique definition for texture. We can characterize a texture by its properties as we perceive them based on visual and tactile senses. For example, we can describe a certain texture with such terms as 'net-like', 'rough' or 'smooth'. Therefore, a good

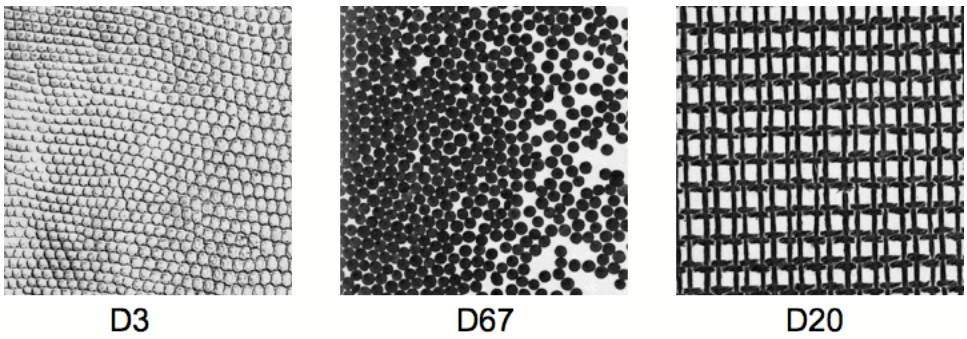


Figure 2.4: Images from [3]. The textural primitives in these textures can be easily described.

approach to quantitative analysis of textures is to first describe a texture in a way that is perceptually and intuitively meaningful and then try to measure these properties in order to approximate visual perception. To be able to measure textural properties, we need to understand the underlying structure of what constitutes texture. There have been traditionally two general approaches to defining texture: structural and statistical.

In the structural view, loosely described, large number of very small well-defined objects or patterns filling an area or a surface can be viewed as texture [15]. Images of surfaces covered with grass, hair or sand are all examples of small objects defining a texture. Spots on a leopard's skin or stripes on a tiger are examples of small patterns defining a texture. This approach to perceiving texture gives rise to *structural* texture analysis, which defines texture as a macroscopic region that is characterized by repetitive patterns of *primitives* (*elements* or *micro-texture*) arranged according to a (not necessarily strict) placement rule [59, 67, 6]. Hence, characterizing a texture using structural texture analysis relies on characterizing the basic patterns or primitives comprising the texture and their interactions. Primitives comprising a texture are characterized by their shapes and sizes [26]. It is believed that these characteristics in the underlying structures are what leads us to perceive texture in terms of 'coarseness', 'uniformity', 'roughness' and so on [37]. Figure 2.4 displays some textures from the Brodatz album [3], which can be easily described by the textural primitives that comprise them. Structural texture analysis seeks to determine the structural primitives in a textured image. However, for most textures, this is a very difficult task. Therefore, structural texture methods are mainly developed for texture synthesis, which is the task of synthetically constructing large regions of texture from samples of small regions [15].

Not all textures can be described by an underlying structure such as repeating patterns or primitives. Some textures have more complex stochastic structures that cannot be simply decomposed into any basic elements. *Statistical* texture analysis methods have been introduced to be able to study these textures. Statistical methods are based on the notion that texture can be characterized by the distribution of gray-levels, their local variations and the relationships between them [59, 67, 6]. Figure 2.5 displays examples of textures from the Brodatz album [3], which cannot be described in

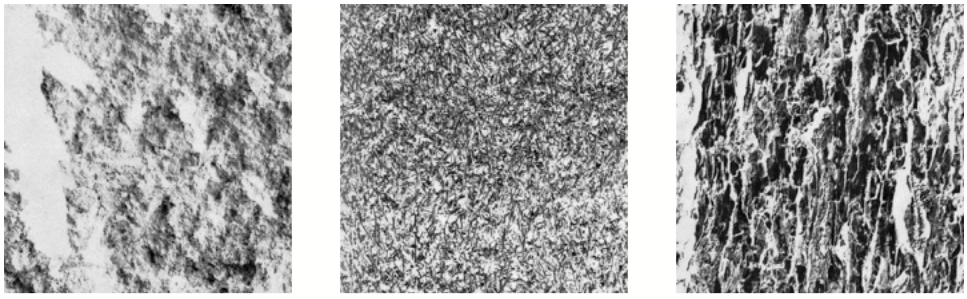


Figure 2.5: Images from [3]. These textures have a more stochastic structure and cannot be easily described by any repeating patterns or primitives.

terms of simple structures.

Structural and statistical texture analysis can be used to described textural properties. For example, coarseness of a texture can be characterized by the size of its textural primitives; the larger the primitives, the coarser the texture and inversely, the smaller the primitives, the finer the texture. Properties such as contrast and homogeneity, on the other hand, can be better characterized by the local variations in gray-levels. Most textures can be described by both structural and statistical properties.

Karu et al. tackle the question of whether a given image has texture in order for texture analysis to be meaningful [26]. They explain that the most important characterizing factor of texture is coarseness. An image must have a certain level of coarseness for it to have meaningful texture and this is determined by the size of its textural primitives or equivalently, the scale of the image. At one extreme, the textural primitives can be so small that they become dots, indicating the highest level of fineness, in which case, the image can be described as white noise. On the other extreme, the textural primitive can be so large that only one can fit in the image. In both cases, there is no meaningful texture in the images. Once we ensure that there is a certain level of texture in an image, the challenge is to characterize the texture in meaningful terms. Tamura et al. specify some basic intuitive textural properties in pairs, which correlate with visual perception of textures and are meaningful when applied to most types of texture [59]. These specifications are as follows with references to examples of textures in Figure 2.6 (D20, D34, D38, D67, D68, D93 and D98):

- Coarse (D98) versus fine (D93). The larger the textural primitives and the more scattered they are, the coarser the texture.
- High contrast (D20) versus low contrast (D38). Contrast relates to how stretched the range of gray levels are. Sharp edges in the image also contribute to higher contrast.
- Directional (D68) versus non-directional (D67). Directionality is a global property over a region, which is based on the shape and placement of the textural primitives.

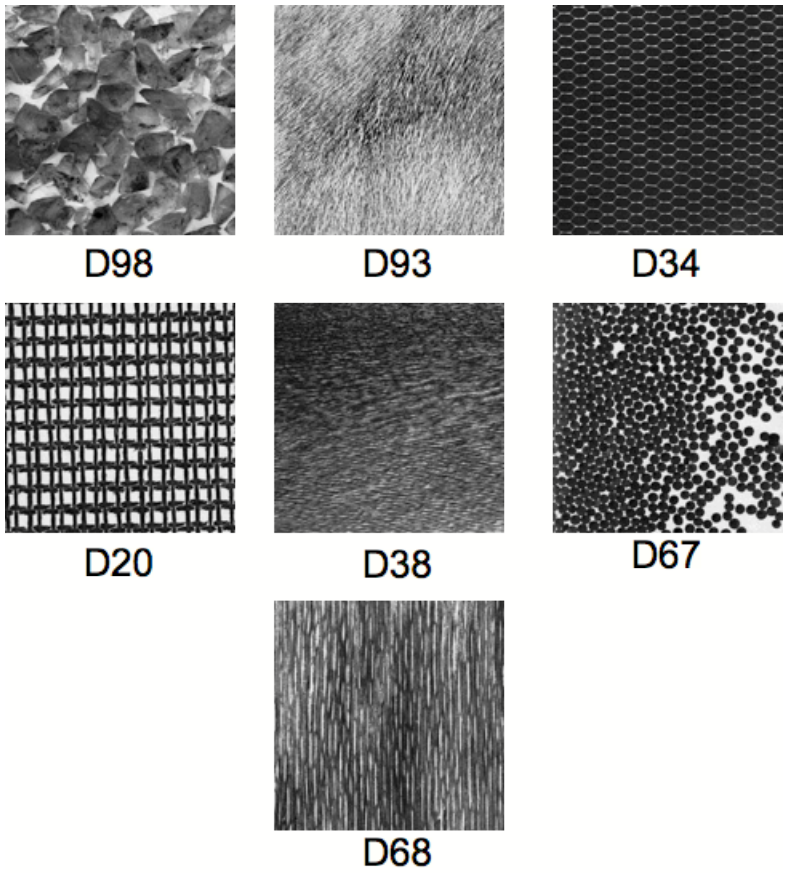


Figure 2.6: Images from [3]. Tamura et al. describe the textural properties of these images based on visual perception [59]. Texture D98 is coarse, irregular and rough, D93 is fine, D20 has high contrast, D38 has low contrast and is smooth, D68 is directional, D67 is non-directional and blob-like, D34 is line-like and regular.

- Line-likeness (D34) versus blob-likeness (D67). This property is characterized by the shape of the textural primitives.
- Regular (D34) versus irregular (D98). Regularity relates to the variations of the placement rule for texture primitives.
- Rough (D98) versus smooth (D38). Roughness is based on what one may perceive a tactile texture by touch rather than by visual senses.

When tested by human subjects who would rate texture images based on these intuitive specifications, some correlations were observed [59]. Tamura et al. note that, for example, contrast and coarseness are somewhat correlated. High contrast resulting from many sharp edges indicates fine texture (picture a checkerboard with very small boxes). They also note that we visually perceive high roughness in coarse textures that have high contrast. Also, line-like textures are visually perceived as directional [59]. Out of all textural properties, regularity is the hardest to describe and implement

quantitatively. It is hard in general to define regularity in such a way that can be applied to any given texture.

Numerous texture analysis methods have been proposed in the past four decades for measuring textural properties. One of the first and most widely used methods is the extraction of second-order statistics based on *pairs* of gray-level distributions in the image [21]. Haralick et al. introduced a method based on Gray-Tone Spatial Dependence Matrices (also known as Gray-Level Co-occurrence Matrices), which assumes that the textural properties of a region can be determined from the overall or average spatial relationship between the gray levels in an image [21]. More specifically, a co-occurrence matrix collects information regarding the distribution of *pairs* of pixels within an image according to a displacement rule, which is defined by a distance and an angle. For a given distance d and angle θ, the entry (i, j) in a normalized co-occurrence matrix $P_{d\theta}$ is the joint probability that a pixel with gray value j appears at a distance d and angle θ with respect to a pixel with gray value i. Haralick et al. propose 14 second-order textural properties, which can be computed from a co-occurrence matrix. Such properties include energy (which measures 'orderliness'), contrast, correlation (which measures gray-level linear dependencies) and more. Many of these properties correlate with each other, thus computing all of them would be redundant. Simple first-order statistical texture properties can also be computed directly from the image. These methods measure basic statistical variations in gray-levels and are mostly based on the histogram of an image, which counts the total number of pixels with a given gray value within the image. Hence, a normalized histogram gives the probability that a given pixel in the image has a certain gray value. Some simple first-order statistics include mean, variance and skewness of the distribution of gray-levels. It is reported that first-order statistics are not effective in capturing textural properties, whereas second-order statistics have a higher correlation with human visual perception [37].

Recent structural texture analysis methods have been proposed by Leung et al. [31] and Varma and Zesserman [61], which tackle the difficult task of determining textural primitives (also called *textons*) for the purpose of texture classification. Textons are more specifically defined as small primitives of pre-attentive visual perception of texture and correspond to the dominant local image structures [31]. Both proposed texture classification methods work by building a universal collection of textons (texton *dictionary*) from many texture images, obtained under varying illuminations and viewpoints, representing textures from different materials, such as leather, marble or rug. Then models are built for each material image by determining the distribution of dictionary textons present in the image. New texture images with novel illumination and viewpoint conditions can then be classified by comparing their texton distributions with the models. Both proposed methods compute the textons for an image by using a popular technique called *linear filtering* [15, 48, 8]. Through linear filtering, every pixel p in the image is represented by a weighted linear combination of the gray values in the region centered at p. The weights in the weighted linear combination are provided by a filter *kernel* (commonly referred to as just a filter). The process of computing the linear combinations

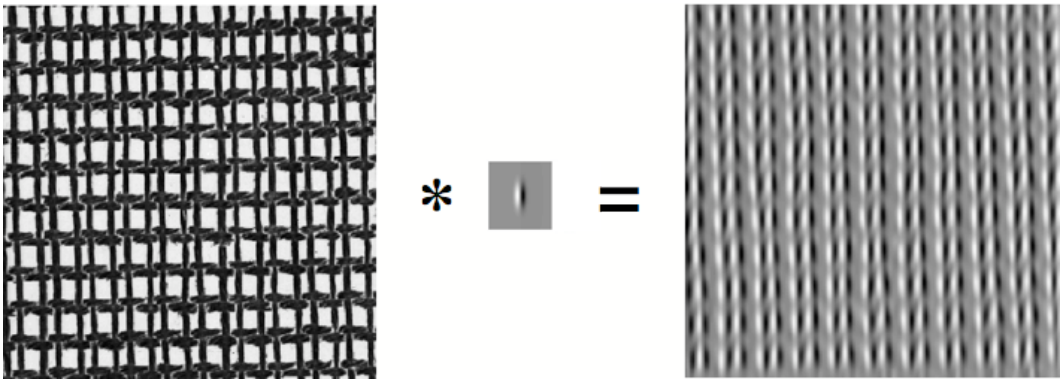


Figure 2.7: Image D20 from [3] convolved with a vertical edge filter.

for every pixel in the image using a filter kernel is referred to as a *convolution* and the resulting value of each linear combination for a pixel is called a filter *response*. Intuitively, the idea behind linear filtering is that given a filter kernel designed with a specific structure in mind, a convolution of an image with the filter kernel results in strong responses in the regions of the image where the local structure is similar to the structure of the filter kernel. For example, in Figure 2.7, a texture image (D20) is convolved (*) with a vertically orientated edge filter and in the resulting image, one can see that vertical edges within the texture image (D20) resembling the edge filter have shown strong responses.

Some of the most commonly used filters are Gaussian filters, and weighted linear sums of multiple Gaussians, which in turn, result in spot and bar filters [15]. A 2-dimensional Gaussian filter is given by:

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2.2}$$

Both Leung et al. [31] and Varma and Zesserman [61] use filter banks (collection of filters with different structures) to build their texton dictionaries. Varma and Zesserman experiment with different filter banks in their texture classification method and report that a filter bank called the Maximum Response 8 (MR8) achieves the best results [61]. The MR8 filter bank is a combination of bar and edge filters at different scales and orientations, whose resulting filter responses are "collapsed" by recording only the maximum responses along different orientations of the same scale of the bar and edge filters (there will be a detailed description of the MR8 filter bank in Section 4.1.4).

In addition to structural and statistical texture analysis methods, another very common class of texture analysis methods are the transform methods. Two of the most important transform methods are the Fourier transform and wavelets [8]. The Fourier transform of an image takes the image from the spatial domain and represents it in the frequency domain. For a given two dimensional image

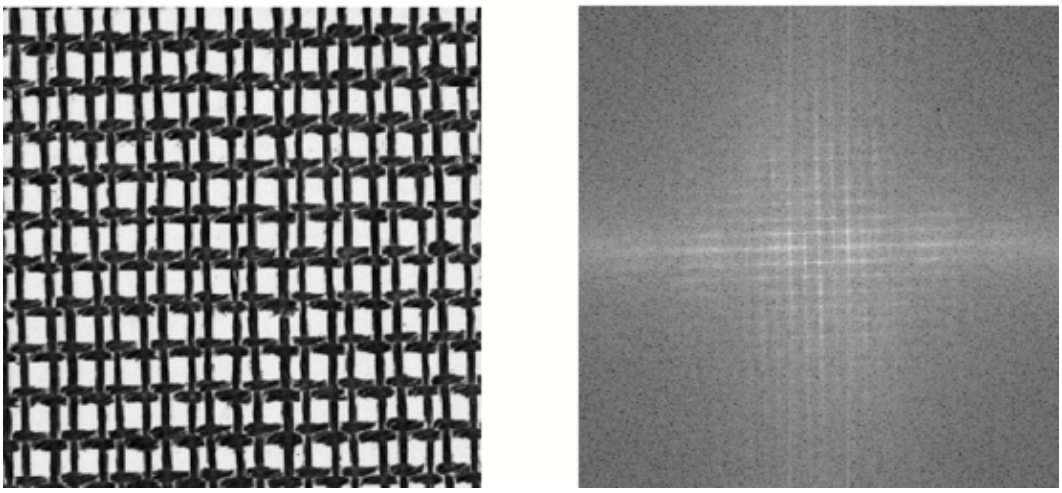


Figure 2.8: Image D20 from [3] and its representation in frequency domain.

$f(x, y)$ where $1 < x < M$ and $1 < y < N$, the discrete Fourier Transform of the image is:

$$F(u, v) = \sum_{x=1}^{M} \sum_{y=1}^{N} f(x, y) e^{-i(\frac{2\pi ux}{M} + \frac{2\pi vy}{N})} \tag{2.3}$$

where, $1 < u < M$ and $1 < v < N$. Given the Euler's formula,

$$e^{i\theta} = cos\theta + isin\theta \tag{2.4}$$

if we let $\theta_{MN}(x, y, u, v) = \frac{2\pi ux}{M} + \frac{2\pi vy}{N}$, then the transform can be re-written as:

$$F(u, v) = \sum_{x=1}^{M} \sum_{y=1}^{N} f(x, y)(\cos\theta_{MN}(x, y, u, v) + i \sin\theta_{MN}(x, y, u, v)) \tag{2.5}$$

As the above equation shows, the Fourier transform of an image represents the image in terms of sinusoids (sines and cosines) with varying frequencies and orientations determined by u and v in the *frequency space*. To better understand what is meant by frequency across a 2-dimensional image, one may draw a straight line, in any direction (orientation), across an image and then treat the gray values along the line as a 1-dimensional signal. Then, one can see that the frequency of the signal results from variations in gray values across the signal. The frequency space of an image describes the spatial frequencies across the entire image in all directions. In terms of texture, high frequencies across an image can be caused by sharp variations in gray values. Figure 2.8 shows a texture image and the log of the magnitude of its Fourier transform. Note that since the Fourier transform is a complex-valued function, it cannot be fully visualized in the frequency domain. Therefore, it is common to instead display the magnitude of the transform (also referred to as the magnitude spectrum), i.e. $|F(u, v)|$.

Pixels in the magnitude spectrum diagram represent magnitudes of the spatial frequencies in the texture image. Pixels closer to the center represent low frequencies and pixels far from the center

represent higher frequencies. The angular relationship of the pixels to the center point represents the orientation of the frequencies. The horizontal and vertical textural patterns in the texture image D20 reflect in the magnitude spectrum as strong horizontally and vertically oriented frequencies (Figure 2.8). Filtering an image is also possible in the frequency domain. This is done by multiplying the Fourier transform of the image by a filter mask. A filter mask can remove certain frequencies from the image by multiplying them by zero in the frequency domain and retaining other frequencies. Low-pass filters allow lower frequencies to stay and remove higher frequencies. On the other hand, high-pass filters allow higher frequencies to stay and remove lower frequencies. A filtered Fourier transform of an image can be converted back into the spatial domain by using the inverse Fourier transform:

$$f(x,y) = \frac{1}{MN}\sum_{u=1}^{M}\sum_{v=1}^{N} F(u,v)e^{i(\frac{2\pi ux}{M}+\frac{2\pi vy}{N})} \tag{2.6}$$

Then textural properties such as first-order statistics can be extracted from this filtered image for texture analysis. One major problem with the Fourier transform is the fact that the spatial information is lost in the frequency domain. It is impossible to know where, in the spatial domain, certain frequencies occur. All that the Fourier transform tells us is what frequencies occur across the image, at what magnitudes and orientations. Computing textural properties in the frequency domain makes it impossible to localize the computed textural properties to certain regions or pixels in the original image. On the other hand, by computing textural properties in the spatial domain, on the original image, we will lose any information about frequencies across the image. As a result, wavelet transforms have been used to resolve the frequency localization problem [8, 48, 6].

Wavelets (small wave filters) characterize a texture image by the frequency content of the image at different directions and different scales of the image. The wavelet transform of an image associates to each pixel, wavelet coefficients, which correspond to different frequency patterns at different scales of the image, describing the frequency pattern of the image at that pixel. Wavelets work in a way similar to linear filter kernels in that at each scale, strong responses (coefficients) are associated with regions in the spatial domain of the image where the local frequency pattern is similar to the wavelet. Wavelets at smaller scales of the image detect high frequency patterns and are more localized. Wavelets at larger scales of the image detect low frequency patterns, which span larger regions. The scale property in wavelet analysis can be related to texture, since low frequencies (at high scales) in the image are associated with the image's overall coarseness, whereas localized high frequencies (at low scales) are associated with the region's fineness. Using wavelet transforms, one can then characterize the frequency patterns of the image at different scales and locations in the image [6].

There are many other texture analysis methods, which utilize different filtering schemes to characterize complex textural properties. Chen et al. compare several filtering methods for the purpose of texture classification [8]. In their experiments, they perform filtering using ring/wedge-like filters

in the frequency space and then compute variances in spatial domains and frequency domains. Ring-like filters in the frequency domain measure coarseness of the image's texture, and wedge-like filters measure directionality. They also use wavelet transforms to produce wavelet coefficients at multiple scales. They compute the variances of the wavelet coefficients in the wavelet transformed images across different scales and directions to capture coarseness and directionality. Their experimental results indicate that the wavelet features produce relatively better accuracies in texture classification. Randen et al. also compare a variety of filtering methods for texture segmentation in multi-textured images [48]. They compare the filtering methods to the classical methods such as second-order statistics from co-occurrence matrices [21]. Their main conclusion is that no one texture extraction method performs well on all types of texture images and that second-order statistics performed mostly poorly.

2.3.1 Texture Analysis of Magnetic Resonance Images

The use of texture analysis on medical images has become prevalent in the past couple of decades. Medical images, and more specifically Magnetic Resonance Images, pose a special challenge in that the local textural properties cannot be easily described by structural texture methods. Upon close inspection on MR images from brain scans, one can see that the textures in tissues are very different than most of the natural texture images discussed in the previous section. One can see that there simply are no textural primitives or repeating patterns within tissues on MR images the same way one can characterize the patterns of the textures in Figure 2.6. Texture in brain tissues has a more complex structure. Moreover, brain tissue (especially diseased tissue) on MRI have more irregular and heterogeneous appearances. As a result, statistical texture methods are more appropriate. In fact, second-order statistics obtained from co-occurrence matrices [21] are one of the most prominent texture analysis methods applied on MRI [51, 27, 23, 35, 6, 52]. However, other more sophisticated texture methods such as transform-based methods [51, 52, 4] and filtering [53, 18, 52] have recently been receiving some attention as well.

Lachmann and Barillot introduced one of the earliest automatic systems for recognition of brain tissue on MRI [29]. They use a method called the spatial gray-level dependence method, which builds co-occurrence matrices similar to ones proposed by Haralick et al. [21], except that they build the matrices for local neighborhoods and compute second-order statistics such as contrast and homogeneity to characterize the textures in those neighborhoods. Kito et al. built an automatic system for segmenting healthy and diseased brain tissue including different classes of tumor, cerebrospinal fluid, white matter and gray matter [51]. They utilize wavelet transforms for localization in the space and frequency domains leading to multi-resolution (i.e. multi-scale, as discussed earlier) analysis. They also compare texture features based on first-order and second-order statistics (GLCMs), extracting all of these statistics locally in small neighborhoods. The authors report that the wavelet-based features consistently performed slightly better than the other texture features. Ko-

valev et al. introduce a more complicated extended version of co-occurrence matrices [27]. They reason that textural differences in brain tissue are faint and not as well-defined as natural textures such as the ones from the Bodatz album [3]. Therefore, much more sensitive texture extraction methods are required to fully capture the subtle characteristics of brain tissue texture. They extend the co-occurrence matrices to higher dimensions by taking into account more than just gray-level co-occurrences between pairs of pixels. In addition to gray-levels, they also include the following relations: local gradient magnitude and relative orientation of gradient vectors between pairs of pixels. This helps their texture methods achieve high sensitivity and specificity. Unlike the Haralick et al. co-occurrence matrices, instead of computing statistics from the matrices, Kovalev et al. simply use the L1-distances between the matrices to compare co-occurrence matrices in order to compare textures. Using their high-dimensional co-occurrence matrices, they succeed in discriminating images from healthy subjects (controls) from images of patients.

