Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

A STUDY OF MACHINE LEARNING AND DEEP LEARNING MODELS FOR SOLVING MEDICAL IMAGING PROBLEMS

by
Fadi G. Farhat

Application of machine learning and deep learning methods on medical imaging aims to create systems that can help in the diagnosis of disease and the automation of analyzing medical images in order to facilitate treatment planning. Deep learning methods do well in image recognition, but medical images present unique challenges. The lack of large amounts of data, the image size, and the high class-imbalance in most datasets, makes training a machine learning model to recognize a particular pattern that is typically present only in case images a formidable task.

Experiments are conducted to classify breast cancer images as healthy or non-healthy, and to detect lesions in damaged brain MRI (Magnetic Resonance Imaging) scans. Random Forest, Logistic Regression and Support Vector Machine perform competitively in the classification experiments, but in general, deep neural networks beat all conventional methods. Gaussian Naïve Bayes (GNB) and the Lesion Identification with Neighborhood Data Analysis (LINDA) methods produce better lesion detection results than single path neural networks, but a multi-modal, multi-path deep neural network beats all other methods. The importance of pre-processing training data is also highlighted and demonstrated, especially for medical images, which require extensive preparation to improve classifier and detector performance. Only a more complex and deeper neural network combined with properly pre-processed data can produce the desired accuracy levels that can rival and maybe exceed those of human experts.

A STUDY OF MACHINE LEARNING AND DEEP LEARNING MODELS FOR SOLVING MEDICAL IMAGING PROBLEMS

by
Fadi G. Farhat

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Data Science

Department of Computer Science

May 2019

APPROVAL PAGE

A STUDY OF MACHINE LEARNING AND DEEP LEARNING MODELS FOR SOLVING MEDICAL IMAGING PROBLEMS

Fadi G. Farhat

Dr. Usman Roshan, Thesis Advisor — Date
Associate Professor of Computer Science, NJIT

Dr. William Graves, Committee Member — Date
Associate Professor of Psychology, Rutgers University - Newark

---

Dr. Fadi P. Deek, Committee Member — Date
Provost and Senior Executive Vice President
Distinguished Professor of Informatics, NJIT

BIOGRAPHICAL SKETCH

Author: Fadi G. Farhat

Degree: Master of Science

Date: May 2019

Undergraduate and Graduate Education:

- Master of Science in Data Science,
  New Jersey Institute of Technology, Newark, NJ, 2019

- Bachelor of Engineering in Electrical Engineering,
  City University of New York, New York, NY, 1995

Major: Data Science

Presentations and Publications:

Yunzhe Xue, Fadi G. Farhat, Olga Boukrina, A. M. Barrett, Jeffrey R. Binder, Usman W. Roshan, William W. Graves, " A multi-path 2.5-dimensional convolutional neural network system for segmenting stroke lesions in brain MRI images", Preprint submitted to Neuroimage Clinical, March 31, 2019.

This Work is Dedicated to

My lovely wife Sabine,

My precious girls Kristen and Kate,

and My adorable little boy George

This Work is also Dedicated to

My amazing mother Leila,

My supportive sisters Rita and Gloria,

and My dearest brother and best friend Roy

In Remembrance of My Father and Eternal Friend

Gerges Youssef Farhat

1940 – 2009

ACKNOWLEDGMENT

First and foremost, I am infinitely grateful to The Lord for giving me the strength, determination and persistence to complete the master's journey.

I am extremely grateful to my advisor, teacher and mentor, Dr. Usman Roshan, for his guidance, his advice and vision. I met Dr. Roshan when I first joined NJIT, and almost immediately, he became instrumental in plotting my academic path. He even helped me make the correct choice every time I was at a crossroads. I thank him for being a great mentor, a patient teacher and a very good friend.

I express my sincere appreciation to Dr. William Graves, Associate Professor of Psychology at Rutgers University - Newark, for always being so involved in my research, and often lending a helping hand in scripting and medical image manipulation. I found Dr. Graves' knowledge in neuroscience, and his agility in computer science to be quite impressive. I learned a lot from Dr. Graves, both about the human brain and the processing of MRI brain scan images. I am grateful to have met him and to have worked with him.

