NO-REFERENCE IMAGE QUALITY ASSESSMENT OF LARGE FLAIR MRI DATASETS

by

Joshua P. Seguin

B.Eng, University of Guelph, 2017

A dissertation

presented to Ryerson University

in partial fulfillment of

the requirements for the degree of

Master of Applied Science

in the program of

Biomedical Engineering

Toronto, Ontario, Canada, 2020

© Joshua P. Seguin, 2020

Author's Declaration For Electronic Submission Of A Thesis

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

NO-REFERENCE IMAGE QUALITY ASSESSMENT OF LARGE FLAIR MRI DATASETS

Joshua P. Seguin

Master of Applied Science

Biomedical Engineering

Ryerson University, 2020

Abstract

The study of neurodegenerative diseases have found promise through white matter lesions best visualized in FLAIR MRI; however, algorithms experience difficulty in generalizing to large multicenter datasets due to the variance of image quality and characteristics. This thesis presents a quality control tool that combines image quality assessment with outlier rejection algorithms; this tool is unique as it is specifically designed for large multicenter FLAIR MRI datasets. An image processing approach evaluates each volume by: intensity-based features, sharpness/blur-based features, signal- and contrast-to-noise ratios, noise field characteristics, motion artifact prevalence and a total IQ score. The performance of this tool was evaluated on labelled ADNI and CCNA data reporting F1 scores of 0.82, and 0.85, respectively. Applications for this tool include potential rescan or longitudinal scanner study alongside the immediate application of outlier removal for large FLAIR datasets.

Acknowledgments

I would like to start by thanking Dr. April Khademi for the time, effort, and expertise that made this thesis a possibility; your guidance in research and life lessons learned along the way, are greatly appreciated.

I also owe thanks to ...

The Image Analysis in Medicine Lab team, for the support inside- but mostly outside- of the lab;

Sergiu, for being by my side through this degree since day one;

and

My parents, Adam, and Alli, for all your love and support.

Table of Contents

Abstract	iii
List of Tables	vii
List of Tables in Appendix	viii
List of Figures	xi
List of Figures in Appendix	xi
List of Abbreviations	1
1 Introduction	1
1.1 Background	1
1.1.1 Neurodegenerative Disease	1
1.1.2 Brain Anatomy	3
1.1.3 Magnetic Resonance Imaging	5
1.1.4 Big Data and the Multicenter Effect	11
1.2 Challenges and Objectives	13
1.3 Previous Works	15
1.4 Proposed Outline	18
1.5 Contributions	22
2 Quality Control Tool: Fields of Focus	24
2.1 Intensity Non-Standardness	24
2.2 Image Sharpness/Blur	28

2.3	Signal-to-Noise Ratio and Contrast-to-Noise Ratio	30
2.4	Noise Field	32
2.5	Motion artifact	34
2.6	Total Image Quality Score	36
2.7	Summary of Approach	36
3	Image Quality Assessment and Quality Control Pipeline: Methodology	39
3.1	Preprocessing	40
3.1.1	Intensity Standardization	40
3.1.2	Mask Generation	40
3.2	No-Reference Image Quality Metrics for FLAIR MRI	43
3.2.1	Intensity-based Features	45
3.2.2	Sharpness/Blur-based Features	47
3.2.3	Signal-to-Noise Ratio and Contrast-to-Noise Ratio	55
3.2.4	Noise Field Features	57
3.2.5	Motion Artifact Feature	59
3.2.6	Total Image Quality Score	61
3.3	Outlier Rejection Overview	63
3.4	Validation Data and Metrics	69
4	Results	75
4.1	Visual Validation	75
4.1.1	Intensity-based Features	77
4.1.2	Sharpness-based Features	79
4.1.3	Noise-based Features	82
4.1.4	SNR and CNR	84
4.1.5	Motion Artifact Feature	85
4.1.6	Total Image Quality Score	87
4.2	Quantitative Validation	89
4.2.1	Quantitative Validation by IQA Category	98
4.3	False Positive Cases	112

5	Conclusions	115
5.1	Summary	115
5.2	Applications	117
5.3	Limitations	118
5.4	Future Work	120
	Appendices	122
	Bibliography	130

List of Tables

3.1	Summary of image quality features	64
3.2	Example format of IQA summary spreadsheet	67
3.3	Example format of outlier rejection spreadsheet	67
3.4	Summary of datasets and metadata	70
3.5	Summary of datasets and metadata	72
3.6	Example confusion matrix	73
4.1	Quality Control Tool's performance results: ADNI	94
4.2	Quality Control Tool's performance results: CCNA	97
4.3	Performance updates after ROC analysis on each feature	111
4.4	Quality Control Tool's performance results before and after individual feature optimization	111

List of Tables in Appendix

1	Beta parameter optimization for total IQ score: ADNI subset	130
2	Beta parameter optimization for motion artifact feature: ADNI subset	131

List of Figures

1.1	Brain anatomy reference displaying tissue classes	4
1.2	Cross-sectional slice of FLAIR MR volume displaying white matter lesions	5
1.3	Example of an MRI volume composed of slices/images	8
1.4	Middle slice of FLAIR MR volume displaying grayscale intensities	9
1.5	Comparison of tissue appearance in T1-weighted, T2-weighted and FLAIR MR im- ages	10
1.6	Example of multicenter effect - ADNI data summary by manufacturer	11
1.7	Example of multicenter effect - ADNI MR images	12
2.1	Example of intensity non-standardness: GE, Siemens, Philips	25
2.2	Philips intensity non-standardness example: image and histogram	26
2.3	Siemens intensity non-standardness example: image and histogram	27
2.4	Example of intensity non-standardness across a whole dataset	27
2.5	Example of edge content representing the sharpness of an image	28
2.6	Edge content comparison between two volumes from the same dataset	30
2.7	Boxplot of CNR by manufacturer	31
2.8	Noise field variability example of two volumes from the same dataset	33

2.9 Bias-field artifact example	34
2.10 Common motion artifact example: Gibb's Phenomenon	35
3.1 Quality control pipeline flowchart	39
3.2 Example of head-mask	41
3.3 Example of brain-mask	43
3.4 Example of CSF-mask	43
3.5 PDF and CDF of volume histogram	46
3.6 Image sharpness calculation example: Sobel edge filter applied	49
3.7 Multi-scale approach using Gaussian blur of varying standard deviation	50
3.8 Spatial energy feature intermediate step: generation of spatial weighted map	53
3.9 Isolation of brain-class and CSF-class via binary masks	55
3.10 Cross-sectional checkpoints of brain content in z-direction of volume	56
3.11 Kurtosis explained through an example	57
3.12 Skewness explained through an example	58
3.13 Flowchart of motion artifact evaluation pipeline	59
3.14 Edge content of image background: no motion artifact example	60
3.15 Edge content of image background: severe motion artifact example	60
3.16 Flowchart of outlier rejection algorithms	64
3.17 Boxplot example of outlier rejection; from [1].	66
3.18 Outlier rejection: accounting for the MCE	68
4.1 Example of a good-quality volume	76
4.2 Outlier case-examples detected by intensity-based features	77
4.3 Outlier case-examples detected by intensity-based features: ADNI	78
4.4 Outlier case-examples detected by sharpness/blur-based features	79
4.5 Outlier case-examples detected by sharpness-based features: ADNI	81
4.6 Outlier case example detected by noise field: bias-field	82
4.7 Outlier case-examples detected by noise-based features: ADNI	83
4.8 Outlier case-examples detected by CNR	84
4.9 Outlier case-examples detected by motion artifact	85

4.10 Outlier case-examples detected by motion artifact feature: ADNI	86
4.11 Outlier case-examples detected by total image quality score: ADNI	88
4.12 Intensity-based feature histograms with cut-off thresholds for ADNI data by manu- facturer	90
4.13 Blur-based feature histograms with cut-off thresholds for ADNI data by manufacturer	91
4.14 Noise-based feature histograms with cut-off thresholds for ADNI data by manufacturer	92
4.15 SNR and CNR feature histograms with cut-off thresholds for ADNI data by manufac- turer	92
4.16 Motion Artifact and total IQ score histograms with cut-off thresholds for ADNI data by manufacturer	93
4.17 ROC Analysis for motion artifact feature cutoff threshold optimization: ADNI	95
4.18 ROC Analysis for total image quality score cutoff threshold optimization: ADNI . . .	96
4.19 ROC Analysis CCNA	97
4.20 Intensity scaled examples of data for validation of intensity-based features	99
4.21 Intensity-based features correlate with intensity scaling validation	100
4.22 Intensity-based metrics validation: features extracted from original and standardized data	100
4.23 Blurred examples of data for validation of sharpness-based features	101
4.24 Multi-scale image sharpness feature validation with varying levels of artificial blur .	102
4.25 Sharpness-based features correlate with varying levels of artificial blur	102
4.26 Additive noise examples of data for validation of SNR and CNR features	103
4.27 SNR and CNR features correlate with varying levels of Gaussian noise	104
4.28 Additive noise examples of data for validation of noise-based features features	105
4.29 Noise-based features accurately measure varying levels of Gaussian noise	105
4.30 Varying intensity of ringing artifact examples of data for validation of motion artifact feature	106
4.31 Different edge detection applications in motion artifact feature validation by corre- lation with severity of artifact	107
4.32 Total IQ score feature study to evaluating contribution to quality discrimination . . .	108
4.33 Total IQ scores on good- and poor-quality volumes	109

4.34 ROC analysis for each feature	110
4.35 Example cases when motion artifact algorithm fails	112
4.36 Example cases when brain extraction algorithm fails	113
4.37 Example cases when outlier detection can fail	114

List of Figures in Appendix

1 Feature histograms with cut-off thresholds for CCNA data	124
2 Outlier case-examples detected by intensity-based features: CCNA	125
3 Outlier case-examples detected by noise-based features: CCNA	127
4 Outlier case-examples detected by motion artifact feature: CCNA	129

List of Abbreviations

AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
ALS	Amyotrophic Lateral Sclerosis
CCNA	Canadian Consortium on Neurodegeneration in Aging
CDF	cumulative distribution function
CDIP	Canadian Dementia Imaging Protocol
CNR	contrast to noise ratio
CSF	cerebrospinal fluid
FLAIR	fluid attenuation inversion recovery
GE	General Electric
GM	grey matter
IQA	image quality assessment
ISms	image sharpness multi-scale
JPEG	joint photographic experts group
MCE	multicenter effect
MCI	mild cognitive impairment
MS	Multiple Sclerosis
MR	magnetic resonance
MRI	magnetic resonance imaging
NIA	National Institute on Ageing
NPV	negative predictive value
PDF	probability density function
PPV	positive predictive value
QC	quality control
SNR	signal to noise ratio
WM	white matter
WML	white matter lesion

Chapter 1

Introduction

The focus of this thesis is to automatically assess image quality of brain scans using quality control measures in fluid attenuated inversion recovery (FLAIR) magnetic resonance imaging (MRI). The objectives are to consistently discriminate between good- and poor-quality volumes in a dataset, and provide a supplemental image quality assessment (IQA) summary. This section outlines the necessary background and motivation to study the field of quality control and MRI. This section also features a description of the objectives of this project, previous works developed in this field, and brief explanation the proposed pipeline.

1.1 Background

The background for this thesis addresses the motivation to study neurodegenerative disease, provides an overview of applicable brain anatomy, shares the necessary context of the field of MRI, and explains the challenges associated with studying medical images in large patient cohorts.

1.1.1 Neurodegenerative Disease

There are many neurological disorders that have a significant impact on the well-being of patients and carry a heavy economic burden on healthcare systems. Many of these disorders have no cures at the present time, yet they are the reason for 9.3 percent of hospitalizations in Canada [2]. The economic burden for all neurological disorders in the United States of America alone, is estimated

at \$277 billion for 2018 [3]. On a global level, a lack of efforts to research solutions could cause the estimated population of 35 million living with neurodegenerative disease to rise to 135 million by 2050 [4]. The repercussions of these diseases are growing at an alarming rate; reports show that dementia actually has a greater economic impact than cancer and cardiovascular disease [5] [6]. This projection is supported by a healthcare strategy, focused on prevention and early intervention, that claims there is opportunity to save over \$7.9 trillion dollars in medical and care costs [3]. Ultimately, the research community needs to study the mechanisms, causes, and progression of neurological disorders so that treatment and intervention can be applied early; subsequently, this will decrease mortality rates, reduce long-term disabilities, and alleviate economic burden.

To provide context, neurodegenerative disease describes a range of conditions that primarily affect the neurons in the human brain. As these neurons do not regenerate, damage to them can be detrimental. Some common examples of neurodegenerative disease include: Alzheimer's Disease (AD), Parkinson's Disease, Multiple Sclerosis (MS), Amyotrophic Lateral Sclerosis (ALS) and ischemic disease [7]. These diseases can affect the body's ability to function properly and in many cases cause systems to fail. Patients typically exhibit slow degeneration of body function through: loss of balance, loss of memory, loss of ability to speak, difficulty breathing, and heart failure (depending on which disease they are affected by).

Scientists recognize the combination of genetic and environmental factors that contribute to the risk of developing neurodegenerative disease [8]. The most highly correlated demographic with these diseases are the elderly; some diseases report to double in frequency per five years over the age of 65 [5]. As the baby-boom generation ages, the healthcare industry hopes to develop a solution as soon as possible.

Global efforts have been recognized to fund research with an objective to identify mild forms of cognitive impairment and pre-clinical dementia such that prevention and treatment can be implemented early on; these goals are set to be achieved by 2025 [9] [10]. The scientific community emphasizes the importance of imaging and analytics in personalized medicine as a promising route to discover solutions to neurodegenerative disease [11] [12]. One specific example is the National Institute on Aging (NIA); they participate in a public-private funding initiative to further study

neurodegenerative disease through medical images of the brain in the form of magnetic resonance imaging (MRI) [13]. This trend of studying MRI yields promise as many researchers believe that precursors found in the brain correlate with disease and that investigating MRI will lead to deeper understanding of neurodegenerative disease progression and etiology.

A very interesting discovery relating imaging of the brain to neurological disorders is a disease marker identified on Fluid-Attenuation Inversion Recovery (FLAIR) MRI. These markers, referred to as white matter lesions (WML), have been found to be expressions of vessel disease and are associated with ischemic disease [14], Alzheimer's Disease [15], and Multiple Sclerosis [16]. There have been some preliminary research studies conducted on FLAIR MRI data to identify the etiology, causes, and progression of disease, but further studies are required [17]. The following section provides the details, from an anatomical and physiological perspective, of commonly used terms surrounding neurodegenerative disease and MRI brain scans.

1.1.2 Brain Anatomy

The brain is the central organ of the human nervous system which serves as the scaffolding for which the complex mind operates. The brain is protected by the skull and is suspended in cerebrospinal fluid (CSF), providing mechanical and immunological defense so it can control all functions of the body. It is important for the brain to remain protected and properly functioning, to avoid altering cognition and body function all-together.

Over time the structure and function of the brain is susceptible to change through the body's process of ageing. Natural ageing of the brain causes a decrease in brain volume, vasculature, and cognition, while increasing vulnerability to neurodegenerative disease, stroke, and memory impairment [18] [19]. In the case of neurodegenerative disease, the brain undergoes progressive degeneration of neurons which can cause atrophy of the affected region [7]. These areas of demyelinated cells found in the white matter of the brain, present hyper-intense regions referred to as WML.

For the scope of this thesis it is important to understand the underlying structures and tissue classes

associated with neurodegenerative disease that are visible through MR images. A visualization is provided in figure 1.1, through a cross-sectional image of a MR brain scan and isolation of the white matter, grey matter, and CSF.

Figure 1.1: Brain anatomy reference displaying tissue classes; from [20].

Figure 1.1 displays a cross-sectional image of the brain in (a), which is then segregated into binary masks of CSF in (b), grey matter in (c) and white matter in (d). These binary masks are annotated images that have labelled pixels that describe the location of tissue classes; for this case these masks allow for the understanding of what each tissue class looks like. The fundamental difference between the grey- and white-matter tissue of the brain is that the grey matter contains cell bodies, axon terminals, and dendrites while the white matter mainly contains myelinated axons [21]. In some MRI images, these subtle neuronal structure differences can appear as different classes. It should be noted that while this image portrays the white matter and grey matter of the brain separately, the FLAIR MR images that are used to study WML do not; the details and explanation can be found in Section 1.1.3. An image to explain how WML appear in a brain scan is presented in figure 1.2; these images are acquired using FLAIR MRI.

Figure 1.2 displays a red-highlighted mask that is overlaid on top of the original image. This image is a severe case where the WML are abundant in the area surrounding the dark CSF filled

Figure 1.2: Cross-sectional slice of FLAIR MR volume displaying white matter lesions; from [22].

cavities referred to as ventricles. These two lateral ventricles are accompanied by two smaller ventricles, and as a system, they are responsible for the regulation and production of CSF. As mentioned earlier, the CSF plays an important role in the support and protection of the brain. Another noteworthy structure that provides protection of the brain is the white border surrounding the brain in both images of figure 1.2, which is the skull.

The following section provides background on MRI in general and why FLAIR MRI is of specific interest in studying WML.

1.1.3 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is the most-widely used modality for studying the anatomy of the brain. In comparison to other imaging modalities, MRI emits no radiation because it uses a strong magnetic field and radiofrequency (RF) pulses to study the composition of the subject. This imaging technique is superior in resolution and in terms of capturing tissue characteristics because it measures the combination of proton density, and T1 and T2 relaxation constants; these factors contribute to the contrast of the image [23]. This section provides a brief explanation of the physics of MRI and how the aforementioned terms can affect the reconstructed image.

MRI focuses on capturing the energy released primarily from lipids and water that compose the tissue being studied, which for this application is the brain. The primary element of study is

hydrogen ions (protons) because they give strong signal intensity but also because their quantity varies depending on the tissue or molecules they constitute [24]. The fundamental idea is to excite these protons using a combination of magnetic fields and detect the energy released (image signal) from the same protons such that the tissues characteristics can be understood; then an image can be constructed to show the contrast between different structures of the brain. Naturally, areas or tissues that are more proton-dense produce a different signal than that of an area with less protons [25]. A specific example is the difference between lipid molecules and water molecules; fatty brain tissue is more proton dense so it will produce a different signal than that of cerebrospinal fluid (CSF) which is mostly water and contains much smaller proton structures.

To provide more detail, the MR coil, that surrounds the subject's head, produces a powerful magnetic field (M_O) that creates an alignment of protons, which spin (precess) in the direction of the field [23]. A secondary magnetic field (M_{XY}), perpendicular to the first, is temporarily applied via an RF pulse, such that the protons rotate to a perpendicular orientation and then they decay back to their original alignment [23]. There are two fundamental constants to describe the rate at which the RF pulse alignment dissipates and the original magnetic field alignment is regained; these are referred to as the T1- and T2-relaxation constants. The dissipation of magnetic field strength is described as follows:

$$M_z = M_o(1 - e^{\frac{-t}{T1}}), \quad (1.1)$$

where, M_z represents the magnetic field strength with respect to the original magnetic field (M_o) produced by the coils at time, t , in *ms*. Similarly, this can be done with respect to T2 as follows:

$$M_{xy} = M_o(e^{\frac{-t}{T2}}). \quad (1.2)$$

Here, M_{xy} represents the magnetic field produced by the RF pulse, which is perpendicular to M_z . It should be understood that the rate at which M_o is regained is not equal to the decay of the transverse alignment, M_{xy} . T1 represents the rate at which the original field, M_o , is regained,

which varies depending on the proton interaction with nearby nuclei, atoms and molecules (spin-lattice interactions). T2 represents the rate at which the transverse magnetization, M_{xy} , decreases, which occurs through the dephasing of these protons from their initial RF pulse excitement. This dephasing can occur via precessing protons interacting with other protons or charged particles that change their rotational frequency. Any T1 interactions with the environment also contribute to dephasing of the protons from their RF pulse, which is why T2 is always shorter than T1 and why more complex structures have shorter constants [25].

For reference, some common values to describe the T1 constant for gray matter of the brain, CSF and WML have been reported at approximately 1170 *ms*, 3340 *ms* and 1120 *ms*, respectively [26] [27]. Common T2 constants for gray matter, CSF and WML have been reported at approximately 90 *ms*, 2560 *ms* and 140 *ms*, respectively [26] [27]. While these are estimated values, they are still helpful in conceptualizing the relaxation rates of different tissues.

These relaxation constants are used in designing specific RF pulse sequences that are repeated throughout the scan to create cross-sectional images of the brain, until an entire brain volume has been scanned¹. The final output from an MRI scan is a series of two-dimensional cross-sectional images (slices) that combine to display a volume. Figure 1.3 displays an example of a brain-scan volume through sampled cross sectional images of the head. This particular figure displays a sample image every four slices throughout the scan.

While figure 1.3 provides context on an image or slice relative to a volume for a single MRI brain scan, the fundamentals of each image are further explained through figure 1.4. This single image is made up of pixels (“picture elements”) that contain grayscale intensities. The following equation 1.3 explains mathematically how a standard grayscale image is represented,

$$I(m, n) = i, \tag{1.3}$$

where variables of m and n represent the x- and y-pixel location within the image, I . For example,

¹This explanation of MRI is tailored to include the necessary information for this thesis. More information can be found at <http://mri-q.com>

Figure 1.3: Example of an MRI volume composed of slices/images

a $M \times N$ image of size 256×256 contains 65,356 pixels, each with a specific intensity value that corresponds to a grayscale level, i . These grayscale levels, for this 10-bit image example, range from 0 to 1023. Observing figure 1.4 it can be seen that the lower intensities represent black or darker gray, while high values represent white or lighter gray; this contrast allows for high

resolution representation of tissue information.

Figure 1.4: Middle slice of FLAIR MR volume displaying grayscale intensities

Figure 1.4 displays a colorbar next to the FLAIR MR image that relates each pixel's intensity value to a grayscale colormap. Each intensity within the image represents the information from the subject's head/brain and the how the corresponding protons, on a molecular-tissue level, react to specific RF pulse sequences. Moreover, these pulse sequences can be manipulated to provide contrast for specific tissues that can be used for varying applications, like T1-weighted images and T2-weighted images. An example of how this can be useful is explained by the dephasing that is associated with the T2-relaxation constant. T2-weighted images produce a fairly restricted range of brain tissue intensities; for this reason, T2-weighted images are particularly useful in detecting degraded or abnormal brain tissue as it would report a higher constant than expected. One setback associated with the T2-weighted images is that the abnormal brain tissue or previously mentioned WML appears bright just as the CSF does. This similarity causes difficulty in differentiation between the two.

Fortunately, an inversion recovery sequence was developed to apply an additional RF pulse with the intention of nulling the CSF; this type of MRI sequence is called fluid attenuation inversion recovery (FLAIR). This sequence allows for scans to produce images that can further the study of WML. Examples of a T1-weighted, T2-weighted and T2-FLAIR images can be found in figure 1.5.

Figure 1.5: Comparison of tissue appearance in T1-weighted, T2-weighted and FLAIR MR images

Figure 1.5 highlights the fundamental differences in terms of appearance and contrast for each of the discussed imaging techniques. The T1-weighted images have the brain class appearing brightest, with substantial difference between gray matter and white matter. The T2-weighted images display the brain class much darker creating opportunity for the study of WML or degraded brain structure but the CSF class appears brighter, which introduces issues for studying WML. The FLAIR image displays the darker brain matter with nulled CSF, creating an ideal environment for studying degraded brain structures and WML; an example of WML at higher intensities can be found in figure 1.4.