Herlidou et al. use simple histogram-based first-order statistics along with second-order statistics (GLCM) such as contrast, correlation, homogeneity and entropy [23]. They also include a higher order statistical method called run-length matrices. Run-length matrices count for each gray-level, the total number of consecutive runs of pixels having the same gray value for a given length and direction in the image. Using these texture features in a hierarchical clustering method, they build a framework, which seeks to discriminate tissue type on MRI of patients with intracranial tumors. Their framework can achieve relative success in discriminating certain tissue types, e.g. white matter regions from other healthy regions on T2-weighted MRI and furthermore in discriminating regions in the tumor from the surrounding edema [23]. Mahmoud et al. introduce another extension to GLCMs [35], which differs from the version proposed by Kovalev et al. [27] in that the displacement rule is extended to three-dimensions (spans across slices) for texture analysis on three-dimensional MRI volumes. The co-occurrence matrices themselves are still two-dimensional, which means they still characterize pair-wise spatial co-occurrences between pixels, much like the original version proposed by Haralick et al. [21] but now they include pairs of pixels across slices as well. Similarly, they compute second-order statistics such as energy and correlation from the matrices in the conventional way. The authors compare their new extended co-occurrence matrices with the original GLCMs in discriminating brain tissue based on texture. They report good discrimination between white matter surrounding the tumor and white matter far away from the tumor using their extended GLCMs [35].

Sasikala et al. combine the traditional GLCMs and wavelet transforms to segment glioblastomas on MRI. [52]. They build GLCMs for local neighborhoods and then compute second-order statistics such as energy, homogeneity and entropy. Then they use a genetic algorithm to perform feature reduction to reduce the number of texture features, to produce the set of "optimal" features. Using their texture extraction method, Sasikala et al. report good segmentation results. They also report successful discrimination between healthy and diseased multiple sclerosis tissue. Ghazel et al. in-

troduce a supervised segmentation framework, where regions of interest containing both multiple sclerosis and healthy tissue are used to create optimal texture filters, which can then discriminate healthy tissue from multiple sclerosis tissue [18]. They build optimal filters maximizing feature separation between two textured regions, which in turn can be discriminated simply by thresholding. This framework, which includes optimal filters, was used in the comparative study of filter methods by Randen et al. [48], which reported it performed well.

Finally, in a recent work, Brown et al. used texture analysis to detect a certain genetic feature with favorable prognosis in oligodendroglioma (a low grade glioma) from MR images [4]. It has been a well-known fact that co-deletion of chromosomes 1p and 19q is associated with a favorable prognosis for oligodendroglioma patients. Brown et al. use S-transforms to analyze local spatial-frequency patterns in texture of the tumors. On MR images, strong low frequency patterns appear as homogeneous whereas strong high frequency patterns appear as heterogeneous. S-transforms, much like wavelet transforms, are a localized transform-based method that describe local frequency patterns for each pixel in the image. The most significant differences between tumors with and without the genetic features were observed on T2-weighted MRI (compared with T1-weighted and FLAIR images) and moreover, they discovered that only medium range (as opposed to high or low) spatial frequencies in the texture of tumor tissue are highly predictive of the favorable genetic feature. Unfortunately, this genetic feature is not necessarily associated with favorable prognosis in glioblastoma patients.

Our work is different from most of the works mentioned here in that we aim to use texture analysis of MRI data to predict prognosis for glioblastoma patients, whereas, all the works (except for the work of Brown et al. [4]) aim to segment tumors or discriminate tissue types. Our goal is to discover whether texture of certain regions of the brain on MRI can discriminate low-survival patients from high-survival ones. For our work, we choose Gray-Level Co-occurrence Matrices because, they are the most widely known and used texture methods on MRI and the features extracted from co-occurrence matrices are the most intuitive. We also choose the MR8 filter bank because they encompass multi-scale, multi-orientation texture extraction, which can measure local textural properties at different scales producing short-length features. Schmidt also used the MR8 filter bank in a tumor segmentation framework in addition to many other features and reported that this filter bank, despite being simple, performed remarkably well [53].

Chapter 3

Survival Prediction Overview

3.1 Motivation

As discussed in Chapter 2, there are a variety of factors used by clinicians and researchers to make prognostic assessment of glioblastoma patients. Clinicians primarily consider clinical data such as age and KPS, which have been shown to strongly predict survival (Section 2.2). Genetic markers are also potentially useful prognostic factors, but they are mostly studied in laboratory analysis by researchers (Section 2.1.1). There are very few genetic markers that have been widely and independently confirmed to be of prognostic value and therefore, genetic markers are not routinely used by clinicians for prognosis. However, clinical imaging is a routine procedure in clinical practices for diagnosis and treatment planning. While there are a few widely confirmed markers such as MGMT methylation, there has not been any studies to find possible relations between such genetic mutations and the appearance of glioblastoma tumors on MRI scans. If such relations held, then we could design features based on MRI to detect these genetic mutations and use them in prognosis. This would save clinicians and surgeons time and avoid unnecessary and risky surgical procedures and costly and time consuming laboratory work for tissue analysis.

However, there are certain appearance characteristics of glioblastomas that are known to be of prognostic value. One important example is when the glioblastoma invades the other cerebral hemisphere through the corpus callosum (Figure 3.1), producing a symmetric shape resembling a butterfly [49]. Such glioblastoma appearance is usually associated with higher tumor aggressiveness and poor prognosis.

Another important appearance characteristic is the presence of mass effect caused by the glioblastoma and the edema surrounding it. The presence of mass effect is indicative of increased growth. As the tumorous mass grows, it exerts pressure on the brain tissue surrounding the growing mass against the skull. This intracranial pressure is a danger to the blood supply of the surrounding brain tissue. The presence of mass effect can be observed by the deformation and blurriness of the nearby sulci. *Sulci* (single: *sulcus*), as seen in Figure 3.2, are thin furrows or wrinkles close to the outter edge of the brain. The texture of the sulci under pressure from mass effect can be described as blurry

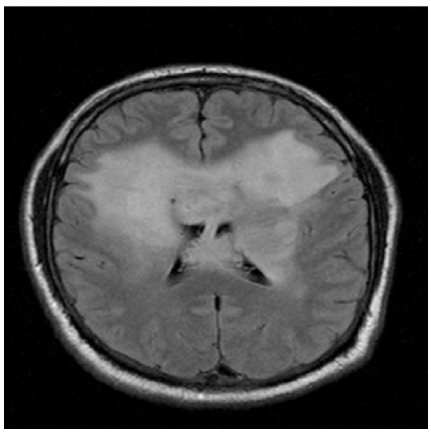


Figure 3.1: Tumor infiltrating through the corpus callosum

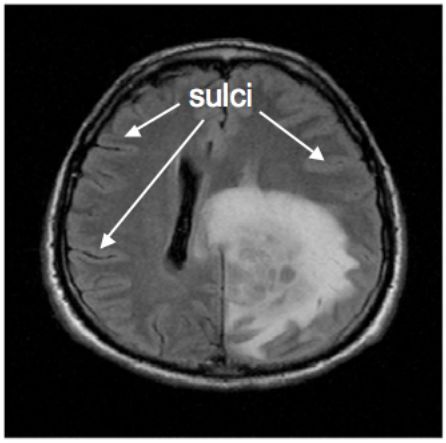


Figure 3.2: Sulci in a brain image. The sulci near the tumor is under pressure due to mass effect.

or smudged, properties that can be measured using texture extraction tools on MRI.

Also, as mentioned earlier (Section 2.1), a few studies have investigated the relations between certain appearance characteristics of glioblastomas and patient survival. The measurements of appearance characteristics were manually and visually measured by experts rather than by automatic extraction through image analysis techniques. Hammoud *et al.* and Pierallini *et al.* each claim that the extent of the necrotic region, which is the central region within the tumor containing dead cells, is related to patient survival [20, 44]. Pope *et al.* explained that the presence of a non-enhancing tumor is indicative of a favorable prognosis in older patients and an increased amount of edema is indicative of poor prognosis [46].

Based on these observations, we lay out a machine learning framework for prognosis in glioblastoma patients. The general idea is to be able to predict the survival category of a given patient based on information from other patients and their survival categories. We achieve this using mostly features extracted from axial brain MR images to measure textural properties, combined with other parameter and image based features.

3.2 Framework Overview

In this section we give a brief overview of our prognosis framework and its different parts. We leave detailed description of the parts for the next chapter (Chapter 4). In the following sections, we first describe the central idea in our approach, Supervised Learning. Then we describe our feature extraction scheme to be used in supervised learning. Finally, we will describe the raw data that we use in our prognosis framework for testing and validation purposes.

3.2.1 Supervised Learning

The central idea of our prognosis framework is to first build a model given a set of patients and their clinical data along with their survival times, then use this model to predict survival times of future, previously *unseen* patients. This method of prediction is called *Supervised Learning*. Intuitively, supervised learning is to get a machine to "learn" a task, such as prediction or recognition, from a set of provided examples and then have the machine perform the task autonomously in the future. The set of provided examples is called the *training set*. The training set is composed of instances, where each instance is paired with a *label*. Each instance is described by a set of *features* called a *feature vector*, where each feature is a value encoding a certain property about the instance it is representing. Each instance in the training set has a label, which is the value or property that we expect the model to be able to predict. A model, which has been trained on the training data, is a mapping from the feature vectors (or equivalently, instances) to the labels, which can then be used to predict the labels of previously unlabeled data (instances). The set of data of unknown labels is called the *test set*. When the labels are categories (or classes), *e.g.* healthy or diseased, then the supervised learning method is referred to as *classification* and the learned model is called a *classifier*. If the number of classes (i.e. number of unique labels) is two, then we call the model a *binary classifier*. There are two phases to every classification problem. The first phase is the training phase, where the feature vectors and their labels are used to build a classifier. The second phase is the prediction (or testing) phase, where the classifier is used to predict the labels of unseen data. Figure 3.3 demonstrates the process of training a classifier on data using paired feature vectors and labels, and then using the classifier to predict labels.

As a simple example of a binary classifier, consider the simple task of distinguishing between images of hand-written digits 1 and 2 using a machine. The training data is composed of images of handwritten numbers and their labels are the corresponding values: 1 and 2. The feature vector for each image can be built by measuring properties based on the image of the handwritten number. The task of building relevant features from instances is called *feature extraction*. For example, one can measure the curvature of the lines or the distortions, amongst many other possibilities. The main goal is to learn a model (binary classifier) based on the labeled data (paired feature vectors and labels), and then be able to use this model to distinguish between unlabeled handwritten images of 1 or 2 as demonstrated in Figure 3.4.

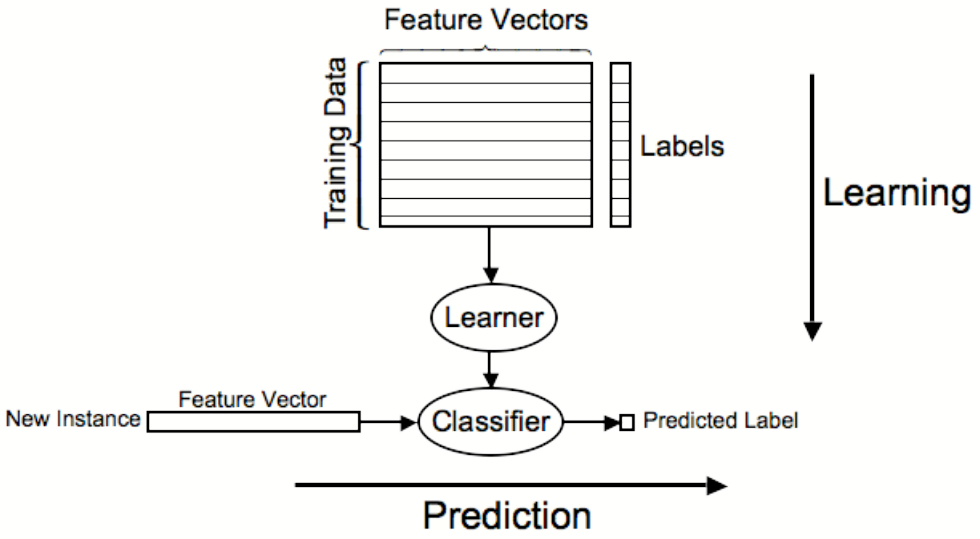


Figure 3.3: The classifier is trained on (learned from) the feature vectors and labels in the training phase. Then the classifier is used to predict labels from feature vectors in the prediction phase.

In this thesis, we build a binary classifier that maps features extracted from raw data of glioblastoma patients to survivals. The raw data (Section 3.3) contains volumes of axial, pre-surgical MRI scans of the FLuid Attenuated Inversion Recovery (FLAIR) modality. Each patient has one volume of scans, which is composed of a number of slice images. Image features are extracted for each patient from the images and then combined with other simple raw data, including clinical data such as age and sex. The label for each patient is the patient's survival category. We define two survival categories, low survival versus high survival. The survival categories are determined based on the distribution of the survival times in our data set. The survival time for each patient is measured in weeks, from the date the MRI scan was taken to the date of death (or date of data collection if date of death is not available).

The feature extraction process of the framework mostly involves extracting textural features from the MR images. As stated before, the primary goal is to measure and use certain textural properties of the glioblastomas as they appear on MR images. We use a number of texture extraction methods (Section 4.1) on a number of pre-defined regions (Section 4.2.1) in the brain images and then combine the results with non-texture features to form a final feature vector for each patient. In the next section, we give an overview of the different texture extraction methods that we use in our framework.

3.2.2 Texture Extraction

For a given axial (Figure 3.5) MR image containing a cross section of the tumor, we first define a set of regions. For example, we define the region *inner* as the area within the segmented tumor. We believe that the textural properties of these regions are of prognostic value. The regions are

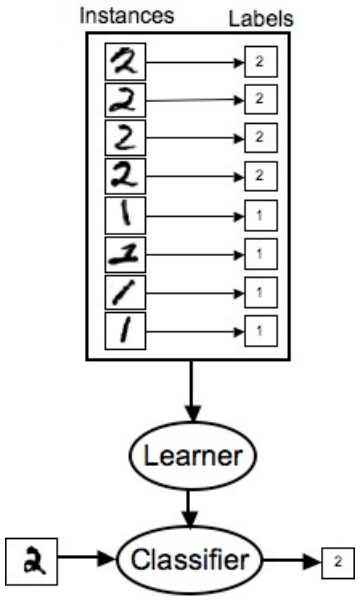


Figure 3.4: Learning digit recognition from a set of images as training data, then classifying a new image of a digit.

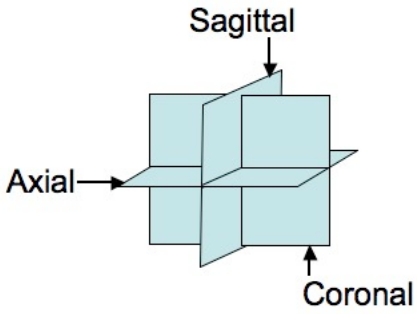


Figure 3.5: An axial image of a brain is parallel to a horizontal plane.

described in more detail in Section 4.2.1. For each region, we extract textural properties using a set of texture extraction methods. Each texture method generates either a resulting texture image or a set of matrices, which represent certain textural properties of the region on which the texture method was used. However, to be able to utilize the results of each texture method effectively and efficiently, we use simple statistical operations on the resulting texture image or the matrices to obtain fewer values for use in our classification framework. Therefore, as shown in Figure 3.6, if each texture method is seen as a black box, then the input to the black box would be a region of the MR image and the output would be a single vector with two or three values depending on the type of the texture method.

For a given region, we combine all the values obtained from the texture methods used on the region to form a feature vector for that region as shown in Figure 3.7. This texture extraction process is repeated for every region resulting in multiple region texture feature vectors. Eventually,

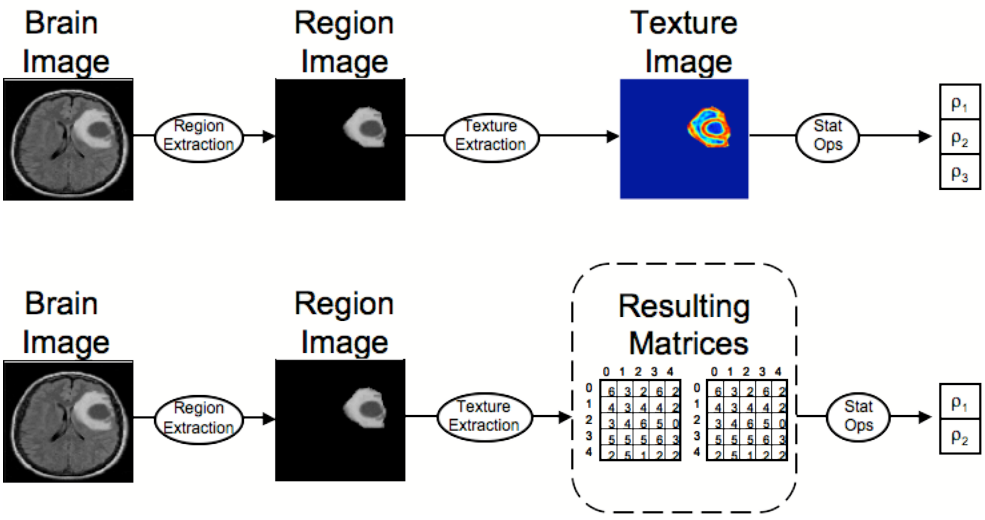


Figure 3.6: Texture extraction produces either a texture image or a set of matrices. In either case, statistical operations are used to reduce the resulting values.

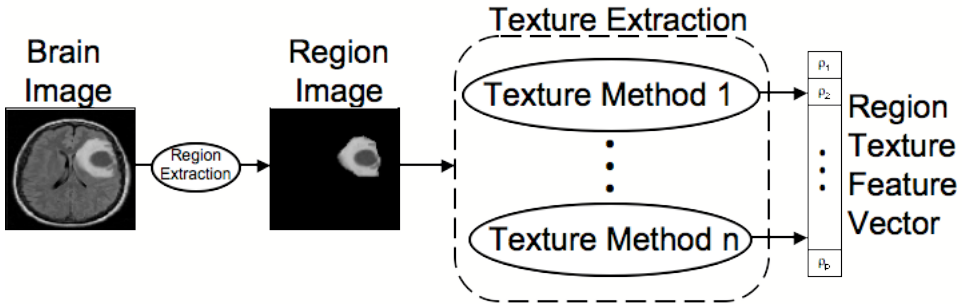


Figure 3.7: For each region, a set of texture methods are used to obtain a set of values, which we combine to form a feature vector for that region.

all the region texture feature vectors are consolidated to form one texture feature vector, called the combined-regions texture feature vector, as shown in Figure 3.8.

After region texture extraction is performed, we also compare the texture values between certain regions in the MR image. This is done to use textural differences between certain regions as prognostic factors in our prognostic framework. For example, significant texture difference between the sucli region on the hemisphere of the brain, where the tumor is located (*ipsi*) versus the opposite sulci region (*contra*) is indicative of presence of mass effect on the ipsi side of the brain. We do not compare all regions pair-wise, but rather use only a select few regional comparisons; see full list in Section 4.2.3. We call these types of features *within-image region comparison features*. As shown in Figure 3.9, in within-image region comparison, we obtain the resulting difference vector from the texture feature vectors of both regions. Then the values of this difference vector are consolidated with the combined-regions texture feature vector from Figure 3.8 to represent the complete texture feature vector for the MR image. The overview diagram for the complete MR image texture feature

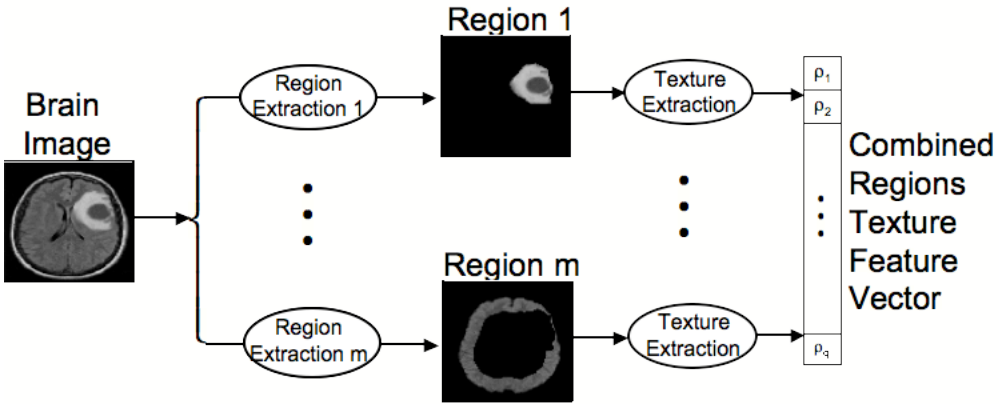


Figure 3.8: Texture extraction is performed on each region and the results are combined into one feature vector representing texture properties for the MR image.

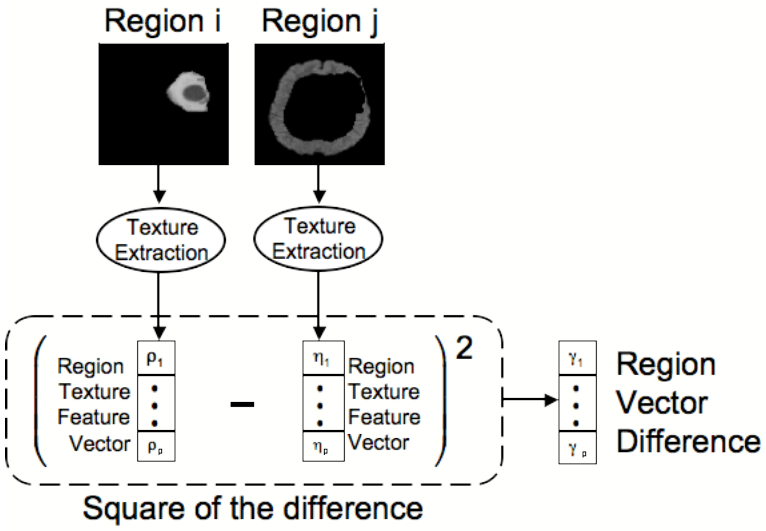


Figure 3.9: Region texture feature vectors from two regions are used to make one difference feature vector.

extraction is shown in Figure 3.10.