I cannot put in words my thanks to my longtime friend and main influencer Dr. Fadi P. Deek, Provost and Senior Executive Vice President and Distinguished Professor of Informatics at NJIT. To put it simply, Dr. Deek is the main reason why I returned to school twenty-one years after finishing my undergraduate studies. He was adamant about my pursuing a higher education. After so many years of being away from academia, everything seemed a bit complicated and even intimidating at times. Dr. Deek made it all

simple and easily approachable. I will forever be thankful for his unwavering support, for believing in me, and for guiding me and making sure I make it this far.

I would be remiss if I did not thank my colleague and teammate Yunzhe Xue, who is currently a candidate for a doctorate in Computer Science at NJIT. I appreciate all the help Yunzhe gave me during my research, and I thank him for his patience and kindness.

I also would like to thank my dear friend Peter Rizk, Senior Director of Technical Marketing and Solutions Architecture at Infoblox, for supporting me and giving me access to the state-of-the-art compute cluster at their facility in San Francisco.

Most of my research would not have been possible without the facilities and compute resources that were made available to me at NJIT throughout my thesis work. My thanks and appreciation to all staff at NJIT's High Performance Clusters.

Breast cancer mammogram data used in this research was provided by the Digital Database for Screening Mammography (DDSM), which is a collection of publicly available mammograms from the following sources: Massachusetts General Hospital, Wake Forest University School of Medicine, Sacred Heart Hospital, and Washington University of St Louis School of Medicine.

Breast cancer biopsy data used in this research was obtained from the Breast Cancer Histopathological Image Classification (BreakHis) database which is a collection of microscopic biopsy images of cancer tissue built by the Laboratory of Vision, Robotics and Imaging (LVRI) of the Federal University of Parana (UFPR) in collaboration with the Prevention & Diagnosis Laboratory, both located in Parana, Brazil.

Brain MRI and lesion data used in this research was provided in part by the Open Access Series of Imaging Studies (OASIS) Brains Project, which is committed to making neuroimaging datasets freely available to the scientific community (Principal Investigators: D. Marcus, R, Buckner, J, Csernansky J. Morris), by the Kessler Foundation, the Medical College of Wisconsin and ATLAS (Anatomical Tracings of Lesions After Stroke) Release 1.1, an open-source dataset consisting of T1-weighted MRIs with manually segmented diverse lesions and metadata.

Last but not least, I would like to recognize my siblings Roy and Gloria for always pushing me and encouraging me and offer my deepest gratitude to my wife and true partner Sabine for her relentless support and encouragement during this challenging journey.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF SYMBOLS

©	Copyright
$\int$	Integration
$\sum$	Summation
μ	Mean of population
σ	Standard Deviation
σ^2	Variance
®	Registered
$\approx$	Approximately
Δ	Change / Difference
∂	Partial Differential
π	Constant = 3.14159
$\rho_{X,Y}$	Correlation
ε	A very small number

LIST OF DEFINITIONS

Adversarial Attack	Input to machine learning models that an attacker has intentionally designed to cause the model to make a mistake.
Artificial Intelligence	The theory and development of computer systems able to perform tasks that normally require human intelligence.
Artificial Neural Network	A computational model with multiple layers which is based loosely on the structure and functions of biological neural networks.
AUC or AUC/ROC	A performance measurement curve for classification problems at various thresholds settings. ROC (**R**eceiver **O**perating **C**haracteristics) is a probability curve, and AUC (**A**rea **U**nder the **C**urve) represents the degree or measure of separability. It tells how much a model is capable of distinguishing between classes. A Higher AUC indicates a better model.
Bayes' Theorem	Describes the probability of an event, based on prior knowledge of conditions that might be related to the event.
BIRADS or BI-RADS	The **B**reast **I**maging **R**eporting and **D**ata **S**ystem was established by the American College of Radiology. It is a scheme for organizing the findings from mammogram screening (for breast cancer diagnosis) into well-defined categories.
Brain Lesion Mask	A binary graphical representation of brain damage. A value 1 is assigned to lesion voxels, and a value 0 for other voxels.
Clustering	The grouping of similar objects into a set known as cluster.
Convolutional Neural Network	A type of artificial neural network used in image recognition and processing that is specifically designed to process pixel data.
Decision Tree	A graph that represents a predictive model based on a branching series of Boolean tests that use specific facts to make more generalized conclusions.