This thesis focuses on the use of good-quality FLAIR MR images such that WML can be further studied against neurodegenerative disease. To better understand the relationship between neurodegenerative diseases and markers (WML), images need to be analyzed and correlated with clinical data to model disease progression, causes, and to identify new risk factors [28]. It is important that this relationship of disease and markers be studied in large patient cohorts to accurately draw conclusions with statistical significance and power, and to get an accurate representation of the population; consequently, this will provide greater clinical utility that can translate to more informed decisions for personalized and predictive medicine [11] [12].

1.1.4 Big Data and the Multicenter Effect

The purpose of emphasizing the requirement for analysis of big data is to bring attention to the challenges associated with studying these markers found in images at a large scale. A current method of manual analysis of images and the associated clinical data is subjective and laborious [29]. To expect a radiologist or an expert in the field to analyze millions of images manually is far too intensive and expensive. A natural solution would be to automate the process using algorithms designed to quantify what a radiologist would otherwise have to visually identify. Many algorithms have been designed to automate segmentations, lesion-load measures and other applications to relate markers to clinical data; however, these algorithms have only been proven successful on synthetic images or very small databases [30] [31] [32]. The reason that these algorithms have not succeeded in generalizing to multi-institutional datasets is because of the variability across manufacturers referred to as the multicenter effect (MCE).

The MCE can be defined as the variability in image properties as a function of vendor type, acquisition parameters, software and hardware differences between institutions; all of which are common for multi-institutional studies. Figure 1.6 shows the MCE through an example of the contrast-to-noise ratio (CNR), the average intensity value and a measure of image sharpness. These metrics are explained throughout the methodology section in detail; however, for the purpose of this example it can be observed that these metrics have great variation, despite being extracted from the same dataset (Alzheimer’s Disease Neuroimaging Initiative) [33].

Figure 1.6: Example of multicenter effect - ADNI data summary by manufacturer

As the acquisition setup determines the noise, intensity and contrast of the images, there is variability

ity in the properties and characteristics of the reconstructed images. Figure 1.6a displays a lower CNR for this particular dataset for the volumes produced by the General Electric (GE) scanners and figure 1.6c shows a lower sharpness measure for GE as well. These lower values of contrast and sharpness could potentially be problematic for a segmentation algorithm that prefers clean borders between classes.

A visual example of the MCE can be observed through figure 1.7. The intensity ranges, when compared on the same gray scale, display how significantly different an image can be. Figure 1.6b displays this phenomenon by summarizing the mean intensity of every volume within the dataset and plotting it by manufacturer, while figure 1.7 shows an example from each manufacturer to further prove the point.

Figure 1.7: Example of multicenter effect - ADNI MR images

Figure 1.7a displays an image from a volume produced by a GE scanner which yields high intensity values for the skull, brain and WML. Figure 1.7b shows an image produced from a Siemens scanner that has a much lower range of intensities that becomes apparent when compared on the same gray level scale. The skull and brain appear much darker than that of the GE scan. Figure 1.7c displays an image that was produced by a Philips scanner and it too has a much lower intensity range than that of GE, but is different from the Siemens scan as the skull appears much brighter. This type of variation is consistent throughout large multicenter datasets, which is why it is important to quantify and understand this difference, known as the MCE prior to implementing algorithms.

To provide context, the most common methods used to glean information in this field include segmentation and classification algorithms; these specific types of image processing algorithms are quantitative. This implies that small changes in intensity values have a negative impact on the

reliability of the results. By observing the disparity in these metrics shown across the above figures it can be understood that designing algorithms for multicenter data is more challenging [34].

The current algorithms that have been developed using synthetic images or a single database will either require an alternative design or require intensive preprocessing in order to generalize to multicenter data [31] [32]. Since computational power and propagation of errors prevents a one-size-fits-all approach for preprocessing, the characteristics of each dataset will need to be understood to allow efficient selection of preprocessing techniques or intelligent design of algorithms [34].

A solution to overcoming the MCE can be accomplished by first understanding the characteristics of any large multicenter dataset by consistently evaluating the dataset, quantitatively. Theoretically, this can be accomplished from an image quality assessment (IQA) perspective that evaluates multiple aspects of the images within the dataset. This thesis presents a quality control tool that uses an all-encompassing image quality assessment that intends to do just-that; it considers the following categories: contrast- and signal- to-noise ratios, intensity non-standardness, image sharpness or blur, noise field, prevalence of artifact and total IQ score. Each of these categories are further explained in Chapter 2. The following section clearly outlines the objectives of this project and the associated challenges.

1.2 Challenges and Objectives

The first objective of this project is to design a quality control tool that can provide a useful IQA summary of large multicenter FLAIR MRI datasets. This IQA summary will include information about the nature of the dataset, such that intelligent and efficient preprocessing could be done prior to the use of said dataset. This IQA summary could also be used for insights into the composition of a dataset prior to algorithm design. The details of a proposed outline are found in section 1.4.

The second objective of this project is to incorporate an outlier rejection algorithm into the IQA

pipeline. This quality control tool will reject poor quality data or outliers from large multicenter FLAIR MRI datasets.

These objectives are unique because the pipeline is designed specifically for FLAIR MRI and tested on FLAIR MRI. As the following section will discuss, it is also a novel approach to include an IQA summary with the outlier rejection. This pipeline should leave no uncertainties or questions remaining about the data in general or reasons for rejecting data.

The challenges associated with accomplishing these objectives can be summarized into three categories: choosing appropriate features, accurately validating the pipeline and achieving accuracy in outlier detection.

The most difficult challenge is to design the appropriate feature set that can define any FLAIR MR image's quality, quantitatively. The features selected should provide an all-encompassing IQA summary of the data, implying that multiple aspects of image quality should be considered. These features should also be discriminative of good- and poor-quality images. The selection of features, sets a base for the success of the classification of pass or fail.

While the IQA summary is very important, ultimately if the pipeline is hypothetically unable to quantify what it intends to describe in terms of image quality then it cannot be deemed a reliable quality control tool. The field of quality control of medical images, is constantly facing the issue of accurately labelling image quality based on a qualitative assessment. It is difficult for a human to visually analyze an image and quantify the quality accurately and consistently. Mortamet *et al.* describes the common challenge of defects or poor image quality that is not visually obvious to a rater [35]. Issues of inter-rater variability through subjective visual analysis is potentially another challenge to overcome. The idea of including a professional's opinion introduces an issue of cost for a radiologist to provide labels. If only some of these aforementioned issues with accurately labelling image quality do come up, the issue of a large enough labelled validation set is then a major challenge to overcome.

The final major challenge is to achieve accuracy in outlier detection; one common issue for similar projects is access to a large, labelled dataset for validation, as it is difficult to find a dataset that

contains data of good- and poor-quality that is also graded. Another challenge, regarding the success of the pipeline, is the ability to account for the MCE, as the pipeline should be designed accordingly. Additionally, special attention should be paid to the computational efficiency of the algorithms; this is an important challenge, especially if this tool has plans for commercialization.

1.3 Previous Works

The understanding for quality control for MR has been developing recently, given the need for automated methods to assess medical images on a large scale. Unfortunately, most of the efforts are generic for all MR or are specifically designed for T1-weighted or T2-weighted images; FLAIR MRI is fundamentally different from these types of imaging and is gaining more attention from researchers and clinicians to further the study of neurodegenerative diseases [36] [37] [38]. While there has been some success in correlating features with quality for FLAIR MRI data, these methods have not shown promise on volumetric applications, on large datasets, and/or multicenter datasets [39] [40]. This section outlines the current quality control methods in the field and addresses their strengths and weaknesses.

The idea of automated quality control in MRI has seen many different applications of varying strengths and weaknesses in terms of performance and generalization. The following works take different approaches to quality control, and help frame the optimal method to study large scale datasets. The first works discussed here distinguish the difference between quality control and quality assurance. Quality assurance focuses more on avoiding the occurrence of problems in image acquisition whereas quality control is more focused on finding problems in the output of that process. Glover *et al.* and Friedman *et al.* designed their studies to be more focused on evaluating the output of the hardware such that the variation could be reduced [41] [42]. For example, they do include measures of signal-to-noise ratio and an evaluation of the noise field but their focus is on the performance of the hardware (scanner) and the specifics of the protocols for imaging and less about quantifying artifacts or outliers. It is important for a quality control tool to assess data with intentions to objectively quantify good- versus poor-quality. Marcus *et al.* focuses

efforts of quality control on evaluation the output of protocols for image acquisition by manual post-scan quality scores [43]. This approach, although successful as a quality control tool, is not feasible for large scale studies, as manual quality control is expensive, subjective and laborious.

Early efforts of evaluating pass and fail images using quality metrics from Woodard and Carley-Spencer began studying T1-weighted MR images with a no-reference approach; this effort had some success discriminating between good quality images and synthetic noisy images or images affected by intensity non-uniformity [44]. The focus of this study is not necessarily applicable to non-synthetic clinical data or FLAIR MRI datasets.

Mortamet *et al.* followed a similar approach in evaluating the quality of images by studying the noise field and by studying artifactual voxels [35]. The output of this study includes two quality predicting indexes that were proven successful on sagittal T1-weighted MRI images. A major contribution from this work is the idea of a quality index or rating that is automatically generated and correlates with visually perceived quality in MR images. These metrics unfortunately do not generalize to FLAIR MRI because intensity thresholds are specific to T1-weighted MR images and a raw evaluation of the noise field can be inconsistent as the background of an image varies from scanner to scanner. This method only considers a portion of the background noise and does not evaluate any information within the head region of the scan. Also, this application is designed specifically for sagittal images, as the background isolation algorithm does not generalize to transaxial images. This application, does present an idea of evaluating quality measures on a volumetric level, which is not common amongst quality control approaches; this method is similar to the work presented in this thesis in that regard.

The UK Biobank pipeline, introduced by Alfaro-Almagro *et al.*, present their quality control tool, which is designed primarily for T1 MR images but reports results of using FLAIR MR images as well [45]. They explain that their pipeline has success for T1 MR images but had worse results in attempting to apply the pipeline to FLAIR MRI. It is recommended that future studies avoid the current method of non-linear warping FLAIR to the T1 space, as the warped images are different from original T1 images and it is difficult to compensate for these differences. Future studies of this work plan to expand to other modalities like FLAIR but at the current point in time the use of

registration techniques, an atlas/template, and segmentation algorithms are specific to T1 images.

Another implementation of a quality control tool for multi-center data was proposed by Esteban *et al.*; their pipeline evaluates MRI scans on a pass-fail basis using supervised machine learning. A major contribution from this paper is the fields/families of focus and their corresponding groupings of quality metrics that help describe image quality with purpose. While this tool did prove to be successful, their features are optimized from non-FLAIR MR images and require labelled data [39]. Their limitations address the need for labelled data from more sites and more scanners; they also admit that their training data had some form of intensity manipulation and image smoothing that was unplanned prior to attempting to classify pass or fail.

An approach that had more success on FLAIR MRI comes from Jang *et al.* who suggested a no-reference IQA method that involves multi-directional filters. This method correlates their no-reference features with full-reference features applied to different levels of synthetically distorted images [40]. The idea of correlating no-reference features with full-reference features does not generalize to clinical applications of sorting volumes with respect to quality. Further, their method evaluates one image per volume, which is addressed as a concern as the algorithm is prone to errors in cases where the slice position is not consistent from the training set to the testing set, or there is variation in contrast or resolution between training and testing data. Though this approach reports great results of correlation of their features with full-reference features on synthetic data, it has yet to be validated on a large multicenter dataset in a quality control application.

A series of studies in the field of MR IQA includes three applications produced by Osadebey *et al.* [46] [47] [48]. The first work considered is an IQA approach that utilizes spatial statistics to describe quality [47]. The focus of this study is to correlate artificial distortion with image quality features. Unfortunately, this approach requires assumptions of similar human anatomy between training and testing sets, which is an unrealistic assumption for clinical data. While this study did report correlation of features with quality of FLAIR data, it also lacked a quality classification approach. The next work considered focused on describing image quality from multiple perspectives, similar to the work in this thesis [46]. The shortcomings of this work were the erroneous quality measures in cases where images presented artifacts, the requirement for bias-field correction prior

to evaluation, and the lack of generalization beyond images. The last work considered, focused on brain and cardiac images from multicenter datasets, which has a similar approach to the work in this thesis [48]. This work intended to correlate quality measures with quality-labelled data. Unfortunately, the results were only evaluated on synthetic degradation of the clinical data and applied in T1 specific applications. Notably, none of these approaches attempt to classify good- or poor-quality as in a quality control application, and none of these methods generalize beyond singular images to volumetric measures of quality.

Two common shortcomings of these aforementioned works is the lack of classification between good- and poor-quality and the designs of these tools are specific to non-FLAIR MR images. Many applications applied synthetic distortions to correlate with their quality features; however, no applications were applied to volumetric classification of quality on clinical data. These previous works provide an excellent description of success, limitations, and how to improve moving forward in the field of MR. These works combine to lay a foundation that is translatable to quality control of large FLAIR MRI datasets. Further discussion of similarities of the work in this thesis and some of these works is described after the proposed outline is explained in section ??.

1.4 Proposed Outline

The envisioned pipeline for the development of this quality control tool involves an all-encompassing summary of the dataset in addition to the rejection of poor quality images or corrupt data. This tool intends to be applicable to any large FLAIR MR dataset. The vision of this tool inspired the following list of constraints:

- The tool must be automated
- No-reference
- Describe image quality in FLAIR objectively
- Unsupervised

- Reject outliers and poor-quality data
- MCE-invariant
- Provide an IQA summary

Automation

The purpose of including automation in the constraints is because the nature of quality control is shifting to that direction; manual visual analysis is time-consuming and expensive for large datasets. As the research community continues to increase these large datasets toward the trend of Big data, it becomes unreasonable to process thousands of volumes or hundreds-of-thousands of images manually. Thus, there is a need for algorithms to be self sufficient.

No-reference

The tool is required to be no-reference because the assessment of each individual image will not have a gold-standard to be compared against. No-reference methods are most useful in quality control for measures of distortion from an original image. For this application, the tool should be able to assess any image for any dataset, without reference to any specific image. This will allow the tool to increase its generalization and eliminate the constraint of a reference image that traditional reference-based methods require.

Ability to describe image quality in FLAIR MRI objectively

The purpose of designing this tool with the ability to describe image quality specifically for FLAIR MRI is to support the algorithms that surround the study of WML and neurodegenerative disease. Many existing methods of quality control in MRI design their features based on T1-weighted MRI, which as previously mentioned is fundamentally different from FLAIR MRI. The major differences to beware of for FLAIR include the unimodal brain class (white matter and grey matter combined),

the nulled CSF class and the high intensity WML. All features that concern these characteristics, should be calculated with caution. The description of image quality must also be objective in attempt to predict a human observer's perception of quality.

Unsupervised

The quality control pipeline should be able to describe image quality and reject outliers without the requirement of labelled data to learn from or train on. In clinical applications, it is not realistic to obtain manually labelled data that can depict the quality of large datasets. This point is supported by the information explained by justifying the purpose of automation; subjective visual analysis is laborious and expensive, thus, it is difficult to obtain labelled data. Further, the labelled data would have to encompass multiple centers, manufacturers, scanner parameters, software and hardware differences, to provide the machine learning algorithm with adequate information for generalization.

Reject outliers and poor-quality data

The rejection of outliers and poor quality is an essential part of this quality control pipeline. The reason this constraint is included, is because the envisioned tool should be able to not only describe the image quality but it should be able to discriminate between volumes that should be kept or included in a study and volumes that should not be kept or included. The value in this part of the pipeline is that a user can apply it to their dataset and avoid the inclusion of outliers and poor-quality data that may affect their algorithms, analysis and conclusions. As the need for confidence in conclusions calls for larger datasets, the opportunity for poor-quality data to corrupt a study increases as well. It is understood that a poor-quality scan could accidentally skew results and cause for incorrect conclusions to be made in a study and it is important to avoid this type of error.

MCE-Invariant

Invariability to the multicenter effect is another constraint that is crucial for the development of any algorithms working with big data as this provides the ability to generalize and succeed on more than one dataset. This tool must also evaluate the image quality and reject outliers with consideration of the MCE; it is important to account for the variability of the data in the outlier rejection decision, just like any other algorithm should account for this variability. An oversight of the MCE could result in wrongful rejection of outliers. For example, a dataset comprised of volumes scanned by Siemens that included a subset of volumes scanned by Philips, could perceive the Philips volumes as outliers since they might differ from the average (Siemens) volume. This type of error in outlier rejection would potentially reject good-quality volumes. Thus, it is important to consider the MCE when evaluating whether a volume is an outlier or not.

Provide an IQA summary

The last constraint of having an IQA summary is the most informative piece of this pipeline; the benefit of the IQA summary is to provide the user with an understanding of the nature of the data, such that intelligent preprocessing steps could be made to combat specific irregularities or issues within the dataset. For example, preprocessing may include contrast enhancement or standardization of intensities to improve an algorithm's performance; the requirement of these steps can be understood using a CNR summary or the intensity non-standardness summary that this tool provides.

The proposed fields of focus

Literature review of previous works and standard image quality assessment examples in other fields have provided motivation for the study of the following fields to capture an accurate description of image quality for FLAIR MR:

- Intensity non-standardness
- Image sharpness or blur
- Signal-to-noise ratio and contrast-to-noise ratio
- Noise field
- Common artifacts
- Total image quality score

Each of these categories is discussed and justified in the following section to explain how studying these fields improves the pipeline. On a general level, the purpose of studying each of these categories is to solve the issue of algorithms reporting less success in generalizing to large multicenter datasets. For the field of algorithms studying brain scans, the common issue of generalization can be improved by removing poor quality data and by applying the appropriate preprocessing to manipulate the data. This tool is designed to remove poor quality data and provide quantitative information about the data that can be used to efficiently apply the necessary preprocessing to yield better algorithm performance. Chapter 2 explains the details of the quality control pipeline and justifies the approach to the IQA category selection.

1.5 Contributions

This thesis presents a pipeline that encompasses an image quality assessment summary and a quality control algorithm specifically designed for FLAIR MRI. This thesis also presents a literature review of the current and previous works in the field of quality control in MRI (with emphasis on FLAIR). This advancement provides opportunity to process large datasets robustly, with a tool that is specifically designed for FLAIR MRI.

The following particular quality metrics are novel contributions to the study of FLAIR MRI brains scans: the first is a **multi-scale image sharpness measure** that uniquely encompasses fine and

course edge content within an image to a single output feature; the second is a **motion artifact feature** that utilizes wavelet decomposition to study vertical-edge ringing artifact common to FLAIR MRI; and the third novel feature is the **total image quality metric** that encompasses contribution from 5 image quality categories to summarize an overall quality score for a volume. Each of these are explained in detail in Chapter 3. Extensive validation of each component of the pipeline and the entire pipeline is another major contribution of this thesis, which is included in Chapter 4.

Chapter 2

Quality Control Tool: Fields of Focus

This section investigates the categories of intensity-standardness, sharpness/blur, SNR, CNR, noise distribution, prevalence of motion artifact and overall an overall quality score for FLAIR MRI. Each quality control category, discusses current work in the field and explains the purpose of including each field of focus in the pipeline. This quality control tool aims to quantitatively assesses the image quality of any given FLAIR MR volume within a dataset. It is important that these aspects comprehensively evaluate not only the quality of each volume, but the similarity of each volume to the rest of the dataset from multiple perspectives. Volumes can appear similar in some of these aspects but show great differences in other aspects; this highlights the importance to consider as many details as realistically possible. The following sections explain why each category should be considered.

2.1 Intensity Non-Standardness

Intensity non-standardness refers to the variation across MRI grayscale pixel values from one scan to another. Since the acquisition setup determines parameters like noise distribution and intensities for an image, the trend of intensity non-standardness is common when comparing volumes collected from different scanner vendors. The intensities produced by a scanner depend on the vendor, protocol and patient anatomy, all of which are difficult to control for and standardize [49]. Figure 2.1 provides an example of the three scans produced by different manufacturers. Each

scan produced a different range of intensities to represent each class. Specifically observing the brain class or the skull intensities with respect to the grayscale colorbar displays the true difference of intensities between scans.

Figure 2.1: Example of intensity non-standardness: GE, Siemens, Philips

Observing figure 2.1a, the brain and skull are represented by much higher intensities than that of figure 2.1b and 2.1c. Further, the skull intensities of figure 2.1c are noticeably higher than that of figure 2.1b. These differences are commonly understood to be intensity non-standardness and prove to be difficult for segmentation and quantification algorithms [49] [34].

The most common method to solve this issue is a post-scan approach called intensity-standardization, which can manipulate the intensities on a scan-by-scan basis to represent specific tissue classes [49].

There have been many efforts put forward but none of these methods have yet to emerge as the optimal pipeline, as each method succeeds under different circumstances [49] [50]. A method published as recent as May 2019 claims to preserve the appearance of WML during the transformation, which is promising for correlating lesions with neurodegenerative disease [34]. Nonetheless, it is important to understand the variation of intensities within a dataset, prior to selecting a method for intensity standardization, if it is even required.

Evaluating the intensities of an image, is a simple approach to understand the similarities or differences of classes between volumes within a dataset. The scale and range of intensities are important parameters for designing thresholding-, clustering- and adaptive intensity-based algorithms. Understanding the extent of variation of the intensities within each volume can be accomplished by quantitatively assessing the shape and range of raw volume-histograms. Then,

each volume-histogram's characteristics, could allow for some form of intensity standardization and ultimately improve performance for the aforementioned intensity-based algorithms.

Another example of intensity non-standardness is found in figure 2.2 and 2.3 are examples of intensity histograms from two images of different volumes within the same dataset, produced by **the same person** [51]. Each image is accompanied by a histogram to summarize the gray-level intensity scale and range for each scan.

Figure 2.2: Philips intensity non-standardness example: image and histogram

Figure 2.2a is a sample slice from the middle of the brain volume produced by a Philips scanner. Figure 2.2b is a truncated version of the intensity histogram; this histogram's x-axis represents the intensities present and summarizes the information up to the intensity value of 1000.

Figure 2.3a displays a comparable middle slice of a brain volume produced by a different scanner (Siemens), from the same patient. Figure 2.3b is, again, an intensity histogram that has been truncated for comparison purposes. Noting the scales of these histograms, it can be observed that although each of these images are representing the same information, they are vastly different. While the brain class is represented by the range of intensities from 450 - 850 in figure 2.2b, the brain class in figure 2.3b is represented by the range of 150 - 220. Again, this is the commonly known issue known of intensity non-standardness. To further prove the existence of intensity non-standardness on a larger scale, the volume histograms from an entire dataset of 700 volumes

Figure 2.3: Siemens intensity non-standardness example: image and histogram

are compiled into figures 2.4a, 2.4b and 2.4c. Each figure contains approximately 33 percent of the volumes in the dataset. Noting the x-axis grayscale intensities, it can be understood how common this issue of variable intensity ranges is and how important it is to understand prior to algorithm implementation or design.