3.2.3 Complete Feature Extraction

Texture extraction is the major part of our feature extraction scheme in our prognosis framework. What has been described so far is the process of obtaining a complete texture feature vector for one MRI slice. There are tens of slices in an MRI volume of a patient. Extracting texture properties on all the images would be time-consuming and unnecessary. Besides, only less than half of the images include cross sections of the tumor. Therefore, we select a few of these slices and extract complete texture features only on these. Then we utilize the complete texture features from the slices to obtain one final texture feature vector representing the whole MRI volume, thus representing the patient, to use in our classification framework. There are, however, other non-texture features, which

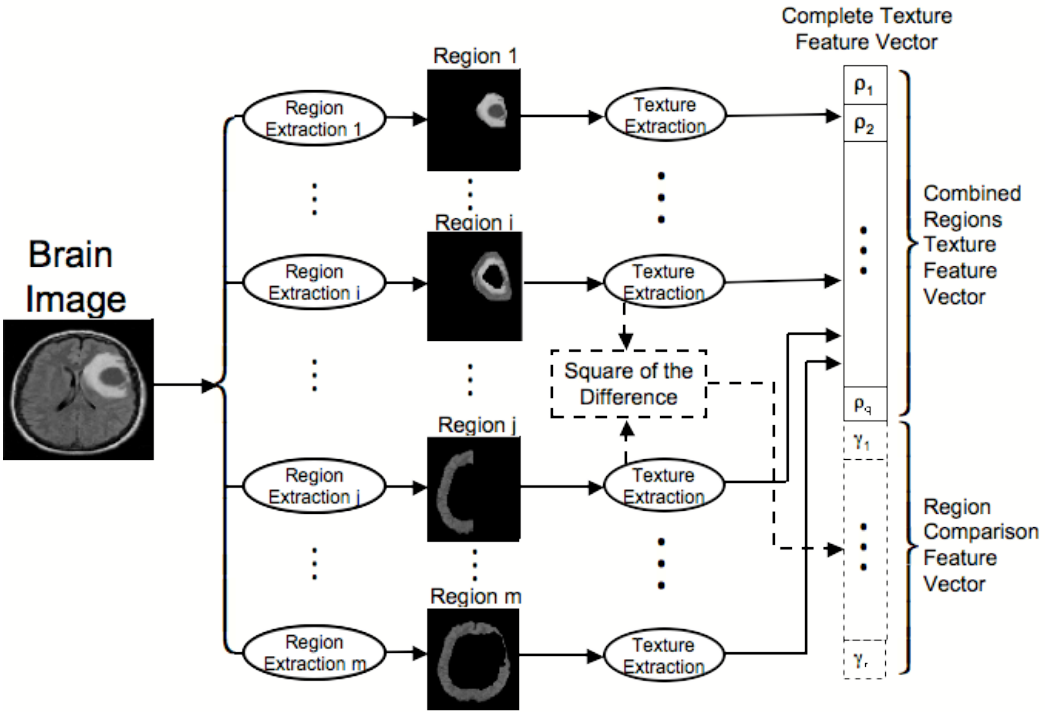


Figure 3.10: Complete texture feature extraction framework on a given image. First regional texture feature vectors are built. Then selected regions are compared by obtaining the difference vector from their region texture feature vectors. Then the results from the regional texture feature vectors and the region comparison vectors are consolidated to obtain the complete texture feature vector for the MR image.

are eventually combined with the final texture feature vector to obtain the final feature vector for a patient. Construction of these features is simpler than the texture features and is described in detail in Section 4.2.5.

3.3 Raw Data

The following data has been provided by the Cross Cancer Institute [24].

3.3.1 Glioblastoma Patients

We work with *Magnetic Resonance* scans of the brains of patients who are later, through histological reports from biopsy, diagnosed to have glioblastoma tumors. Each patient may have several scans, each with a known date. Some of the scans are pre-surgery and some are post-surgery. We chose to use only images that come from raw *pre-surgery* scans, as this avoids surgical artifacts that are introduced in the post-surgery scan images. Most patients have one pre-surgery scan but for those patients with multiple pre-surgery scan, we use only one for our experiments.

In total we have images for **55** patients. The images for the patients were collected in two batches. The first batch consists of patients, who were diagnosed between September 2006 and June

2007; these were collected in July 2008. The second batch consists of patients, who were diagnosed between July 2007 and December 2007; these were collected in December 2008. Most of these patients had passed away at the date of collection and so their *dates of deaths* are known. However, some patients were alive at the date of collection so their dates of deaths are unknown. For the patients with known dates of deaths, we define **survival** to be the time period between the scan date and the date of death, measured in weeks. For the patients with unknown dates of deaths, we take the time period between the scan date and the collection date to be an underestimate for their survival time, also measured in weeks. These patients are said to be *right censored* [12].

The data collected for our patients unfortunately include only images, age and sex. Therefore we cannot include other useful clinical data such as KPS, surgical and treatment parameters.

3.3.2 Brain Images

Every brain scan includes between 19 to 22 *axial* slices of the brain separated evenly according to the machine settings determined by the technician who took the scans. Each patient has MR scans of a number of modalities, typically a subset of T1-weighted, post-contrast T1-weighted, T2-weighted and FLAIR. Tissues with high fat content appear bright on T1-weighted images. Tissues with high water content appear bright on T2-weighted images. FLAIR images are similar to T2-weighted images but free water regions such as cerebrospinal fluid in the ventricals is supressed. Edema, which is the swelling caused by accumulation of fluid, shows up bright on FLAIR and T2-weighted images. However, not all our patients have all the modalities. As a result, we decided to use only FLAIR images, which is the modality that is most prevalent amongst all the patients. Figure 3.11 shows all 20 slices of a pre-surgery FLAIR volume for one patient. The FLAIR images come with different dimensions and ranges of gray values depending on their scanning and machine settings. Therefore, in order to standardize the images dimensions and gray level ranges, we re-scale the images to be 258-by-258 pixels, where each pixel has a gray level in the range 0 to 255.

To make meaningful assessment about the tumors, we needed tumor-segmented MRI scans. Unfortunately, our images were collected without accurate expert segmentation. We therefore segmented each image ourselves. Therefore, for each FLAIR volume, we manually segmented the abnormal regions of the brain in each axial slice. We define an abnormal region to be an area of visible tumorous tissue, which sometimes includes a necrotic center, in addition to the edema surrounding it. Our manual segmentation may not be accurate, as we believe that perfect segmentation is not crucial to our prognosis method. Nevertheless, we tried to be as consistent as possible. There are a number of automatic segmentation tools available. We believe our manual segmentation is sufficiently consistent, moreover this task gave us an opportunity to familiarize ourselves to the different shapes and appearances of glioblastoma tumors.

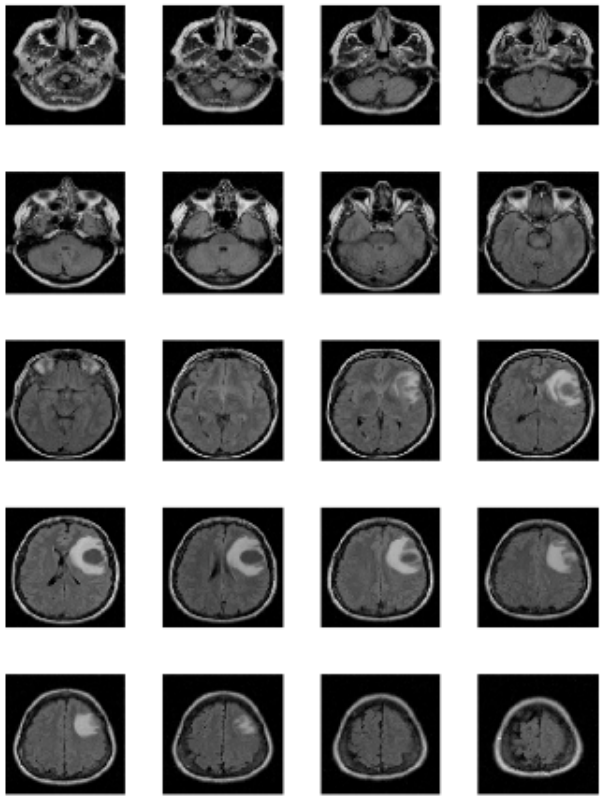


Figure 3.11: Slices of a FLAIR volume.

Chapter 4

Survival Prediction Framework

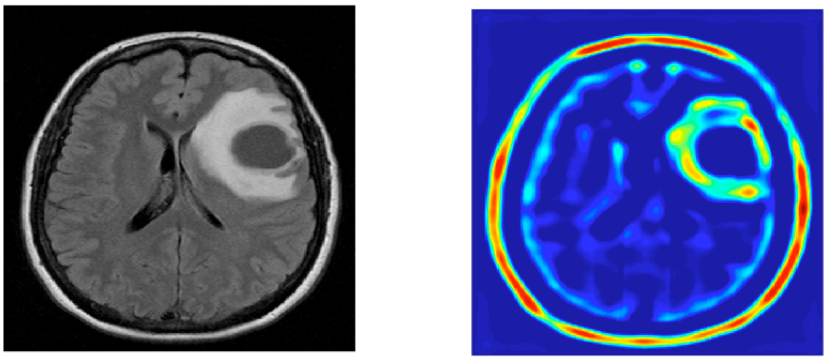
Section 4.1 describes the texture extraction methods used in the texture feature extraction process. Section 4.2 describes our feature consolidation scheme, which involves combining regions from axial brain image slices to extract texture features, and then computing statistics based on the extracted texture, to be used in the subsequent classification framework (Section 4.3).

4.1 Texture Extraction Methods

An MR image is a grayscale image, which can be represented by a matrix of pixels. Different textural properties can be extracted from a region of interest in the image by performing computations that use gray values of the pixels and/or the distribution of gray values within neighborhood(s) of the pixels within the region. The result of the computations can be an image or a matrix, representing either local or global properties of the original image or region of interest.

A **local** texture extraction method is one where for every pixel in the image, the gray values of the pixel's neighboring pixels are utilized to compute the local textural properties around that pixel. The result of a local texture extraction on an image is an image with the same dimensions as the original image. There is a pixel-to-pixel correspondence between the original image and the resulting image. The intensity value of a given pixel in the resulting image represents the local textural properties of the corresponding pixel in the original image. Such resulting image is called a **texture image**. Figure 4.1 shows an example of a raw image and its corresponding texture image obtained by using a Gaussian filter (described in more detail in Section 4.1.4).

If a texture extraction is not specified to be local, then all the pixels in the region of interest can be used collectively to compute textural properties that *pertain to the whole* region. In this case, the result of the texture extraction can be a single value or a matrix. If the result is a matrix, unlike in local texture extraction, this resulting matrix should not be assumed to have the same dimensions as the original image. The resulting value or matrix of a non-local texture extraction is a representation of textural properties of the whole region. In other words, there is no pixel-to-pixel correspondence between the original image and the resulting matrix.



(a) Raw Image (b) Texture Image

Figure 4.1: A raw MR image and its corresponding texture image produced by a Gaussian filter.

4.1.1 Basic Statistics

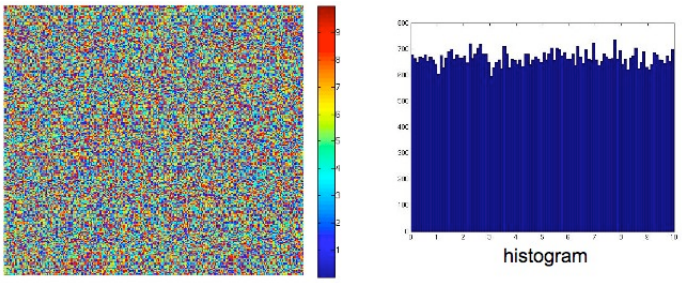
Basic statistics are used on regions of interest (which will be defined in Section 4.2.1) on raw or texture images to characterize their textural properties. For any region of interest, the *mean* and the *standard deviation* of the gray values in the region can be used to measure the spread of gray values of the pixels within that region. For example, a relatively *dark* region with a texture that can be characterized as *homogeneous* has a relatively low mean and a low standard deviation, assuming that the lowest gray value is black and the highest is white on the gray color spectrum. The mean and standard deviation represent the gray level distribution in the region of interest.

Another useful statistic is **entropy** [32], which can be used on regions of interest from both raw and texture images. Given a region of interest in a grayscale image, entropy is a function of pixel intensities (or probabilities), which measures *uncertainty* in the region of interest. If the histogram of the region, which describes the frequency distribution of the gray values (e.g. Figure 4.2), is taken to be a probabilistic distribution, then the entropy computed using the histogram is a measure of the region's *randomness*. Let $h = h_1, \ldots, h_n$ be a normalized histogram of an image, where h_i for $i = 1, \ldots, n$ is the frequency of gray values that fall into bin i. Then the entropy for the image is given by:

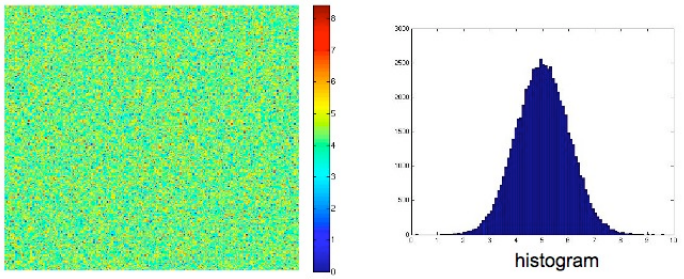
$$\text{Entropy} = -\sum_{i=1}^{n} h_i \log h_i \tag{4.1}$$

In our experiments, we chose a granularity of $n = 100$ for the histograms in the entropy computations. We chose this level of granularity arbitrarily, subject only to the condition that it be sufficiently large.

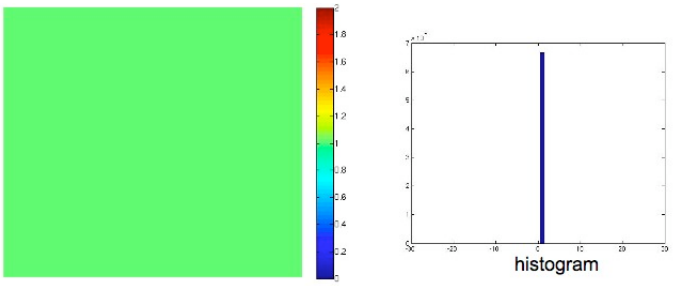
An image with a *uniform* distribution of gray values has a high rate of randomness. In other words, the probability that a given pixel has a certain gray value is equal to the probability that the pixel has any other gray value. In this case the uncertainty is maximum. As an example, consider an image whose pixels only have binary values: 1 or 0. Also assume that half the pixels in the image



(a) Uniform noise: entropy 6.6



(b) Gaussian noise: entropy 5.6



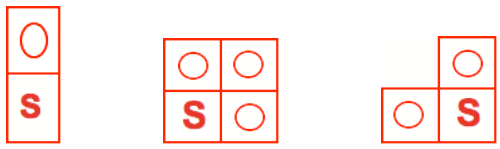
(c) All pixels equal 1: entropy 0

Figure 4.2: Images generated with different types of noise and their histograms. The one with uniform noise has the highest entropy.

are 1 and the rest are 0 (in which case the distribution is uniform). In this case, given a random pixel in the image, one cannot say with a high level of certainty that this pixel has a value of 1 (or 0), since the probability of any value occurring is 0.5. Therefore, the entropy, or equivalently the randomness, is highest. On the other hand, if an image only has values of 1, then one can say with 100% certainty that any given pixel in the image will have the value 1. Therefore, the uncertainty for such an image is minimized and the entropy is 0. Figure 4.2 shows three images with their corresponding entropy measures; one generated by uniform noise, one generated by Gaussian noise and one where all pixels have the value one.

4.1.2 Gray Level Co-occurrence Matrices: Second Order Statistics

The basic statistical tools, introduced earlier, extract *first order* statistics. First order statistics are measures that do not take into account the location of gray values relative to each other. Therefore, if



(a) (b) (c)

Figure 4.3: In the above neighborhood structures, the **S** indicates the pixel being referenced and the **O** indicates the offset.

the pixels in a region of interest were to be scrambled, the statistical results would remain the same. **G**ray **L**evel **C**o-occurrence **M**atrices (GLCM), first introduced in [21], use *second order* statistics. The central idea behind GLCMs is that gray values of pairs of pixels and their relative positions characterize certain textural properties.

The first step in building co-occurrence matrices is to specify a neighborhood structure, which in turn is used to construct the co-occurrence matrices from the region of interest in the grayscale image. Then second order statistics are computed on the co-occurrence matrices to characterize certain textural properties in the region of interest.

Neighborhood Structure

A gray level co-occurrence matrix is an n-by-n matrix, where n is the total number of gray levels in an image. It is similar to an adjacency matrix, except that gray levels need not necessarily occur adjacent to each other in the image; they can occur at any adjacent or distant arrangements, which are defined by the **neighborhood structure**. In other words, the purpose of a gray level co-occurrence matrix is to count, in a grayscale image, the number of times a certain gray level occurs together with another gray level in a neighborhood defined by the neighborhood structure. The neighborhood structure specifies which neighboring pixels or **offsets** are to be used in constructing the co-occurrence matrices. For example, an offset of *one above* as a neighborhood structure, shown in Figure 4.3(a), means that for every gray value pixel being referenced, the gray level of the pixel immediately above it is used in constructing the co-occurrence matrix. Therefore, assuming that there are n gray levels, $1, \ldots, n$, in a certain grayscale image, then an offset of *one above* as the neighborhood structure will count for each gray level g_i ($i = 1, \ldots, n$), the number of times each gray level g_j ($j = 1, \ldots, n$) appears immediately above g_i, or equivalently, the number of times the following pattern or ordering appears in the region of interest: $\begin{array}{|c|}\hline g_j \\ \hline g_i \\ \hline \end{array}$

A neighborhood structure can contain as many different offsets as considered suitable. For example, the neighborhood structure in Figure 4.3(b) has three offsets: one above, one to the right and one diagonal. In our experiments, we use the neighborhood structure shown in Figure 4.3(c), which considers adjacent pixels above and to the left of the referenced pixel.

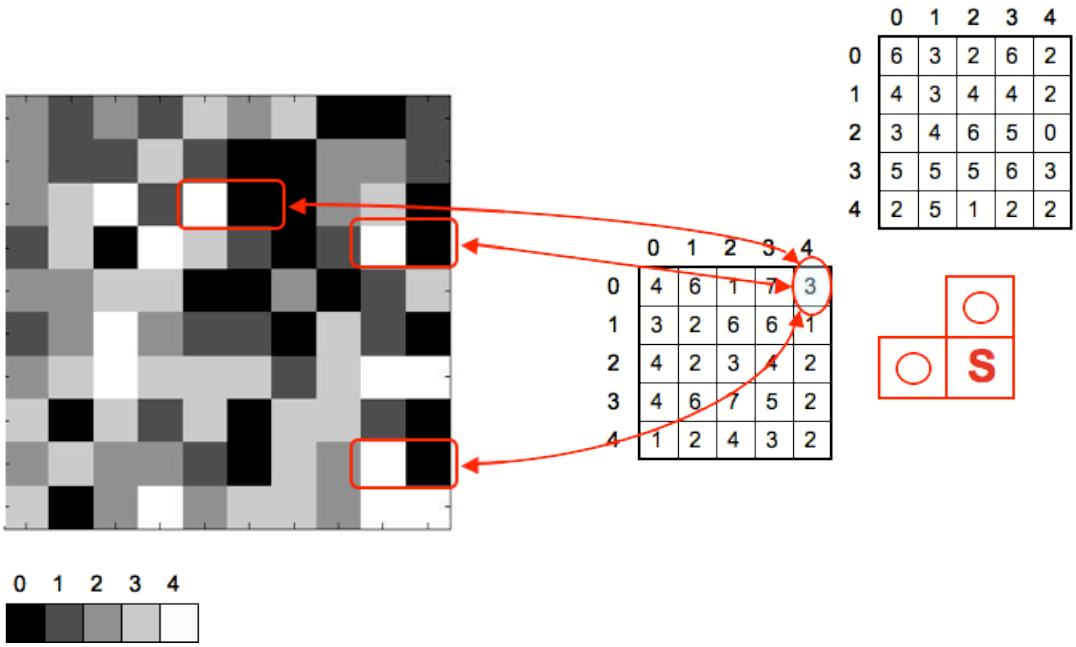


Figure 4.4: Construction of GLCMs. On the left, the grayscale image has 5 gray levels with values 0 to 4. Therefore, each co-occurrence matrix is 5-by-5. The neighborhood structure has 2 offsets and there is one co-occurrence matrix for each offset. In each matrix, the rows represent the gray levels of reference pixels and the columns represent the gray levels of offset pixels. For example, in the matrix for the offset left, as demonstrated in the Figure, gray level 0 appears three times with gray level 4 to its left.

GLCM Construction

For a given region of interest in the grayscale image, if there are n gray levels in total, then the dimensions of each co-occurrence matrix is n-by-n. The number of co-occurrence matrices is equal to the number of offsets in the neighborhood structure. Each row of a co-occurrence matrix represents the gray level of a pixel being referenced and the columns represent the gray levels of pixels that are offset to the reference pixel. Therefore, the number k_{ij} located at row i and column j of the co-occurrence matrix representing offset O, indicates the number of times gray level g_i appears with gray level g_j offset by O. Figure 4.4 shows the co-occurrence matrices built for a sample gray scale image using a neighborhood structure that has two offsets, each located at a distance of one from the reference pixel.

Once the gray level co-occurrence matrices are constructed, then each matrix M is normalized to transform the values M_{ij} from number of co-occurrences to probabilities (P_{ij}) of co-occurrences:

$$P_{ij} = \frac{M_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij}} \tag{4.2}$$

Second Order Statistics

Given normalized co-occurrence matrices, certain statistical properties can be measured that describe certain textural properties of the image. For example, the co-occurrence values appearing along the diagonal of a co-occurrence matrix represent the frequency at which pixels with the same gray levels occur together in the image. If for a certain image, the values along the diagonals of its co-occurrence matrices are large, then this image must have little contrast as this means adjacent pixels have similar values. On the other hand, if the values farther away from the diagonal of the co-occurrence matrices are more significant, then the image must have high contrast. The following are some statistical tools used to extract textural properties from a normalized co-occurrence matrix $P = (P_{ij})$:

$$\text{Energy:} \qquad \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} P_{i,j}^2} \tag{4.3}$$

$$\text{Entropy:} \qquad -\sum_{i=1}^{n}\sum_{j=1}^{n} P_{i,j} \log P_{ij} \tag{4.4}$$

$$\text{Contrast:} \qquad \sum_{i=1}^{n}\sum_{j=1}^{n} P_{i,j}(i-j)^2 \tag{4.5}$$

$$\text{Homogeneity:} \qquad \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{P_{i,j}}{1+(i-j)^2} \tag{4.6}$$

Energy is a measure of how uniform the texture is. Entropy is negatively correlated to energy and is a measure of randomness. When entropy is calculated based on co-occurrence matrices, it is a measure of randomness in co-occurrences, as opposed to entropy that is calculated based on the values in a raw image. Contrast is also negatively correlated with homogeneity. Due to these correlations, in our experiments we choose to use only energy and contrast. As mentioned earlier, we use the neighborhood structure shown in Figure 4.3(c) in our experiments. As a result, for each region of interest, we construct two co-occurrence matrices. We then apply each second order statistic to both matrices and then use the average of the two results. Therefore, two values, one for energy and one for contrast, are returned for the GLCM texture results.

Finally, note that for an image with 256 gray levels (as is the case for all our images) the co-occurrence matrices will be 256-by-256, which means the computations will be quite expansive. Also, for such large matrices and for so many gray levels, it is expected that the matrices will be quite sparse. As a result, using the mentioned statistical measures to extract textural characteristics from these sparse matrices would be less effective. Therefore, in order to avoid these issues in our experiments, we quantized the input grayscale images to 32 gray level grayscale images. Quantizing to 16 or 8 gray levels would increase computation efficiency, however, too much information would be lost in the process of quantization.

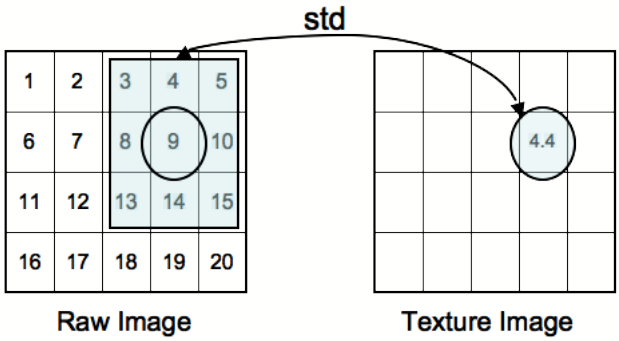


Figure 4.5: Calculation of local standard deviation values to obtain the corresponding texture image. Local entropy is computed in a similar way.