LIST OF DEFINITIONS
(Continued)

Deep Learning	A collection of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
Deep Neural Network	A neural network that has more than two layers, and that uses sophisticated mathematical modeling to process data in complex ways.
Dice Similarity Coefficient	Sørensen–**D**ice (**S**imilarity) **C**oefficient. A statistic used for comparing the similarity of two samples.
DICOM Image Format (2D)	**D**igital **I**maging and **CO**mmunications in **M**edicine is a standard for handling, storing, printing, and transmitting information in medical imaging. It includes a file format definition and a network communications protocol.
Ductal Carcinoma	Also called Invasive Ductal Carcinoma (IDC) or Infiltrating Ductal Carcinoma and is the most common type of breast cancer. About 80% of all breast cancers are invasive ductal carcinomas. Ductal means that the cancer began in the milk ducts.
Ensemble Learning	Uses multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent algorithms alone.
Fibroadenoma	The most common type of benign breast tumor, and most don't increase the risk of breast cancer. Although women of any age can develop fibroadenomas, they usually occur in younger, premenopausal women.
Hyperplane	A subspace whose dimension is one less than that of its ambient space.
Linear Classifier	An algorithm that uses an object's characteristics to identify which class (or group) it belongs to. Decision is based on the value of a linear combination of the characteristics.

LIST OF DEFINITIONS
(Continued)

Lobular Carcinoma	An area (or areas) of abnormal cell growth that increases a person's risk of developing invasive breast cancer later on in life. Lobular means that the abnormal cells start growing in the lobules, the milk-producing glands at the end of breast ducts.
Machine Learning	An application of artificial intelligence that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
Mucinous Carcinoma	Sometimes called colloid carcinoma; a rare form of invasive ductal carcinoma. In this type of cancer, the tumor is made up of abnormal cells that "float" in pools of mucin, a key ingredient in the slimy, slippery substance known as mucus.
Naïve Bayes Classifier	A probabilistic classifier based on applying Bayes' Theorem with strong (naive) independence assumptions between the features.
NIfTI Image Format (3D)	**N**euroimaging **I**n**f**ormatics **T**echnology **I**nitiative is new Analyze-style data format, proposed by the Data Format Working Group as a short-term measure to facilitate inter-operation of functional MRI data analysis software packages.
Non-Linear Classifier	An algorithm used when the combination of characteristics cannot be represented using a linear function.
Papillary Carcinoma	A rare form of breast cancer, accounting for less than 1-2% of cases. In most cases, these types of tumors are diagnosed in older women who have already been through menopause. A papillary carcinoma usually has a well-defined border and is made up of small, finger-like projections.
Phyllodes Tumor	Rare type of benign tumor, accounting for less than 1% of all breast tumors. The name "phyllodes," which is taken from the Greek language and means "leaflike," refers to that fact that the tumor cells grow in a leaflike pattern.

LIST OF DEFINITIONS
(Continued)

Precision	Also known as positive predictive value and is the fraction of relevant instances among the retrieved instances. $P = TP / (TP + FP)$.
Random Forest	An ensemble of simple decision trees, which are used to determine the final outcome. In classification, the ensemble vote for the most popular class.
Recall	Also known as sensitivity or true positive rate and is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. $R = TP / (TP + FN)$.
Repetition Time & Time to Echo	In MRI, Repetition Time (TR) is the amount of time between successive pulse sequences applied to the same slice. Time to Echo (TE) is the time between the delivery of the RF pulse and the receipt of the echo signal.
Statistical Parametric Mapping (SPM)	**S**tatistical **P**arametric **M**apping refers to the construction and assessment of spatially extended statistical processes used to test hypotheses about functional imaging data. These ideas have been instantiated in a software package called SPM, which has been designed for the analysis of brain imaging data sequences.
Supervised Learning	An algorithm that can apply what has been learned in the past to new data using labeled examples to predict future events.
Support Vector Machine	An algorithm that defines a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or outlier detection.
T1 MRI Scan	T1 images are produced by using short TR and TE times.
T2 MRI Scan	T2 images are produced by using longer TR and TE times.

LIST OF DEFINITIONS
(Continued)

Tubular Adenoma	Also known as pure adenoma, is a rare epithelial (relating to the thin tissue that lines the surfaces of the body) tumor of the breast. Only a few cases have been reported in the literature, especially in young women of reproductive age. Postmenopausal women are very rarely affected.
Unsupervised Learning	An algorithm used when the information available is neither classified nor labeled. A function to describe a hidden structure from unlabeled data is inferred.