Figure 2.4: Example of intensity non-standardness across a whole dataset

This vast difference of grayscale intensities draws concern for quantitative algorithms studying non-standardized, multicenter data. Ultimately, the field of intensity non-standardness should aim to quantitatively describe the variation of intensities of any FLAIR MR dataset such that a user could recognize how and why a particular algorithm may fail; thus, intelligent preprocessing measures could be implemented as a solution. Further, the measures that study the scale and range of intensities should be able to serve as a validation tool, as it can be used before and after preprocessing of the data. The user can apply their standardization or normalization to the

data and evaluate their progress to understand if they have successfully accomplished what was intended or not. This type of familiarity with the data instills confidence in the user and can prevent commonalities of poor algorithm performance or failure.

2.2 Image Sharpness/Blur

Image sharpness or blur refers to the quantification of edge contrast present in an image. Blur may arise in images due to variable imaging software and hardware differences as well as artifacts. An image quality assessment should consider the blur within an image to ensure an image has the ability to discern one tissue from another or the ability to clearly define borders prior to using said data. It is important to use reliable quantitative measures to describe the fine- and coarse-edge content within an image; these measures should be independent of the previously discussed intensity non-standardness. In addition to measuring the edge content, the spread of information throughout the image should be considered as well.

Figure 2.5 displays an example the simplest way to visualize the borders between classes that represent the edges within an image. Figure 2.5a displays a sample image that has well defined borders, thus good edge-content or image sharpness. Figure 2.5b displays an edge map, which yields higher values or appears brighter in regions that contain sharper edges.

Figure 2.5: Example of edge content representing the sharpness of an image

In the case of FLAIR MRI, a sharper image is often preferred as it may depict the size of specific lesions/regions of the brain or help distinguish between particular classes; this contrast and accuracy is particularly useful in diagnostics. For example, an image with high contrast between a white matter lesion and the surrounding area of brain matter will be more accurate in determining the borders and thus result in a more precise total volume reading. In cases where the edge content appears low for a dataset or for scans produced by a manufacturer, there are solutions available that can enhance edges and reduce noise; one example titled *Image enhancement and noise suppression for FLAIR MRIs with WML* was designed by Khademi *et al.* in 2010 [52]. Despite the solutions being available, the edge content or blur must first be understood in order to know that this type of preprocessing should be done to the data. An example of the variable edge content between two volumes from the same dataset can be found in figure 2.6 [33]. Figures 2.6a and 2.6d represent the middle cross-sectional brain slice of two different volumes that contain different sharpness/blur. Figures 2.6b and 2.6e show the ventricles from figures 2.6a and 2.6d, respectively. These images display the contrast in edge content between the two volumes; to further support this claim, figures 2.6c and 2.6f show the edge map of these images and their ability to separate the brain class and the ventricles.

The edge content in figure 2.6c displays a sharper border between the ventricles and brain class than that of figure 2.6f. Extrapolating this information to an entire dataset, it can be understood that this much variation of sharpness/blur exists for multicenter studies that include data from more than one scanner. This type of disparity is problematic for quantitative algorithm design, as it is difficult to account for such a wide range of edge content within a single dataset. Further, by including such poor-quality data in a study, the results can be affected and incorrect conclusions could potentially be drawn. As the focus of this thesis surrounds FLAIR MRI and cleaning datasets such that WML can be studied, it is important that this tool considers abnormally blurred data as poor quality to avoid inaccurate WML segmentations.

Figure 2.6: Edge content comparison between two volumes from the same dataset

2.3 Signal-to-Noise Ratio and Contrast-to-Noise Ratio

The signal-to-noise ratio (SNR) of an image is a great descriptor of image quality as it assesses the average signal or image intensities in ratio to the standard deviation of the noise. Less noise present in an image corresponds to less unwanted random variation of pixel intensities and thus a higher SNR. A high SNR is important in imaging, especially for the quantitative algorithms that perform better when there is less unwanted random variation of intensities. Common measures of SNR evaluate the signal present in an image with respect to the background noise; this ratio is often the mean of a patch or class like the brain divided by the spread or standard deviation of the background [53]. The issue with the traditional SNR measures for today's MRI scans, is that some scanners null the background in an attempt to inflate the SNR; by nulling the background pixel intensities, the standard deviation decreases which in turn drives up the SNR [53]. The solution is

to consider the average brain intensity against the standard deviation of the brain class [53].

The contrast-to-noise ratio (CNR) of an image is another crucial measurement that describes the image's quality through quantifying the differences in intensities between classes. This metric provides a simple and accurate representation of the contrast between classes in comparison to the amount of noise present in the image. The CNR, similar to the SNR, avoids the use of the background noise to avoid the artificial inflation; this calculation uses the brain and CSF classes instead. CNR directly affects the performance of segmentation and classification algorithms [53]. Higher CNR corresponds to high contrast between classes with low noise present; this implies that finding borders of classes with accuracy should be simpler. In addition to the possibility that data produced by a scanner may result in low CNR, a more common case for multicenter data is that the CNR (and SNR) can be adequate for some volumes and poor for others; this type of variation in the data should be understood prior to designing algorithms. Figure 2.7 displays an example of a CNR summary from a multicenter dataset [54].

Figure 2.7: Boxplot of CNR by manufacturer

In figure 2.7, the difference between the highest average CNR and lowest average CNR by manufacturer is approximately 28 percent; this type of variation explains the difficulty of designing an algorithm for large multicenter datasets and stresses the importance of understanding the data prior to designing. A SNR and CNR summary for all volumes within a dataset would allow the user to make informed decisions about algorithm development and preprocessing methods. For example,

a FLAIR MRI dataset that reports a particularly low CNR per volume could be contrast-enhanced, especially in cases where WML volume is being studied and the border affects accuracy [55]. It might also be useful to evaluate the data before and after contrast enhancement to see if the data has been preprocessed appropriately and if the intended purpose was accomplished.

2.4 Noise Field

MRI has a reputation for providing high-resolution images, which relates directly to the noise present during acquisition and reconstruction of the images. The phenomenon of the noise field is studied so intensely in MRI because it is a limiting factor in the speed of acquisition; a typical scan requires long acquisition time to reduce the noise present, thus increasing the resolution [56]. To provide context, the noise in MRI is an additive property that *often* takes a Gaussian distribution by the central limit theorem. The intensities for a class within an image vary randomly, but surround a mean value to describe said class. The noise field can vary depending on the on geometry of the MR receiver coil resistance, number of coils, and the reconstruction algorithm [57] [58]. As these variables change with respect to scanner, it is difficult to have an ideal noise distribution that is consistent throughout a dataset; thus it is important to understand the types of noise distributions prior to implementing an algorithm. This is especially true for algorithms that assume an underlying Gaussian noise distribution, as this may not always be the case [57].

The shape of the distribution of intensities of a volume histogram, is a difficult phenomenon to account for in algorithms. The shape of the volume histogram describes the varying classes present in the scan, which may include: background, skull, CSF, brain and WML. These classes create a distribution that varies from scan-to-scan. It is important to study this variation in a dataset to avoid including outliers and incurring issues with algorithm performance due to the MCE.

The shape of any class of the volume histogram can be represented by the noise present in the scan. Variation of this shape, highlights a different distribution or relationship of tissue classes. Figure 2.8 provides an example of volume histograms of the noise field from two different volumes extracted from the same dataset [33].

Figure 2.8: Noise field variability example of two volumes from the same dataset

The difference in noise field may not be perceivable from a subjective visual analysis of each volume; however, the isolation of the noise histograms displays the difference. Figure 2.8a yields a flatter, left-tailed distribution than that of figure 2.8b, which has a greater peakedness and more right-tailed distribution. One issue that arises in this situation is the dissimilarity between the two distributions and another issue is that neither appear naturally Gaussian. Significant variation of these distributions implies that model-based methods, such as those that fit Gaussians to a histogram, may not be applicable to large multicenter datasets; especially those that have different distributions.

One major artifact common to MRI is called the bias-field artifact or intensity inhomogeneity [59]. Essentially, this artifact is mainly caused by improper acquisition process that results in a misrepresentation of grayscale intensities for each class. Bias-field artifact can be caused by the imaging coils having variable sensitivity throughout a single image/slice acquisition [60]. This inconsistency becomes apparent through the intensity representation in the reconstructed image. Figure 2.9 displays an example of this intensity inhomogeneity throughout the brain class, as the brain pixels that surround the ventricles are of much lower intensity than the brain pixels toward the bottom of the image. This type of artifact can cause an unusual noise distribution; thus, it is important to study the noise distribution to characterize these types of artifacts.

Figure 2.9: Bias-field artifact example

This type of artifact can affect clinical interpretation of important tissue boundaries and can be problematic for automated analysis of affected images [60]. If bias field artifact affects any volumes within a dataset, it should be known, such that it can be corrected by using existing methods that serve this purpose [61] [62]. Ultimately, it is important to consider the general shape of the noise field distribution throughout a dataset to understand when issues of non-Gaussianity or bias field artifacts exist, and it is also important to consider how the noise field varies from volume-to-volume within a dataset.

2.5 Motion artifact

Motion artifact can be described by captured movement of tissue or fluid during an scan that results in unnatural borders and incorrect representation of the patient in the reconstructed image. MRI is particularly vulnerable to motion artifacts due to the nature of the signal acquisition and the longevity of the scan. As mentioned in the introduction section about MRI, the T1 and T2 relaxation constants are the limiting factors in signal acquisition; thus, it will be difficult to design hardware or software solutions to speed up scans or reduce motion artifacts. While there are many types

of motion artifact, some common motion artifacts can manifest themselves into the reconstructed image and appear as blurring or ringing in the phase direction [63]. This movement is commonly caused by the patient's head tilting left and right while lying in the supine position [64]; an example of the resultant ringing motion artifact, referred to as Gibb's Phenomenon, can be found in figure 2.10. These two images appear sequentially in a volume, yet the orientation of the patient's head is not consistent, despite the short period of time in which these images are captured; this occurs as the patient's head moves drastically during the scan and the result is severe motion artifact.

Figure 2.10: Common motion artifact example: Gibb's Phenomenon

Strictly through visual observation it can be understood that the quality of this image in figure 2.10a and 2.10b is impaired by the motion that took place during the acquisition. It should also be noted that these artifacts appear as ringing parallel lines to the intense edge of the skull repeated in the horizontal direction; this artifact is portrayed in the phase-encoding direction of the acquisition sequence [65] [66]. Should the phase-encoding direction of the acquisition change to the vertical direction, then these artifacts may be portrayed differently; however, this direction of artifact is common to the data used in this study and in works referenced by this thesis.

Unfortunately, there is no generic solution to all motion artifacts; instead, there are many partial solutions [64]. Several methods can be used to reduce motion artifacts, including prospective motion correction (PMC) [67], patient immobilization, cardiac and respiratory gating, and signal

suppression of the tissue causing the artifact [68] [69]; however, these solutions are required to be applied in advance of the signal acquisition. Some post-scan solutions are being developed for motion artifact of large FLAIR MRI datasets; for example, Ryu *et al.* designed a deep-learning effort to correct motion artifact on synthetic FLAIR images [70]. Ultimately, it is important to understand the severity of motion artifacts, if they exist within a dataset, prior to evaluating the potential solutions in terms of artifact correction that should be applied. For this reason, the motion artifact feature is included in this quality control pipeline.

2.6 Total Image Quality Score

An overall quality score is an effective method to quantitatively evaluate how each volume relates to the entire dataset. Many attempts at deriving the ideal image quality score/index have left multiple examples for varying applications that all share the same goal: to simplify an image quality assessment to one singular quantitative figure. Notable examples include: a universal image quality index designed by Wang *et al.* in a generic application that considers the the image distortion [71]; a lightness-contrast measure combined with a texture-contrast measure to output a total quality score for clinical MRI by Osadebey *et al.* [48]; and Mortamet *et al.* who proposed two quality indices that evaluate artifactual voxels and noise distribution [35]. While none of these methods are specific to FLAIR MRI, it is important to still consider them as many are designed as generic image quality scores.

The total image quality scores designed by each of these works are unique and efficient in describing the quality of an image. The quality score for this pipeline should summarize the findings of the IQA and output a single metric, as the previous works in this field have done.

2.7 Summary of Approach

To summarize the approach for this quality control pipeline, the fields of focus to describe a FLAIR MR volume's image quality include: intensity-based features, sharpness-based features, SNR

and CNR, noise-based features, a motion artifact feature and a total image quality score. This feature set is specifically selected to detect/reject outlier volumes from large multicenter clinical FLAIR datasets. Specific artifacts of corrupt data, bias field, blur, noise and motion artifacts were considered in the selection of this feature set. In this section, this approach is compared to similar works in the field of quality control of structural MRI.

The work of Mortamet et al. had a similar application of sorting volumes on a pass/fail basis using quality measures. This work was particularly similar because it generalized to a volumetric application and was applied to clinical data. This work also similarly applied a threshold technique (further explained in Chapter 3) to its quality measures using receiver-operator-characteristic analysis. These quality measures only evaluated artifactual voxels in the background and the noise distribution of the background of T1 sagittal images for two output quality metrics. The work described in this thesis, is applied to FLAIR MRI volumes in a transaxial orientation and it describes image quality by evaluating whole images instead of just a portion of the background of each image.

The work of Jang et al. utilized multi-directional filters and image statistics to evaluate image quality. This approach is addressed here because it was applied to FLAIR MR images [40]. This study focused on correlating feature values with full-reference features on different levels of distorted data. While this study was one of few to be applied to FLAIR data, it did require the data to be of consistent contrast and resolution for the features to work properly. The work in this theses is similarly applied to FLAIR data, however, the application is to sort data with respect to pass/fail labels of quality and the work in this thesis can generalize to unseen data of varying qualities and resolutions. The work in this thesis also considers multiple IQA aspects and generalizes to volumes not just images, as the objective is to provide quality control for large multicenter databases.

The most influential quality assessment approach was that of Osadebey [46]. This work utilized features of: noise, lightness, contrast, sharpness and texture alongside a quality measure that incorporates each feature in a weighted sum. This work correlates quality with synthetically distorted T1 data without any quality control or sorting of pass/fail application; however, it did provide an excellent framework for feature extraction and a total image quality score. Although, these features did not cover the same space as described in this thesis, it has a very similar

approach with consideration of aspects of noise, contrast, sharpness and an overall quality rating in common. It should be noted that this work did require preprocessing of bias field correction before assessing quality; has an erroneous sharpness measure in the presence of artifacts; and its quality score did not generalize beyond a singular image for a volume (score varies with position within volume). The work in this thesis is applied to FLAIR volumes, incorporates features from overlapping image quality aspects, has a quality control outlier rejection application, and does not require manipulation of data prior to feature extraction, as the features are extracted from raw data.

Additionally, the work described in this thesis incorporates an IQA summary that provides the raw extracted data information in a spreadsheet. This information can be used to understand the variance of each IQA category and can be used to suggest preprocessing techniques that may be required prior to algorithm implementation. It should also be noted that this work is specific to FLAIR MRI volumes due to the methods of extracting features of: SNR, CNR, and motion artifact features (further explained in Chapter 3). These methods could not generalize to T1 applications in comparison to many of the aforementioned "structural" MRI quality assessments have been designed.

Chapter 3

Image Quality Assessment and Quality Control Pipeline: Methodology

This chapter explains the methodology of the entire IQA and quality control pipeline. The methodology has been divided into sections of preprocessing, feature extraction, total image quality score and outlier rejection. Then the validation datasets and validation metrics are explained just prior to the results section. To provide insight into the methodology of this quality control tool the overview of this pipeline is summarized in figure 3.1 and the details can be found in the following sections.

Figure 3.1: Quality control pipeline flowchart

3.1 Preprocessing

The preprocessing that is done to the original dataset includes the generation of an intensity standardized version of the data and the generation of class masks. The purpose of standardizing and generating binary class masks is to intentionally isolate regions of the scan to be used in further examination. Features will be extracted from these isolated regions that will ultimately be used to depict the quality of an image/volume.

3.1.1 Intensity Standardization

The original data is affected by the MCE and has variable ranges of intensities for each class depending on acquisition software and hardware differences between scanners. Thus, it is difficult to isolate any class using a consistent intensity threshold. For the purpose of class mask generation, predictable ranges of intensities for each class is advantageous. The intensity standardization of the original data produces a version of the data that has consistent intensity ranges for each class. The method of intensity standardization developed by Khademi *et al.* for FLAIR MRI is used to generate this predictable standardized version of the data to be used further in mask generation [34].

3.1.2 Mask Generation

The mask generation is an important step performed prior to any metric extraction. The binary masks include a head-mask that is used to isolate the background; a brain-mask for isolation of the brain; and the CSF mask which is used to isolate the CSF class. Each of these isolated regions are used in feature extraction for the following features: SNR, CNR, noise field characteristics, and the prevalence of motion artifact. In each respective section throughout the methodology, an explanation is provided for how these masks are utilized in the extraction of each feature.

3.1.2.1 Head Mask Generation

The head-masks are generated by denoising and smoothing each image prior to thresholding the background class from the foreground; then post-processing is applied to create the binary head-mask [34]. The purpose of creating a head-mask is to be used in feature extraction from the isolation of the background class. By applying a head-mask or an inversed-head-mask to an image, the background and foreground can be studied separately. To provide more detail each step is explained below.

The first step in generating the head mask is to apply a median-filter to each original image as a denoising technique. The median filter is determined to be of size $0.08M \times 0.08N$ for an $M \times N$ image. The next step is to generate each image's intensity histogram from the standardized version of the data and use Otsu's thresholding method to generate two thresholds [72]. The lower threshold separates the background class from the foreground. Any pixels that report an intensity lower than this threshold are deemed background class. Then post-processing is applied to the rough head-mask in the form of morphological operations. The combination of dilation, filling, and erosion, removes the holes and finalizes the shape of the mask. This final output mask is then dilated again, to include the immediate surrounding area of the skull. For clarification these binary masks contain 1's that mark pixel location within the image to be preserved and 0's that mark pixel location to be deleted. Figure 3.2 displays an example of a headmask produced for a middle-slice of a volume. Note that the binary head mask will appear black (0) and white (1).

Figure 3.2: Example of head-mask

Figure 3.2a displays the original mid-slice image; figure 3.2b displays the inversed-head-mask that is generated to be slightly larger than the head; and figure 3.2c displays the resultant isolation of the background by point-wise multiplication of the original image with the inversed-head-mask.

The purpose of generating a head-mask that is slightly larger than the actual head, is because of the application; the head-mask is applied inversely to isolate the background to study motion artifact. It is important for this application that there are no hard edges from the skull accidentally included in the background image, as it could falsely appear as motion artifact. This topic is further discussed in the limitations section of Chapter 5 of this thesis.

3.1.2.2 Brain and CSF Mask Generation

The first step in generating the brain-masks was to apply an intensity standardization pipeline such that the tissue classes were in predictable ranges of intensities [34]. This step is useful for distinguishing the brain class from the CSF class later on.

The next step is to extract total brain-masks, such that a brain class mask and a CSF class mask could be created. The method used for brain extraction and skull stripping was made available by Khademi *et al.* [73]. This framework is based in machine learning and was designed to specifically handle the contrast mechanisms associated with FLAIR MRI and multicenter data. Brain tissue is segmented on a per pixel basis using an intuitive feature set paired with a random forest classifier. The total brain-masks contain the brain class and some CSF class primarily via the ventricles. The next step to generate the final brain- and CSF-masks required the intensity standardized data and the use of a simple threshold. The output is a binary brain-mask and a binary CSF-mask. Figures 3.3 and 3.4 display an example of a brain-mask and a CSF-mask, respectively.

When observing figure 3.3b it can be seen that this binary brain-mask generated from figure 3.3a can be used to isolate the brain class, which is achieved in figure 3.3c. Similarly an example is provided for the CSF-mask in figure 3.4.

Observing the binary CSF-mask in figure 3.4b it can be seen that it isolates the CSF from figure

(a) Original image

(b) Brain-mask

(c) Isolated brain

Figure 3.3: Example of brain-mask

(a) Original image

(b) CSF-mask

(c) Isolated CSF

Figure 3.4: Example of CSF-mask

3.4a, which is most clearly outlined in figure 3.4c. The isolation of these classes must be accurate to maintain the integrity of the feature extractions that they are used for. The SNR, CNR, and noise field calculations utilize these masks; thus, any errors in the generation of these masks will propagate through in the IQA of each volume.

3.2 No-Reference Image Quality Metrics for FLAIR MRI

The quality control pipeline uses the feature extraction of eleven image quality metrics for two applications. The first application is to provide an IQA summary of each volume with respect to the dataset; the second application is to label good- and poor-quality (outliers) with justification

for each rejected volume. Features were selected, as justified in Chapter 2, to not only describe general image quality, but also to detect artifacts specific to FLAIR MRI. Recall from Chapter 2, the purpose of the **intensity-based metrics** is to evaluate the intensity non-standardness of the dataset, with a particular ability to detect significant variation or problematic volumes that contain extreme intensity values. The **sharpness-based metrics** are included for an evaluation of the variability of the edge content throughout the volumes in the dataset. The **SNR and CNR** are included to depict the general quality and contrast of the images throughout the dataset. Particular cases like bias-field artifacts can be detected thanks to the inclusion of the **noise-field features**. The **motion artifact feature** is included to detect specific artifact cases like Gibb's phenomenon. Most importantly, recall that the **total image quality score** is included to discern good-quality data from poor-quality data and used as a false-negative reduction method. Again, each field is included in the IQA summary to instill knowledge of potential preprocessing or elimination of data that should be done prior to the use of a dataset.

The details for each of these categories of image quality are included throughout the methodology section and are summarized at the end in table 3.1.

Prior to an explanation of the details of this pipeline, it is important to understand the structure of a potential dataset that could be assessed. Recall section 1.1.3 as it explained the relation of an image to a volume to a dataset. Any given FLAIR MRI dataset is comprised of brain-scan volumes produced by scanners that subject the dataset to the MCE. An equation to provide mathematical notation of this explanation of a volume is,

$$V(x, y, z) = I_z(x, y). \quad (3.1)$$

The above equation shows how each 3-dimensional volume, V , is comprised of 2-dimensional images, I , denoted by $z \in [1, Z]$, where Z represents the number of slices. Each image is then comprised of pixels, denoted by x and y , where $x \in [1, M]$ and $y \in [1, N]$ for an $M \times N$ image; these pixels are assigned grayscale intensity values. A volume has many 2-dimensional cross-sectional images sorted by z , that represent the brain/head at different locations in a cranial/caudal direction.

The organization and relationships of intensities throughout an image and volume is the important information that represents the head-scan and ultimately discerns the quality of the data.

The quality control pipeline evaluates each volume through volumetric measures, followed by a slice-by-slice extraction of metrics from the images. The volumetric measures can be extracted without evaluating each individual image, but instead from volumetric classes or from the volume histogram, like the intensity-based metric, for example, utilize the volume histogram. Whereas the image-based feature extraction requires a feature to be extracted from each individual image and either averaged or summed across the volume; these measures are more computationally expensive. This information is summarized into a file that contains a total data summary that lists each volume and its respective feature values. The image quality assessment (IQA) summary produced by this pipeline is unique for the field of quality control of FLAIR MRI. This approach encompasses the following fields: intensity-based features, sharpness/blur-based features, SNR and CNR, noise-field, prevalence of motion artifact, corrupt data detection and a total image quality score.

3.2.1 Intensity-based Features

Section 2.1 discussed the importance of evaluating a dataset from the perspective of intensity non-standardness. The metrics selected to quantify the variation of the scale and range of intensities include: mean, standard deviation and maximum intensity. The mean intensity represents the average of the intensities, which corresponds to the approximate center of the histogram. The standard deviation talks to the spread of the intensities; this metric works in tandem with the mean to describe the volume histogram. The maximum intensities explain the length of the histogram, or the highest intensity present, as an additional metric to the mean and standard deviation.