4.1.3 Local Statistics

We use the basic statistics of standard deviation and entropy to measure local standard deviation and local entropy in a given region of the image. To obtain local standard deviation measures for a region of interest, we compute for every pixel, the standard deviation of the neighborhood of that pixel and record the result in a texture image as shown in Figure 4.5. Local entropy is computed the same way, except that entropy is instead measured and recorded in the texture image. In our experiments, we use a square neighborhood of 9-by-9 for both local standard deviation and local entropy. The choice of window size is rather arbitrary but it includes enough of each pixel's neighborhood in the computations to capture the local texture. If the window size is too large, it will include pixels too far away from the center pixel of the window, which in turn may include too much information losing the locality of the measure. If the window is too small, it will not include enough information and the computations will be meaningless.

4.1.4 MR8 Filter Bank

Another useful group of texture extraction tools are linear filters. Similar to the local statistics method (Section 4.1.3), a linear filter is used to extract local textural properties. Therefore, filtering an image results in a texture image. For every pixel in a given region of interest in the image, the neighboring pixels of that pixel are used in conjunction with a **filter** to extract local properties. A filter is simply a matrix of scalars, which are used as weights in a linear combination of the pixels neighboring the given pixel. As an example, consider the filter in Figure 4.6 and a matrix representing an image in Figure 4.7.

For a given pixel, the result of this filtering centered on that pixel is a weighted sum of all the pixels (including the center pixel itself) in the area, where the filter is superimposed. The scalar values in the filter are used as weights in this weighted summation. Therefore the value stored in the corresponding location in the texture image is:

3	5	6
-5	2	4
-6	-4	1

Figure 4.6: Sample 3-by-3 filter

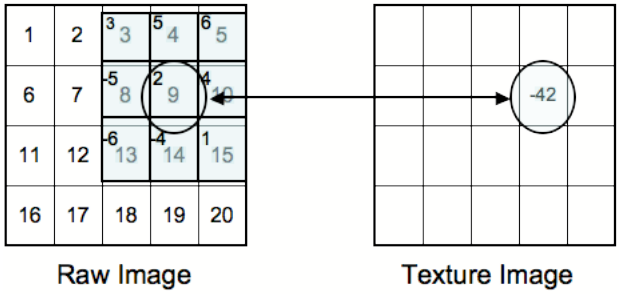
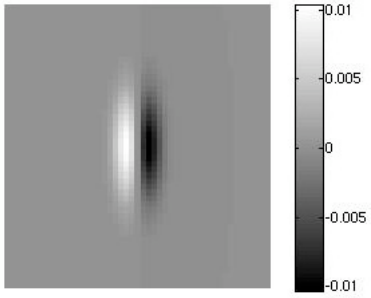


Figure 4.7: Sample filter used to create the texture image.

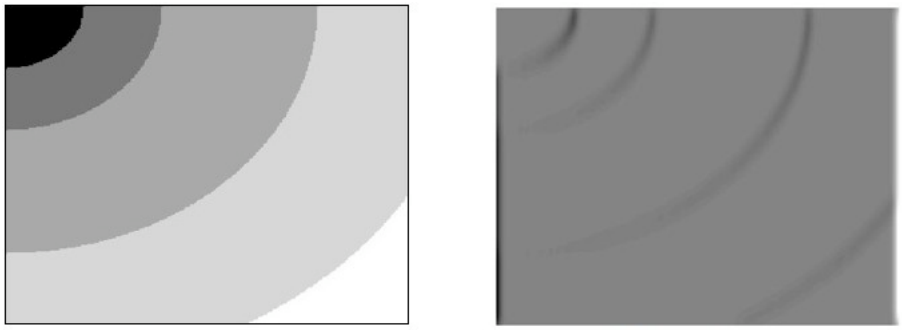
$$3 \cdot 3 + 5 \cdot 4 + 6 \cdot 5 - 5 \cdot 8 + 2 \cdot 9 + 4 \cdot 10 - 6 \cdot 13 - 4 \cdot 14 + 1 \cdot 15 = -42$$

Filters, much like GLCMs, are used to extract certain textural properties from images. For example, if a vertically oriented *edge* filter is used on an image, then there will be strong responses in areas of the resulting texture image, which correspond to areas in the original image containing vertical edges. Figure 4.8 demonstrates how a vertically oriented filter induces strong responses on certain areas of an image with mostly vertically oriented edges. Note that edges that are less vertically oriented induce weaker responses as seen in the texture image. One can understand how a filter induces responses by looking at the image of the filter itself. In the vertical edge filter, Figure 4.8, the bright and dark values indicate weights with higher magnitudes (positive and negative) and the values surrounding these correspond to weights with diminishing magnitudes. When an area of an image with a similar pattern as the filter is aligned with the filter, then most of the pixels in the pattern area are aligned with the heavier weights in the filter. Therefore, the heavier weights cause the pixel values within the pattern to contribute high amounts to the weighted sum while the smaller weights cause the areas outside the pattern to contribute minimal amounts. This results in a higher sum and thus a stronger response. When an area of an image with a dissimilar pattern as the filter is aligned with the filter, then most of the pattern is aligned with the lighter weights of the filter, resulting in a smaller weighted sum and therefore, a weaker response.

A *filter bank* is collection of filters used to create a collection of texture images, which then can be used as features for classification purposes. Filter banks can extract more complex textural properties from images than single filters. The **Maximum Response 8** (MR8) [61] is a filter bank that consists of a collection of edge and bar filters with varying scales and orientations, as well as a



(a) An enlarged view of a 49-by-49 edge filter



(b)

Figure 4.8: The resulting texture image when a 258-by-258 image is filtered with a vertically oriented 49-by-49 edge filter. Note that areas where edges tend to be somewhat vertically oriented induce stronger signals.

Gaussian filter and a *Laplacian of a Gaussian* filter. The MR8 filters are shown in Figure 4.9. The first three rows of Figure 4.9 consist of edge filters and the next three rows consist of bar filters. The bottom row shows the Gaussian and the Laplacian of a Gaussian filters. For each group of the edge or the bar filters in Figure 4.9, filters sharing rows have the same scale, and filters sharing columns have the same orientation.

Although there are 38 filters in the MR8 filter bank, we only obtain 8 texture images. This is because in the MR8 filters, for each scale (row in Figure 4.9), only the maximum response produced by an orientation is recorded for each pixel, hence the term 'maximum' in the name of the filter bank. As a result, 6 texture images are produced from the edge and bar filters, and 2 from the Gaussian and the Laplacian of a Gaussian filters.

An advantage of the MR8 filter bank is that, due to its use of varying orientations for bar and edge filters and also its use of the symmetric filters, Gaussian and Laplacian of Gaussian, the filter bank is rotation invariant. Another advantage of this filter bank is that it is small and efficient since only 8 texture images are produced. Figure 4.10 shows the resulting texture images from filtering a FLAIR image with the MR8 filters.

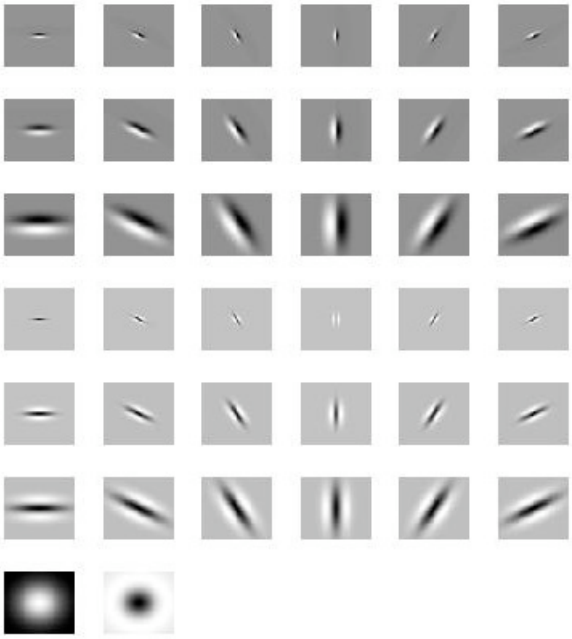


Figure 4.9: Maximum Response 8 Filter Bank: Rows 1,2 and 3 are edge filters with varying scales and orientations. Rows 4 to 6 are bar filters with varying scales and orientations. Row 7 contains the Gaussian and the Laplacian of a Gaussian filters.

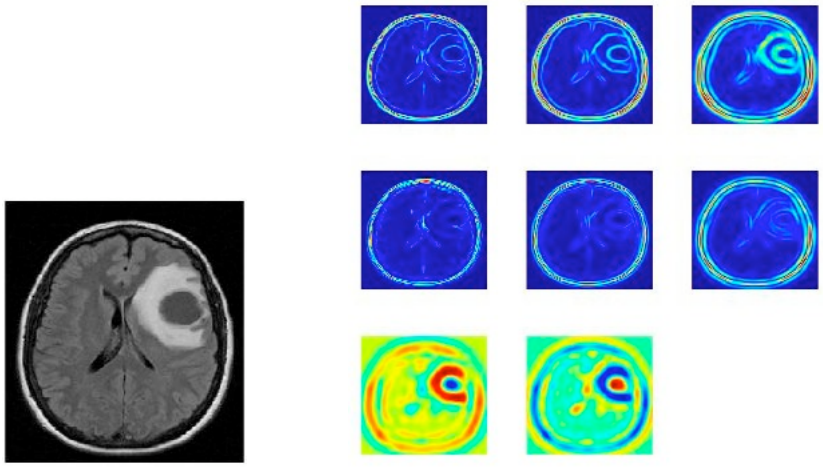
4.2 Texture Feature Extraction

As mentioned earlier in Section 3.2.2, we first define a set of brain regions for the purpose of texture extraction. In this section we describe how the regions of interest are determined and finally how the extracted textural measurements are combined with other patient information to form the features.

4.2.1 Brain Regions

When examining axial MR images of a brain tumor patient, certain regions of the brain are of prognostic interest. For example, lack of prominence in the sulci near the edema indicates a higher malignancy and degree of progression in the tumor since it is indicative of high levels of pressure under mass effect. Lack of prominence in the sulci farther away from the edema may be even more dangerous to the patient since this shows extreme levels of pressure caused by the tumor, which can block blood flow to critical areas of the brain and cause death if not treated immediately.

When comparing the MR image of a healthy brain to one with a malignant tumor, these sulci under pressure often have a more ‘blurry’ or ‘smudged’ texture characterized with fading edges and possibly lower contrast and higher homogeneity. These properties can be measured using our texture extraction tools. Therefore, we define a variety of regions for texture extraction purposes in order to capture and measure such textural properties in abnormal sulci. We are also interested to see if there are other textural properties, perhaps pertaining to other areas and tissues in the brain, that can be of prognostic relevance. For example, are sharp boundaries around the tumorous region



(a) Raw FLAIR (b) MR8 Texture Images

Figure 4.10: The resulting texture images from filtering a FLAIR image with the MR8 filters. The first row of the texture images is the result from edge filtering with varying scales, the second row is bar filtering with varying scales and the last row is the result of filtering with the Gaussian and the Laplacian of a Gaussian filters.

correlated to survival? Figure 4.11 displays a FLAIR image (original image) and a list of regions from which we extract textural properties. We now define these brain regions in detail. Note that the depth for regions *border*, *outer* and *rim* were chosen based on visual inspection so that they include just enough area from inside the segmented tumor (in *border*) or the non-tumor tissue in the brain (*outer* and *rim*) but not too much to dominate the regions.

1. **brain:** The brain region is obtained by removing the *skull*, which leaves only the *cerebrum*. The purpose for using this region is to investigate if there are any general textural properties pertaining to the whole brain tissue that can be of prognostic value. **Note: The regions that follow are all sub-regions of the brain region. No skull tissue is included in any of the regions.**

2. **inner:** The region within the edema (containing the tumor), which is segmented by hand as mentioned in Section 3.3.2. The purpose for this region is to investigate whether certain textural properties related to the edema, tumor and the *necrotic* region (region containing only dead cells) can be of prognostic value.

3. **brain&NOtum:** The region obtained by removing the segmented tumor from the brain. The purpose for this region is to investigate whether possible effects of the tumor on the texture of the rest of the cerebrum can be of prognostic value.

4. **border:** The strip containing the border of the segmented tumor, including the outer rim of the segmented tumor as well as the area immediately outside the tumor. If d_{max} is defined to be the maximum depth of the tumor measured from its border in pixel units, then the *border*

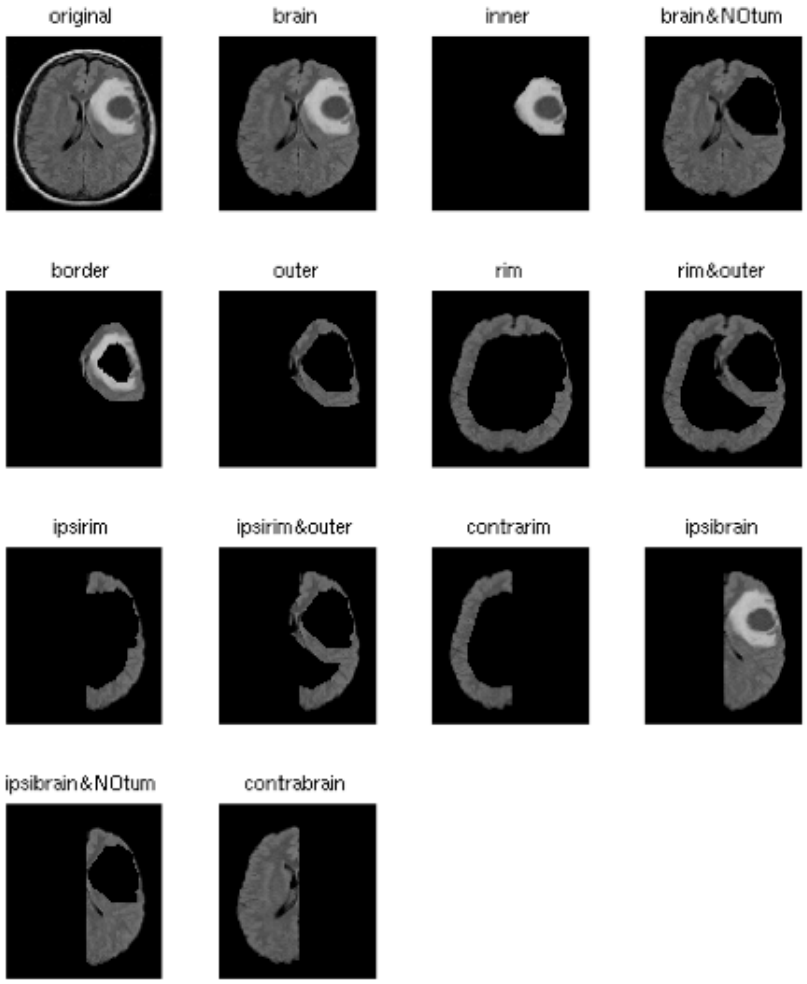


Figure 4.11: Original MR image on the first row and the brain regions, which we use in texture extraction. The original image is an axial FLAIR image of a glioblastoma tumor surrounded by edema.

region is taken to include a distance equal to 40% of d_{max} inward and 40% of d_{max} outward from the border. The purpose of including this region is to see whether a sharp and well defined border versus a fading one can be of prognostic value.

5. **outer:** The strip immediately outside of the border of the segmented tumor. If d_{max} is defined to be the maximum depth of the tumor measured from its border in pixel units, then the *outer* region is taken to include 40% of d_{max} outward from the border. The purpose for this region is to see if the texture of the tissue immediately surrounding the tumor, whether it includes sulci or ventricles or other tissue, can be of prognostic value.

6. **rim:** The outer rim of the cerebrum or the brain tissue excluding the segmented tumor. If b_{max} is defined to be the maximum depth of the *brain* region measured from its border in pixel units, then the *rim* region includes 30% of b_{max} inward from the border of the brain.

This region contains the majority of the sulci in the brain and we hypothesize that the texture of the sulci is of prognostic value.

7. **rim&outer:** The union of the *rim* and the *outer* regions. We would like to see if the texture of the *edges* of the soft tissue, which is under the pressure caused by the tumor (i.e. edges of tissue pushing against the tumor and edges of tissue pushing against the cranium) can be of prognostic value.

8. **ipsirim:** The outer rim of the brain that is on the same hemisphere (left versus right) that contains the tumor. If the tumor is a butterfly glioma, then one side is chosen arbitrarily. This region contains sucli closest to the tumor and we hypothesize that the texture of this region is of prognostic value.

9. **ipsirim&outer:** The union of the *ipsirim* and the *outer* regions. We would like to see if the texture of the *edges* of the soft tissue, which is under the pressure caused by the tumor, and is mostly located on the same hemisphere (left versus right) that contains the tumor, can be of prognostic value.

10. **contrarim:** The outer rim of the brain that is on the hemisphere (left or right) opposite the one containing the tumor. This region contains sucli further away from the tumor and we hypothesize that the texture of this region is of prognostic value.

11. **ipsibrain:** The full hemisphere of the brain containing the tumor. We would like to see if the contrast and other textural effects introduced by the tumor on the hemisphere containing it can be of prognostic value.

12. **ipsibrain&NOtum:** The *ipsibrain* region excluding the segmented tumor. This region contains the tissue relatively close to the tumor and since the texture of this region can be affected by the amount of pressure exerted by the tumor, we hypothesize that its texture is of prognostic value.

13. **contrabrain:** The full hemisphere of the brain opposite where the tumor is located. We would like to see if the textural effects caused by the tumor pushing against the opposite hemisphere can be of prognostic value.

4.2.2 Feature Construction

This section describes how we use a set of texture extraction tools (Section 4.1) to obtain features for a set of brain regions (Section 4.2.1). For each brain region, we extract texture information from basic statics, local statistics, Gray Level Co-occurrence Matrices, and Maximum Response 8 filter bank as follows. First, let B_i for $i = 1, \ldots, 13$ represent a brain region of a slice of an axial MR image for a patient. Then the texture feature vector F_{B_i} for the brain region is composed of the following:

- The basic statistics of mean, standard deviation and entropy calculated on the raw region and represented by single real values:
  $\{\mu_{B_i}(\text{raw}), \sigma_{B_i}(\text{raw}), H_{B_i}(\text{raw})\}$

- Local entropy, local standard deviation, and each of the MR8 filters produce a texture image from the raw region. Then for each of these texture images, the basic statistics of mean, standard deviation and entropy are calculated. This results in $(2 + 8) \cdot 3 = 30$ single real values:
  $\{\ \mu_{B_i}(\text{localent}), \sigma_{B_i}(\text{localent}), H_{B_i}(\text{localent}), \mu_{B_i}(\text{localstd}), \sigma_{B_i}(\text{localstd}), H_{B_i}(\text{localstd}),$
  $\mu_{B_i}(\text{MR1}), \sigma_{B_i}(\text{MR1}), H_{B_i}(\text{MR1}), \ldots, \mu_{B_i}(\text{MR8}), \sigma_{B_i}(\text{MR8}), H_{B_i}(\text{MR8})\}$

- Second order statistics of energy and contrast are computed from GLCMs on the raw region. As mentioned in Section 4.1.2, since two co-occurrence matrices are constructed (one for each offset in the neighborhood structure), we arbitrarily choose to take the average of the two values produced by the second order statistic to obtain one real value. Therefore, we obtain two more values:
  $\{\text{eng}_{B_i}, \text{cont}_{B_i}\}$

4.2.3 Within-Image Region Comparison

Sometimes, it is possible for a tumor mass to deform the appearance of nearby tissue but leave untouched the tissue farther away. Furthermore, we would like to investigate whether differences in texture between certain regions inside the same brain image can be of prognostic value. Therefore, we introduce textural comparisons between regions within the same image.

Let F_{B_i} and F_{B_j} be texture features obtained for regions i and j as described in the previous section. Then, the within-image texture comparison features are obtained by taking the component-wise square of their differences:

$$F_{B_i,B_j} = (F_{B_i} - F_{B_j})^{\cdot 2} \tag{4.7}$$

$$= [(F_{B_i}(1) - F_{B_j}(1))^2, \ldots, (F_{B_i}(n) - F_{B_j}(n))^2] \tag{4.8}$$

where n is the total number of features in the feature vector. We hypothesize that the difference in texture between the sulci closer to the tumor and the sulci further away from the tumor is of prognostic value. We also hypothesize that the difference in texture between the tissue closer to the tumor and the tissue further away is of prognostic value. The regions we have chosen to be compared against each other are:

- inner-vs-outer
- inner-vs-brain&NOtum
- brain-vs-brain&NOtum

- ipsibrain-vs-contrabrain
- ipsibrain&NOtum-vs-contrabrain
- ipsirim&outer-vs-contrarim
- ipsirim-vs-contrarim

Here, *inner-vs-outer*, *inner-vs-brain&NOtum*, *brain-vs-brain&NOtum* and *ipsibrain-vs-contrabrain* represent differences in texture between healthy tissue and areas containing tumorous tissue. Additionally, the feature, *ipsibrain-vs-contrabrain* along with *ipsibrain&NOtum-vs-contrabrain*, *ipsirim&outer-vs-contrarim* and *ipsirim-vs-contrarim* represent the differences in texture between the tissue located on different hemispheres. Moreover, *ipsirim&outer-vs-contrarim* and *ipsirim-vs-contrarim* represent the differences in texture between the edges of soft tissue (including the sulci) affected by the pressure from the tumor.

4.2.4 Slice Selection

As mentioned earlier, each MRI volume has 19 to 22 axial slices of the brain. A tumor usually spans across many slices. Texture extraction from every slice of a patient repeated over all patients can be computationally intensive and thus time consuming and unnecessary. So it is helpful to wisely select a few slices to represent each patient. In order to ensure that the slice selection is not biased by human interference towards specifically choosing slices that may seem to improve performance, we decided to only consider methods that use generic automatic measurements. We experimented with different ways of selecting slices such as choosing the slices with the largest tumor area, or slices containing the middle section of the tumor, or slices that contain the upper and the lower sections of the tumor. But we finally decided to choose, for each scan, the slice that contains the largest area of the segmented tumor, together with the slice immediately above and the slice immediately below.

Once a set of three slices are selected for a patient, texture features (as described in Chapter 3) are extracted on each slice. This results in three slice-specific texture features F_{below}, $F_{\text{largesttum}}$ and F_{above}. We define the final texture feature vector $\mathbf{F}$ for a patient as the vector containing the component-wise maximum of the three slice-specific texture features:

$$\mathbf{F} = \max\{F_{\text{below}}, F_{\text{largesttum}}, F_{\text{above}}\} \tag{4.9}$$

$$= [max\{F_{\text{below}}(1), F_{\text{largesttum}}(1), F_{\text{above}}(1)\}, \ldots, max\{F_{\text{below}}(m), F_{\text{largesttum}}(m), F_{\text{above}}(m)\}] \tag{4.10}$$

where m is the total number of texture features. We experimented with minimum and average vectors as well, but we chose the maximum vector as it produced better results.

4.2.5 Inclusion of Non-texture Features

Along with texture features, we also include some other patient specific information in each patient's feature vector. These include *age*, *sex*, and general tumor properties such as *tumor volume*, *mass center invade* and *maximum diameter* of the tumor. To compute the tumor volume, we use the total number of pixels making up the tumor across all the slices in FLAIR volume. To compute the maximum diameter, we use Principal Component Analysis (PCA). The tumor (spanning across multiple slices) is treated as a three dimensional object. The longest diameter of this object is the diameter parallel to the pricipal component of the object obtained by the PCA method. *Mass center invade* is a measure of invasion through corpus callosum by the tumor and is performed on all three slices. For each slice, it is the percentage of tumor mass that invades the hemisphere that is opposite of where the center mass of the tumor is located. In cases, where there is a full butterfly appearance (the tumor is symmetric across the center line), this invasion is 50%. Once this measure is obtained for all three slices that have been selected as described in Section 4.2.4, their maximum is used as the mass center invade feature.

We use these features as clinical data for the patients and to compare how our prognosis performs using texture features with or in comparison with these clinical features. As mentioned in Section 3.3, age and sex are the only raw clinical data in our dataset.