CHAPTER 1

INTRODUCTION

1.1 Objective

The objective of this study is to present applications of machine learning and deep learning methods on medical imaging to solve abnormality classification and segmentation problems. The ultimate goal is to develop systems that can aid in the diagnosis of disease, to automate the difficult and time-consuming tasks of reading and analyzing medical images, and to facilitate treatment planning. We want to be able to apply machine learning methods to automatically classify medical images (for example, breast x-ray or biopsy images) as healthy (non-cancerous) or not healthy (cancerous). Having isolated the unhealthy images, the next objective is to identify the regions in the image that have damaged tissue using segmentation, which is largely done by human experts, costing tremendous resources. Finally, diagnosis and treatment decisions can be made based on what is learned about the images. The hope is that automatic classification and segmentation can be performed using new and innovative deep learning techniques, and high levels of accuracy can be achieved.

1.2 Background

Deep learning methods have been used in image recognition for quite some time now. A popular dataset in this field of study is ImageNet, a database of over 15 million images that belong to 22,000 categories of common objects, fruits, vegetables, animals, and even

persons. While very high levels of accuracy (well over 80%; e.g., Jeremy Howard, fast.ai, 2018, 93%) have been achieved using models trained on the ImageNet data, medical images are hard to obtain and present unique challenges.

Machine learning models perform better if trained on a large amount of data. Medical studies typically offer a very small number of images per study, usually under 100, which makes machine learning more difficult. Medical images are also usually much larger than object images used in training recognition models. A typical object image is 32x32 pixels, or 64x64 pixels, but medical images can be 100 times that size. To add to the complexity, many medical studies produce 3D images, which increases the size of each sample drastically (200 to 300 times), requiring a lot more storage and compute resources. Additionally, classifiers in medical imaging applications are tasked with finding very subtle differences between almost identical images, which is harder than identifying a dog or a cat in a photo. It should also be noted that medical studies, not unlike other studies, generally produce highly imbalanced data, where the number of controls is larger than the number of cases. This also presents an added challenge when training a machine learning model to recognize a particular pattern that is typically present only in case images.

In the past year, there has been a great increase in applying deep learning methods to medical imaging problems. While conventional algorithms like Support Vector Machine (SVM) and Random Forest (RF) do well for larger areas of damaged tissue, Deep Learning tends to do better in general, especially on smaller, harder to detect damaged areas, such as small lesions in brain MRI (**M**agnetic **R**esonance **I**maging) scans (Yunzhe Xue, et al., Neuroimage Clinical, 2019).

Manual image annotation and abnormality tracing is complex and time consuming, so there is also a growing research interest in unsupervised anomaly detection methods. Although supervised deep learning is achieving the best results, it is often limited by the accuracy and reliability of the ground truth, which is provided by a human rater. A number of studies in this area (for brain segmentation) that came out in the past few months are worth noting:

1. **Deep Learning vs. Conventional Machine Learning: Pilot Study of WMH Segmentation in Brain MRI with Absence or Mild Vascular Pathology.**

   Compares deep learning algorithms, namely the deep Boltzmann machine (DBM), convolutional encoder network (CEN) and patch-wise convolutional neural network (patch-CNN), with two conventional machine learning schemes: SVM and RF, for white matter hyperintensities (WMH) segmentation on brain MRI with mild or no vascular pathology (Rachmadi, M.F., et al., 2017).

2. **Robust Image Segmentation Quality Assessment without Ground Truth.**

   Proposes a new method to protect neural networks from robustness problems (e.g. vulnerability to adversarial attacks), by utilizing the difference between the input image and the reconstructed image, which is obtained from the segmentation to be assessed. The deep-learning-based reconstruction network (REC-Net) is trained with the input image masked by the ground truth segmentation against the original input image as the target (Zhou, Leixin, et al., 2019).

3. **Unsupervised Brain Lesion Segmentation from MRI using a Convolutional Autoencoder.**

   Presents a novel unsupervised segmentation approach to address the problem of variability in lesion load, placement of lesions, and voxel intensities, using a convolutional autoencoder, which learns to segment brain lesions as well as the white matter, gray matter, and cerebrospinal fluid by reconstructing Fluid Attenuated Inversion Recovery (FLAIR) images as conical combinations of Softmax layer outputs (dense layer in a neural network) generated from the corresponding anatomical and FLAIR MRI images (Atlason, Hans E., et al., 2018).