The intensity based metrics of the mean, maximum intensity and standard deviation of the intensities were extracted from each entire volume. The maximum intensity is extracted from the raw volume histogram. The mean and standard deviation are extracted from the 98th percentile of the volume histogram to prevent an extreme maximum intensity from skewing the description of the

range of the data. This is calculated by taking the cumulative distribution function (CDF) of the probability density function (PDF) and finding the length at which the CDF is equal to 0.98. Figure 3.5 displays a visual of a typical PDF and the corresponding CDF for a volume histogram of a FLAIR MR volume. The PDF in figure 3.5a displays the shape of the PDF in red and the excess trimmed portion of the PDF in blue. Figure 3.5b displays the threshold of the 98th percentile in red, where the CDF is trimmed. The spike at the beginning of the PDF in figure 3.5a is just the overwhelming amount of background pixels that exist within the volume.

Figure 3.5: PDF and CDF of volume histogram

It can be seen that trimming the PDF at the 98th percentile avoids the inclusion of the few extreme intensities while maintaining a proper representation of the volume. Once this step has been completed the maximum, mean and standard deviation can be calculated. The maximum intensity is simply calculated by evaluating the length of the original PDF, which is represented by K , or it can be obtained by comparing each pixel value in every image throughout the volume to find the largest grayscale intensity value. Thus the first image quality metric can be calculated through,

$$IQ_1 = \max\{V(x, y, z)\}. \quad (3.2)$$

The equation to calculate the mean or the second image quality metric is accomplished by use of

the PDF,

$$IQ_2 = \sum_{x=1}^{K^*} x \cdot p(x). \quad (3.3)$$

In the equation 3.3, IQ_1 represents the mean intensity extracted from the truncated PDF, $p(x)$, of the volume's intensity histogram. The length of the truncated $p(x)$ is represented by K^* . Once the mean intensity, also represented as μ_{int} , has been extracted, it can then be used to calculate the standard deviation, which is described by equation 3.4, as follows:

$$IQ_3 = \sqrt{\sum_{x=1}^{K^*} (x - \mu_{int})^2 \cdot p(x)}. \quad (3.4)$$

Equation 3.4, takes a similar approach in using $p(x)$ to calculate the standard deviation of the intensities, IQ_3 , that represent any given volume.

3.2.2 Sharpness/Blur-based Features

The importance of quantifying the variation of image sharpness and blur within a dataset addressed in Section 2.2 yielded the selection of three metrics in total; two of these metrics to describe the sharpness and one to describe the blur.

The metrics selected to describe the edge content in an image include the novel multi-scale image-sharpness metric and what is referred to as a measure of blockiness. These measures are calculated from each image in a scan and then summarized to represent the whole volume. It should be noted that these metrics are calculated on a normalized scale as each image was resized to a standard 256x256 pixel image and normalized by the maximum intensity within the volume.

Multi-scale Image Sharpness

The multi-scale image sharpness was extracted by combining the magnitude of the image's edge content in the vertical and horizontal directions across multiple filtered versions of the image to ultimately obtain one feature value. The step-by-step process of extracting this feature is explained in detail below. The first step was to select an operator to evaluate the edge content present in the image. The Sobel, Prewitt, Canny and Laplacian were considered. The Canny and Laplacian operators were ruled out due to their sensitivity to noise and requirement for a smoothing filter to be applied prior to extraction. The Sobel and Prewitt operators were considered and evaluated on experimental data to understand which operator should be selected. The results included in section 4.1, reveal redundancy between the two operators, and so the Sobel operator was selected as it is more commonly used in these applications. The next step is accomplished through the use of the operator's filter in x- and y-direction. To generate the edge information in the x-direction, a Sobel operator is generated and applied to the image [74].

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}$$

This kernel above represents the Sobel operator, G_x , can be convolved with the image to generate the edge content in the x-direction (for a Prewitt operator the 2's would be replaced with 1's) [75]. This kernel can then be transposed and convolved with the image to generate the edge content in the y-direction. The kernel below, G_y , is the Sobel operator for the y-direction (again, replacing the 2's with 1's results in the Prewitt operator).

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix}$$

The product of convolving each of these kernels, G_x and G_y , with the image results in an isolation of the edge information in the x- and y-directions, respectively. An example of the resultant images can be found in figure 3.6.

Figure 3.6: Image sharpness calculation example: Sobel edge filter applied

Figure 3.6b displays the edge content in the x-direction; the higher intensities on the left and right sides of the skull and on the medial sides of the ventricles shows the edge content in the horizontal direction. This image can be referred to as I_x . Figure 3.6c highlights the y-directional edge content, which can be best observed through the top and bottom of the skull; this image can be referred to as I_y . This information is then combined by taking the magnitude of the x- and y-directional edge content, explained in equation 3.5,

$$IS_k = \sum_{m=1}^M \sum_{n=1}^N \sqrt{I_x(m,n)^2 + I_y(m,n)^2}. \quad (3.5)$$

In equation 3.5, IS_k represents the image sharpness of an $M \times N$ pixel image. Within the images, m and n represent a pixel in the x- and y-directions, respectively. This metric, IS , is then repeated at

15 scales of Gaussian blurred versions of the original image, where $\sigma \in [0, 13]$; then each result is combined into a single output measure, weighted by σ . The purpose of including information from each of these scales is so the metric can represent a range of fine edges to coarse edges within the image. An example found in figure 3.7 displays the original image, representing fine edge content, and two blurred examples representing more coarse edge content.

Figure 3.7: Multi-scale approach using Gaussian blur of varying standard deviation

The equation to describe the combination of the original and each blurred image's output, can be found in the following equation,

$$IQ_4 = \frac{1}{M \cdot N} \sum_{k=1}^K \frac{IS_k}{\sigma_k}. \quad (3.6)$$

Equation 3.6 shows the combination of features from each blurred image, IS_k , through a sum that is scaled by each respective blur, denoted by σ_k . In this case K is equal to 16, as the first iteration is the original and the following 15 iterations are Gaussian blurred. Although each image for this pipeline is normalized to a consistent 256x256, this metric is still normalized by the size of the $M \times N$ image. Each volume has every image go through this process, and the final output for the volume is reported as the average image-sharpness, IQ_4 , for the volume.

Blockiness

The next metric to represent the edge content in the image is a blockiness (B) measure; this feature is a patch based assessment of the vertical and horizontal magnitude of edge content of the raw image, as used in the *JPEG image quality score* designed by Zhou Wang *et al.* [76]. The approach has similarity to the image sharpness metric; however it is a more local measure, as it is patch based. This comparison of average differences across block boundaries can be explained in the equations below. Equation 3.7 explains the differencing signal, d_h , that evaluates the edge information in the x-direction, which is then used in the blockiness estimation in equation 3.9. Here,

$$d_h(m, n) = I(m, n + 1) - I(m, n), \quad (3.7)$$

describes the differencing signal, where m and n represent pixel location for $m \in [1, M]$ and $n \in [1, N - 1]$, respectively. The horizontal difference signal, d_h , is calculated in the x-direction of the image, I . To be thorough the vertical equation is,

$$d_v(m, n) = I(m + 1, n) - I(m, n). \quad (3.8)$$

Equation 3.8 simply evaluates the difference signal, d_v , in the y-direction. These horizontal and vertical difference signals are used separately to calculate patch-based blockiness, B , in each direction; the horizontal metric, B_h , is found in equation 3.9,

$$B_h = \frac{1}{M(\lfloor N/8 \rfloor - 1)} \sum_{m=1}^M \sum_{n=1}^{\lfloor N/8 \rfloor - 1} |d_h(m, 8n)|. \quad (3.9)$$

In equation 3.9 the horizontal blockiness, B_h , is extracted using the d_h , in localized regions of length 8. Again, m and n represent the pixel location of an $M \times N$ image. This process is repeated

in the vertical direction,

$$B_v = \frac{1}{N(\lfloor M/8 \rfloor - 1)} \sum_{n=1}^N \sum_{m=1}^{\lfloor M/8 \rfloor - 1} |d_v(8m, n)|. \quad (3.10)$$

These equations output the horizontal and vertical difference information, which is then combined to output a single metric, B ,

$$IQ_5 = \frac{B_h + B_v}{2}. \quad (3.11)$$

Equation 3.11 shows that the blockiness, IQ_5 , is obtained by averaging B_h and B_v . Once the metric has been extracted from each image within the volume, the process is repeated for the same image at a downsampled scale of one then two; this approach, similar to the image sharpness feature, provides a multi-scale representation of the edge content within the image. An average is then taken across these scales and across the volume to report a single metric per volume.

Spatial Energy

The selected metric to describe the blur or dispersion of information throughout an image, is the spatial energy; this metric is also useful for detecting extraordinarily noisy images. The process of extracting the spatial energy feature begins with the generation of a spatial weighted map; this map contains the smallest weights in the center and the weights increase moving outward to fill each pixel in the 256x256 image. This is implemented to put emphasis on the energy in the background of the image, which is useful for counter-balancing the higher intensities found within the center or head region of the image. Figure 3.8, displays an example of a spatial weighted map, I_{map} .

This process of generating a spatial weighted map can be described by the following sequence of equations. First, an artificial image is created with column-wise pixels weighted by assigning a value corresponding to the vertical pixel location. Thus, the top row has a value of 1 and the

Figure 3.8: Spatial energy feature intermediate step: generation of spatial weighted map

bottom row has a value of 256. Equation 3.12 provides supplemental explanation of how to obtain this image referred to as I_{col} ,

$$I_{col}(m, n) = m, \quad m, n \in [1, 256]. \quad (3.12)$$

This operation is repeated in the horizontal direction, such that the row-wise pixels are assigned a value corresponding to the horizontal pixel location. Equation 3.13 explains how this artificial image, I_{row} , is created,

$$I_{row}(m, n) = n, \quad m, n \in [1, 256]. \quad (3.13)$$

The next step is to obtain the average column and row value with consideration of the original image,

$$\mu_{col} = \frac{1}{\sum_{m=1}^M \sum_{n=1}^N I(m, n)^2} \sum_{m=1}^M \sum_{n=1}^N [I(m, n) \cdot I_{col}(m, n)]^2, \quad (3.14)$$

where I refers to the original image, I_{col} refers to the column-wise weighted artificial image, and m

and n correspond to pixel location. Again, M and N are both normalized to 256. The normalization constant in front of the equation is the total energy of the original image. Naturally, this is repeated for the row-wise weighted image,

$$\mu_{row} = \frac{1}{\sum_{m=1}^M \sum_{n=1}^N I(m, n)^2} \sum_{m=1}^M \sum_{n=1}^N [I(m, n) \cdot I_{row}(m, n)]^2. \quad (3.15)$$

The last step of generating the spatial weighted map, I_{map} , is to combine the magnitude of the weighted horizontal and vertical artificial images, as seen in the following equation,

$$I_{map} = \sum_{m=1}^M \sum_{n=1}^N [(I_{col}(m, n) - \mu_{col})^2 + (I_{row}(m, n) - \mu_{row})^2]. \quad (3.16)$$

Again, the final produce can be observed in figure 3.8. Once the spatial weighted map is generated, the original image can be multiplied pixel-wise with the spatial map. Then a simple calculation of the energy can be accomplished by summing the squares of the weighted intensity values of the image. It should be noted that the image is normalized by the maximum intensity of the volume. The spatial energy calculation is

$$IQ_6 = \frac{1}{M \cdot N} \sum_{m=1}^M \sum_{n=1}^N [I(m, n) \cdot I_{map}(m, n)]^2, \quad (3.17)$$

where the original image, I , is point-wise multiplied by the spatial weighted map, I_{map} . After extracting this metric, the same downsampling process performed in the blockiness methodology is repeated for the spatial energy as well. Again, the average is taken across these scales and across the volume to report a single metric per volume.

3.2.3 Signal-to-Noise Ratio and Contrast-to-Noise Ratio

The justification for inclusion of the general image quality descriptors of SNR and CNR were provided in Section 2.3. The SNR and CNR values were extracted from each volume with the use of a binary brain-mask and CSF-mask to isolate the classes prior to calculations. These masks were multiplied pixel-wise by the original image to isolate each class. Figure 3.9b and 3.9c display an example of isolated brain and CSF classes from the original image 3.9a.

Figure 3.9: Isolation of brain-class and CSF-class via binary masks

The mean and standard deviation are extracted from each isolated class such that the ratios could be calculated; these classes are volumetric. Equation 3.18 explains how the SNR is calculated, with inspiration from Yan *et al.* [53]. For clarification of how the mean and standard deviation are obtained, refer to equations 3.3 and 3.4 in Section 3.2.1. Here, the SNR is calculated as follows:

$$IQ_7 = \frac{\mu_{brain}}{s_{brain}}, \quad (3.18)$$

where μ_{brain} represents the mean of the volume's brain class and s_{brain} represents the standard deviation of the brain class. Equation 3.19 explains how the CNR is calculated, as used in Metcalf

et al. [77],

$$IQ_8 = \frac{|\mu_{brain} - \mu_{csf}|}{\sqrt{s_{brain}^2 + s_{csf}^2}}. \quad (3.19)$$

In equation 3.19, the variables μ_{csf} and s_{csf} represent the mean and standard deviation for the volume's CSF class. Again, this only includes the pixels after the binary CSF mask was applied to the original image.

It should be noted that to quantify the most accurate SNR and CNR on a volume-basis, the middle 60 percent of images from the volume are considered. The theory behind this consideration is that the majority of important information within the scan is held where the brain becomes visibly apparent. Figure 3.10 displays a snapshot of checkpoints in the z-direction of a scan; the figures in sequential order represent the first slice, 20 percent depth, 50 percent depth, 80 percent depth, and the last slice. Figure 3.10b represents the point at which the brain becomes a visible class and information becomes important; this is where the inclusion of information begins. Figure 3.10d represents the point at which the brain class fades and the inclusion of information ends; information after this slice is less crucial to the scan.

Figure 3.10: Cross-sectional checkpoints of brain content in z-direction of volume

It was experimentally determined that this approximate range consistently represents the most important data within the scan.

3.2.4 Noise Field Features

Given the need to assess the noise field described in Section 2.4, it was decided that the noise characteristics of kurtosis and skewness describe the shape of each tissue class, quantitatively. Kurtosis is able to describe a Gaussian shape, relating the information in the tails relative to the peak of the distribution. The kurtosis is also referred to as the peakedness of a curve, and can best be described through figure 3.11. It can be understood that a higher kurtosis relates to a steeper curve (orange), referred to as leptokurtic, and a lower kurtosis relates to a flatter curve (yellow), referred to as platykurtic, while a Gaussian shaped curve is referred to as mesokurtic (blue) [78].

Figure 3.11: Kurtosis explained through an example

While kurtosis describes the shape of the distribution in the vertical direction, skewness is able to describe the asymmetry of the distribution by explaining which side of the distribution has the larger tail. Skewness can most easily be understood by comparing the mean, median and mode of a distribution. If the mean is greater than the median and the median is greater than the mode, then the distribution is right-tailed or positively skewed. If the mean is less than the median and the median is less than the mode for a distribution, then it is left-tailed or negatively. Figure ??, does an excellent job of visualizing this phenomenon [78].

In combination, these metrics can describe whether a noise distribution appears Gaussian or not, which is of most concern to a user attempting to use some sort of Gaussian mixture model. In addition to determining the level of agreement to a Gaussian distribution, the supplemental information of the weight in the tails and the sidedness of the distribution allows for thorough volume-wise comparison of the dataset.

Figure 3.12: Skewness explained through an example

The study of the noise of any given volume is accomplished by taking a histogram of one tissue class in isolation; for this application, the brain class is extracted using the brain-masks as explained in Section 3.1.2.2 [73]. The use of the brain class is unique to FLAIR MR because the brain class is unimodal. These shape-measuring metrics are calculated from a probability density function of the isolated class, p_x . The equation for kurtosis is calculated as follows [79]:

$$IQ_9 = \sum_{n=1}^N \frac{(p_x(n) - \mu_x)^4 / N}{s^4}. \quad (3.20)$$

In the equation above, $p_x(n)$ represents each probability in the probability density function, while N is the length of the probability density function. The μ_x represents the mean and s represents the standard deviation of the noise field or isolated brain class, x [79]. It should be noted, that this calculation of kurtosis does not shift/subtract by three, as some calculations do to provide positive and negative relation to the mesokurtic distribution, which in this case occurs when kurtosis is equal to three. The skewness is calculated as follows:

$$IQ_{10} = \sum_{n=1}^N \frac{(p_x(n) - \mu_x)^3 / N}{s^3}. \quad (3.21)$$

Equation 3.21, uses the same terms described above for equation 3.20 [79].

3.2.5 Motion Artifact Feature

Motion artifacts can be detected by evaluating each individual slice's background for repeated vertical edges in the horizontal direction, which is a common trait observed across the data mentioned in section 3.4 and recreated in experiments on images in the transaxial orientation [80]. These artifacts are commonly observed as ringing lines parallel to intense edges (like the skull) due to patient motion during acquisition [81] [80]. As mentioned in Chapter 2, the phase-encoding direction of the reconstruction of the brain-scan can affect the portrayal of the artifacts, but the concern of this work involves a horizontal phase-encoding direction resulting in the aforementioned characteristics of motion artifact [65] [66].

The feature extraction begins with a normalization of intensity by the maximum intensity present in the volume; then, an inversed head-mask is applied to the volume to isolate the background intensities. Then a gradient of the edge content (primarily vertical edges) in the horizontal direction is evaluated, and finally an energy metric is extracted from the resultant image to report a prevalence of motion artifact through horizontally repeated vertical edginess. Figure 3.13 summarizes these steps in a flowchart.

Figure 3.13: Flowchart of motion artifact evaluation pipeline

Once an image has been normalized and had the inverse head-mask applied, the edge content is extracted from the background. A few edge detection methods were considered for this application, including the Sobel operator, Prewitt operator, Canny operator, and a subband of a wavelet decomposition of the image. Deeper investigation ruled out the use of the Canny operator as it is sensitive to noise and requires a filter to smooth the image prior to isolating edge content, which could jeopardize the intensity of edge content in the background due to motion artifacts. The use of wavelet decomposition was ruled out as it was deemed too computationally intense for this application. The Sobel and Prewitt operators were both considered and in comparison of their performance, for this application, they were declared redundant. Some results are included in

section 4.1 to show their identical performance. For this reason the more commonly used Sobel operator was selected to describe the ringing artifact edges in the background of the image. An example of what this motion artifact and process looks like can be found in figure 3.14 and figure 3.15, where figure 3.14 displays an example of absence of artifact and figure 3.15 displays an example of severe motion artifact.

Figure 3.14: Edge content of image background: no motion artifact example

Figure 3.15: Edge content of image background: severe motion artifact example

The original images each have their background information isolated in figures 3.14b and 3.15b. This isolation of background information then has the Sobel operator applied such that the edge content of the background is presented in figures 3.14c and 3.15c. These last two images have their edge content represented with positive and negative integers. Zero (gray) represents no edge content, while black and white represent the most extreme edge content. These cases are examined under the same normalized scale. It can be understood that a more black/white edge representation correlates with a higher presence of vertical edges in the form of a ringing motion

artifact. Comparing figure 3.15c with extreme edge content to figure 3.14c with minimal edge content, it becomes clear that the ringing artifact presents itself through vertical edges. Simply, taking the energy of the vertical edge information present in the image will summarize the approximate severity of the motion artifact. Notably, the Sobel operator yields a gradient edge map which is able to capture the information that is not perfectly represented as a vertical line, which is ideal for this application because the ringing of the skull is not perfectly vertical. Equation 3.22 explains how the energy (of motion artifact), IQ_{11} , is taken from the gradient edge map of the image after background isolation (I_b), as follows:

$$IQ_{11} = \frac{1}{M \cdot N} \sum_{m=1}^M \sum_{n=1}^N I_b(m, n)^2, \quad (3.22)$$

where, m and n represent the pixel location of an $M \times N$ image. This energy of the vertical edges in the background of the image, IQ_{11} , is also normalized by to a standard sized image at 256x256. To quantify the approximate severity of motion artifact on a volume-basis, the middle 60 percent of images are considered, and the maximum value reported within this range is taken as the representative metric. The theory described above in Section 3.2.3 and supplemental explanation by figure 3.10 are shared for this approach. This algorithm begins evaluating at the point at which the brain becomes a visible class and information becomes important. The point at which the brain class fades is where the motion artifact algorithm ceases its evaluation.

Prior to the implementation of this reduced range, the motion artifact algorithm predicted a fail volume due to a minor motion artifact in the first few slices, while the remainder of the volume was of good-quality. Since, this occurrence was more common than anticipated, the range of data considered was refocused.

3.2.6 Total Image Quality Score

The image quality score, justified in Section 2.6, is a key descriptor of the image quality in a dataset and a necessary additional inclusion for this pipeline. This feature is generated as an alternative

measure designed to find outliers; it is extracted through the use of the z-score from each metric across the dataset and essentially summarizes the overall general quality of a volume. This inclusion of multiple IQA features allows the rating to be capable of detecting when any combination of artifacts are occurring simultaneously, in potentially more mild cases. Equation 3.23 explains how the z-score is obtained for any given feature:

$$Z_{kn} = 0.6745 \frac{IQ_{kn} - Med_n}{MAD_n}. \quad (3.23)$$

Equation 3.23 shows how the z-score (Z_{kn}) is obtained for any volume (k) and feature (n). The normalized z-score is calculated with respect to the feature's median (Med_n) and median absolute deviation (MAD_n), where

$$MAD_n = median(|IQ_{kn} - Med_n|).$$

This median-based z-score is used against the traditional mean-based z-score because it does not require a strong normality assumption. The z-score is then used in calculating the rating; the sum of the absolute value across metric's z-scores provides a single output: *likeliness-of-rejection* rating. This accumulation of z-score values reported by every volume with respect to feature can be indicative of cases where a volume appears distant from the average (or good quality) if a high value is reported. It should be understood that this rating is counter-intuitive as a high rating represents a poor-quality volume. An abnormal value for a feature will be accounted for by a larger magnitude of z-score for that specific feature, which then contributes to the overall quality score. For example, a volume that reports an abnormally low CNR and a high maximum intensity feature, both close to their respective cut-off thresholds but not outside the acceptable data range, would yield a high magnitude in total quality score through the sum of z-scores. The purpose for implementing this type of quality score is to generally describe image quality and specifically identify cases that report feature values that are abnormal, but were not necessarily rejected by one

specific feature. The calculation of the total image quality metric (IQ_T) is completed as follows:

$$IQ_T = \sum_{n=1}^N Z_{kn}. \quad (3.24)$$

This equation above displays an example calculation for a volume, k , as calculated by summing the z-score across N features (n). It should be noted that not all features are guaranteed to contribute positively, thereby requiring ablation study to determine which features should be included. This study is further explained in section 4.1. A higher score corresponds to a volume that is far from the standard of the dataset and it is deemed poor quality, while a lower score correlates with better quality volumes. The ratings are also generated with respect to manufacturer to avoid the influence of the MCE.