4.3 Classification

Given a feature vector for each patient, the goal of our prognosis method is to predict a survival category for the patient. The label for each patient is the number of weeks to date of death (if known). We label each right-censored patient with the minimum time period for survival in weeks. In order to define categories of survival, we use the finite mixture-models implementation of Expectation Maximization (EM) [14] to find clusters of patients according to their survivals (both known and right-censored survivals combined). Figure 4.12 shows the histogram of the survivals and the distribution mixtures produced by the EM method. Based on the result of EM clustering, we choose two categories of survivals, and the cutoff is set at 30 weeks. Therefore, patients who have passed away before 30 weeks from their scan date are included in the *low survival* category (a total of 25 patients), and the patients who live longer than 30 weeks are included in the *high survival* category (a total of 30 patients). Note that with this categorization, all the right-censored patients are included in the long survival category.

As discussed in Section 2.2, censored data pose a particular challenge when determining survival categories. For two or more survival categories, if the label of a right-censored patient happens to be in a category other than the longest survival category, then this patient is usually removed from the dataset [13]. This is because the label of a right-censored patient determines the last time the patient was known to be alive and so whether the patient died at that date or lived longer is undetermined.

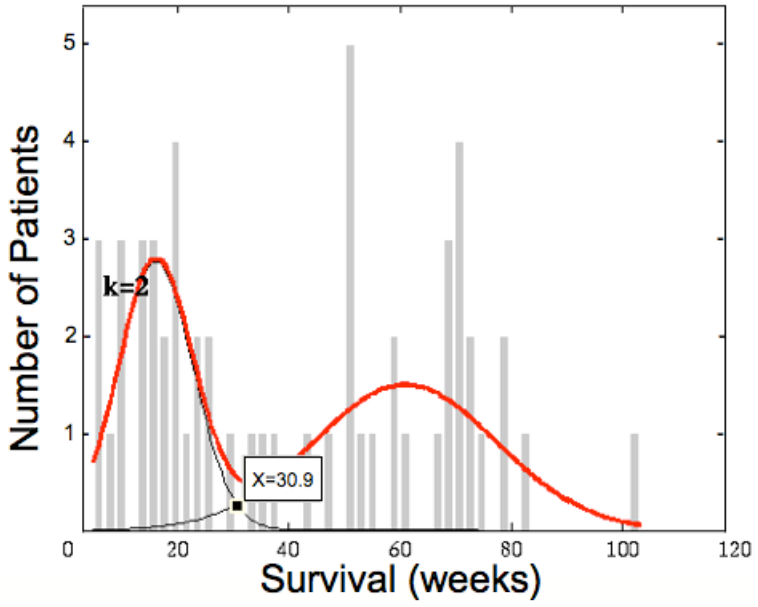


Figure 4.12: Histogram for patient survivals for all the patients, including both known and the right-censored survivals. Distribution mixtures produced by finite mixture-models (Expectation Maximization) determine two clusters.

As a result, if the label of a right-censored patient falls in a survival category other than the longest survival category, then one cannot assume that the patient died thus belonging to that category, nor can one assume that the patient survived long enough to be included in the longest surviving category. Therefore, the patient needs to be removed. In our case, all the right-censored patients fall in the longest surviving category. Therefore, we do not need to deal with this challenge.

Now, given a set of features for each patient, we use his/her survival category as the classification label in a binary classification. For our binary classifier, we use a C4.5 decision tree with pruning [47] implemented in Weka [64]. However, since each feature vector has 700 features, we use a feature selection scheme implemented in Weka called *Correlation-based Feature Subset Selection* [19]. It chooses a subset of the features where each feature is highly correlated with the class label while between-feature correlations are low. A resulting decision tree from a C4.5 model is intuitive and similar to Recursive Partitioning Analysis trees (Section 2.2). However, there are many other powerful classifying schemes commonly used for outcome prediction. According to Cruz and Wishart [10], aside from decision trees, some of the most widely used machine learning schemes in cancer prediction are Naive Bayes, Neural Networks and Support Vector Machines.

4.4 Summary

As described in the previous sections, we consider 13 regions for each brain MR slice image. For each brain region B_i, we compute a total of 35 texture features in its texture feature vector, F_{B_i}.

Also, we have 7 within-image region comparisons, each producing a texture difference vector of length 35. Therefore, each brain MR slice image will be represented by a complete texture feature vector F, which is the combination of all the F_{B_i} for $i = 1, \ldots, 13$ in addition to 7 within-image texture comparison vectors for a total of $35 \cdot 13 + 35 \cdot 7 = 700$ texture based features.

To be able to conveniently refer to certain texture features in our feature vector, we adopt the following naming scheme for each feature:

texture extraction method - first order or second order statistic - region

For example, *glcm-eng-inner* represents the texture feature value obtained from extracting GLCMs on the *inner* region using the second order statistic *energy*. As another example, *localent-std-border* represents the texture feature value obtained by extracting the local entropy on the *border* region using the first order statistic, standard deviation.

For each patient, three slices are considered and the complete texture feature vectors are computed for each slice. Then maximum values across the slices are chosen to form the final texture feature vector for a patient. Additionally, 5 non-texture features are included to form the final feature vector for each patient. Therefore, each patient has a feature vector of length 705. Then our learner uses a feature selection method to reduce the total number features before building a classifier. The complete feature extraction process is visualized in Figure 4.13. Once the features are extracted, then the learning process follows, as demonstrated in Figure 4.14.

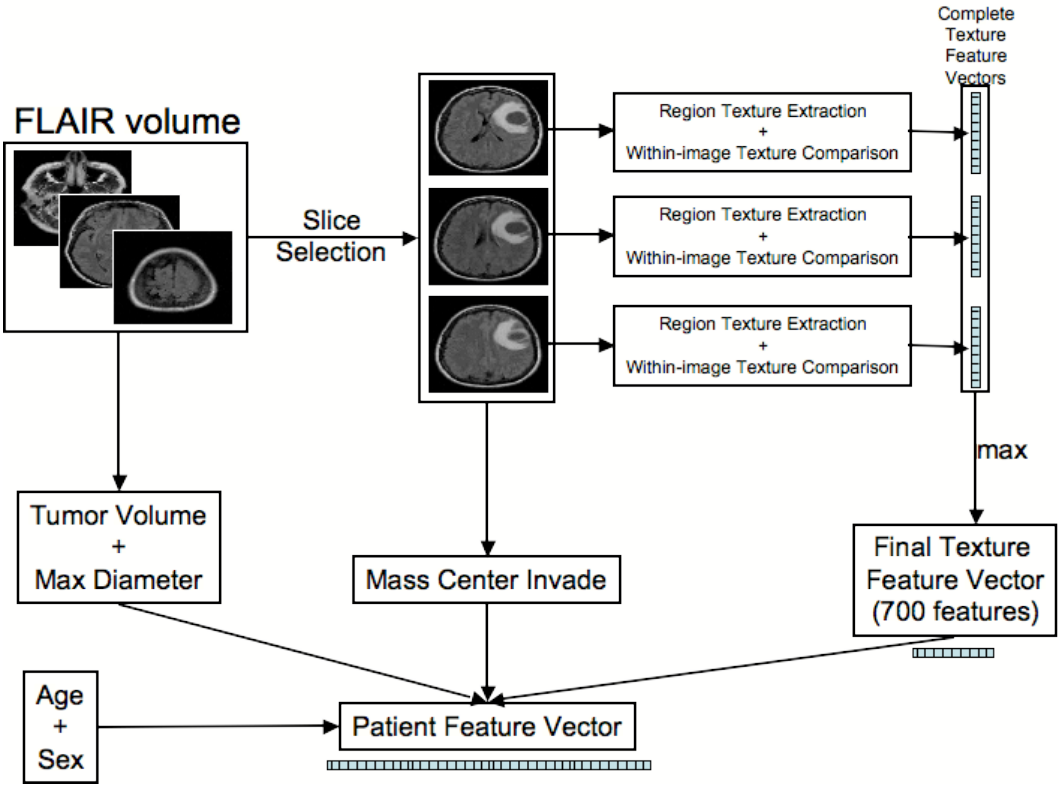


Figure 4.13: The complete feature extraction process for each patient. The image features are combined with age and sex to form a patient's feature vector for use in classification.

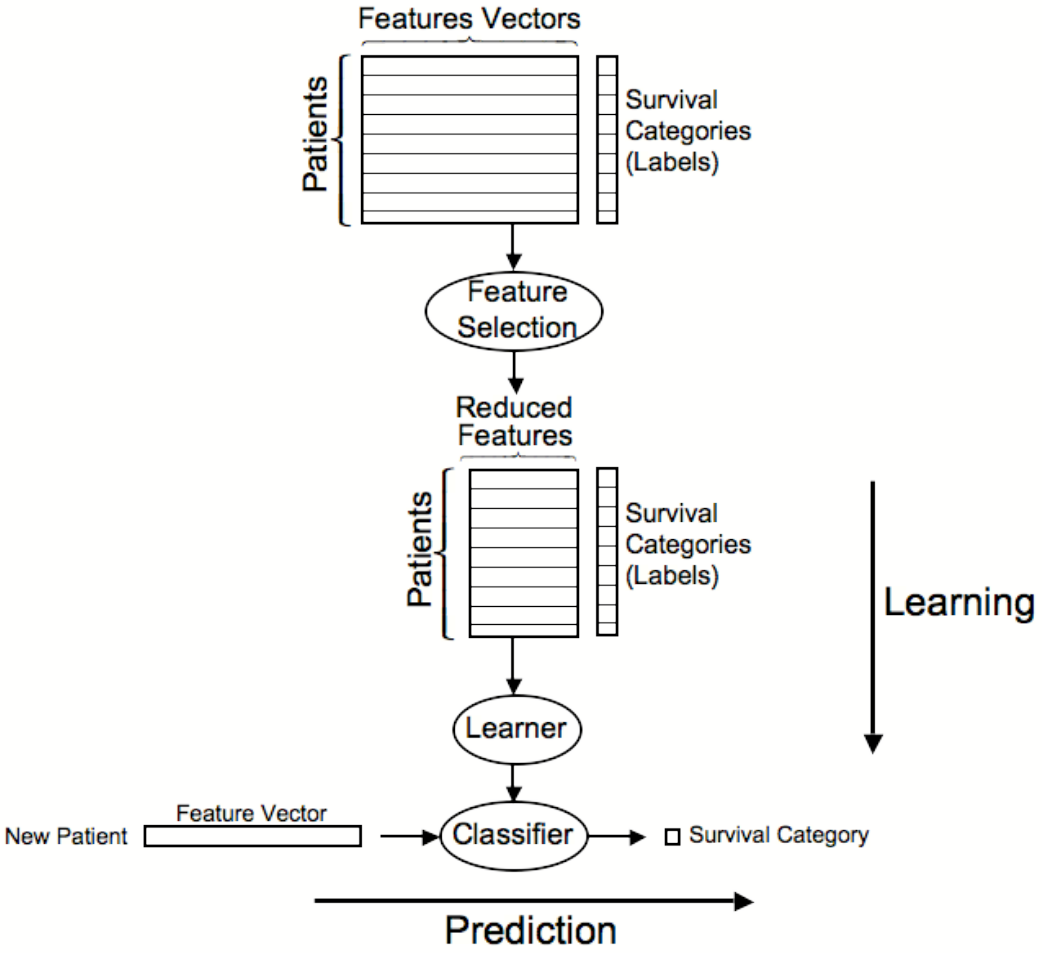


Figure 4.14: The complete classification process. First, a feature selection method is used to reduce the number of features used in learning. Then the features and the labels in the training set are used to build a classifier. Then this classifier is used to predict the label for a new patient.

Chapter 5

Results and Discussion

In this chapter we use our full prognosis framework as laid out in the previous chapters for survival prediction. In Section 5.1, we describe the evaluation measures we use in our evaluation scheme, cross-fold validation, which we describe in Section 5.2, for assessing the performance of our prognosis framework. Then, we test the performance on the data introduced in Section 3.3 using cross-fold validation and the evaluation measures discussed in Section 5.1. Finally, we display the results and comparisons in Section 5.3 and Section 5.4 and follow up with discussions in Section 5.5.

5.1 Evaluation Measures

For a given binary classifier model, we define its *accuracy* to be the proportion of correctly classified instances out of all instances. A perfect classification would be 100%. An accuracy of 50% would be considered the worst performance, since it would be equivalent to a classifier that ignores the training data and assigns class labels randomly at a flip of a fair coin. Accuracy, however, is not always sufficient to truly assess performance. Consider using a binary classifier to predict the labels of a set of data that has an imbalanced class distribution, e.g. where there are 90 instances of class A and 10 instances of class B. A trivial classifier model, which always predicts a label of A for every instance, would achieve an accuracy of 90% on this imbalanced dataset. But this classifier can hardly be considered useful. Therefore, in addition to accuracy, we use other measures to fully assess the prediction power of a binary classifier.

A *confusion matrix* (*contingency table*) is commonly used to encode the classification distribution of a classifying model. The confusion matrix for a binary classifier, which is built to predict positive instances (as opposed to negative instances) in a dataset, is a 2-by-2 table as shown in Table 5.1. Each row in this table represents the true labels and each column represents the labels predicted by the binary classifier. TPs is the total number of True Positive instances (i.e. positive instances that were correctly classified as positives); TNs is the total number of True Negative instances (i.e. negative instances that were correctly classified as negatives); FNs is the total number of False Negatives (i.e. positive instances that were incorrectly classified as negatives); and FPs is the

Classified as:	Positive	Negative
Positive	TPs	FNs
Negative	FPs	TNs

Table 5.1: Generic confusion matrix. TPs: Positive instances correctly classified as positive. FNs: Positive instances falsely classified as negative. FPs: Negative instances falsely classified as positive. TNs: Negative instances correctly classified as negative.

Classified as:	Positive	Negative
Positive	8	2
Negative	8	82

Table 5.2: Confusion matrix example: low precision, high accuracy and high recall.

total number of False Positives (i.e. negative instances that were incorrectly classified as positives). The sum of each row in Table 5.1 is the total number of instances in the class that is represented by the row. The sum of the diagonals (TPs + TNs) is the total number of correctly classified instances.

Precision, *recall* and the *F-measure* are also commonly used to assess how well instances of each class are classified by the classifier.

$$\text{Precision:} \quad \frac{TPs}{TPs + FPs} \tag{5.1}$$

$$\text{Recall:} \quad \frac{TPs}{TPs + FNs} \tag{5.2}$$

$$\text{F-measure:} \quad \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}} = \frac{2 \cdot TPs}{2 \cdot TPs + FPs + FNs} \tag{5.3}$$

Precision is the proportion of correctly classified positive instances out of all instances that were classified as positive. In other words, precision is the proportion of instances classified as positive that are actually positive instances. A classifier can have high accuracy, but poor precision. Consider a case where there are 10 positive instances and 90 negative instances in a dataset. Assume that a binary classifier results in the confusion matrix in Table 5.2. Here, the accuracy for this classifier is 90%, whereas its precision is 50%. Recall is the proportion of correctly classified positive instances out of all positive instances. In other words, recall is the proportion of positive instances that were correctly classified. In the case in Table 5.2, recall is 80%, which is quite high. Assume in another dataset, we get the confusion matrix show in Table 5.3. Here, recall is 50%, despite the fact that both precision and accuracy are high. F-measure is a performance measure that combines precision and recall, assigning equal importance to both measures. In fact, F-measure is the *harmonic mean* of precision and recall, assigning equal weight to both measures. The best possible F-measure is 1 and the worst is 0. For example, the F-measure for the classifier in Table 5.2 is 0.62 and the F-measure for the classifier in Table 5.3 is 0.67.

Classified as:	Positive	Negative
Positive	5	5
Negative	0	90

Table 5.3: Confusion matrix example: low recall, high accuracy and high precision.

In the case of our prognosis framework, which has two classes of low and high survival categories, we can treat either class as the positive class and then assess the resulting values for the measures precision, recall and the F-measure for that class.

5.2 Evaluation Schemes

Thus far, we have discussed evaluation measures commonly used for performance assessment. Now, we describe evaluation schemes, which use the discussed measures to produce reliable assessments and estimations of the prediction accuracy for a given machine learning model.

To test the predictive power of our prediction model, we must use an evaluation scheme that does not provide an overly optimistic assessment. With a small dataset such as ours, over-training is an important issue that we need to deal with. Over-training may occur when the classifier model fits the collection of training data too well producing misleading results. The main goal of a prediction scheme is to be able to correctly classify new data. Too often a classifier model is considered accurate when it was tested only on the same data on which it was trained [13]. A true assessment of a classifier is achieved when it is tested on previously unseen data. Over-training can be detected when a classifier, which obtains accurate results when tested on the training data, performs poorer on new test data than another apparently worse classifier. Therefore, the process of testing on new data is a requirement for an evaluation scheme to be considered reliable.

Dupuy and Simon [13] state that many outcome prediction studies in microarray analysis make the mistake of testing their prediction model on data that was used in building the model. This issue renders many results reported in such studies useless and unreliable. Another common mistake, according to [13], is the misplaced use of feature selection during the evaluation process. In many studies, relevant features are selected based on the whole dataset first. Then, the prediction model is trained on a subset of the data, then tested on another subset. Although the model is tested on presumably new data, the process of feature selection was performed on the whole dataset, including the test set, and therefore, the evaluation result may be too optimistic. Evaluation of a prediction model cannot be reliable if any part of the test data was involved in building the prediction model.

There are quite a few reliable evaluation schemes commonly used to test the performance of machine learning models. However, they differ in their levels of rigor. The simplest scheme is *split-sample* (*hold-out*) validation. A percentage of the dataset is set aside for training and the rest for testing. The classifier model, including feature selection as well as learning, is built on the training

set. Then the prediction power of the classifier is tested on the test set using the performance measures discussed earlier. Data may be randomly sampled for each split. A model can also be built on a dataset from one institution, and then tested on data from a different institution. However, this evaluation process may still produce misleading results, since the training set may not be representative. To overcome this representation problem, one may repeat the hold-out procedure several times, each time choosing a set of random samples for training and the rest for testing. The accuracy for each hold-out repetition is recorded and at the end the average of the accuracy values is reported.

A more rigorous and reliable evaluation scheme is *cross-validation*. In *k-fold* cross-validation, the set of patients is divided into k disjoint sets (folds) of equal size. For each $i = 1, \ldots, k$, fold i is set aside, then the rest of the dataset is used as the training set to build the classifier model (this includes feature selection as well). Then, fold i is used as the test set. This process is repeated for all folds. Note that for each fold, the feature selection scheme may choose a different set of features in building the classifier. There are two ways to report the accuracy of a classifier using k-fold cross-validation. One way is to record the accuracy reported for testing on each fold and at the end report the average and standard deviation across all the folds. This way of reporting the k-fold cross-validation results tells us how robust a prediction model is, since the validation model uses several test sets within the dataset. Another way to report the performance uses the fact that in cross-validation, each instance is used in some fold as a test instance exactly once. Consequently, each instance is classified exactly once. Therefore, the predicted labels can be used to produce a confusion matrix, which is based on the whole dataset.

For both split-sample and cross-validation schemes, the training and the test data may be chosen via *stratification*. Stratification is when the proportions of the classes in the whole dataset are preserved in each split (fold). For example, if 40% of a dataset belongs to class A and the rest to class B, then a randomly sampled subset used for training should contain 40% class A instances and the rest class B instances. This way, the training set is representative of the whole dataset.

5.3 Survival Prediction Results

In this section we describe the results of our survival prediction framework on our dataset of patients. As mentioned earlier, we use a C4.5 decision tree with pruning [47] using Correlation-based Feature Subset Selection. We describe each of our 55 patients with a feature vector of length 705, which contains both texture features (based on FLAIR images) and non-texture features. Each patient belongs to either the low survival (S1) category or the high survival (S2) category.

5.3.1 Decision Tree

The pruned decision tree resulting from the C4.5 algorithm run on the *whole* dataset is shown in Figure 5.1. Based on this tree, the texture features, *mr7-entropy-contrarim*, *mr3-std-inner-vs-brain&NOtum* and *mr7-entropy-rim* are predictive of survival categories. When tested on the whole dataset, this

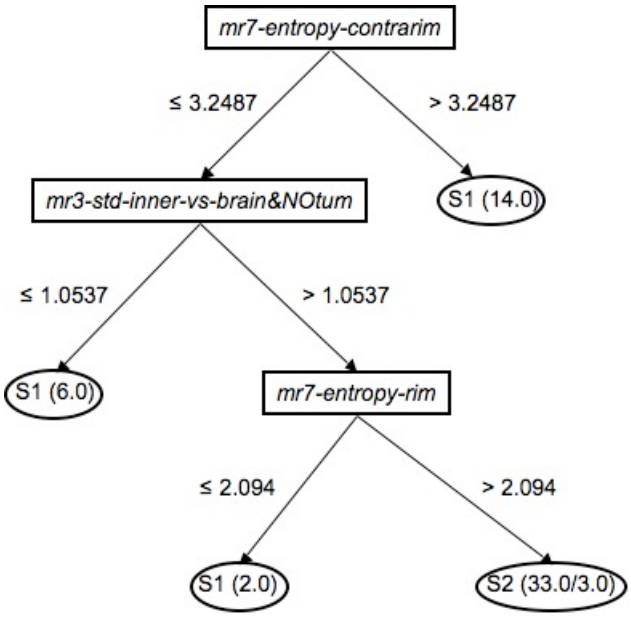


Figure 5.1: Pruned C4.5 decision tree built on the whole dataset for survival prediction

model misclassifies 3 low survival patients as high survival patients; see the bottom right leaf – Figure 5.1.

The *mr7-entropy-contrarim* feature is the entropy of the Gaussian filter responses in the FLAIR image computed over (essentially) the sulci region on the hemisphere opposite the one containing the tumor. This is a measure of "blurriness" of the sucli further away from the tumor. The feature *mr7-entropy-contrarim* is the entropy of the Gaussian filter responses computed over the entire outer rim of the cerebrum, where all the sulci are located. The feature *mr3-std-inner-vs-brain&NOtum* is the difference in variation of edge filter responses between the region within the tumor and the rest of the brain tissue. Figure 5.2 displays texture images for patients that are extreme cases with respect to these three texture features. A **" * "** next to the feature value of a patient indicates that the patient received the highest value for that feature amongst all the other patients and a **" v "** indicates the lowest value for the feature amongst all the other patients.

Patient **a** is an S1 patient, whose *mr7-entropy-contrarim* and *mr3-entropy-rim* values were the highest over all patients. Patient **b** is also an S1 patient, whose *mr7-entropy-contrarim* value was the lowest over all patients. Patient **c** is an S2 patient, whose *mr3-std-inner-vs-brain&NOtum* value was the highest over all patients. Patient **d** is an S1 patient, whose *mr3-std-inner-vs-brain&NOtum* and *mr7-entropy-contrarim* values were the lowest over all patients. Using these feature values, our decision tree in Figure 5.1 correctly classifies these patients.

5.3.2 Cross-Validation

We chose $k = 10$ in our k-fold cross validation and sampled the folds with stratification. The results of our survival prediction model for 10-fold cross validation are displayed in Table 5.4.