It should be noted that there are three features that are exceptions to the standard z-score calculation: SNR, CNR and the motion artifact feature. These features are one-sided and thus only one-side of the z-score should contribute value. For example, a high SNR and CNR should not be penalized in the total quality score the same way a low SNR and CNR should. Thus any values above the median do not contribute in the rating. The opposite is true for the motion artifact feature, as lower z-score values should not contribute to a higher z-score sum; the values below the median for this feature are nulled as well. The reason why the extremes on either end contribute to the total IQ score is because the absolute value of the z-scores are used.

3.3 Outlier Rejection Overview

Each volume is evaluated from three perspectives prior to determining whether it is good- or poor-quality. Every volume is tested by a metric-wise comparison for the first ten IQ metrics, then motion artifact and total image quality score thresholds are applied. Table 3.1 displays a summary of these twelve IQ metrics and their corresponding information with regard to IQA category, equation and outlier rejection method.

Image Quality Categories	Feature	Equation	Outlier Rejection Test
Intensity-based	IQ1	$IQ_1 = \max\{V(x, y, z)\}$	Two-sided
	IQ2	$IQ_2 = \sum_{x=1}^{N^*} x \cdot p(x) (3.3)$	Two-sided
	IQ3	$IQ_3 = \sqrt{\sum_{x=1}^{N^*} (x - \mu_{int})^2 \cdot p(x)} (3.4)$	Two-sided
Sharpness-based	IQ4	$IQ_4 = \frac{1}{M \cdot N} \sum_{k=1}^K \frac{IS_k}{\sigma_k} (3.6)$	Two-sided
	IQ5	$IQ_5 = \frac{B_h + B_v}{2} (3.11)$	Two-sided
	IQ6	$IQ_6 = \frac{1}{M \cdot N} \sum_{m=1}^M \sum_{n=1}^N [I(m, n) \cdot I_{map}(m, n)]^2 (3.17)$	Two-sided
SNR and CNR	IQ7	$IQ_7 = \frac{\mu_{brain}}{s_{brain}} (3.18)$	One-sided
	IQ8	$IQ_8 = \frac{ \mu_{brain} - \mu_{csf} }{\sqrt{s_{brain}^2 + s_{csf}^2}} (3.19)$	One-sided
Noise-based	IQ9	$IQ_9 = \sum_{n=1}^N \frac{(p_x(n) - \mu_x)^4 / N}{s^4} (3.20)$	Two-sided
	IQ10	$IQ_{10} = \sum_{n=1}^N \frac{(p_x(n) - \mu_x)^3 / N}{s^3} (3.21)$	Two-sided
Motion Artifact	IQ11	$IQ_{11} = \frac{1}{M \cdot N} \sum_{m=1}^M \sum_{n=1}^N I_b(m, n)^2 (3.22)$	One-sided*
Total IQ Score	QIT	$IQ_T = \sum_{n=1}^N Z_{kn} (3.24)$	One-sided*

Table 3.1: Summary of image quality features

For a volume to be deemed good-quality it must first be within the upper- and lower-limits for each two-sided feature, and on the correct side for each one-sided feature; secondly, the volume must yield a value below the adaptable thresholds for the motion artifact feature; and lastly, the volume must have a total image quality score below an optimized cutoff. If a volume reports a feature value outside of the defined limits for any singular feature, then the volume is deemed a *fail* and the reason for failure is documented in the outlier-rejection file. It is possible for a volume to be rejected for multiple reasons/features as well. Figure 3.16 displays a flowchart highlighting the general algorithms contained within the outlier rejection pipeline.

Figure 3.16: Flowchart of outlier rejection algorithms

After feature-extraction, a volume must first pass a feature-wise comparison to the average of the

dataset (or subdataset sorted by manufacturer). For this method, a non-parametric approach was taken due to the non-Gaussian distributions of many features; this implies the use of median-based approaches over mean-based approaches. If a volume's metrics are found to be 1.5 times the inter-quartile range away from their respective scanner's upper or lower quartile value, then the volume is deemed an outlier [82] [83]. Thus, an acceptable range is developed for each feature that is dataset (and manufacturer) dependent. A volume must register each feature value within this acceptable range in order to be deemed good-quality by feature-wise comparison. A volume produced by GE, for example, that registers a maximum intensity, (IQ_1), that is three-fold of the GE's average volume's maximum intensity (an extreme value), would result in that volume being labelled as an outlier. This method accounts for the MCE provided the scanner information is available with each volume. In the event that this information is not available, the dataset would be assessed as a whole. The theory behind this approach of utilizing the upper and lower cutoffs stems from the idea that 99.3 percent of data is contained within 1.5 times the inter-quartile range of the upper and lower whiskers of a normal dataset [82]; a non-normal dataset or a dataset that contains multiple outliers (right/left skew) will report more data outside of this range, rightfully so. The purpose of using the median-based approach with the quartiles and inter-quartile range is to avoid restriction of requiring the distribution to be perfectly normal. The traditional method of using the mean and standard deviations to define an outlier is subject to skewing the mean and standard deviation by outliers contributing to each of these values; whereas the median-based approach is less influenceable [84]. Figure 3.17 displays a labelled example of outlier rejection through the use of a boxplot [1]. In this figure we can visualize the median, lower quartile (Q1), upper quartile (Q3), the minimum limit and the maximum limit. It is now possible to understand how far a value must be reported from the median to be classified as an outlier.

Thus, statistically, any volumes that register values outside this range for any or multiple features are deemed outliers. The only exception we make to this rule is for the SNR, CNR and motion artifact and total image quality score features, as a higher value for the SNR and CNR features is indicative of better quality, while a lower value for the motion artifact and total IQ features are indicative of better quality. For these features, a one-sided rejection is applied to avoid accidentally rejecting data for the wrong reason. This logic was applied to all the features to determine whether

Figure 3.17: Boxplot example of outlier rejection; from [1].

they should be one-sided rejections or two-sided rejections. Now, while the lower limit of the SNR and CNR features use the method of 1.5 times the interquartile range less than the first quartile as a cutoff threshold, the motion artifact and total IQ score features use a slightly different approach.

The motion artifact and total IQ feature's upper limit threshold was experimentally evaluated using the 1.5 times the IQR greater than Q3 threshold and found many false negative labels. Since the volumes that reported values near these thresholds could be evaluated for visually obvious artifacts and could be correlated with labels, it was determined that these thresholds required an ROC analysis to determine the optimal cut-off threshold. This same strategy was not applied to the other one-sided features because a visual evaluation of quality for optimization of a threshold was not an option, as outliers by SNR and CNR are difficult to observe visually. Thus, the motion artifact feature and total IQ score required an improved cut-off threshold that used a similar median-based approach but instead of applying the suggested 1.5 value, a β parameter was introduced, as follows:

$$T_{cutoff} = Q3_{IQ_k} + \beta \cdot IQR_{IQ_k}.$$

This adaptable cut-off threshold is used to segregate the good- and poor-quality data and can be optimized per feature through ROC analysis, achieved in section 4.2 of Chapter 4.

Example format of Data Summary and Outlier Rejection Tables

This subsection describes the output provided by the quality control tool once the outlier rejection scripts have been executed. First, table 3.2 displays an example output IQA summary this pipeline would produce. Each feature has a value assigned to every volume, written alongside the file name; for the purpose of this example, dummy-file names were used. This example only shows information for 5 example files, when in reality the length of the file would be equal to the size of the dataset.

Filename	Max	Mean	Std	Isms	Blockiness	SpatialE	Kurtosis	Skewness	SNR	CNR	MotionArtifact	Rating
A	4669.0	475.0	552.4	102677	4.03514E-05	131.166	4.0478	0.08517	5.7236	2.2197	0.0017705	5.0554
B	4447.0	444.8	518.4	105534	3.93549E-05	114.056	4.1457	0.02519	5.7707	2.2267	0.0014765	4.1970
C	4734.0	452.5	530.0	101428	3.88183E-05	122.025	4.1356	0.08461	5.7173	2.2111	0.0015600	5.2454
D	4405.0	388.7	462.4	103649	3.71395E-05	93.424	4.3234	-0.18942	5.6981	2.2769	0.0015874	3.9176
E	4444.0	386.6	465.2	99593	3.5362E-05	121.926	4.1699	-0.19378	5.5113	2.2327	0.0015974	4.3402

Table 3.2: Example format of IQA summary spreadsheet

Table 3.3 does not pertain to the information found in table 3.2; instead, it is an arbitrary example of an outlier rejection spreadsheet provided by this pipeline. Each metric displays whether a volume is an outlier or not on a binarized scale. Only outliers are included in this spreadsheet.

Filename	Max	Mean	Std	ISms	Blockiness	SpatialE	Kurtosis	Skewness	SNR	CNR	MotionArtifact	Rating
Outlier1	0	0	0	0	0	0	0	0	0	1	0	0
Outlier2	0	0	0	0	0	0	1	0	0	0	0	0
Outlier3	1	1	1	0	0	1	0	0	0	0	0	1
Outlier4	0	0	1	0	0	0	0	0	1	0	1	1
Outlier5	0	0	0	0	0	0	0	0	0	0	1	0

Table 3.3: Example format of outlier rejection spreadsheet

Ability to overcome the MCE

Another notable design criteria that was included in the quality control tool is an implementation of multicenter generalization through the grouping of volumes with respect to scanner manufacturer. Evaluating a volume with respect to other volumes produced by the same scanner reduces the unintentional rejection of volumes because of the MCE. This concept is best explained through a generic example of a feature histogram that can be found in figure 3.18. The original distribution of

the feature is found in figure 3.18a. The distribution of each manufacturer is plotted in figure 3.18b using blue, green, and red to represent the volumes produced by the scanners of GE, Siemens, and Philips, respectively.

Figure 3.18: Outlier rejection: accounting for the MCE

Observation of figure 3.18b actually displays a multi-modal distribution of the histogram due to the MCE. Originally, the majority of the left side outliers for the Philips (red) distribution in figure 3.18b was contained within the whole-dataset distribution in figure 3.18a; thus, outliers for Philips containing low CNR may have fit into the regular distribution of another manufacturer, like Siemens (green). By separating the data by manufacturer, these hidden distributions are revealed and can lead to better segregation of the data by application of more efficient thresholds. Although figure 3.18b displays these distributions overlaid, the quality control pipeline would evaluate each of the three distributions separately.

It can be understood here that the MCE can drastically affect the results of not only this quality control pipeline but any algorithm attempting to generalize to multicenter data. This example feature was the most prominent example of the MCE displaying a multi-modal histogram distribution, which is why it was included as an example.

3.4 Validation Data and Metrics

To prove the ability of this image quality assessment and quality control pipeline, validation datasets and outcome measures are required. These are defined below in their respective sections and the results are found in the following chapter.

Validation Data

Validation datasets in the field of medical imaging and quality control are difficult to obtain because the methods to generate them are subjective and laborious. While there was significant amounts of data available to help design the pipeline and experimentally determine thresholds, much of this data was not originally labelled. This thesis utilized three key datasets in the design of the pipeline that comprise 4,720 volumes (approximately 225,000 images). The purpose of considering multiple datasets of multicenter nature to study thousands of volumes, is to increase statistical power and to have assured validation of the quality control pipeline. The three datasets considered are described below. Table 3.4 summarizes the information pertaining to each dataset presented below.

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset is a longitudinal multicenter observational study of healthy elders, mild cognitive impairment (MCI), and Alzheimer’s disease [33]. While ADNI composes multiple types of medical images, the subset of data this thesis is concerned with are the FLAIR MRI volumes. The FLAIR MRI dataset consists of 4263 volumes (approximately 200,000 images), which is the largest dataset considered in this thesis. This dataset consists of a 53:47 ratio of men to women with an age range of 73.5 ± 7.4 . Manual observation of most of these volumes gave insight to the range of quality the data possesses; since there are many volumes of poor image quality, this dataset can be used for validating the performance of the quality control pipeline. It is uncommon for a dataset to contain a substantial amount of poor quality volumes for study, which makes this dataset unique.

Another dataset used in this thesis is the Canadian Dementia Imaging Protocol (CDIP) dataset

produced exclusively by one individual; this data is referred to here as the Human Phantom Dataset [51]. This dataset is unique as it consists of one patient scanned at multiple centers across Canada 78 times; this is important because the *human phantom* mitigates the person-to-person anatomical variation typically found in a dataset, which has been studied and proven to be significant [85].

The final dataset used in this thesis is the Canadian Consortium on Neurodegeneration in Aging (CCNA), which also follows CDIP [51]. CCNA released The Comprehensive Assessment of Neurodegeneration and Dementia (COMPASS-ND) Study. The goal of COMPASS-ND is to study dementia in all its forms, including cases where there are multiple pathologies involved, such as Alzheimer’s disease and cerebrovascular disease, referred to as mixed dementia [86]. This study includes FLAIR MRI scans that can be used in validating and testing the quality control pipeline¹.

Dataset	Volumes	Images	TR (ms)	TE (ms)	TI (ms)	Multicenter	Reference
ADNI	4,263	204,000	9505 $\pm$ 870	106 $\pm$ 27	2434 $\pm$ 137	yes	[33]
Human Phantom	78	3,700	9000 $\pm$ 0	125 $\pm$ 6	2471 $\pm$ 80	yes	[51]
CCNA	379	18,000	9015 $\pm$ 119	124 $\pm$ 7	2478 $\pm$ 71	yes	[86]

Table 3.4: Summary of datasets and metadata

The data listed here in table 3.4 encompasses the data that was used in the designing and experimentation to arrive at the current iteration of this quality control pipeline. This data was not used for quantitative validation and testing of the pipeline; however, two subsets with binary (pass/fail) quality labels have been made available to test the performance of the quality control pipeline.

For consistency and repeatability, it was decided that a set of criteria that clearly define a pass/fail volume should be outlined. The criteria used to visually assess the quality of a volume included a checklist of the following statements:

- There is no motion artifact of any form present throughout the brain in any image within the volume.
- There are no significantly noisy images within the volume.
- There are no blank or unnatural artifacts in any image within the volume.

¹More information about CCNA and access to the data can be found at <http://ccna-ccnv.ca/>

- The volume did not cause any algorithm within the pipeline to fail.

If the volume satisfies all of these criteria, it was considered a pass; otherwise it was labelled a fail. The most important region of the image evaluated was the brain.

For the ADNI subset, a set of professionally developed quality labels were used as each volume's evaluation was performed by an expert in the field, Dr. Clifford Jack [87]. For this dataset four classes were used to describe the quality of any given volumes. Class 1 is described as good-quality, as a volume should have no artifacts present, and no coverage problems. [87]. Class 2 is described good-quality as well, however, some mild to moderate artifacts might be present but should have no significant impact on the images, artifacts can be local. Class 3 is described as poor-quality with moderate to severe artifact presence, and requires caution when running automated analysis. Class 4 is regarded as unusable, where artifacts are severe and classes may be indistinguishable. Of these labels, class 1 aligns with the criteria of good-quality; class 3 and 4 align with the criteria of poor-quality; and class 2 does not consistently present volumes of good- or poor-quality, so it is not used to avoid the use of controversial labels. To reiterate, the volumes that presented moderate and severe artifacts were deemed a *fail* totaling 607 volumes, while the volumes that were void of artifacts were deemed good quality, totaling 135 volumes [87]. These criteria align with the criteria listed above.

The next subset of labelled data to be generated using these criteria, was for done using the CCNA dataset; a subset of volumes were blindly evaluated by two experienced neuroimaging members of the Image Analysis in Medicine Lab (IAMLAB). The resulting labels were averaged to obtain a final set of labelled data; the Cohen's Kappa coefficient was calculated for each pair of raters [88]. For reference, this is a common labelling validation method as it measures the reliability of the agreement between raters by accounting for the possibility of agreement occurring by chance [88]. These labels yielded an 80 percent raw agreement and a Cohen's kappa coefficient of *substantial agreement*.²

The ADNI and CCNA subsets provided the opportunity to validate the quality control tool, as

²Substantial agreement implies that the raters' agreement of labelling is validated for this application [88].

good- and poor-quality cases could be isolated and studied manually. This combination of subsets yielded 820 labelled volumes (approximately 49,000 images), equally distributed across the three manufacturers of GE, Siemens and Philips; these volumes were selected to test the performance of the quality control tool. This information is summarized in table 3.5.

Additional validation datasets were generated for specific intentions of optimizing particular metrics that pertain to visible artifacts. The first dataset was sampled from professionally labelled ADNI data, but was further analyzed for presence or absence of motion artifact. The purpose of generating this subset was for the optimization of the motion artifact feature. The *pass* volumes were sampled from class 1 if the following criterion was satisfied: the volume does not present any image that displays any degree of motion artifact (ringing) in the brain matter of the scan. The fail volumes were sampled from Class 3 and 4 if the above criterion was not met. This clearly defined criterion allowed for pass/fail labels to be generated on 75 ADNI volumes equally distributed across manufacturers; 45 volumes were labelled *fail* and 30 volumes were labelled *pass*.

The second dataset was also sampled from professionally labelled ADNI data. The purpose of this subset of data was for the optimization of the total IQ score. Random samples were taken from class 1 to generate good-quality examples and random samples were taken from class 3 and 4 to represent poor-quality examples. This subset of 75 ADNI volumes had an equal distribution across manufacturers; 45 volumes were labelled *fail* and 30 volumes were labelled *pass*. Table 3.5 summarizes the information for these data subsets.

Dataset	Volumes	Pass	Fail	Labelled by	Number of Raters	Purpose
ADNI	742	135	607	Dr. Clifford Jack	1	Total pipeline validation
CCNA	78	44	34	IAMLab	2	Total pipeline validation
ADNI subset	75	30	45	Dr. Clifford Jack	1	Total IQ feature validation
ADNI subset	75	30	45	Dr. Clifford Jack	1	Motion feature validation

Table 3.5: Summary of datasets and metadata

Validation Metrics

The validation measures are included to summarize the ability of the quality control tool to depict good- and poor-quality. The selected performance metrics surround the output of the confusion matrix. Here, a true positive (TP) is a volume that is detected by the pipeline as a *fail* and ground truth labelled as a *fail*; a false positive (FP) is a volume detected a *fail* but ground truth labelled as a *pass*; a true negative (TN) is a volume detected as a *pass* and ground truth labelled as a *pass*; and a false negative (FN) is a volume detected as a *pass* but ground truth labelled as a *fail*. Table 3.6 summarizes this information in an example confusion matrix. These metrics can also be displayed as percentages of the total number of volumes considered.

Dataset		Ground Truth Labels	
		Pass	Fail
Quality Control Tool	Pass	TN	FN
	Fail	FP	TP

Table 3.6: Example confusion matrix

The performance measures that utilize this information include the sensitivity (recall), specificity, precision/positive predictive value (PPV), and negative predictive value (NPV). The following equations describe how each measure of success is calculated using true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN):

$$Sensitivity = \frac{TP}{TP + FN}, \quad (3.25)$$

$$Specificity = \frac{TN}{TN + FP}. \quad (3.26)$$

Equations 3.25 and 3.26 speak to the true positive rate and true negative rate, respectively. Two additional measures of positive predictive and negative predictive values were also included in this

evaluation:

$$PPV = \frac{TP}{TP + FP}, \quad (3.27)$$

$$NPV = \frac{TN}{TN + FN}. \quad (3.28)$$

Also included is the F1 score, as it uses the precision and recall to create an additional accuracy measure:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \quad (3.29)$$

Chapter 4

Results

The results section of this thesis is categorized into two major approaches. The first approach is to prove that the quality control pipeline is successful through visual analysis and examples. The second approach is to validate the quality control pipeline through quantitative analysis. These objectives are achieved for every IQA category individually and for the entire pipeline.

4.1 Visual Validation

The visual analysis of the outliers detected by this quality control pipeline is organized into examples for each IQA category. The purpose of mapping the rejected volumes back to each IQA category it was rejected by is to ensure that a volume is rejected for the correct reason. This section includes examples of rejections for the intensity-based features, sharpness-based features, noise-based features, motion artifact features and the total image quality score. For this application the ADNI dataset is used to generate these examples.

In order to appreciate the severity of the artifacts presented across each IQA category, a comparative good-quality example volume is provided. Figure [4.1](#) displays this volume from the ADNI dataset.

Figure 4.1: Example of a good-quality volume

4.1.1 Intensity-based Features

The intensity-based metrics are particularly useful in detecting extreme pixel values, abnormal ranges of intensities and corrupt data. These types of outliers, although potentially disruptive for algorithms, may not always be visually obvious; however, an example that can be understood through subjective visual analysis is a case where the maximum intensity is found to be significantly larger than the rest of the data. This type of artifact can be especially an issue for automated processes that rely on the common technique of normalization by maximum intensity, or any methods that depend on reliable intensity ranges and distributions. This example occurs in cases where data is corrupted and pixels are assigned extreme intensity values as seen in figure 4.2a. Figure 4.2b is another example of corrupt data that was detected by extreme pixel intensity values; this image displays a streak of high intensity pixels. Another visually obvious example is portrayed through an extraordinarily noisy volume; although the noise is consistent throughout the image in figure 4.2c, it can best be observed within the skull region via the random variation of intensities. This specific example was detected using the standard deviation feature that detected an unusually high value for the ADNI dataset.

Figure 4.2: Outlier case-examples detected by intensity-based features

These examples in figure 4.2 were handpicked to represent just a few of the visually-obvious extreme outliers detected by the intensity-based metrics from the ADNI dataset.

Further validation is done through the visual inspection of labelled ADNI volumes. A random

selection of outliers detected by the intensity-based features can be visually compared against the good-quality volume in figure 4.1. Figure 4.3 displays outliers detected by the maximum intensity feature to represent the most visual outliers for the intensity based metrics. Figures 4.3a to 4.3c are found as outliers from GE scanners. Figures 4.3d to 4.3f are outliers from Philips scanners. Figures 4.3g to 4.3i are outliers from Siemens scanners.

Figure 4.3: Outlier case-examples detected by intensity-based features: ADNI

The images in figure 4.3 display examples of extreme intensities detected by the intensity-based features. By comparing these images to the images in figure 4.1 that represent a good-quality volume, it can be seen that these images contain abnormal bright spots or streaks. These types of visual outliers, could quantitatively affect automated algorithms, and ultimately skew results and conclusions of studies that involve them. It is important that these types of outliers are flagged within a dataset.

4.1.2 Sharpness-based Features

The sharpness-based metrics are useful in detecting extraordinarily sharp images and extremely blurred or noisy images where information is unusually dispersed; again, most of the outliers detected by these metrics are not easily visible on a grayscale image. Some visual examples of outliers detected by the sharpness/blur-based features are provided in figure 4.4.

Figure 4.4: Outlier case-examples detected by sharpness/blur-based features

The first edge-based example was detected by the image sharpness and the blockiness measures, as an unusually sharp edge was introduced in the image seen in figure 4.4a. This case occurs when half of a slice is misaligned, presenting a significant edge and causing an abnormally high value for each edge metric. An example of a case detected by a high spatial energy feature value and low sharpness is found in figure 4.4b. This image displays a poorly defined ventricle/brain

border, indicative of a blurred image. In many cases, this type of blurred (non-sharp) image is not perceivable to the human eye when evaluating the image as a whole; however, quantitatively, it can be accurately measured and flagged when extremely low sharpness levels present themselves. The spatial energy metric can also detect unusual information near the extremity of the image, like in cases where the image is cutoff on the bottom or top; this type of image or volume could skew conclusions within a study or cause algorithms to fail by presenting such unusual positioning of the head.

Further validation through visual analysis is provided for the sharpness-based features, similar to the intensity-based metrics. Figure 4.5 displays the most visual examples of outliers detected by the image sharpness metric. Figures 4.5a to 4.5c are found as outliers from GE scanners. Figures 4.5d to 4.5f are outliers from Philips scanners. Figures 4.5g to 4.5i are outliers from Siemens scanners.

Comparing the images in figure 4.5 to the images of the good-quality example volume in figure 4.1, it can be seen that these images appear abnormal. The outliers in figures 4.5a, 4.5b, 4.5c, and 4.5g were selected for presenting significantly lower sharpness/blockiness measures and higher spatial energy measures. Particular attention should be paid to the ventricle borders with the brain. Although some images may not appear qualitatively as an outlier, they are still outliers quantitatively. Alternatively the other images presented significant edge content due to a misalignment error. Either of these types of extremes can be problematic for automated processes that rely on consistent and similar data.

Figure 4.5: Outlier case-examples detected by sharpness-based features: ADNI

4.1.3 Noise-based Features

The most common form of an unusual distribution of the noise field occurs in bias-field artifacts. This common artifact yields an image where the intensities within the same tissue class are represented by a gradient of values. Figure 4.6 shows an example of an image deemed as an outlier by the kurtosis metric. This image appears abnormally darker in the brain region in the middle drifting upward toward the right of the brain class, and has a gradient that increases brightness in regions further toward the bottom of the brain class; this is an example of a bias-field artifact.

Figure 4.6: Outlier case example detected by noise field: bias-field

Figure 4.6 displays a strong bias field that can be detected using the features of kurtosis and skewness after isolating the noise field.

A measure of validation is provided through figure 4.7, as each image is an example of an outlier detected by the noise-based features from the ADNI dataset. Figures 4.7a to 4.7c are found as outliers from GE scanners. Figures 4.7d to 4.7f are outliers from Philips scanners. Figures 4.7g to 4.7i are outliers from Siemens scanners.

Again, compared to the good-quality example volume in figure 4.1, it can be understood how this common bias-field artifact can be observed visually. Quantitatively, the kurtosis and skewness features combine to find these types of abnormal noise distributions.

Figure 4.7: Outlier case-examples detected by noise-based features: ADNI

4.1.4 SNR and CNR

The SNR and CNR are excellent descriptors of general image quality. For this application the SNR and CNR are concerned with the brain class and the contrast the brain class has with the CSF class. Given the nature of these metrics, they are one-sided rejections only. Of course the outliers are detected from volumes reporting ratios that are too low. Figure 4.8a provides an example of a low CNR through the poor contrast between the brain class and the CSF within the ventricles. This case is further studied through figures 4.8b and 4.8c as they isolate the brain and CSF classes respectively. A data marker has been selected to represent intensities from each class to show how similar these two classes actually are. Again, the CNR metric compares the contrast between the brain class intensities and the CSF class intensities; thus, when a volume has difficulty differentiating between these two classes, it presents an outlier.

Figure 4.8: Outlier case-examples detected by CNR

These types of outliers presented by the SNR and CNR metric are not always visually observable as they rely on quantitative comparison of two classes. While this contrast might not be perceivable from the perspective of a whole image, the quantitative measure of each class in isolation provides confidence in rejection of poor-quality volumes.

4.1.5 Motion Artifact Feature

The motion artifact category is actually the most common type of artifact found in the ADNI dataset. An example of detected motion artifact within a dataset can be seen in figure 4.9a. The image in figure 4.9a clearly presents severe motion artifact that can be detected throughout the image, including the background. These significant motion artifacts are highlighted using the motion artifact detection algorithm within the quality control tool pipeline. An example of a less intense case is found in figure 4.9b. This example is supplemented with an isolation of the background using an inversed head-mask to display the ringing artifact. To be clear, figure 4.9b displays the original image to which figure 4.9c shows the isolated background. The ringing of the motion artifact can be seen clearly in figure 4.9c; the visibility improves in this image because the grayscale contrast only spans the range of intensities in the background class.

Figure 4.9: Outlier case-examples detected by motion artifact

Figure 4.10 provides a supplemental validation of this feature through visual examples from the ADNI dataset. Figures 4.10a to 4.10c are outliers from GE scanners. Figures 4.10d to 4.10f are outliers from Philips scanners. Figures 4.10g to 4.10i are outliers from Siemens scanners.

Figure 4.10: Outlier case-examples detected by motion artifact feature: ADNI

4.1.6 Total Image Quality Score

Since the total image quality score (IQ_T) encompasses multiple IQA features, it is difficult to describe one type of outlier it can consistently detect. Despite the unpredictable nature of outliers detected by the IQ_T , some examples have been provided in figure 4.11. The majority of these types of outliers are visually represented by mild motion artifact from volumes that presented a motion artifact feature value that was near the cutoff threshold but not above it; this z-score value contributed to the higher rating and thus a rejection by the (IQ_T) threshold.

The majority of these images in figure 4.11 represent the aforementioned case of mild motion artifact. By observing each image closely, the vertical ringing lines of motion artifact can be seen. For the ADNI dataset, this is the most common rejection by the total IQ score when only considering rejections that were done solely by this metric. In other words, these cases would have been missed without this IQ_T .

Figure 4.11f is an outlier that yielded an extreme maximum intensity, but was just below the cutoff threshold for the maximum intensity feature. These types of detection are important because it allows the total image quality score to reduce the number of false negative cases this pipeline would otherwise produce.

Figure 4.11: Outlier case-examples detected by total image quality score: ADNI

4.2 Quantitative Validation

This section outlines the capability of the quality control pipeline when using all quality metrics and the total IQ score together from a quantitative performance perspective. The previous section that contained visual analysis of outliers, retrieved all cases from volumes that reported feature values significantly far from the feature average (median). This concept of outlier rejection is best conveyed through feature histograms with marked upper and lower limits to show what portions of the data are rejected. The feature histograms for the ADNI dataset can be found, grouped by IQA category, below.

The first features displayed in figure 4.12 are the intensity-based features plotted with respect to manufacturer. Figure 4.13 displays the blur-based features. Then, figure 4.14 provides the kurtosis and skewness feature histograms. SNR and CNR feature histograms are found in figure 4.15. Lastly, the motion artifact and total IQ score histograms are included in figure 4.16. These last two metrics are plotted with respect to the previously discussed upper cutoff threshold.

It should be noted that these IQA metrics are plotted with respect to manufacturer to mitigate the effects of the MCE. The blue histograms represent General Electric scanners, red histograms represent Philips scanners, and green histograms represent Siemens scanners. The vertical bars represent the cutoff thresholds.

(a) Max: GE

(b) Max: Philips

(c) Max: Siemens

(d) Mean: GE

(e) Mean: Philips

(f) Mean: Siemens

(g) Std: GE

(h) Std: Philips

(i) Std: Siemens

Figure 4.12: Intensity-based feature histograms with cut-off thresholds for ADNI data by manufacturer

(a) Sharpness: GE

(b) Sharpness: Philips

(c) Sharpness: Siemens

(d) Blockiness: GE

(e) Blockiness: Philips

(f) Blockiness: Siemens

(g) Spatial: GE

(h) Spatial: Philips

(i) Spatial: Siemens

Figure 4.13: Blur-based feature histograms with cut-off thresholds for ADNI data by manufacturer

Figure 4.14: Noise-based feature histograms with cut-off thresholds for ADNI data by manufacturer

Figure 4.15: SNR and CNR feature histograms with cut-off thresholds for ADNI data by manufacturer

(a) Motion Artifact: GE

(b) Motion Artifact: Philips

(c) Motion Artifact: Siemens

(d) Rating: GE

(e) Rating: Philips

(f) Rating: Siemens

Figure 4.16: Motion Artifact and total IQ score histograms with cut-off thresholds for ADNI data by manufacturer

To quantify how successful the quality control pipeline is at detecting outliers, the ADNI validation dataset described in Chapter 3 from table 3.5 was used. The confusion matrix summary is combined with the performance metrics below to describe the results of the quality control pipeline in comparison to the labelled ground truth. The metrics of sensitivity, specificity, PPV and NPV are expressed as percentages. The ADNI dataset (N = 742) utilized the labels (607 fail and 135 pass) provided through Dr. Clifford Jack and the results can be found in table ??.

	TP	FP	TN	FN	Sensitivity	Specificity	PPV	NPV	F1
QC Pipeline	77%	27%	73%	23%	77%	73%	93%	41%	0.82

Table 4.1: Quality Control Tool's performance results: ADNI

Most rejections by the quality control pipeline occur from volumes labelled as fail found by the motion artifact feature and/or the total IQ score. In fact the percentage of volumes rejected by the intensity based features, sharpness-based features, SNR and CNR, noise-based features, motion artifact and total IQ score were: 8%, 1%, 1%, 3%, 69%, and 58%, respectively. It should be understood that a volume could be rejected for multiple reasons. Thus, adjusting the optimized thresholds to be lower, would cause a higher sensitivity but potentially a lower specificity. For this application, the receiver operator characteristic (ROC) analysis optimized the cutoff thresholds on training subsets of data for the motion artifact feature and total IQ score.

Receiver Operator Characteristic Analysis

To determine the optimal performance and prove that the quality control pipeline can distinguish between good- and poor-quality, a ROC analysis was performed for the two most influential features involved in outlier rejection. The motion artifact feature and the total IQ score feature have adaptable linear thresholds that determine good- and poor-quality labels. It is important to ensure that these thresholds are optimized such that the total pipeline can operate at peak performance. It should also be known that these are the only two features that have labels specific to their direct measurements, which is the reason why these features can be optimized.

First the motion artifact feature analysis was performed on a subset of the ADNI validation dataset

that contains 75 labelled volumes, produced by Dr. Clifford Jack. For this ROC analysis, the 75 volumes were isolated with respect to manufacturers of General Electric, Philips and Siemens (25 each) to generate the optimal threshold that generalizes across the MCE. Figure 4.17 summarizes the results performed for the ADNI validation dataset in varying the β parameter in the adaptive cutoff threshold. Recall the equation to be:

$$T_{cutoff} = Q3_{IQ_k} + \beta \cdot IQR_{IQ_k}.$$

In this case, $k = 11$ to represent the motion artifact feature, and $\beta \in [-1, 1]$. Of course Q3 is the third quartile and IQR represents the inter-quartile range, both for the same respective feature ($k = 11$).

Figure 4.17: ROC Analysis for motion artifact feature cutoff threshold optimization: ADNI

It was found that for the area under curve (AUC) for this analysis was 0.99. The optimal performance for this pipeline on the ADNI dataset found the adaptive threshold to require $\beta = -0.2$, where the sensitivity is approximately 96 percent and the specificity is approximately 100 percent. These results are excellent for the application to the subset of data that is differentiating between excellent quality volumes and moderate-to-severe motion artifact affected volumes.

Another ROC analysis was performed to determine the optimal cutoff threshold for the total image quality score feature. Again, the objective is to optimize the cutoff threshold that utilizes the

equation:

$$T_{cutoff} = Q3_{IQ_k} + \beta \cdot IQR_{IQ_k}.$$

In this equation $k = T$ to represent the total IQ score, and the parameter to optimize is again, β ($\beta \in [-2, 2]$). Again, the ADNI validation dataset is used for this analysis, except this subset includes 75 labelled volumes with varying artifacts for the analysis of the total IQ score. These 75 volumes are equally distributed across the aforementioned manufacturers and contain both examples of good- and poor-quality. Figure 4.18 displays the results of the ROC analysis.

Figure 4.18: ROC Analysis for total image quality score cutoff threshold optimization: ADNI

The AUC for this analysis was found to be 0.83. The optimal performance for this feature on this particular dataset found the adaptive threshold to require $\beta = -0.1$. The optimal sensitivity and specificity were found to be approximately 0.62 and 0.87, respectively. The philosophy behind this selection was to allow for more false negatives through a lower sensitivity, such that other features could detect them instead of introducing more false positives that cannot be reversed.

The last piece of information to digest alongside this analysis is the overall quality of the ADNI dataset, which is exceptionally poor. The typical clinical dataset, of which the tool wishes to generalize to, should not yield such a common presence of motion artifact, thus the parameter, although optimized for ADNI, should be tested to ensure that although trained on an abnormal dataset, it still yields the ability to generalize.

Generalization to other datasets

The ability of the quality control pipeline must be tested on a different dataset, like the CCNA validation dataset described in section 3.4. The β parameter that dictates the cutoff threshold, which was optimized for the ADNI validation dataset, was applied on the CCNA dataset. The results yielded a sensitivity of 97 percent but a specificity of 66 percent. In analyzing why the sensitivity is so high yet the specificity is so low, it became clear that the threshold was too low for the CCNA dataset and it was including too many false positives.

The CCNA validation dataset had an ROC analysis performed to determine an optimal threshold, such that the values/results from ADNI and CCNA could be compared. The results are found in figure 4.19 below.

Figure 4.19: ROC Analysis CCNA

The AUC for this analysis was 0.89, implying that the CCNA volumes were more easily separated into good- and poor-quality. The optimal β parameter to determine the cutoff thresholds was determined to be 0.9 for both motion artifact and total image quality score, as they were optimized simultaneously. The performance measures completed for the ADNI validation dataset was repeated for the CCNA dataset ($N = 78$) after optimization. The results can be found in table 4.2.

	TP	FP	TN	FN	Sensitivity	Specificity	PPV	NPV	F1
Total Pipeline	88%	18%	82%	12%	88%	84%	79%	90%	0.85

Table 4.2: Quality Control Tool's performance results: CCNA

For the CCNA dataset, the percentage of volumes rejected by the intensity based features, sharpness-based features, SNR and CNR, noise-based features, motion artifact and total IQ score were: 18%, 24%, 0%, 44%, 29%, and 53%, respectively. Again, it should be understood that a volume could be rejected for multiple reasons. This composition of the dataset is vastly different from that of ADNI, where most of the rejections were due to motion artifact.

In the investigation of the ADNI dataset, and by evaluation of the labels provided it became clear that the average quality of a FLAIR MRI ADNI volume was abnormally poor. Investigation of the CCNA dataset yielded an average volume quality of *good* instead of *poor*. Also, the most common artifact for ADNI was a motion artifact while noise-based artifacts were most common in CCNA. Since the CCNA dataset exhibited a more common representation of a typical clinical dataset, in terms of general quality and what this tool might be used for, the finalized adaptive threshold parameter β is suggested to be approximately 0.5 for a dataset of decent quality.

The difficulty in selecting an optimal threshold that can generalize across every dataset is a common challenge for these types of quality control pipelines; Mortamet *et al.* explained that, despite excellent feature engineering, an optimal threshold can vary from dataset to dataset, regardless of adaptability [35].

4.2.1 Quantitative Validation by IQA Category

This section includes a validation study performed for each individual IQA aspect to prove that each feature is capable of measuring what it intends to measure. It is also important that these features can describe common distortions and artifacts that arise within FLAIR MRI. Each feature is validated using either a subset of the ADNI labelled data or a known synthetic variation to that data. Similar techniques of evaluating feature correlation with artificial distortions have been executed in the works of Mortamet [35], Jang [40] and Osadebey [48] [46] [47].

Intensity-based Features

The intensity based metrics are validated by extracting each feature from the original data, and two linearly scaled versions of the original data. In this case the original data is 135 ADNI volumes labelled as excellent quality (Class 1). One dataset is generated by scaling the original data by a factor of two (labelled scale 1), and another dataset is generated by scaling the original data by a factor of three (labelled scale 2). The purpose of generating these datasets is to have a quantitatively known outcome that these features will attempt to describe.

An example image of the original data is seen in figure 4.20a; an example of the same data scaled by a factor of two is found in figure 4.20b; and an example of the original data scaled by a factor of three is found in figure 4.20c. These images are displayed on a consistent grayscale range to highlight the differences of intensities values between datasets.

Figure 4.20: Intensity scaled examples of data for validation of intensity-based features

Figure 4.21 displays the results of how the maximum, mean and standard deviation of intensities or IQ_1 , IQ_2 , and IQ_3 describe each of the three datasets: original, scale 1 and scale 2. As anticipated, each of the intensity-based features increased linearly with scale, exemplifying ability to describe the intensities within a volume and a dataset.

Additionally, these features are compared on the Human Phantom dataset before and after intensity standardization. Intensity standardization maps each tissue class to a specific and consistent range. It can be expected that the standardized data will yield much more restricted results for the intensity

Figure 4.21: Intensity-based features correlate with intensity scaling validation

based metrics.

Observing figure 4.22, it is clear that the standardized volumes have less variation, as expected, and thus a reduced MCE for the human phantom dataset. It should be noted that the y-axis scale represents the z-score of these metrics.

Figure 4.22: Intensity-based metrics validation: features extracted from original and standardized data

The maximum intensity metric in figure 4.22a displays a lower range of values with the standardized data, which was expected, again, considering that the original intensities are mapped to a predictable standardized range [34]. Similar results were observed in figures 4.22b and 4.22c for the mean and standard deviation; this proves that these metrics are truly intensity based and are measuring the maximum, mean and spread of the intensities present within the dataset.

Sharpness-based Features

The sharpness-based metrics are validated by extracting each feature from the original data, and two blurred datasets created from the original data. Again, in this validation study, the original data is 135 ADNI volumes labelled as excellent quality (Class 1). The first artificially blurred dataset is generated by applying a Gaussian blur of standard-deviation 1 with kernel size 5×5 (labelled level 1). The second dataset is generated by applying a Gaussian blur of standard-deviation 3 with the same kernel size (labelled level 2). The purpose of generating these datasets is to have a quantitatively known outcome that these features will attempt to describe. Similar approaches have been taken in Osadebey's works, where features show correlation with different artificial blurs [46] [47] [48]. This synthetic validation is common for the field of quality control in MRI.

An example image of the original data is seen in figure 4.23a; an example of the blur level 1 data is found in figure 4.23b; and an example of the blur level 2 data is found in figure 4.23c. Observing these images from left to right it can be seen that the blur increases from original to blur level 1 and again increases from blur level 1 to blur level 2.

Figure 4.23: Blurred examples of data for validation of sharpness-based features

Figure 4.24 displays the results of how the multi-scale image sharpness feature (IQ_4) is validated. This approach includes the comparison of the Sobel operator and the Prewitt operator, as mentioned in the methodology section. The results show that these two operators in this application were redundant; the Sobel operator was selected for the multi-scale image sharpness metric.

(a) Sharpness: Sobel

(b) Sharpness: Prewitt

Figure 4.24: Multi-scale image sharpness feature validation with varying levels of artificial blur

Observing figure 4.24 it can be seen that as the level of artificial blur increases, the sharpness measure decreases, as expected. This synthetic validation of the multi-scale image sharpness metric proves that this feature is able to quantify the edge content and conversely blur within an image or volume. Figure 4.25 displays the other two sharpness/blur feature of Blockiness and Spatial Energy. Again, these features are extracted from the original data, level 1 blurred data and level 2 blurred data to show a correlation and prove that they are able to describe either edge content or blur within a volume.

(a) Blockiness

(b) Spatial Energy

Figure 4.25: Sharpness-based features correlate with varying levels of artificial blur

The blockiness, and spatial energy, or IQ_5 , and IQ_6 describe each of the three datasets: original, blur level 1 and blur level 2. Naturally, the blockiness feature decreased as the blur level increased, proving its ability to describe the sharpness or edge content within a volume and a dataset. Con-

versely, the spatial energy or blur metric increased as the blur level increased, which was also expected.

SNR and CNR

The signal- and contrast-to-noise ratios are validated by comparing features values across an original dataset, and two datasets with different levels of artificial noise added. The original data is 135 ADNI volumes labelled as excellent quality (Class 1). The first noisy dataset is generated by adding zero-mean two-percent Gaussian noise (labelled level 1). The second noisy dataset is generated by adding zero-mean five-percent Gaussian noise (labelled level 2). The purpose of generating these datasets is to have a quantitatively known outcome that these features will attempt to describe. Similar approaches of artificial distortion through additive noise have been applied in the aforementioned works of Osadebey and also by Jang [40].

An example image of the original data is seen in figure 4.26a; an example of the noise level 1 data is found in figure 4.26b; and an example of the noise level 2 data is found in figure 4.26c. Observing these images from left to right it can be seen that the noise increases from original to noise level 1 and again increases from noise level 1 to noise level 2.

Figure 4.26: Additive noise examples of data for validation of SNR and CNR features

Figure 4.27 displays the results of how the SNR and CNR, or IQ_7 and IQ_8 describe each of the three datasets: original, noise level 1 and noise level 2. As expected, a higher level of noise present

in the image (or volume) decreases the ratios. The trend of a decreasing feature value with increase of noise level is true for both the SNR and CNR.

Figure 4.27: SNR and CNR features correlate with varying levels of Gaussian noise

Noise-based Features

The noise-based features are validated by evaluating the ability of these features to describe known noise distributions. The kurtosis and skewness are compared on original data, and on data that has had additive noise artificially induced. For this validation study, the original data is 135 ADNI volumes labelled as excellent quality (Class 1). The first noisy dataset is generated by adding zero-mean two-percent Gaussian noise (labelled level 1). The second noisy dataset is generated by adding zero-mean five-percent Gaussian noise (labelled level 2). The purpose of generating these datasets is to have a quantitatively known outcome that these features will attempt to describe. Again, this approach is common for the field of quality control of MRI.

An example image of the original data is seen in figure 4.31a; an example of the noise level 1 data is found in figure 4.31b; and an example of the noise level 2 data is found in figure 4.28c. Observing these images from left to right it can be seen that the noise increases from original to noise level 1 and again increases from noise level 1 to blur level 2.

Figure 4.29 displays the results of how the kurtosis and skewness, or IQ_9 and IQ_{10} , describe each of the three datasets: original, noise level 1 and noise level 2. Naturally, as the noise level increases,

(a) Original

(b) Level 1

(c) Level 2

Figure 4.28: Additive noise examples of data for validation of noise-based features features

the kurtosis and skewness begin to surround the Gaussian characteristics of a distribution, which is a kurtosis value of 3 and a skewness of 0 [79]. The expected values of the noise field for an ideal Gaussian distribution are approached as the additive noise begins to dominate the image. Should a volume exhibit a substantial amount of noise, or a noise distribution that may be indicative of poor-quality, these features will report values that are abnormal for said volume; thus, the volume will be flagged as an outlier.

(a) Kurtosis

(b) Skewness

Figure 4.29: Noise-based features accurately measure varying levels of Gaussian noise

Motion Artifact

The motion artifact feature is validated by evaluating its ability to discriminate between volumes with and without a ringing motion artifact present. This feature should correlate with the severity

of motion artifact, such that extreme cases yield higher values and mild cases yield lower values. To prove that this feature can indeed measure the presence of ringing artifact in a volume, a subset of data has been selected from the ADNI dataset. The labels of class 1 represent excellent quality where no motion artifact is present, class 2 represents mild-to-moderate motion artifact, and class 3 represent severe motion artifact present in the scan. There were 30 labelled volumes equally distributed across these classes that were selected from the professionally generated labels done by Dr. Clifford Jack. The motion artifact feature is extracted from each volume and plotted with respect to severity of motion artifact to show that the feature correlates with the artifact.

An example image of a volume labelled with no motion artifact is seen in figure 4.30a; an example of the moderate motion artifact is found in figure 4.30b; and an example of the severe motion artifact is found in figure 4.30c. Observing these images from left to right it can be seen that the severity of ringing artifact increases.