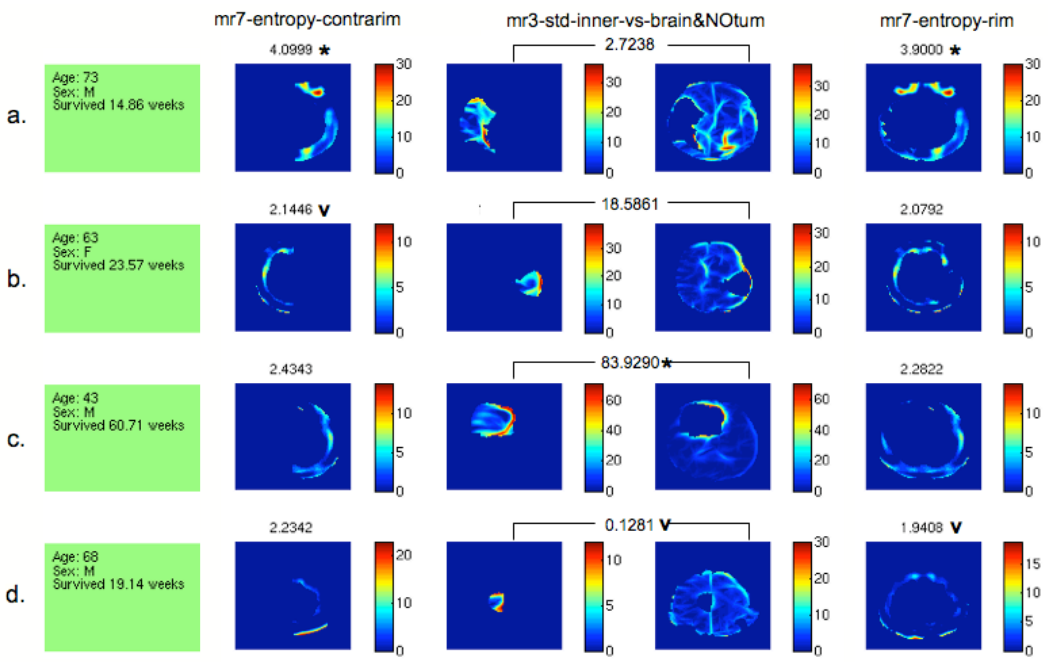


Figure 5.2: Extreme cases for each feature in the decision tree in Figure 5.1. A **" * "** indicates the highest feature value amongst all patients and a **" v "** indicates the lowest. Since the *mr3-std-inner-vs-brain&NOtum* feature is a within-image comparison feature of the two regions, *inner* and *brain&NOtum*, we display the texture images for both features.

This produced an accuracy of 80%, which may seem high. But the classifier had a more difficult time correctly classifying patients with low survivals, since 7 out of 18 low survival patients were misclassified as high survival patients. Hence, this results in a lower recall value (0.72) and consequently, a lower F-measure (0.766) for the class of low survivals.

The standard deviation of the accuracies across all 10 folds is 19.69. This high standard deviation indicates that our prediction model is far from robust. A prediction model is robust only if it can maintain high levels of accuracy with low variations when tested on any given datasets. In fact, our model obtained accuracies as low as 50% on one fold in the cross-validation.

Table 5.7 shows the mid-slice FLAIR image and the slices below and above for every patient that was classified as S1 in our 10-fold cross validation test. Table 5.8 shows the mid-slice FLAIR image and the slices below and above for every patient that was classified as S2 in our 10-fold cross validation test. Misclassified patients are labeled in both Table 5.7 and Table 5.8. Note that extreme-case patient **a** in Figure 5.2 is patient **6** in Table 5.7. Extreme-case patient **b** is patient **7** in Table 5.8, which is listed as misclassified in the 10-fold cross validation test. Note that this patient is correctly classified by the decision tree that was built on the *whole* data, whereas the decision tree that misclassified this patient in 10-fold cross validation was built on part of the data. Extreme-case patient **c** in Figure 5.2 is patient **22** in Table 5.8 and extreme-case patient **d** is patient **12** in Table 5.7.

Correctly Classified	**44/55**	**80** %
Classified as:	S1	S2
S1 = Low Survival	18	7
S2 = High Survival	4	26

Class	Precision	Recall	F-Measure
Low Survival	0.818	0.72	0.766
High Survival	0.788	0.867	0.825

Table 5.4: Prediction results for 10-fold cross validation.

Correctly Classified	**40/55**	**72.7** %
Classified as:	S1	S2
S1 = Low Survival	16	9
S2 = High Survival	6	24

Class	Precision	Recall	F-Measure
Low Survival	0.727	0.64	0.681
High Survival	0.727	0.8	0.762

Table 5.5: Prediction results for 10-fold cross validation with SVM as the classifier.

5.3.3 Decision Tree versus Support Vector Machine

Now we test the predictive power of our decision tree algorithm against another powerful classification algorithm called the Support Vector Machine (SVM) [45]. We keep the feature extraction and the feature selection part of our framework the same as before, but we replace the C4.5 decision tree learner with an SVM learner. The 10-fold cross validation results for our SVM classifier are shown in Table 5.5. Our SVM classifier achieves an accuracy of 72.7%. Based on the training data, an SVM classifier assigns weights to every feature, which can be indicative of the significance of the feature in predicting class labels. The weights assigned to the features by the SVM are shown below. Note that these features were the only ones chosen by our feature selection algorithm (Correlation-based Feature Subset Selection [19]), and thus were the only features considered by our SVM classifier:

2.2718 $*$ (normalized) mr8-entropy-inner-vs-outer

- **1.2373** $*$ (normalized) mr7-mean-ipsirim&outer-vs-contrarim

- **1.0634** $*$ (normalized) mr3-mean-ipsibrain&NOtum

- **1.025** $*$ (normalized) maxdiam

- **1.0065** $*$ (normalized) mr7-entropy-contrarim

+ **0.913** $*$ (normalized) mr3-entropy-inner-vs-brain&NOtum

+ **0.8868** $*$ (normalized) mr3-entropy-brain-vs-brain&NOtum

+ **0.7998** $*$ (normalized) mr7-mean-inner-vs-brain&NOtum

+ **0.1668** $*$ (normalized) mr3-mean-inner-vs-brain&NOtum

- **0.0656** $*$ (normalized) mr3-std-inner-vs-brain&NOtum

+ **0.043** $*$ (normalized) mr7-entropy-rim

+ 0.2232

The SVM classifier has assigned the highest weight to the texture feature, *mr8-entropy-inner-vs-outer*. The texture feature, *mr7-entropy-contrarim*, which was chosen as the first node by our C4.5 decision tree, has been assigned a relatively high weight by the SVM, as well. However, the non-texture feature, *maximum diameter*, has been assigned a higher weight than *mr7-entropy-contrarim*.

Our SVM classifier can achieve an accuracy of 72.7% in 10-fold cross validation, which seems to be lower than the accuracy achieved by our C4.5 decision tree classifier (80%). However, to confirm that our C4.5 decision tree performs statistically significantly better than our SVM classifier, we compare the results by performing 10-fold cross validation over 10 runs, where on each run, the data is randomly partitioned into stratified folds (so that the folds are different for each run). The average accuracy of our C4.5 decision tree classifier in 10 runs of 10-fold cross validation tests is 79.2% (std 17.55) and the average accuracy of our SVM classifier is 63.9 (std 20.15). Our decision tree significantly outperforms our SVM (paired t-test, $p < 0.05$).

5.3.4 Kaplan Meier Plots

In the process of designing stratified clinical trials, researchers often use the Kaplan-Meier Product-Limit method, usually with the logrank test, to verify that risk groups obtained by the RPA method, or any other statistical method, are statistically significantly different. Here we show that our decision tree survival prediction framework can produce survival groups that are statistically significantly different. Figure 5.3 shows the Kaplan-Meier plots for the patients that were predicted to be low-survival (S1) and patients predicted to be high-survival (S2). The plots are based on the 10-fold cross validation results in Section 5.3.2. The logrank test indicates that the difference between the two plots is significant ($p = 0.00118$).

5.4 Standardized Texture Image Statistics

Here we slightly modify our prognosis model in the texture extraction phase. When we look at the texture images produced by the local statistics and the maximum response fitlers (MR8) (as described in Section 4.1), we notice that the texture response values are not intuitive. In other words, the units for these responses are not known and it is not possible to tell whether a given response value, on its own, should be considered high or low. For example, what does a value of 1.0537 really mean for *mr3-std-inner-vs-brain&NOtum*, in Figure 5.1? One way to make these response values intuitive is to consider their values with respect to all the other values in the same image. Since every region described in Section 4.2.1 is a sub-region of the *brain region*, we can perform texture

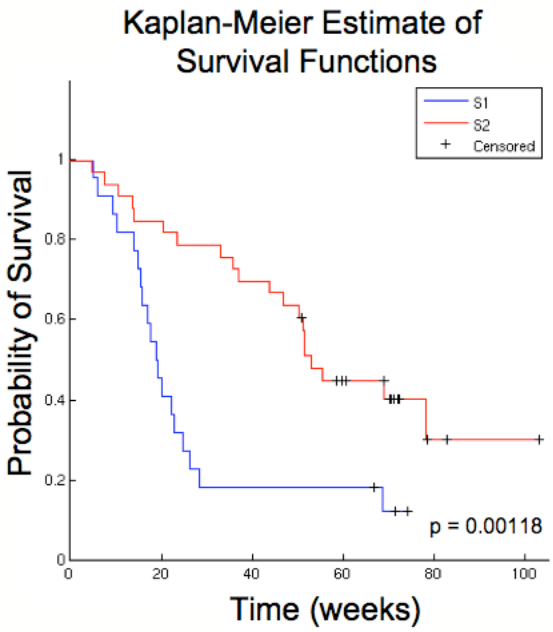


Figure 5.3: Kaplan-Meier plots of the predicted S1 and predicted S2 labels. The plots are based on the 10-fold cross validation results in Section 5.3.2

extraction on the full *brain region* first, then standardize the resulting brain texture image and finally proceed with computing basic statistics on the specified sub-region of interest. To standardize the brain texture image, $T = (t_{ij})$, each pixel, t_{ij}, in the brain region is converted to a standard score value, z_{ij}, by subtracting the mean of the brain region and then dividing by the standard deviation of the brain region:

$$\text{Standard Score:} \qquad z_{ij} = \frac{t_{ij} - \mu_{\text{brain\_region}}}{\sigma_{\text{brain\_region}}} \tag{5.4}$$

Note that we only use the mean and standard deviation of the brain region rather than the full image, since the full image contains many background pixels, which are irrelevant to our computations. Standardization re-adjusts the mean and standard deviation of the brain region in the texture image to 0 and 1 respectively. Thus, the resulting standard scores in the texture image are without dimension, and the units are now in the number of standard deviations a score is apart from the mean of the brain region. This slight modification and the original diagram from Figure 3.6 (Section 3.2.2) are shown in Figure 5.4.

There are a few points we need to address before we continue to the experimental results. First, note that the *brain region* itself is one of the regions used in feature extraction. It would not make sense to standardize the *brain region* if the whole brain region is to be used to compute basic statistics. This way, the resulting basic statistics measures of mean and standard deviation would always be 0 and 1 for every slice of the patient's FLAIR volume. Therefore, we apply this standardization step for every region described in Section 4.2.1, except for the *brain region*.

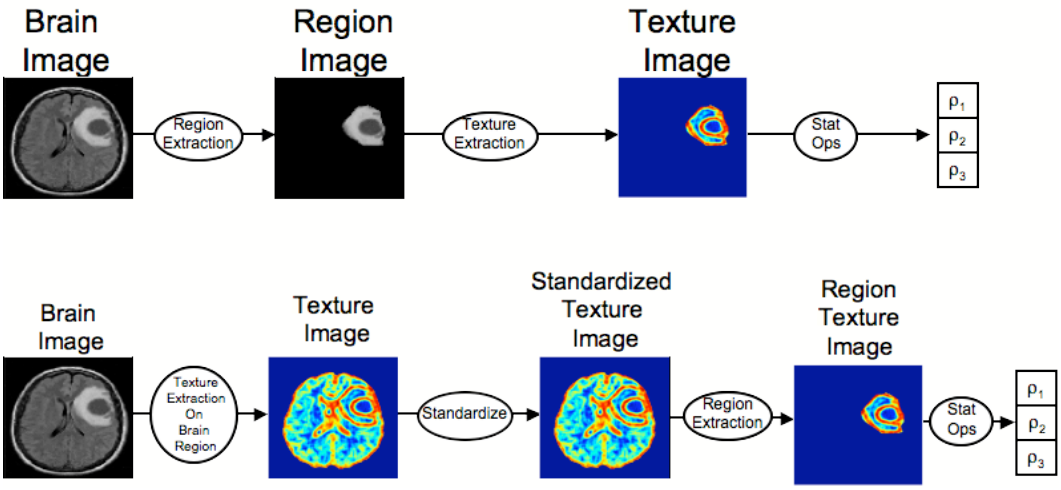


Figure 5.4: A modification to the texture extraction methods that produce texture images. Before basic statistics are computed on a sub-region of the *brain region*, texture extraction is performed on the whole *brain region* first, then the resulting brain texture image is standardized.

Secondly, note that this standardization step does not change the results for the basic statistics measure of entropy, as described in Section 4.1.1, since entropy deals with the shape of a distribution and is not affected by standardization. Finally, the co-occurrence matrices obtained for the GLCM features are already normalized to obtain probability values as described in (Section 4.1.2), and much like standard scores, probability is without dimension. Therefore, we use the second order statistic features obtained from co-occurrence matrices as before. We refer to this modified model as survival prediction with standardized texture images. The following sections describe the results of testing this model on our data.

5.4.1 Decision Tree

The pruned decision tree obtained based on the whole data is shown in Figure 5.5. Similar to the original survival prediction model, the texture feature *mr7-entropy-contrarim* is the most predictive feature for survival categories. Note that even though this is a feature built using a texture image (maximum response Gaussian filtered image), since the basic statistics measure of entropy is used, the standardization step does not affect the texture responses. Therefore, the root node of the tree in Figure 5.5 and the splitting value of 3.2487 are exactly the same as the decision tree (Figure 5.1) built with the original model. When tested on the whole data, this model misclassifies only 1 patient, who was predicted to have a high survival but really has a low survival.

5.4.2 Cross-Validation

The results of the modified survival prediction model with standardized texture images for 10-fold cross validation are displayed in Table 5.6. The modified model achieves a worse accuracy of 65.45%, while misclassifying 36% of the low survival patients. As a result, this model has a low

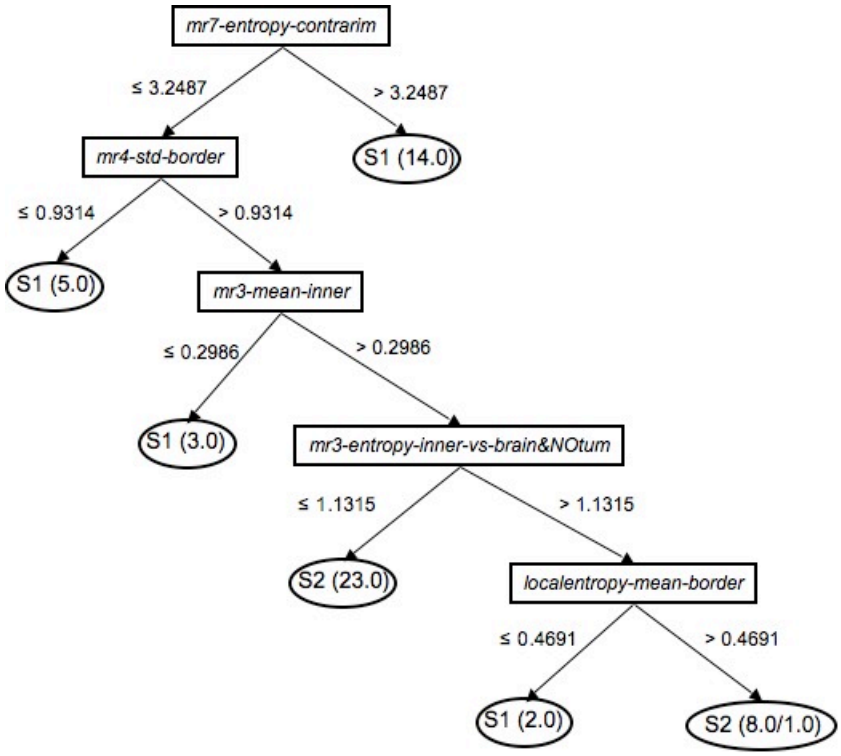


Figure 5.5: C4.5 decision tree built on the whole dataset for survival prediction with standardized texture images.

recall value for the class of low survival patients. The standard deviation of the accuracies across all 10 folds for the modified model is 18.07 and the highest accuracy achieved on a fold is 83.3%. Even though the additional standardization process made the texture image features more intuitive, it degraded the performance of our prediction model.

5.5 Discussion

According to the cross-validation results with the C4.5 decision tree classifier, both the original prediction model and the modified version with standardized texture images misclassify too many low survival patients as high survival patients. The original C4.5 model performs somewhat better with a higher recall for the low survival category. As there are 25 low survival patients as opposed to 30 high survival patients, there are fewer low survival patients to train the model on. Deciding which patients should belong to the low survival category and which ones belong to the high survival category can be a subjective matter. If we had raised the cut-off time above 30 weeks, we could have had more low survival patients. However, we wanted to divide our set of patients into survival categories in a meaningful manner. Therefore, we used mixture models to find a meaningful cut-off time in our data as demonstrated in Figure 4.12. Our original C4.5 decision tree classifier significantly outperformed our SVM classifier. SVMs are powerful classifiers, however, the C4.5 decision tree

Correctly Classified	**36/55**	**65.45** %
Classified as:	S1	S2
S1 = Low Survival	16	9
S2 = High Survival	10	20

Class	Precision	Recall	F-Measure
Low Survival	0.615	0.64	0.627
High Survival	0.69	0.667	0.678

Table 5.6: Prediction results for 10-fold cross validation. The model is the modified model with standardized texture images.

classifier is easier to understand by clinicians and performs better in our case.

Out of 705 features, our decision tree models chose the texture feature *mr7-entropy-contrarim* as the most predictive feature for survival. This feature is built by measuring the entropy of the *contrarim region*, which is passed through a Gaussian filter. This may imply that the blurriness of the sulci located on the hemisphere opposite the one containing the tumor is predictive of the patient's survival. Parameters such as age, maximum diameter and mass center invasion, which are commonly used by clinicians in prognosis assessment were not chosen by our decision tree prediction model. However, maximum diameter was a significant feature when the classifier was an SVM.

Figure 5.6 shows the distributions of the non-texture parameters in our data. As the figure shows, very high values in both maximum diameter and mass center invasion are indicative of low survival categories. A mass center invasion of 40% and higher is indicative of a butterfly glioma, which, as expected, corresponds to low survival. However, neither of these two features were chosen by our decision tree as the texture features were more significant predictors. In our data, as shown in Figure 5.6 and Figure 5.7, age does not correlate well with survival. In fact, the correlation coefficient between age and survival (in weeks) is -0.2252. However, this is not representative of much larger datasets since according to many studies, age is the most predictive parameter for survival in malignant brain tumors (see Section 2.1).

And finally, second-order statistical features based on GLCMs, although popular in the literature for segmentation and recognition tasks on MRI, do not exhibit much predictive power in our framework. The linear filtering MR8 features are the main features that are chosen by our feature selection method. This means that, out of all the features considered, the MR8 features have the highest correlation with patient survival in our database.

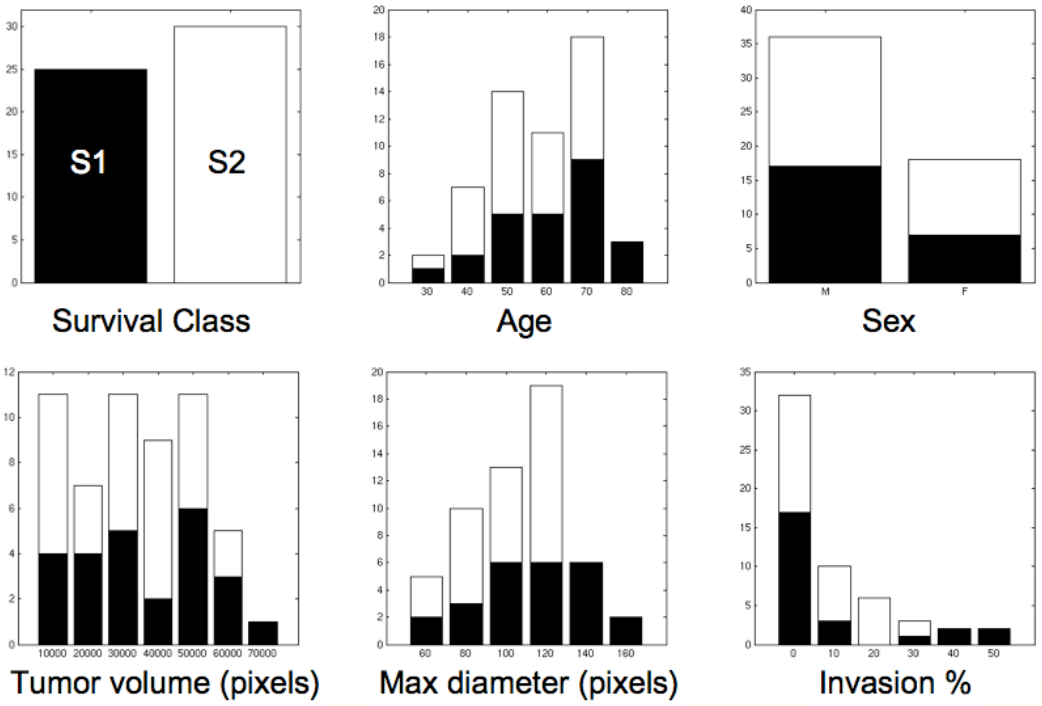


Figure 5.6: Distribution of non-texture parameters in our data. The distribution of maximum diameter and mass center invasion indicate that very high values correspond to low survival times.

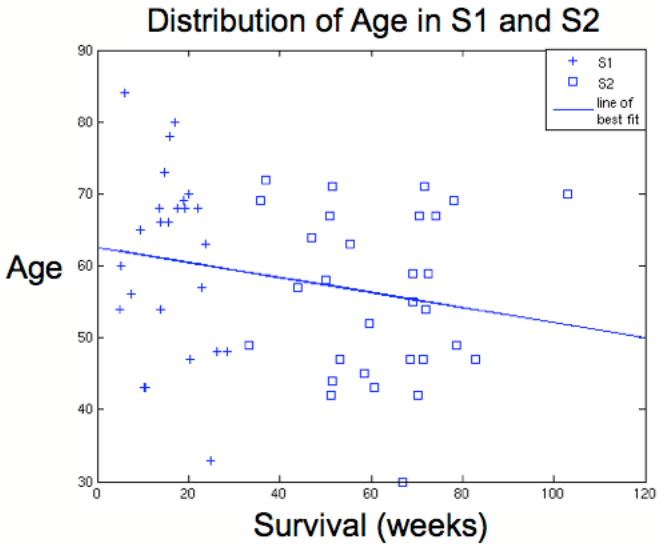


Figure 5.7: Distribution of age in our dataset and the line of best fit. Many population studies indicate that there is a high negative correlation between age and survival time. However, in our dataset, a correlation coefficient of -0.2252 indicates that there is very little correlation between age and survival time. Moreover, both low-survival (S1, i.e. below 30 weeks) and high-survival (S2, i.e. above 30 weeks) patients in our dataset appear to be scattered uniformly over a wide age range.

Table 5.7: The mid-slice and the slices below and above for every patient predicted to be in **S1** in the 10-fold cross validation test.