Figure 4.30: Varying intensity of ringing artifact examples of data for validation of motion artifact feature

Figure 4.31 displays the results of how the motion artifact feature or IQ_{11} discriminates between different levels of ringing artifact present in data. For this approach, as mentioned in Chapter 3, two edge detection methods were evaluated in the feature extraction pipeline. The Sobel operator and Prewitt operator were originally considered in analyzing the edge content of the background of an image to highlight ringing artifacts.

It can be observed that these two box-plot comparisons are extremely similar and were declared redundant, which led to the selection of the Sobel operator to be used in the motion artifact feature

Figure 4.31: Different edge detection applications in motion artifact feature validation by correlation with severity of artifact

extraction. Naturally, this feature increases with severity of artifact, proving ability to describe the presence of ringing artifact within a transaxial volume and a dataset.

Total Image Quality Score

Prior to the validation of the total IQ score, the optimization of features selected to contribute to the score is explained. As mentioned in the methodology, the z-score for each feature is first taken, then specific features are chosen to contribute their z-score magnitude to an overall summation for a given volume. To ensure that the IQ score is using positively contributing features, ablation studies were conducting evaluating the z-score magnitudes with respect to quality label. The results of plotting each feature's z-score magnitude by pass or *fail* label can be found in figure 4.32.

The selected features to compose the total IQ score include: maximum intensity, SNR, CNR, kurtosis, skewness, and motion artifact features. The purpose for selecting these features to be included in the total IQ score is due to the higher z-score amplitudes of these features amongst the poor-quality data in comparison to the good-quality data. The features that were ignored or not included had box-plots too similar between good- and poor-quality or did not yield higher z-score magnitudes for poor-quality data. Although the feature of motion artifact had the largest discrimination of z-score magnitude between good- and poor-quality, it was not given a higher weight to avoid redundancy in this application with the motion artifact feature that has already

Figure 4.32: Total IQ score feature study to evaluating contribution to quality discrimination

been optimized.

The total image quality score is then validated by evaluating its ability to discriminate between good- and poor-quality volumes. To prove that this IQ score can indeed discriminate between good-

and poor-quality, the labelled ADNI dataset is utilized. The labelled data is drawn from Class 1, Class 3 and Class 4, to represent good-quality, poor-quality and unusable quality, respectively. Recall that poor-quality volumes can be labelled so because of multiple different artifacts like extreme pixel intensities, bias field artifacts, motion artifacts, or even combinations of artifacts.

Figure 4.33 displays the results of how the total IQ score correlated with the quality of labelled data. The design of this total IQ score yields higher values for poor quality and lower values for better quality volumes. Observing the box plot of figure 4.33 proves the positive correlation with poor-quality. However, this feature does not perfectly discriminate between good- and poor-quality, as it can be observed that there is some overlap of quality scores between the Class 1 and Class 3 groups. For this application, an optimized linear threshold will still result in some false positive and false negative cases; this topic is further discussed in limitations and future work in Chapter 5.

Figure 4.33: Total IQ scores on good- and poor-quality volumes

Feature Analysis

Given the validation of each feature above, a retrospective analysis was done to evaluate the ability of each feature to detect outliers. Originally the statistical approach was taken for all features except for the motion artifact feature and the total IQ score, where the cutoff thresholds were defined to be one-sided or two-sided based on a range defined as 1.5 times the interquartile

range above or below Q3 and Q1, respectively. These thresholds were reevaluated, with the same approach as the total IQ score, using a ROC analysis except with the focus on maintaining the sensitivity to avoid introducing false positives while increasing the number of true positives.

Recall the intention of each feature is not necessarily to predict good or poor quality but to detect outliers from the dataset and describe the image quality characteristics of the dataset. That being said, there is opportunity to use labelled data to optimize the cutoff thresholds with respect to good and poor-quality data. This can help define a more optimal threshold to reject outliers for each feature. Figure 4.34 displays the results of the ROC analysis completed on the labelled ADNI data.

Figure 4.34: ROC analysis for each feature

It can be observed in figure 4.34, that none of these features are particularly good at discriminating between good- and poor-quality as each ROC curve essentially shows a line of equality; however, this analysis allows for each feature to have optimized cutoff thresholds to ultimately increase the true positives detected without major increase to the false positives detected. Table 4.3 summarizes how each feature improves the performance of the pipeline by contributing true positives or eliminating false positives, compared to the statistical threshold rejection approach (recall there are 607 fail volumes and 135 pass volumes).

Features	Max	Mean	Std	ISms	Blockiness	Spatial Energy	Kurtosis	Skewness	SNR	CNR
Beta	1.1	1.5	1.5	1.7	1.3	1.1	1.1	1.5	1.5	1.1
TP change	+3.3%	0%	0%	0%	+1.2%	+2.1%	+5.6%	0%	0%	+2.5%
FP change	0%	0%	0%	-1.5%	0%	0%	+0.7%	0%	0%	0%

Table 4.3: Performance updates after ROC analysis on each feature

It should be noted that these true positives and false negatives might be redundant amongst features, and the true change could only be reflected in the total pipeline performance. To quantify the change, the pipeline performance was reevaluated and the results can be found in table 4.4.

	TP	FP	TN	FN	Sensitivity	Specificity	PPV	NPV	F1
QC Before	77%	27%	73%	23%	77%	73%	93%	41%	0.82
QC After	79%	27%	73%	21%	79%	73%	93%	44%	0.82

Table 4.4: Quality Control Tool's performance results before and after individual feature optimization

It can be seen that the performance marginally increases through the addition of the true positives with minimal change to the false positives. Analysis of this result, reveals that many of the true positives that are now being detected by new metrics were redundant; however, some new true positives were identified with the new optimization approach. This results in an small increase in sensitivity but not much change overall for the pipeline. Although the optimization of each feature's threshold improved performance, the optimization of each individual threshold biases the approach to the composition of the labelled dataset and reduces generalization of the pipeline, which is a limitation to this quality control tool. If the cutoff thresholds are to be optimized instead of statistically defined, we have a marginal gain in performance in this case, but we subject this study to bias and reduce the ability to generalize.

4.3 False Positive Cases

The shortcomings and future work of this quality control pipeline are discussed in the limitations and future work sections in Chapter 5. However, some cases of failure for this pipeline have been recognized and should be presented alongside the performance of the tool. Mainly, some instances of mask-failure result in the wrongful misclassification of a good-quality volume as a fail.

In the case of the head-mask generation failure, the motion artifact algorithm recognizes a major edge, similar to the ringing produced by head motion during a scan when the head-mask applied to the original volume is too small and leaves portions of the skull in the resultant background image.

Figure 4.35: Example cases when motion artifact algorithm fails

Figure 4.35c displays this major edge as does figure 4.35f. In the case of figure 4.35c, the head-mask appears too small but in the case of figure 4.35f the head-mask has misclassified portions along the

left border, which is another example of how the motion artifact feature can produce a false positive. This shortcoming of the tool yielded a retrospective implementation into the mask generation that involves a dilation of the headmask to attempt to include the immediate surrounding area of the skull, but also to avoid these types of errors.

The other case is the brain-mask failure, which is less frequent than the head-mask failure, but it is still a limitation to the pipeline and potential source of false positives. In any case that the brain-mask fails, it will affect the SNR, CNR and noise field calculations by accidentally including information from an undesired class. An example of this case can be found in figure 4.36, where figure 4.36a represents the original image, figure 4.36b represents the binary brain-mask and figure 4.36c represents the resultant isolation of the brain class.

Figure 4.36: Example cases when brain extraction algorithm fails

Figure 4.36c displays the incorrect isolation of the brain class as the brain-mask in figure 4.36b has failed to accurately depict solely the pixels that compose the brain tissue. Visual observation of figure 4.36b draws attention to the incidental inclusion of the ventricles. The brain extraction algorithm may fail in cases where the images present particularly poor quality through motion- or bias field artifacts, or in cases where there are poorly defined borders between the brain and CSF. One reason why there might be a lack of concern for this particular type of error, is because many of these volumes should be classified as poor-quality anyway. Further discussion is provided in section 5.4.

Aside from the mask-malfunctions, occasionally a good-quality volume can be labelled as a *fail*

(FP), due to the nature of linear thresholds. Two particular features have been selected to display examples for further investigation (for the ADNI dataset). The first feature is the maximum intensity that occasionally produces outliers that are due to a bright spot found in the background of the image. This type of outlier could be problematic for quantitative algorithms; however, the bright spot is technically not a part of the head. While this issue may stem from the subjective determination of whether this volume is good- or poor-quality it could also be perceived as an issue. Figure 4.37a displays an example.

(a) Outlier: bright spot

(b) Outlier: Abnormally large skull

Figure 4.37: Example cases when outlier detection can fail

These types of scans might be good-quality, despite this minor bright spot, but they are labelled as a fail. Another case in which a good-quality volume might be predicted as a failure, is in the case of the spatial energy metric, where the skull appears abnormally large and close to the border of the image. Figure 4.37b displays an example. Despite the thickness of the skull and proximity to the border in figure 4.37b, this volume is of good-quality, yet the pipeline cannot differentiate this case from a typical outlier by the spatial energy. In cases, where an image is not presenting a high value due to a high detection of blur and the volume is not cutoff by the border of the image, this feature can present false positives. While the case in figure 4.37b, does present information close to the border, it is technically not an outlier. Further discussion of solutions or optimization of the linear threshold to account for these types of outliers is discussed in section 5.4 in Chapter 5.

Chapter 5

Conclusions

To conclude this thesis, this chapter begins with a summary of the major contributions. Then, the potential applications for this pipeline are described, prior to discussing the limitations of this work and the future directions for further research in the field of quality control and MRI.

5.1 Summary

This thesis presented a universal IQA summary and an outlier rejection algorithm referred to in conjunction as the quality control pipeline; this tool was explained to be unique as it was specifically designed for large multicenter FLAIR MRI datasets. Three novel image quality measures were introduced as a part of the image quality assessment approach: the multi-scale image sharpness metric, motion artifact feature, and total image quality score. Each of these metrics alongside every aspect of the quality control pipeline was justified, explained and validated.

The necessary background of neurodegenerative disease, FLAIR MRI and the common ground of studying WML, provided the motivation in the background of Chapter 1. The problem statement of big data and the multicenter effect was outlined as a challenge for the field to overcome. Further limitations of quality control in MRI were discussed through previous works in the field. Using this direction, the pipeline was outlined through the constraints in the proposed methods defining the tool to be: automated, no-reference, objective, unsupervised, invariable to the MCE, accurate in outlier rejection, and intelligent through IQA summary.

The fields of focus considered in Chapter 2 were addressed and justified for the inclusion of each category in the IQA tool. To recap, these categories include: intensity non-standardness, sharpness/blur, SNR and CNR, noise field, motion artifact, and total image quality score. Although throughout these sections, each aspect discussed preprocessing techniques that could form as solutions to combat poor quality, it is unrealistic for every dataset to undergo each possible preprocessing step. In theory, it would be ideal to perform intensity standardization, contrast enhancement, noise suppression, bias field reduction, and multiple artifact corrections; however, this brute-force method is cumbersome and computationally intense. This further justifies why it is advised to approach each dataset intelligently by understanding the data using the IQA summary, such that necessary preprocessing methods can be applied efficiently.

Chapter 3, the methodology, discussed the quality control pipeline from start to finish. First the preprocessing that included standardization and generation of class masks was explained. Then the IQA section discussed which metrics/features were selected to measure what each IQA category intended to, through equations and figures. The outlier rejection algorithm and adaptable thresholds were explained alongside a summary of the features that comprise each IQA category. The last discussion of Chapter 3 regarded the validation data and metrics. Each of the IQA categories and features, alongside the entire pipeline were then validated, visually and quantitatively, in the results section of Chapter 4. The tool was evaluated on its ability to label pass and fail volumes from a labelled subset of volumes. The pipeline reported success across different multicenter datasets. These results were discussed to be variable depending on the dataset and general quality of the dataset.

Given the motivation provided for this thesis, justification for the approach, validation of the pipeline and the success of the results, it can be understood that this quality control tool is prepared for clinical applications. The following section discusses the potential uses for this pipeline.

5.2 Applications

The potential applications for this quality control tool are categorized into four main categories at this point in time. First, the tool provides the ability to summarize and visualize the data from various aspects that describe image quality. Second, the tool can highlight or reject specific volumes within the dataset due to poor image quality. The third application is for longitudinal study of scanner performance as a scanner evaluation tool; this can be applied to any FLAIR MR scanners to measure the consistency and longevity of the equipment. The fourth application can be applied as a single volume assessment to be performed immediately after the scan is acquired; this application serves as a recall or rescan quality assurance step that could be introduced as the standard FLAIR MRI protocol. This section describes these applications in detail.

These first two applications have been discussed and explained throughout this thesis as this pipeline is designed to provide the IQA summary and outlier rejection for large FLAIR MRI datasets. The first use of this tool is intended to provide a data summary that includes an output of every aspect of IQA including the intensity-based metrics, sharpness/blur-based metrics, SNR, CNR, noise distribution, prevalence of motion artifact and total image quality score for each volume within the dataset. This data can be accompanied with a visualization in the form of box plots and feature plots of every metric to determine the characteristics of the data. The summary and visualization of the data will allow for a greater understanding for development of algorithms or preprocessing techniques to implement. The second application of this tool is the outlier rejection algorithm that can be applied generally to two cases: data issues and image quality. As previously mentioned, the outliers are defined by data that report values outside the upper and lower limits on a boxplot for each respective feature [83]. Together these applications will allow for cleaner FLAIR MRI datasets and an understanding of the nature of the data such that algorithms can be designed to overcome the MCE and ultimately, further the study of neurodegenerative disease.

The third application of evaluating scanner performance over time is a simple method to determine longevity of scanners or requirements for an upgrade. By continuously applying the quality control pipeline features to the output volumes from a scanner, this tool can detect when the quality of

images becomes poor or appears to change over time; this indication could allow for parameters to be adjusted or an upgrade to be made semi-automatically.

The fourth application of evaluating the quality of an MRI post-scan to detect rescan-cases is similarly feasible by incorporating a protocol to input a volume into this quality control pipeline immediately after a scan is complete. The quality control tool will be able to assess the FLAIR MR image quality of the scan and determine if it contains corrupt slices, motion artifact or appears as an outlier (using previous data). This protocol can automate radiologists' procedures of evaluating scan quality and manual detection of rescan cases, which in turn will allow for more scans to be studied by radiologists [89]. This protocol could also allow for immediate rescans of subjects by reporting a binary pass or fail to the technician conducting the scan. Ultimately, this application can reduce costs associated with radiologists manually evaluating quality of data and reduce costs for the hospital and patient in cases of rescans on separate dates [89].

While each of these applications seem feasible and ready to implement, there is always room for improvement in algorithm design. The following section discusses the limitations of this pipeline.

5.3 Limitations

Despite the success reported by this quality control tool, there are some limitations throughout the pipeline. The greatest limitation of this pipeline is that it requires intensity standardization and class mask generation; this opens up opportunity for a propagation of failure throughout the quality control pipeline, should the class mask generation perform with less than 100 percent accuracy.

One example of a failure, is the head-mask generation, which can lead to a false detection of an outlier; two specific examples can be found in figure 4.35 discussed in Chapter 4. These cases highlight the accidental inclusion of a major edge due to a head-mask that is too small or makes misclassifications along the border between the head and background. The head-mask generation algorithm fails and produces irregular borders occurs when the Otsu's threshold cannot decisively separate the foreground from the background; this case is common for intense bias-field artifacts

where the intensities of the volume are inhomogeneous. One reason why this case is less concerning is because these failures would often be classified as outliers by the noise field anyway as they exhibit intense bias-field artifacts.

The next limitation to discuss is also a mask generation, but it is for the brain-mask. An example is provided in Chapter 4 through figure 4.36. Essentially, in rare cases the brain-mask generation does not represent the entire brain class, and thus information from one-class (CSF) is included in another class (brain) which propagates error in calculations of SNR, CNR and noise-based metrics. These types of error, although uncommon, can affect outlier rejection and ultimately the accuracy of this quality control pipeline. Though, the failure of the mask for these examples, may be due to poor quality and thus the volume would be detected as an outlier anyway; however, this failure can still be concerning. A positive note for the brain-mask generation is found in the low frequency of occurrence of failures which will be further mitigated as the brain extraction algorithm is currently being improved [73]. While these cases in figures 4.35 and 4.36 are anomalies, the solution to further mitigating these types of errors, can be accomplished by improving the mask generation pipeline.

Another limitation that this pipeline has is the linear classification of data. While there is an ability to adjust a buffer depending on the optimized ROC analysis for the ADNI and CCNA datasets, it is difficult to increase the accuracy of the tool without sacrificing either false negatives or false positives. Some features, although discriminative between good- and poor-quality do not clearly separate the data and the linear threshold is not optimized per dataset. To combat this potential limitation, a figure could be generated that displays the metrics with respect to each volume, allowing the user to visually adjust the thresholds depending on the distribution of data; this additional step aims to increase accuracy for the tool by optimizing the threshold should there be any obvious multimodal distributions for any features prior to outlier rejection. Mortamet *et al.* experienced similar problems when pioneering a quality control method in structural MRI with novel IQA metrics [35]. It was concluded that more research could optimize this tool for a specific application, which is also true for this method in FLAIR MRI.

The last limitation mentioned for this quality control pipeline is from the perspective of computa-

tional efficiency; this tool has not been optimized, as it was designed in Matlab [90]. There are opportunities to improve on the efficiency of this pipeline by operating using another language or by allowing the data to be stored in any file format, instead of requiring the use-of/conversion-to .mat files.

5.4 Future Work

The future work for the IQA summary and outlier rejection of large FLAIR MRI datasets, has been roughly outlined from the preceding limitations sections. Essentially, to improve this pipeline a more effective mask-generation pipeline should be considered to report a higher accuracy and lower false positive rate. Another limitation addressed, is the computational efficiency of the pipeline as there is room for improvement through optimization in another language like Python or C/C++.

Although the motivation of this quality control tool is to provide information about the data to be used for intelligent and efficient preprocessing to overcome the MCE; the tool also must be used to depict outliers. The accuracy measures reported in table ??, leave room for improvement for outlier rejection and finding an optimized threshold for multiple datasets of varying quality. While deep learning approaches might strictly provide an output list of outliers, this may be something to consider in conjunction with the IQA summary, if it can improve the accuracy of the tool. Deep learning approaches surrounding applications on brain MRI have been gaining interest as they have been gradually outperforming state-of-the-art classical machine learning methods. [91].

Ultimately, the field of studying neurodegenerative disease through FLAIR MRI has room for improvement; primarily, the focus is attempting to design algorithms to study disease that can overcome the MCE. The use of the quality control tool presented in this thesis shows promise to help in this respect. Further solutions may include the segregation of the data based on manufacturer, model, or location of scan depending on the dataset. This approach introduces multiple subsets from the initial multicenter dataset, which may not be the ideal approach, but could serve as a solution to the MCE. Alternatively, with the generalizability of deep learning, perhaps a suitable model-based approach to methods of segmentation or classification specifically for FLAIR MR

volumes could be the ideal future work. Another perspective to overcoming or reducing the MCE is on the acquisition side, through implementation of protocols; this would reduce variability of data by controllable factors like the scanner parameter settings.

Appendices

Additional Results of testing on CCNA data

Applying the quality control pipeline to any dataset yields the rejection of outliers by the aforementioned three approaches. Outliers can be detected by reporting a feature value outside the upper and lower limits of the boxplot for that feature; by reporting a motion artifact feature above a cut-off threshold; or by reporting a total IQ score higher than a cut-off threshold. The results of applying the outlier rejection pipeline to a dataset can be observed by feature histograms, with the regions of acceptable data highlighted by vertical threshold-lines. Figure 1 displays the results for the 379 volumes within the CCNA dataset.

Figures 1a, 1b and 1c show the range of acceptable data for the metrics of maximum-, mean- and standard deviation of intensities by the values within the red-lines. Again, these vertical lines are 1.5 times the inter-quartile range below Q1 and above Q3 for each feature. The same format of information is presented in figures 1d, 1e and 1f for the sharpness/blur-based metrics. Figures 1g and 1h display the acceptable range for SNR and CNR, with the exception made for these two metrics specifically, that they are one-sided or only have a lower limit. The upper threshold for SNR and CNR has been eliminated to avoid accidentally rejecting high SNRs and CNRs, as these measures increase to indicate better quality. The noise field features and their acceptable ranges are also included in figures 1i and 1j. Figure 1k, displays the motion artifact feature, which had been previously explained to be a cut-off threshold approach, just like the total IQ score in figure 1l. By eliminating the data outside these vertical-lines in figure 1, the quality control pipeline eliminates most outliers and poor quality data; however, the pipeline is not perfect. The data that reports feature values close to any threshold, is the most difficult to label correctly; in rare cases, some poor-quality data reports values slightly less than the threshold and some good-quality data reports values slightly greater than the threshold.¹

A visual representation of the outliers detected by each of the IQA categories for the CCNA dataset are also included below. First, the outliers recognized by the intensity-based features are included in figure 2. A random selection of outliers are included below here that are equally obtained from each scanner vendor of GE, Siemens, and Philips. Figures 2a to 2c are found as outliers from GE

¹Further discussion of how this problem can be approached is discussed in Section 5.4.

Figure 1: Feature histograms with cut-off thresholds for CCNA data

scanners. Figures 2d to 2f are outliers from Philips scanners. Figures 2g to 2i are outliers from Siemens scanners.

Figure 2: Outlier case-examples detected by intensity-based features: CCNA

The next IQA category to provide an example of outlier detection is the noise field of an image. The common visual artifact for that this feature detects is the bias-field artifact. As this type of artifact is an example of a visually observable artifact, a random selection of artifacts detected by the noise-based features is provided from the CCNA dataset in figure 3. Again, the outliers in figure 3 display the common types of images that are found outside the thresholds of acceptable data for the noise-based features. Figures 3a to 3c are found as outliers from GE scanners. Figures 3d to 3f are outliers from Philips scanners. Figures 3g to 3i are outliers from Siemens scanners.

Figure 3: Outlier case-examples detected by noise-based features: CCNA

The next IQA category to provide examples of an outlier rejection is the motion artifact evaluation. This type of artifact is visually observable by the human-eye so some outlier cases have been included below. The outliers in figures 4 display the common types of images that are found above the threshold of acceptable data for the CCNA dataset. Figures 4a to 4c are found as outliers from GE scanners. Figures 4d to 4f are outliers from Philips scanners. Figures 3g to 3i are outliers from Siemens scanners. It should be noted that these CCNA outliers appear less intense than ADNI as seen in figure 4.10, but they are still problematic.

These are just some more visual examples of data issue cases that were highlighted by the quality control pipeline. The other types of data rejection are for data that simply reports metrics that are statistically recognized as an outlier [82]. Although some of these cases, may not be visually obvious, they are still significantly different from the average of the dataset, and thus, problematic. The total IQ score typically identifies the cases in which a volume is different from the dataset average, but not significant enough in any singular metric to be deemed an outlier. For example, a noisy image that is not highlighted by any metric during the outlier detection, may have been close to the cutoff threshold for multiple metrics; this case would be highlighted after evaluating the total IQ score. Again, the total IQ score is the sum of the absolute value of the z-score of each metric.

Figure 4: Outlier case-examples detected by motion artifact feature: CCNA

ROC Analysis Tables

The following tables are included to show the optimization of the β parameter for the cutoff thresholds developed for the motion artifact and total IQ scores on labelled ADNI subsets.

heightBeta	TP	FP	TN	FN	Sensitivity	Specificity
-1	42	17	13	3	0.933	0.433
-0.9	41	14	16	4	0.911	0.533
-0.8	40	13	17	5	0.889	0.567
-0.7	38	11	19	7	0.844	0.633
-0.6	38	11	19	7	0.844	0.633
-0.5	35	9	21	10	0.778	0.7
-0.4	34	7	23	11	0.756	0.767
-0.3	31	7	23	14	0.689	0.767
-0.2	28	4	26	17	0.622	0.867
-0.1	27	3	27	18	0.6	0.9
0	25	3	27	20	0.556	0.9
0.1	23	3	27	22	0.511	0.9
0.2	21	1	29	24	0.467	0.967
0.3	20	0	30	25	0.444	1
0.4	20	0	30	25	0.444	1
0.5	19	0	30	26	0.422	1
0.6	19	0	30	26	0.422	1
0.7	18	0	30	27	0.4	1
0.8	18	0	30	27	0.4	1
0.9	16	0	30	29	0.356	1
1	14	0	30	31	0.311	1

Table 1: Beta parameter optimization for total IQ score: ADNI subset

heightBeta	TP	TN	FP	FN	Sensitivity	Specificity
-1	45	15	15	0	1	0.5
-0.9	45	19	11	0	1	0.633
-0.8	45	21	9	0	1	0.7
-0.7	45	25	5	0	1	0.833
-0.6	45	28	2	0	1	0.933
-0.5	43	29	1	2	0.956	0.967
-0.4	43	29	1	2	0.956	0.967
-0.3	43	30	0	2	0.956	1
-0.2	41	30	0	4	0.911	1
-0.1	40	30	0	5	0.889	1
0	40	30	0	5	0.889	1
0.1	40	30	0	5	0.889	1
0.2	38	30	0	7	0.844	1
0.3	36	30	0	9	0.8	1
0.4	36	30	0	9	0.8	1
0.5	35	30	0	10	0.778	1
0.6	33	30	0	12	0.733	1
0.7	33	30	0	12	0.733	1
0.8	32	30	0	13	0.711	1
0.9	31	30	0	14	0.689	1
1	31	30	0	14	0.689	1

Table 2: Beta parameter optimization for motion artifact feature: ADNI subset

Bibliography

- [1] M. Galarnyk, “Understanding boxplots,” Oct 2019, Last accessed: Aug 2019. [Online]. Available: <https://towardsdatascience.com/understanding-boxplots-5e2df7bcd51>
- [2] C. Tator, G. Bray, and D. Morin, “The cbanch report-the burden of neurological diseases, disorders, and injuries in canada,” *Canadian Journal of Neurological Sciences*, vol. 34, no. 3, pp. 268–269, 2007.
- [3] A. Association *et al.*, “2018 Alzheimer’s disease facts and figures,” *Alzheimer’s & Dementia*, vol. 14, no. 3, pp. 367–429, 2018.
- [4] M. J. Prince, *World Alzheimer Report 2015: the global impact of dementia: an analysis of prevalence, incidence, cost and trends*. Alzheimer’s Disease International, 2015.
- [5] J. Cummings, C. Reiber, and P. Kumar, “The price of progress: Funding and financing Alzheimer’s disease drug development,” *Alzheimer’s & Dementia: Translational Research & Clinical Interventions*, vol. 4, pp. 330–343, 2018.
- [6] M. D. Hurd, P. Martorell, A. Delavande, K. J. Mullen, and K. M. Langa, “Monetary costs of dementia in the united states,” *New England Journal of Medicine*, vol. 368, no. 14, pp. 1326–1334, 2013.
- [7] R. Mayeux, “Epidemiology of neurodegeneration,” *Annual Review of Neuroscience*, vol. 26, no. 1, pp. 81–104, 2003.
- [8] L. Migliore and F. Coppedè, “Genetic and environmental factors in cancer and neurodegenerative diseases,” *Mutation Research/Reviews in Mutation Research*, vol. 512, no. 2-3, pp. 135–153, 2002.
- [9] P. S. Aisen, J. Cummings, C. R. Jack, J. C. Morris, R. Sperling, L. Frölich, R. W. Jones, S. A. Dowsett, B. R. Matthews, J. Raskin *et al.*, “On the path to 2025: understanding the alzheimer’s disease continuum,” *Alzheimer’s Research & Therapy*, vol. 9, no. 1, p. 60, 2017.
- [10] G. Vradenburg, “A pivotal moment in alzheimer’s disease and dementia: how global unity of purpose and action can beat the disease by 2025,” *Expert Review of Neurotherapeutics*, vol. 15, no. 1, pp. 73–82, 2015.
- [11] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” *Health Information Science and Systems*, vol. 2, no. 1, p. 3, 2014.
- [12] A. Madabhushi, S. Agner, A. Basavanhally, S. Doyle, and G. Lee, “Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data,” *Computerized Medical Imaging and Graphics*, vol. 35, no. 7-8, pp. 506–514, 2011.
- [13] D. M. Jones-Davis and N. Buckholtz, “The impact of the Alzheimer’s disease neuroimaging initiative 2: what role do public-private partnerships have in pushing the boundaries of clinical and basic science research on Alzheimer’s disease?” *Alzheimer’s & Dementia*, vol. 11, no. 7, pp. 860–864, 2015.

- [14] M. Brant-Zawadzki, D. Atkinson, M. Detrick, W. G. Bradley, and G. Scidmore, "Fluid-attenuated inversion recovery (flair) for assessment of cerebral infarction: initial clinical experience in 50 patients," *Stroke*, vol. 27, no. 7, pp. 1187–1191, 1996.
- [15] K. Oppedal, T. Eftestøl, K. Engan, M. K. Beyer, and D. Aarsland, "Classifying dementia using local binary patterns from different regions in magnetic resonance images," *Journal of Biomedical Imaging*, vol. 2015, p. 5, 2015.
- [16] J. He, R. I. Grossman, Y. Ge, and L. J. Mannon, "Enhancing patterns in multiple sclerosis: evolution and persistence," *American Journal of Neuroradiology*, vol. 22, no. 4, pp. 664–669, 2001.
- [17] N. Altaf, P. S. Morgan, A. Moody, S. T. MacSweeney, J. Gladman, and D. P. Auer, "Brain white matter hyperintensities are associated with carotid intraplaque hemorrhage," *Radiology*, vol. 248, no. 1, pp. 202–209, 2008.
- [18] C. Coffey, W. Wilkinson, L. Parashos, S. Soady, R. Sullivan, L. Patterson, G. Figiel, M. Webb, C. Spritzer, and W. Djang, "Quantitative cerebral anatomy of the aging human brain: a cross-sectional study using magnetic resonance imaging," *Neurology*, vol. 42, no. 3, pp. 527–527, 1992.
- [19] R. Peters, "Ageing and the brain," *Postgraduate Medical Journal*, vol. 82, no. 964, pp. 84–88, 2006.
- [20] C.-C. Chuang, Y.-T. Lee, C.-M. Chen, Y.-S. Hsieh, T.-C. Liu, and C.-W. Sun, "Patient-oriented simulation based on monte carlo algorithm by using MRI data," *Biomedical Engineering Online*, vol. 11, no. 1, p. 21, 2012.
- [21] J. S. O'Brien and E. L. Sampson, "Lipid composition of the normal human brain: gray matter, white matter, and myelin," *Journal of Lipid Research*, vol. 6, no. 4, pp. 537–544, 1965.
- [22] M. Bilello, J. Doshi, S. A. Nabavizadeh, J. B. Toledo, G. Erus, S. X. Xie, J. Q. Trojanowski, X. Han, and C. Davatzikos, "Correlating cognitive decline with white matter lesion and brain atrophy magnetic resonance imaging measurements in Alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 48, no. 4, pp. 987–994, 2015.
- [23] R. A. Pooley, "Fundamental physics of MRI imaging," *Radiographics*, vol. 25, no. 4, pp. 1087–1099, 2005.
- [24] R. G. Bryant and J.-P. Korb, "Nuclear magnetic resonance and spin relaxation in biological systems," *Magnetic Resonance Imaging*, vol. 23, no. 2, pp. 167–173, 2005.
- [25] C. Westbrook and J. Talbot, *MRI in Practice*. John Wiley & Sons, 2018.
- [26] P. Schmitt, M. A. Griswold, P. M. Jakob, M. Kotas, V. Gulani, M. Flentje, and A. Haase, "Inversion recovery truefisp: quantification of T1, T2, and spin density," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 51, no. 4, pp. 661–667, 2004.
- [27] V. Stevenson, G. Parker, G. Barker, K. Birnie, P. Tofts, D. Miller, and A. Thompson, "Variations in T1 and T2 relaxation times of normal appearing white matter and lesions in multiple sclerosis," *Journal of the Neurological Sciences*, vol. 178, no. 2, pp. 81–87, 2000.
- [28] J. Staals, S. D. Makin, F. N. Doubal, M. S. Dennis, and J. M. Wardlaw, "Stroke subtype, vascular risk factors, and total MRI brain small-vessel disease burden," *Neurology*, vol. 83, no. 14, pp. 1228–1234, 2014.
- [29] A. Khademi and D. Hosseinzadeh, "Type II fuzzy systems for amyloid plaque segmentation in transgenic mouse brains for Alzheimer's disease quantification," in *Medical Imaging 2014: Digital Pathology*, vol. 9041. International Society for Optics and Photonics, 2014, p. 90410T.

- [30] J. M. Wardlaw, M. C. Valdés Hernández, and S. Muñoz-Maniega, “What are white matter hyperintensities made of? relevance to vascular cognitive impairment,” *Journal of the American Heart Association*, vol. 4, no. 6, p. e001140, 2015.
- [31] J. E. Iglesias, C.-Y. Liu, P. M. Thompson, and Z. Tu, “Robust brain extraction across datasets and comparison with publicly available methods,” *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1617–1634, 2011.
- [32] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins, “Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging,” *Medical Image Analysis*, vol. 17, no. 1, pp. 1–18, 2013.
- [33] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward *et al.*, “The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods,” *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 27, no. 4, pp. 685–691, 2008.
- [34] B. Reiche, A. Moody, A. Khademi, A. D. N. Initiative *et al.*, “Pathology-preserving intensity standardization framework for multi-institutional flair MRI datasets,” *Magnetic Resonance Imaging*, 2019.
- [35] B. Mortamet, M. A. Bernstein, C. R. Jack Jr, J. L. Gunter, C. Ward, P. J. Britson, R. Meuli, J.-P. Thiran, and G. Krueger, *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 62, no. 2, pp. 365–372, 2009.
- [36] A. Khademi, B. Reiche, J. DiGregorio, G. Arezza, and A. R. Moody, “Whole volume brain extraction for multi-centre, multi-disease flair MRI datasets,” *Magnetic Resonance Imaging*, 2019.
- [37] J. Knight and A. Khademi, “MS lesion segmentation using flair MRI only,” *Proceedings of the 1st MICCAI Challenge on Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure-MICCAI-MSSEG*, pp. 21–28, 2016.
- [38] I. C. George, P. Sati, M. Absinta, I. C. Cortese, E. M. Sweeney, C. D. Shea, and D. S. Reich, “Clinical 3-tesla flair* MRI improves diagnostic accuracy in multiple sclerosis,” *Multiple Sclerosis Journal*, vol. 22, no. 12, pp. 1578–1586, 2016.
- [39] O. Esteban, D. Birman, M. Schaer, O. O. Koyejo, R. A. Poldrack, and K. J. Gorgolewski, “MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites,” *PLOS One*, vol. 12, no. 9, p. e0184661, 2017.
- [40] J. Jang, K. Bang, H. Jang, D. Hwang, and A. D. N. Initiative, “Quality evaluation of no-reference MR images using multidirectional filters and image statistics,” *Magnetic Resonance in Medicine*, vol. 80, no. 3, pp. 914–924, 2018.
- [41] G. H. Glover, B. A. Mueller, J. A. Turner, T. G. Van Erp, T. T. Liu, D. N. Greve, J. T. Voyvodic, J. Rasmussen, G. G. Brown, D. B. Keator *et al.*, “Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies,” *Journal of Magnetic Resonance Imaging*, vol. 36, no. 1, pp. 39–54, 2012.
- [42] L. Friedman and G. H. Glover, “Report on a multicenter fMRI quality assurance protocol,” *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 23, no. 6, pp. 827–839, 2006.
- [43] D. S. Marcus, M. P. Harms, A. Z. Snyder, M. Jenkinson, J. A. Wilson, M. F. Glasser, D. M. Barch, K. A. Archie, G. C. Burgess, M. Ramaratnam *et al.*, “Human connectome project informatics: quality control, database services, and data visualization,” *Neuroimage*, vol. 80, pp. 202–219, 2013.

- [44] J. P. Woodard and M. P. Carley-Spencer, "No-reference image quality metrics for structural MRI," *Neuroinformatics*, vol. 4, no. 3, pp. 243–262, 2006.
- [45] F. Alfaro-Almagro, M. Jenkinson, N. K. Bangerter, J. L. Andersson, L. Griffanti, G. Douaud, S. N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee *et al.*, "Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank," *Neuroimage*, vol. 166, pp. 400–424, 2018.
- [46] M. Osadebey, M. Pedersen, D. Arnold, and K. Wendel-Mitoraj, "No-reference quality measure in brain MRI images using binary operations, texture and set analysis," *IET Image Processing*, vol. 11, no. 9, pp. 672–684, 2017.
- [47] M. Osadebey, M. Pedersen, D. Arnold, K. Wendel-Mitoraj, and ADNI, "The spatial statistics of structural magnetic resonance images: application to post-acquisition quality assessment of brain MRI images," *The Imaging Science Journal*, vol. 65, no. 8, pp. 468–483, 2017.
- [48] M. Osadebey, M. Pedersen, D. Arnold, and K. Wendel-Mitoraj, "Image quality evaluation in clinical research: A case study on brain and cardiac MRI images in multi-center clinical trials," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 6, pp. 1–15, 2018.
- [49] L. G. Nyul and J. K. Udupa, "On standardizing the MR image intensity scale," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 42, no. 6, pp. 1072–1081, 1999.
- [50] F. Jager and J. Hornegger, "Nonrigid registration of joint histograms for intensity standardization in magnetic resonance imaging," *IEEE Transactions on Medical Imaging*, vol. 28, no. 1, pp. 137–150, 2008.
- [51] S. Duchesne, I. Chouinard, O. Potvin, V. S. Fonov, A. Khademi, R. Bartha, P. Bellec, D. L. Collins, M. Descoteaux, R. Hoge *et al.*, "The canadian dementia imaging protocol: Harmonizing national cohorts," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 2, pp. 456–465, 2019.
- [52] A. Khademi, A. Venetsanopoulos, and A. R. Moody, "Image enhancement and noise suppression for flair MRIs with white matter lesions," *IEEE Signal Processing Letters*, vol. 17, no. 12, pp. 989–992, 2010.
- [53] J. Yan, J. Schaefferkoetter, M. Conti, and D. Townsend, "A method to assess image quality for low-dose pet: analysis of snr, cnr, bias and image noise," *Cancer Imaging*, vol. 16, no. 1, p. 26, 2016.
- [54] J.-C. Tardif, J. D. Spence, T. M. Heinonen, A. Moody, J. Pressacco, R. Frayne, P. L'Allier, B. J. Chow, M. Friedrich, S. E. Black *et al.*, "Atherosclerosis imaging and the canadian atherosclerosis imaging network," *Canadian Journal of Cardiology*, vol. 29, no. 3, pp. 297–303, 2013.
- [55] A. Khademi, A. Venetsanopoulos, and A. Moody, "Automatic contrast enhancement of white matter lesions in flair MRI," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2009, pp. 322–325.
- [56] E. McVeigh, R. Henkelman, and M. Bronskill, "Noise and filtration in magnetic resonance imaging," *Medical Physics*, vol. 12, no. 5, pp. 586–591, 1985.
- [57] J. A. Fessler, "Model-based image reconstruction for MRI," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 81–89, 2010.
- [58] R. Sadleir, S. Grant, S. U. Zhang, B. I. Lee, H. C. Pyo, S. H. Oh, C. Park, E. J. Woo, S. Y. Lee, O. Kwon *et al.*, "Noise analysis in magnetic resonance electrical impedance tomography at 3 and 11 t field strengths," *Physiological Measurement*, vol. 26, no. 5, p. 875, 2005.

- [59] R. Guillemaud and M. Brady, "Estimating the bias field of MR images," *IEEE Transactions on Medical imaging*, vol. 16, no. 3, pp. 238–251, 1997.
- [60] K. Kim, P. A. Habas, V. Rajagopalan, J. A. Scott, J. M. Corbett-Detig, F. Rousseau, A. J. Barkovich, O. A. Glenn, and C. Studholme, "Bias field inconsistency correction of motion-scattered multislice MRI for improved 3D image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1704–1712, 2011.
- [61] W. M. Wells, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz, "Adaptive segmentation of MRI data," *IEEE Transactions on Medical imaging*, vol. 15, no. 4, pp. 429–442, 1996.
- [62] U. Vovk, F. Pernus, and B. Likar, "A review of methods for correction of intensity inhomogeneity in MRI," *IEEE Transactions on Medical Imaging*, vol. 26, no. 3, pp. 405–421, 2007.
- [63] L. Erasmus, D. Hurter, M. Naudé, H. Kritzinger, and S. Acho, "A short overview of MRI artefacts," *SA Journal of Radiology*, vol. 8, no. 2, 2004.
- [64] M. Zaitsev, J. Maclaren, and M. Herbst, "Motion artifacts in MRI: a complex problem with many partial solutions," *Journal of Magnetic Resonance Imaging*, vol. 42, no. 4, pp. 887–901, 2015.
- [65] G. H. Glover and J. M. Pauly, "Projection reconstruction techniques for reduction of motion effects in MRI," *Magnetic Resonance in Medicine*, vol. 28, no. 2, pp. 275–289, 1992.
- [66] E. M. Bellon, E. M. Haacke, P. E. Coleman, D. C. Sacco, D. A. Steiger, and R. E. Gangarosa, "MR artifacts: a review," *American Journal of Roentgenology*, vol. 147, no. 6, pp. 1271–1281, 1986.
- [67] S. B. Vos, C. Micallef, F. Barkhof, A. Hill, G. P. Winston, S. Ourselin, and J. S. Duncan, "Evaluation of prospective motion correction of high-resolution 3D-T2-flair acquisitions in epilepsy patients," *Journal of Neuroradiology*, vol. 45, no. 6, pp. 368–373, 2018.
- [68] A. Elster and J. Burdette, "Introduction to nuclear magnetic resonance," *Questions and Answers in Magnetic Resonance Imaging*, Elster AD and Burdette JH. 2nd ed., St Louis, Mosby, pp. 19–53, 2001.
- [69] W. Huda and R. M. Slone, *Review of radiologic physics*. Lippincott Williams & Wilkins, 2003.
- [70] K. Ryu, Y. Nam, S.-M. Gho, J. Jang, H.-J. Lee, J. Cha, H. J. Baek, J. Park, and D.-H. Kim, "Data-driven synthetic MRI flair artifact correction via deep neural network," *Journal of Magnetic Resonance Imaging*, 2019.
- [71] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.
- [72] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [73] A. Khademi, B. Reiche, J. DiGregorio, G. Arezza, and A. R. Moody, "Whole volume brain extraction for multi-centre, multi-disease flair MRI datasets," *Magnetic Resonance Imaging*, 2019.
- [74] I. Sobel and G. Feldman, "A 3x3 isotropic gradient operator for image processing," *A Talk at the Stanford Artificial Project*, pp. 271–272, 1968.
- [75] J. M. Prewitt, "Object enhancement and extraction," *Picture Processing and Psychopictorics*, vol. 10, no. 1, pp. 15–19, 1970.

- [76] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *Proceedings. International Conference on Image Processing*, vol. 1. IEEE, 2002, pp. I–I.
- [77] M. Metcalf, D. Xu, D. T. Okuda, L. Carvajal, R. Srinivasan, D. A. Kelley, P. Mukherjee, S. J. Nelson, D. B. Vigneron, and D. Pelletier, "High-resolution phased-array MRI of the human brain at 7 tesla: initial experience in multiple sclerosis patients," *Journal of Neuroimaging*, vol. 20, no. 2, pp. 141–147, 2010.
- [78] D. Jain, "Skew and kurtosis: 2 important statistics terms you need to know in data science," Mar 2019, Last accessed: Aug 2019. [Online]. Available: <https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa>
- [79] H. Pham, *Springer handbook of engineering statistics*. Springer Science & Business Media, 2006.
- [80] M. Zaitsev, J. Maclaren, and M. Herbst, "Motion artifacts in MRI: a complex problem with many partial solutions," *Journal of Magnetic Resonance Imaging*, vol. 42, no. 4, pp. 887–901, 2015.
- [81] C. Studholme, V. Cardenas, E. Song, F. Ezekiel, A. Maudsley, and M. Weiner, "Accurate template-based correction of brain MRI intensity distortion with application to dementia and aging," *IEEE Transactions on Medical Imaging*, vol. 23, no. 1, pp. 99–110, 2004.
- [82] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Fundamentals of exploratory analysis of variance*. John Wiley & Sons, 2009, vol. 367.
- [83] R. Dawson, "How significant is a boxplot outlier?" *Journal of Statistics Education*, vol. 19, no. 2, 2011.
- [84] S. S. Sonnad, "Describing data: statistical and graphical methods," *Radiology*, vol. 225, no. 3, pp. 622–628, 2002.
- [85] C. Hawco, J. D. Viviano, S. Chavez, E. W. Dickie, N. Calarco, P. Kochunov, M. Argyelan, J. A. Turner, A. K. Malhotra, R. W. Buchanan *et al.*, "A longitudinal human phantom reliability study of multi-center T1-weighted, dti, and resting state fMRI data," *Psychiatry Research: Neuroimaging*, vol. 282, pp. 134–142, 2018.
- [86] H. Chertkow, M. Borrie, V. Whitehead, S. E. Black, H. H. Feldman, S. Gauthier, D. B. Hogan, M. Masellis, K. McGilton, K. Rockwood *et al.*, "The comprehensive assessment of neurodegeneration and dementia: Canadian cohort study," *Canadian Journal of Neurological Sciences*, vol. 46, no. 5, pp. 499–511, 2019.
- [87] C. R. Jack Jr, M. A. Bernstein, B. J. Borowski, J. L. Gunter, N. C. Fox, P. M. Thompson, N. Schuff, G. Krueger, R. J. Killiany, C. S. DeCarli *et al.*, "Update on the magnetic resonance imaging core of the Alzheimer's disease neuroimaging initiative," *Alzheimer's & Dementia*, vol. 6, no. 3, pp. 212–220, 2010.
- [88] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica: Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.
- [89] J. B. Andre, B. W. Bresnahan, M. Mossa-Basha, M. N. Hoff, C. P. Smith, Y. Anzai, and W. A. Cohen, "Toward quantifying the prevalence, severity, and cost associated with patient motion during clinical MR examinations," *Journal of the American College of Radiology*, vol. 12, no. 7, pp. 689–695, 2015.
- [90] MATLAB, version 9.6.0.1072779 (R2019a). Natick, Massachusetts: The MathWorks Inc., 2019.
- [91] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep learning for brain MRI segmentation: state of the art and future directions," *Journal of Digital Imaging*, vol. 30, no. 4, pp. 449–459, 2017.