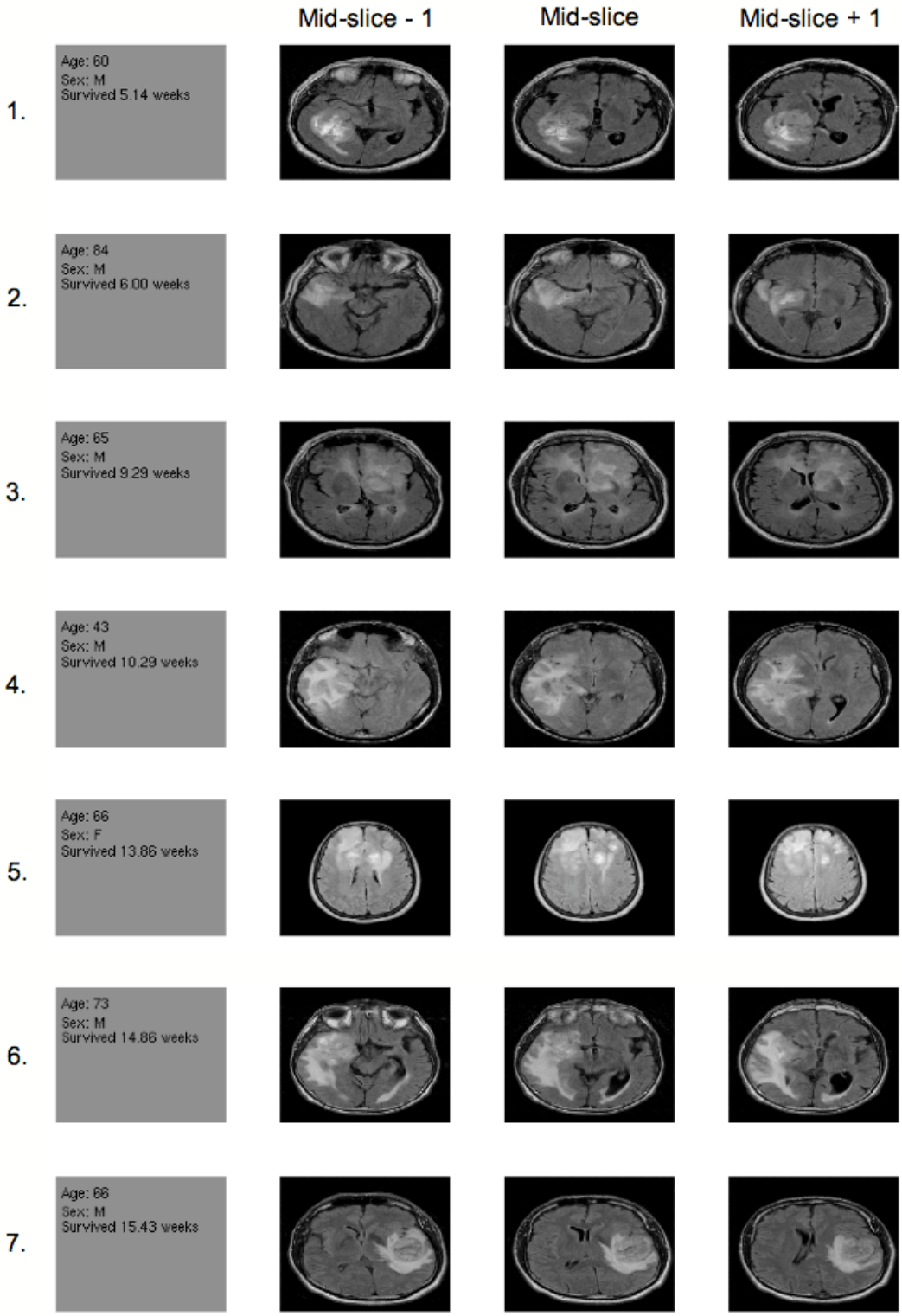


Table 5.7: (continued)

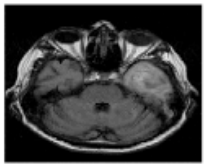
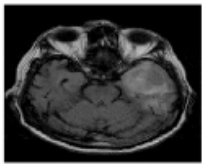
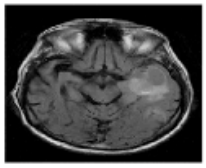
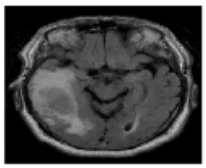
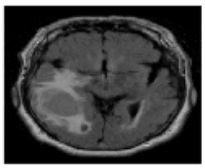
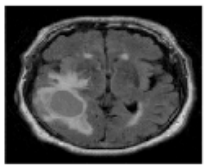
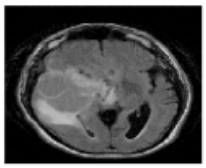
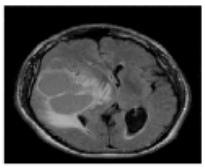
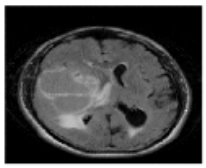
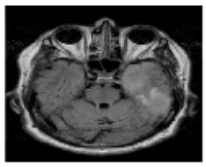
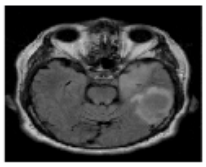
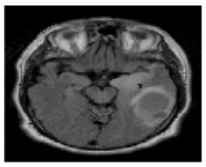
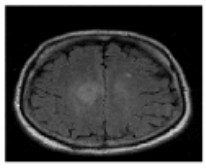
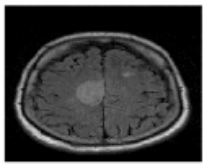
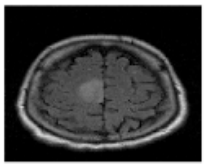
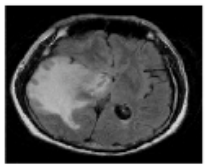
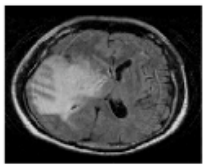
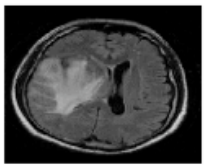
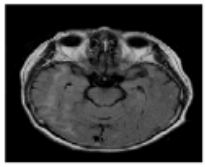
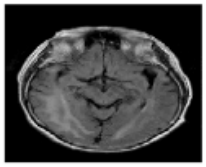
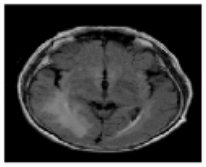
		Mid-slice - 1	Mid-slice	Mid-slice + 1
8.	Age: 78 Sex: Survived 15.86 weeks			
9.	Age: 60 Sex: M Survived 17.00 weeks			
10.	Age: 68 Sex: M Survived 17.57 weeks			
11.	Age: 69 Sex: M Survived 19.00 weeks			
12.	Age: 68 Sex: M Survived 19.14 weeks			
13.	Age: 70 Sex: M Survived 20.14 weeks			
14.	Age: 68 Sex: F Survived 22.14 weeks			

Table 5.7: (continued)

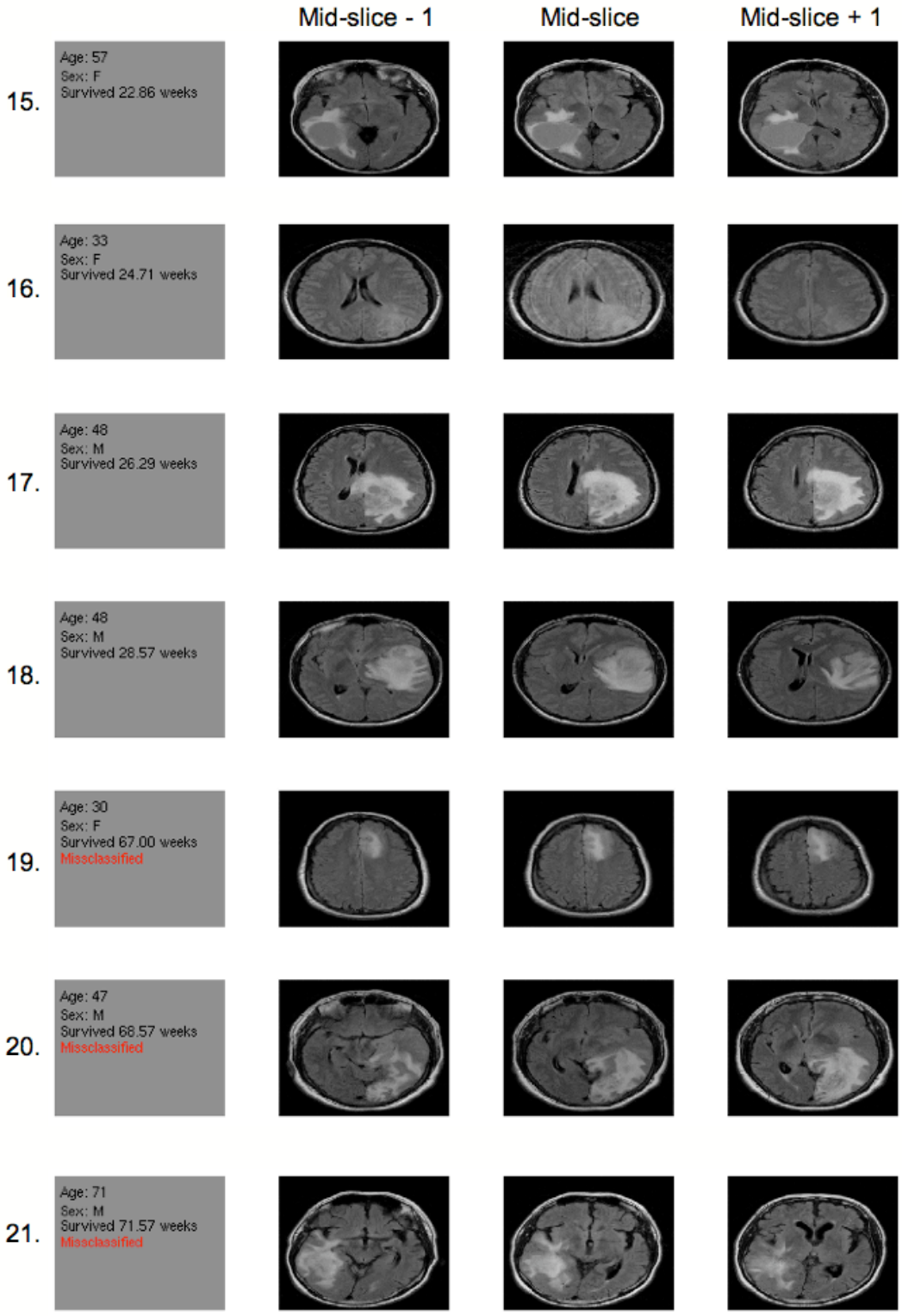


Table 5.7: (continued)

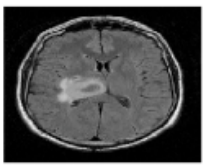
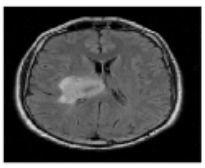
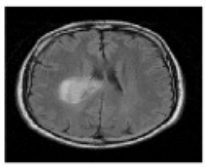
		Mid-slice - 1	Mid-slice	Mid-slice + 1
22.	Age: 67 Sex: M Survived 74.14 weeks Missclassified			

Table 5.8: The mid-slice and the slices below and above for every patient predicted to be in **S2** in the 10-fold cross validation test.

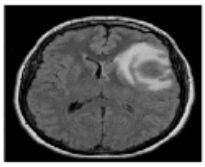
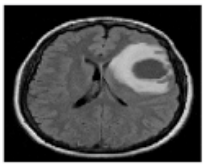
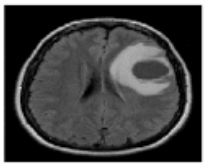
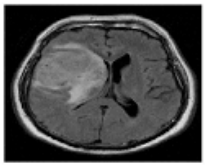
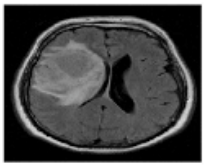
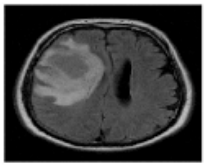
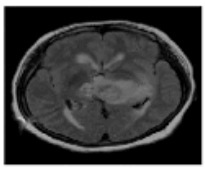
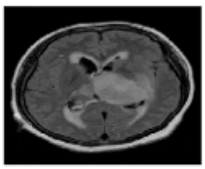
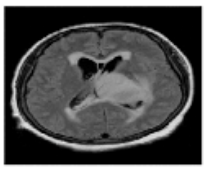
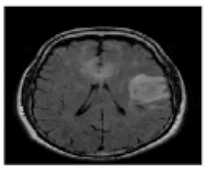
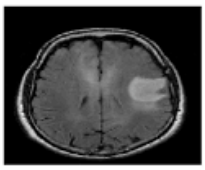
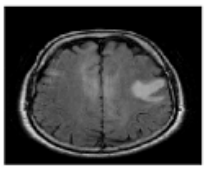
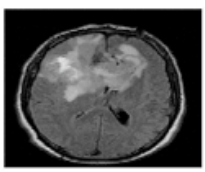
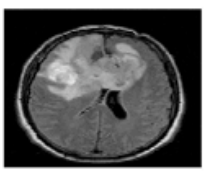
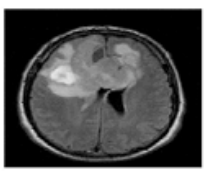
		Mid-slice - 1	Mid-slice	Mid-slice + 1
1.	Age: 54 Sex: F Survived 4.86 weeks Missclassified			
2.	Age: 56 Sex: M Survived 7.43 weeks Missclassified			
3.	Age: 43 Sex: F Survived 10.57 weeks Missclassified			
4.	Age: 68 Sex: M Survived 13.57 weeks Missclassified			
5.	Age: 54 Sex: M Survived 13.86 weeks Missclassified			

Table 5.8: (continued)

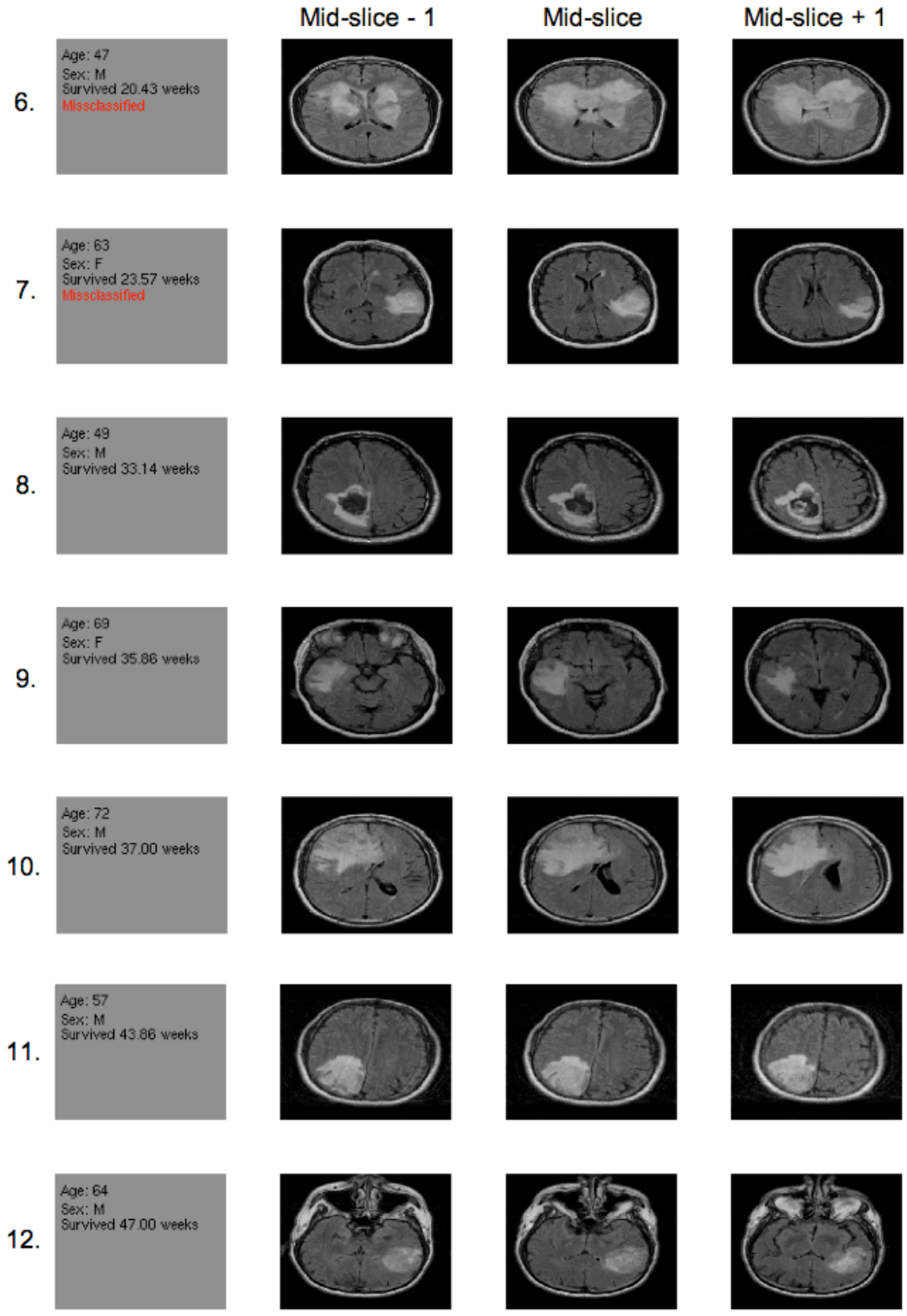


Table 5.8: (continued)

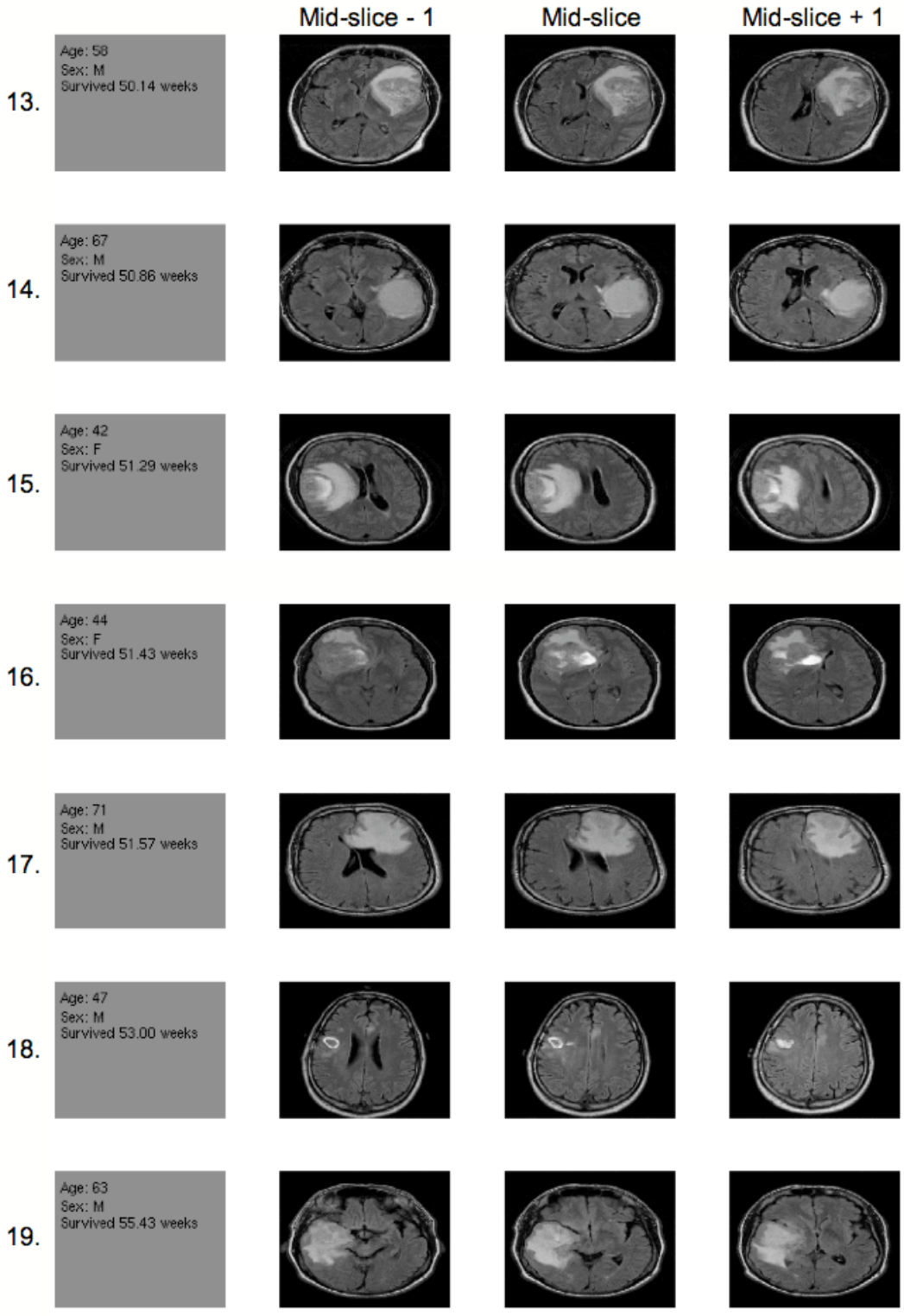


Table 5.8: (continued)

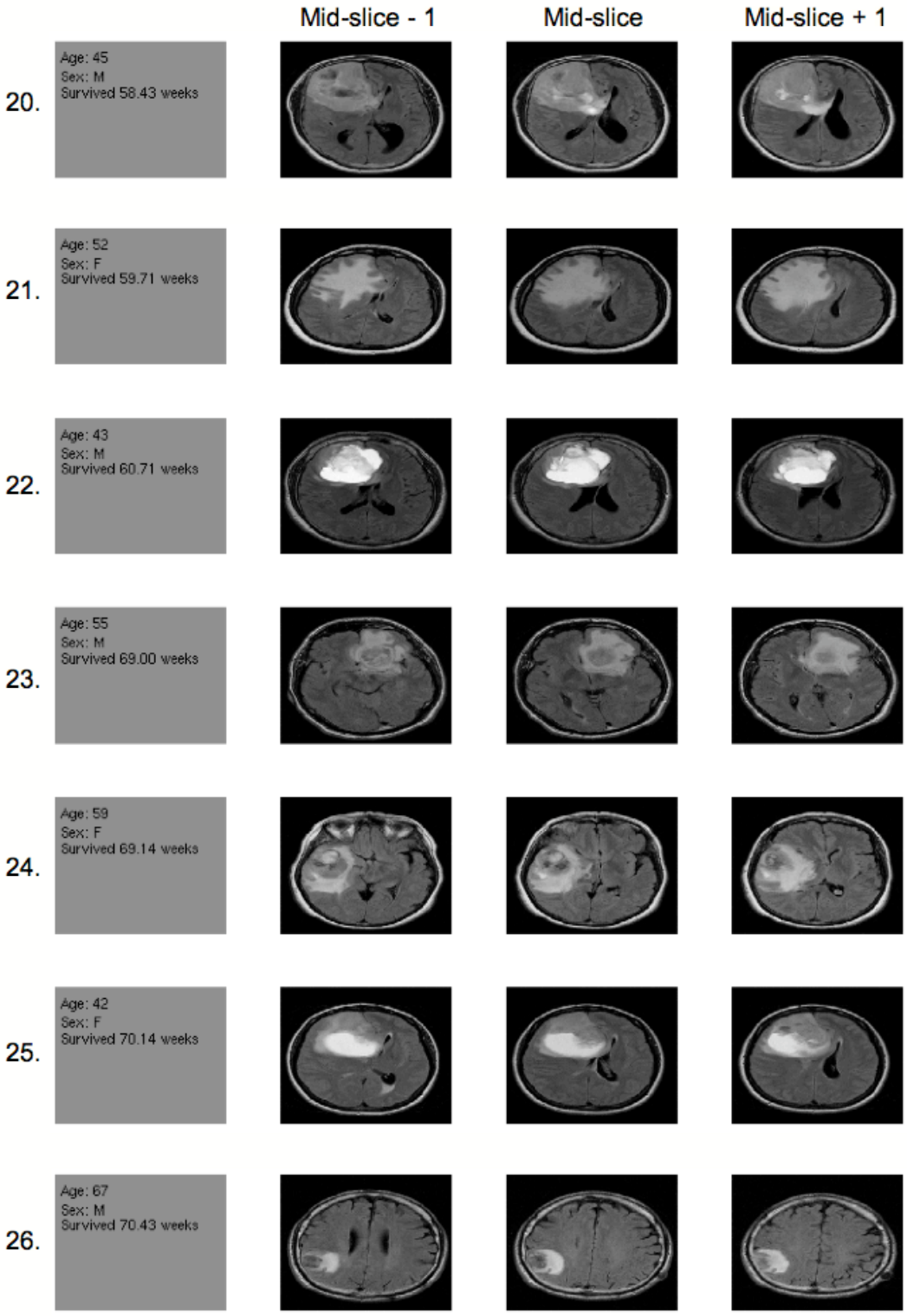
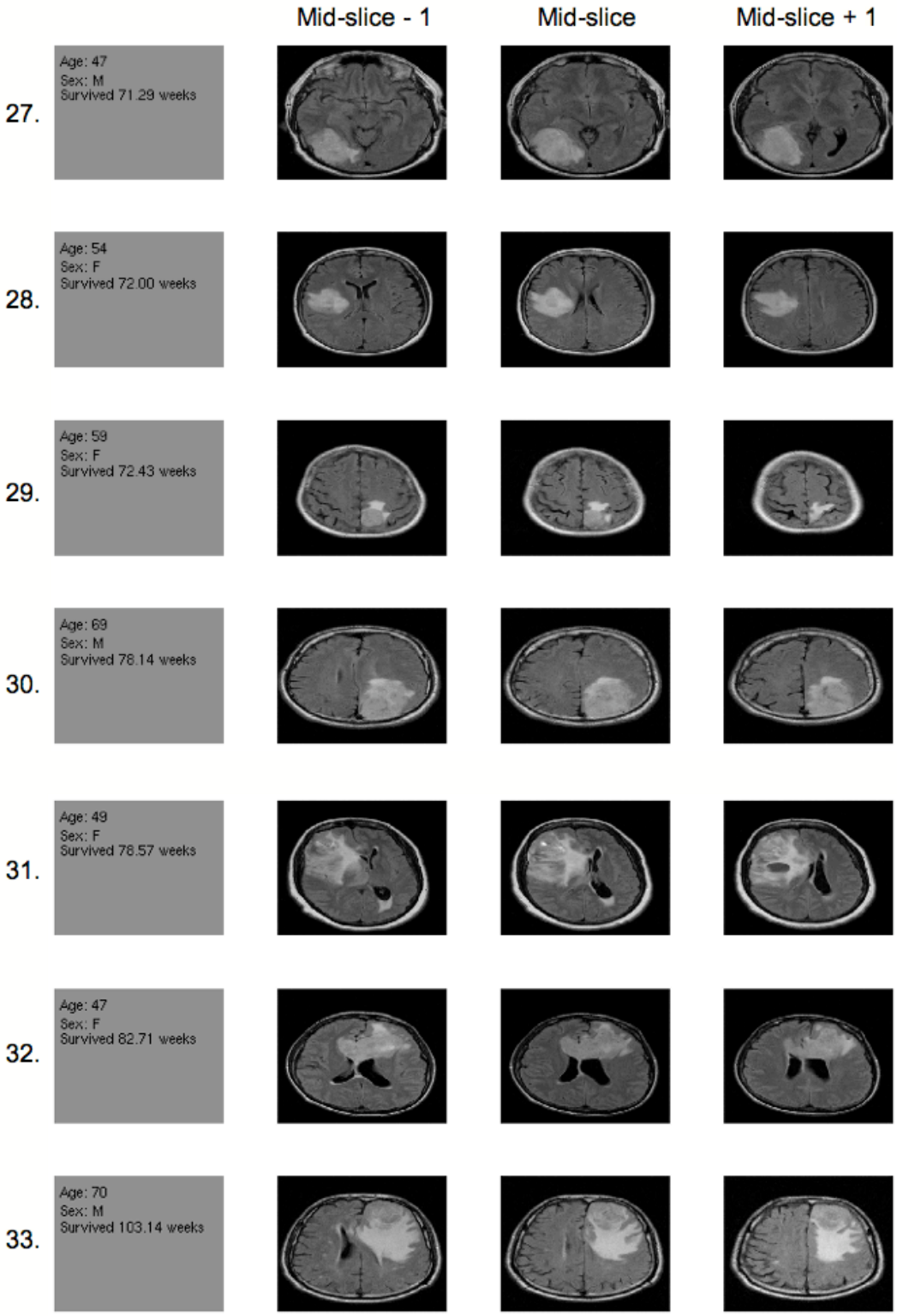


Table 5.8: (continued)



Chapter 6

Conclusion

6.1 Challenges and Future Work

Our original database contained volumes of different modalities: T1, T1-contrast, T2 and FLAIR. However, in choosing useful modalities for our work, we encountered two problems. First, in most post-surgical volumes, the tumor has been partially removed, completely or has been tampered with, introducing artifacts in the image. For our work, we wanted to use images with tumors completely intact. Therefore, we only looked for pre-surgical volumes. The second problem was that not all patients had all four modalities for their pre-surgical scans in our database. In fact, the total number of patients who had pre-surgical T1 , T1-contrast or T2 modalities was too small. On one hand, if we wanted to use a range of modalities, we would have to use a smaller number of patients. On the other hand, if we wanted to include more patients, we would have to use fewer modalities. Eventually, we decided to use more patients at the expense of using only one modality. In fact, most patients had pre-surgical FLAIRs and therefore, we decided to use only FLAIR volumes for our current study. For future studies, we can definitely obtain more patients to augment our current dataset, and also include more modalities. The inclusion of T1-contrast images can definitely improve survival prediction because, as discussed in Chapter 2, contrast enhancing is an important factor in prognosis of glioblastomas. Our database also lacked many clinical information such as KPS and extent of surgery about the patients. We suspect that including these parameters in survival prediction will improve prediction because these parameters have been found to be the most significant predictive factors in prognosis, as we learned in Chapter 2. A preliminary study by members in our group has shown that survival prediction can reach very high accuracies when clinical data and image-based information are combined.

For each texture analysis method in our framework, it is intuitive as to what they individually measure in an image. For example, the second-order statistic, contrast, as indicated by its name, measures the contrast in an image, and an edge filter induces strong response in areas of the image that resemble an edge. However, when these texture analysis methods are combined together with different statistical operations, such as entropy, computed over different brain regions in an elaborate

framework, the intuitive meaning can be easily lost. It is quite difficult to explain to a clinician what a decision tree that has *mr3-std-inner-vs-brain&NOtum* as a node can do. Therefore, instead of using the raw texture features, as we do in our framework, to build a decision tree, we can try to use the texture analysis methods to build clinically intuitive features explicitly, such as "sucli clarity", and then use these features to build the decision tree. For example, we can use second-order statistics with a Gaussian filter on the sulci region of the brain to build one feature, called "sulci clarity", and then use this feature to build our decision tree. Here, the challenge would be to fine-tune these intuitive features to correlate with the judgment of a clinician.

Also, as another future direction, to make our prediction framework more practical to clinical applications, we can extend our prediction framework to multi-class or even regression-based framework. As a preliminary study prior to this work, we did experiment with regression. However, the prediction results were very poor probably because our dataset was small. Any regression analysis on this type of data needs to deal with censored observations from the dataset. Patients whose exact dates of death are not known are problematic in regression analysis (this differs from classification tasks, where we do not need to remove a censored patient if this patient happens to fall into the long survival class). Therefore, regression analysis with survival data is much more challenging, which is precisely why methods like Recursive Partitioning Analysis and Kaplan-Meier estimators have been developed specifically for survival analysis.

6.2 Contributions and Concluding Remarks

Our goal was to use image-based textural information for prognosis of glioblastoma in a way that helps predict patient survival. Earlier studies in clinical trials and prognosis of glioblastoma have used tumor size as the only image-based feature. The framework in this thesis extracts textural features from pre-defined brain regions on FLAIR images and uses them to predict a patient's survival class. Based on 10-fold cross validation, our prediction framework can achieve an average accuracy of 80%. Therefore, we have shown that textural information about glioblastomas can be predictive of patient survival. Our framework also outputs a decision tree, which is an intuitive means of classification, and therefore, can be more easily applied by clinicians. We also showed, using Kaplan-Meier plots and the logrank test, that our decision tree survival prediction framework can produce significantly different survival groups, which can be useful in designing stratified clinical trials. However, our results show that our framework is not completely robust. Therefore, more work needs to be done in finding more powerful feature extraction methods that can yield a robust prediction system, which in turn can be reliably applied to prognosis by clinicians.

Bibliography

[1] Merriam-webster medical dictionary. http://www.merriam-webster.com/medical/prognosis. Accessed July, 2009.

[2] Svetlana Borovkova. Analysis of survival data. http://www.math.leidenuniv.nl/ naw/serie5 /deel03/dec2002/pdf/borovkova.pdf. Accessed April, 2009.

[3] P. Brodatz and A. Textures. *A photographic album for artists and designers*. 1966. Images downloaded in July 2009 from: http://www.ux.uis.no/ tranden/brodatz.html.

[4] Robert Brown, Magdalena Zlatescu, Angelique Sijben, Gloria Roldan, Jay Easaw, Peter Forsyth, Ian Parney, Robert Sevick, Elizabeth Yan, Douglas Demetrick, David Schiff, Gregory Cairncross, and Ross Mitchell. The use of magnetic resonance imaging to noninvasively detect genetic signatures in oligodendroglioma. *Clin Cancer Res*, 14(8):2357–2362, April 2008.

[5] Marc R.J. Carlson, Whitney B. Pope, Steve Horvath, Jerome G. Braunstein, Phioanh Nghiemphu, Cho-Lea Tso, Ingo Mellinghoff, Albert Lai, Linda M. Liau, Paul S. Mischel, Jun Dong, Stanley F. Nelson, and Timothy F. Cloughesy. Relationship between survival and edema in malignant gliomas: Role of vascular endothelial growth factor and neuronal pentraxin 2. *Clin Cancer Res*, 13(9):2592–2598, May 2007.

[6] G. Castellano, L. Bonilha, L.M. Li, and F. Cendes. Texture analysis of medical images. *Clinical Radiology*, 59(12):1061–1069, December 2004.

[7] Arnab Chakravarti, Gary Zhai, Yoshiyuki Suzuki, Sormeh Sarkesh, Peter M. Black, Alona Muzikansky, and Jay S. Loeffler. The prognostic significance of phosphatidylinositol 3-Kinase pathway activation in human gliomas. *J Clin Oncol*, 22(10):1926–1933, May 2004.

[8] Chien-Chang Chen and Chaur-Chin Chen. Filtering methods for texture discrimination. *Pattern Recognition Letters*, 20(8):783–790, August 1999.

[9] A Ciampi, S A Hogg, S McKinney, and J Thiffault. RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. i. methods and program features. *Computer Methods and Programs in Biomedicine*, 26(3):239–256, June 1988. PMID: 3383562.

[10] Joseph A. Cruz and David S. Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, 2, February 2006.

[11] Walter J. Curran, Charles B. Scott, John Horton, James S. Nelson, Alan S. Weinstein, A. Jennifer Fischbach, Chu H. Chang, Marvin Rotman, Sucha O. Asbell, Robert E. Krisch, and Diane F. Nelson. Recursive partitioning analysis of prognostic factors in three radiation therapy oncology group malignant glioma trials. *J. Natl. Cancer Inst.*, 85(9):704–710, May 1993.

[12] Beth Dawson and Robert Trapp. *Basic & Clinical Biostatistics: Fourth Edition*. McGraw-Hill Medical, 4 edition, March 2004.

[13] Alain Dupuy and Richard M. Simon. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl. Cancer Inst.*, 99(2):147–157, 2007.

[14] M.A.F. Figueiredo and A.K. Jain. Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396, 2002.

[15] D.A. Forsyth and J. Ponce. *Computer vision: A modern approach*. Prentice Hall Professional Technical Reference, 2002.

[16] William A. Freije, F. Edmundo Castro-Vargas, Zixing Fang, Steve Horvath, Timothy Cloughesy, Linda M. Liau, Paul S. Mischel, and Stanley F. Nelson. Gene expression profiling of gliomas strongly predicts survival. *Cancer Res*, 64(18):6503–6510, September 2004.

[17] John P. Geisler, Michael C. Wiemann, Zhen Zhou, Greg A. Miller, and Hans E. Geisler. Markov texture parameters as prognostic indicators in endometrial cancer. *Gynecologic Oncology*, 62(2):174–180, August 1996.

[18] Mohsen Ghazel, Anthony Traboulsee, and Rabab K. Ward. Optimal filter design for multiple sclerosis lesions segmentation from regions of interest in brain MRI. In *Signal Processing and Information Technology, 2006 IEEE International Symposium on*, pages 1–5, 2006.

[19] Mark A. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, University of Waikato, 1998.

[20] Maarouf A. Hammoud, Raymond Sawaya, Weiming Shi, Peter F. Thall, and Norman E. Leeds. Prognostic significance of preoperative MRI scans in glioblastoma multiforme. *Journal of Neuro-Oncology*, 27(1):65–73, 1996.

[21] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Trans. on Systems Man and Cybern*, SMC-3(6):610–621, 1973.

[22] Monika E. Hegi, Annie-Claire Diserens, Thierry Gorlia, Marie-France Hamou, Nicolas de Tribolet, Michael Weller, Johan M. Kros, Johannes A. Hainfellner, Warren Mason, Luigi Mariani, Jacoline E.C. Bromberg, Peter Hau, Rene O. Mirimanoff, J. Gregory Cairncross, Robert C. Janzer, and Roger Stupp. MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med*, 352(10):997–1003, March 2005.

[23] S Herlidou-Młme, J.M Constans, B Carsin, D Olivie, P.A Eliat, L Nadal-Desbarats, C Gondry, E Le Rumeur, I Idy-Peretti, and J.D de Certaines. MRI texture analysis on texture test objects, normal brain and intracranial tumors. *Magnetic resonance imaging*, 21(9):989–993, November 2003.

[24] Cross Cancer Institute. http://www.cancerboard.ab.ca/Treatment/CancerCareFacilities/CCI/, 2009.

[25] EL Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, june 1958.

[26] Kalle Karu, Anil K. Jain, and Ruud M. Bolle. Is there any texture in the image? *Pattern Recognition*, 29(9):1437–1446, September 1996.

[27] V A Kovalev, F Kruggel, H J Gertz, and D Y von Cramon. Three-dimensional texture analysis of MRI brain datasets. *IEEE Transactions on Medical Imaging*, 20(5):424–433, May 2001.

[28] Dietmar Krex, Barbara Klink, Christian Hartmann, Andreas von Deimling, Torsten Pietsch, Matthias Simon, Michael Sabel, Joachim P Steinbach, Oliver Heese, Guido Reifenberger, Michael Weller, and Gabriele Schackert. Long-term survival with glioblastoma multiforme. *Brain: A Journal of Neurology*, 130(Pt 10):2596–606, October 2007.

[29] Frederic Lachmann. Brain tissue classification from MRI data by means of texture analysis. In *Proceedings of SPIE*, pages 72–83, 1992.

[30] Yohan Lee, Adrienne Scheck, Timothy Cloughesy, Albert Lai, Jun Dong, Haumith Farooqi, Linda Liau, Steve Horvath, Paul Mischel, and Stanley Nelson. Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC Medical Genomics*, 1(1):52, 2008.

[31] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, June 2001.

[32] Seth Lloyd and Paul Penfield. 6.050j / 2.110j information and entropy. http://ocw.mit.edu/OcwWeb/Electrical-Engineering-and-Computer-Science/6-050JSpring-2008/CourseHome/, 2008. published by MIT OpenCourseWare: Massachusetts Institute of Technology, (Accessed February, 2009). License: Creative commons BY-NC-SA.

[33] David Louis, Hiroko Ohgaki, Otmar Wiestler, Webster Cavenee, Peter Burger, Anne Jouvet, Bernd Scheithauer, and Paul Kleihues. The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathologica*, 114(2):97–109, 2007.

[34] E A Maher, F B Furnari, R M Bachoo, D H Rowitch, D N Louis, W K Cavenee, and R A DePinho. Malignant glioma: genetics and biology of a grave matter. *Genes & Development*, 15(11):1311–33, June 2001. PMID: 11390353.

[35] Doaa Mahmoud-Ghoneim, Grégoire Toussaint, Jean-Marc Constans, and Jacques D. de Certaines. Three dimensional texture analysis in MRI: a preliminary evaluation in gliomas. *Magnetic resonance imaging*, 21(9):983–987, November 2003.

[36] Ramon Martinez, Gabriele Schackert, Ricard Yaya-Tur, Iigo Rojas-Marcos, James Herman, and Manel Esteller. Frequent hypermethylation of the DNA repair gene MGMT in long-term survivors of glioblastoma multiforme. *Journal of Neuro-Oncology*, 83(1):91–93, May 2007.

[37] A. Materka and M. Strzelecki. Texture analysis methods - a review. Technical report, Technical University of Lodz, Institute of Electronics, COST B11, Brussels, 1998.

[38] Rene-Olivier Mirimanoff, Thierry Gorlia, Warren Mason, Martin J. Van den Bent, Rolf-Dieter Kortmann, Barbara Fisher, Michele Reni, Alba A. Brandes, Juergen Curschmann, Salvador Villa, Gregory Cairncross, Anouk Allgeier, Denis Lacombe, and Roger Stupp. Radiotherapy and temozolomide for newly diagnosed glioblastoma: Recursive partitioning analysis of the EORTC 26981/22981-NCIC CE3 phase III randomized trial. *J Clin Oncol*, 24(16):2563–2569, June 2006.

[39] Paul S Mischel, Timothy F Cloughesy, and Stanley F Nelson. DNA-microarray analysis of brain cancer: molecular classification for therapy. *Nature Reviews. Neuroscience*, 5(10):782–92, October 2004. PMID: 15378038.

[40] Paul S Mischel, Stanley F Nelson, and Timothy F Cloughesy. Molecular analysis of glioblastoma: pathway profiling and its implications for patient therapy. *Cancer Biology & Therapy*, 2(3):242–7, 2003.

[41] Catherine L. Nutt, D. R. Mani, Rebecca A. Betensky, Pablo Tamayo, J. Gregory Cairncross, Christine Ladd, Ute Pohl, Christian Hartmann, Margaret E. McLaughlin, Tracy T. Batchelor, Peter M. Black, Andreas von Deimling, Scott L. Pomeroy, Todd R. Golub, and David N. Louis. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*, 63(7):1602–1607, April 2003.

[42] Kamalakannan Palanichamy, Michael Erkkinen, and Arnab Chakravarti. Predictive and prognostic markers in human glioblastomas. *Current Treatment Options in Oncology*, 7(6):490–504, November 2006.

[43] Uwe Pichlmeier, Andrea Bink, Gabriele Schackert, Walter Stummer, and the ALA Glioma Study Group. Resection and survival in glioblastoma multiforme: An RTOG recursive partitioning analysis of ALA study patients. *Neuro-oncol*, 10(6):1025–1034, 2008.

[44] A. Pierallini, M. Bonamini, P. Pantano, F. Palmeggiani, M. Raguso, M. F. Osti, G. Anaveri, and L. Bozzao. Radiological assessment of necrosis in glioblastoma: variability and prognostic value. *Neuroradiology*, 40(3):150–153, March 1998.

[45] John C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods: support vector learning*, pages 185–208. MIT Press, 1999.

[46] Whitney B. Pope, James Sayre, Alla Perlina, J. Pablo Villablanca, Paul S. Mischel, and Timothy F. Cloughesy. MR imaging correlates of survival in patients with High-Grade gliomas. *AJNR Am J Neuroradiol*, 26(10):2466–2474, November 2005.

[47] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., 1993.

[48] T. Randen and J.H. Husoy. Filtering for texture classification: a comparative study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(4):291–310, 1999.

[49] JH Rees, JG Smirniotopoulos, RV Jones, and K Wong. Glioblastoma multiforme: radiologic-pathologic correlation. *Radiographics*, 16(6):1413–1438, November 1996.

[50] Jeremy N. Rich, Christopher Hans, Beatrix Jones, Edwin S. Iversen, Roger E. McLendon, B.K. Ahmed Rasheed, Adrian Dobra, Holly K. Dressman, Darell D. Bigner, Joseph R. Nevins, and Mike West. Gene expression profiling and genetic markers in glioblastoma survival. *Cancer Res*, 65(10):4051–4058, May 2005.

[51] Kito S., Iritani A., Bavister B.D., and Busch C.[1]. Wavelet based texture segmentation of multi-modal tomographic images. *Computers and Graphics*, 21:347–358, May 1997.

[52] M. Sasikala, N. Kumaravel, and L. Subhashini. Automatic tumor segmentation using optimal texture features. In *Advances in Medical, Signal and Information Processing, 2006. MEDSIP 2006. IET 3rd International Conference On*, pages 1–4, 2006.

[53] M. Schmidt. Automatic brain tumor segmentation. Master's thesis, University of Alberta, 2005.

[54] C B Scott, C Scarantino, R Urtasun, B Movsas, C U Jones, J R Simpson, A J Fischbach, and W J Curran. Validation and predictive power of radiation therapy oncology group (RTOG) recursive partitioning analysis classes for malignant glioma patients: a report using RTOG 90-06. *International Journal of Radiation Oncology, Biology, Physics*, 40(1):51–5, 1998.

[55] E. G. Shaw, W. Seiferheld, C. Scott, C. Coughlin, S. Leibel, W. Curran, and M. Mehta. Reexamining the radiation therapy oncology group (RTOG) recursive partitioning analysis (RPA) for glioblastoma multiforme (GBM) patients. *International Journal of Radiation Oncology*Biology*Physics*, 57(2, Supplement 1):S135–S136, October 2003.

[56] N. Shikama, S. Sasaki, and A. Shinoda. Prognostic factors of patients with glioblastoma (Recursive partitioning analysis: RPA classes 5-6). *International Journal of Radiation Oncology*Biology*Physics*, 69(3, Supplement 1):S252, November 2007.

[57] James G. Smirniotopoulos, Frances M. Murphy, Elizabeth J. Rushing, John H. Rees, and Jason W. Schroeder. From the archives of the AFIP: patterns of contrast enhancement in the brain and meninges. *Radiographics*, 27(2):525–551, March 2007.

[58] Spotswood L. Spruance, Julia E. Reid, Michael Grace, and Matthew Samore. Hazard ratio in clinical trials. *Antimicrob. Agents Chemother.*, 48(8):2787–2792, aug 2004.

[59] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *Systems, Man and Cybernetics, IEEE Transactions on*, 8(6):460–473, 1978.

[60] Los Angeles Neuro-Oncology University of California. A comprehensive brain tumor program.

[61] M. Varma and A. Zesserman. Classifying images of materials: Achieving viewpoint and illumination independence. In *Lecture Notes in Computer Science*, volume 2352, pages 255–271, 2002.

[62] Patrick Y. Wen and Santosh Kesari. Malignant gliomas in adults. *N Engl J Med*, 359(5):492–507, July 2008.

[63] Barbara Weyn, Gert Van De Wouwer, Marek Koprowski, Andr Van Daele, Karl Dhaene, Paul Scheunders, Willem Jacob, and Eric Van Marck. Value of morphometry, texture analysis, densitometry, and histometry in the differential diagnosis and prognosis of malignant mesothelioma. *The Journal of Pathology*, 189(4):581–589, 1999.

[64] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Managemen. Morgan Kaufmann, second edition, June 2005.

[65] R Yamanaka, T Arao, N Yajima, N Tsuchiya, J Homma, R Tanaka, M Sano, A Oide, M Sekijima, and K Nishio. Identification of expressed genes characterizing long-term survival in malignant glioma patients. *Oncogene*, 25(44):5994–6002, May 2006.

[66] K Yogesan, T Jrgensen, F Albregtsen, K J Tveter, and H E Danielsen. Entropy-based texture analysis of chromatin structure in advanced prostate cancer. *Cytometry*, 24(3):268–276, July 1996.

[67] Jianguo Zhang and Tieniu Tan. Brief review of invariant texture analysis methods. *Pattern Recognition*, 35(3):735–747, March 2002.