4. **Unsupervised Detection of Lesions in Brain MRI using Constrained Adversarial Auto-encoders.**

   Detection of lesion regions is studied in an unsupervised manner by learning data

distribution of brain MRI of healthy subjects using auto-encoder based methods. The Human Connectome Project dataset is used to learn distribution of healthy-appearing brain MRI and report improved detection, in terms of AUC (**A**rea **U**nder the **C**urve), of the lesions in the BraTS (**Bra**in **T**umor **S**egmentation) challenge dataset (Chen, Xiaoran & Konukoglu, Ender, 2018).

The number of studies that address medical imaging problems using machine learning and deep learning methods is still relatively limited, but that seems to be changing. In 2018, the first edition of the International conference on "Medical Imaging with Deep Learning" (MIDL) was held in Amsterdam. The number of abstracts submitted was not huge, but interest is clearly growing, and as of the drafting of this document, the second MIDL conference is already scheduled for July 2019.

CHAPTER 2

BREAST CANCER IMAGE CLASSIFICATION

2.1 Description of Breast Cancer Datasets

"Breast cancer is the second leading cause of cancer death among women. Each year it is estimated that over 252,710 women in the United States will be diagnosed with breast cancer and more than 40,500 will die" (National Breast Cancer Foundation, 2019). Analysis of breast cancer imaging data by humans is time consuming and inevitably ties up expert human resources that could otherwise focus more on patient care and treatment. Having reliable automated breast cancer diagnosis tools, such as cancer image classification, can help streamline analysis, reduce human error and speed up treatment.

The breast cancer images used in this study are from two publicly available datasets obtained from the Digital Database for Screening Mammography (DDSM) and the Laboratory of Vision, Robotics and Imaging (LVRI) of the Federal University of Parana (UFPR) in Brazil. DDSM is a collection of mammograms from the following sources: Massachusetts General Hospital, Wake Forest University School of Medicine, Sacred Heart Hospital, and Washington University in St Louis School of Medicine. The Breast Cancer Histopathological Image Classification (BreakHis) database is a collection of microscopic biopsy images of cancer tissue and was built by LVRI in collaboration with the Prevention & Diagnosis Laboratory, in Parana, Brazil (prevencaoediagnose.com.br).

Our first dataset, the CBIS-DDSM (**C**urated **B**reast **I**maging **S**ubset of DDSM), includes decompressed images in DICOM (**D**igital **I**maging and **CO**mmunications in

Medicine) format. The data set contains 753 calcification cases and 891 mass cases. Multiple x-ray images are provided per patient, for a total of 6671 samples, out of which 3103 samples are actually full mammograms. Only these images are included in the final dataset we use for classification. The samples are pre-split into training and testing sets. 80% of the data is designated for training, and the remaining 20% make up the testing set. The data split was performed in such a way to provide equal level of difficulty in the training and test sets. Table 2.1 below shows the number of benign and malignant cases for each set, and how the data is distributed. It is worth noting each type of abnormality found in the images (calcification or mass) can be benign or malignant.

Table 2.1 Cases and Abnormalities in the Breast Cancer Mammogram X-Ray Images

	Benign Cases	**Malignant Cases**
Calcification Training Set	329 cases (552 abnormalities)	273 cases (304 abnormalities)
Calcification Test Set	85 cases (112 abnormalities)	66 cases (77 abnormalities)
Mass Training Set	355 cases (387 abnormalities)	336 cases (361 abnormalities)
Mass Test Set	117 cases (135 abnormalities)	83 cases (87 abnormalities)

Source: Lee, R. S., et al. (2017). "A curated mammography data set for use in computer-aided detection and diagnosis research." Scientific Data 4: 170177. https://www.nature.com/articles/sdata2017177 (accessed in April 2019)

The Breast Imaging Reporting and Data System (BIRADS) was used to categorize the samples and create the train/test data splits. BIRADS was established by the American College of Radiology as a means to help radiologists organize the findings from

mammogram screening (for breast cancer diagnosis) into a small number of well-defined categories. The samples were split for mass cases and calcification cases separately for machine learning research purposes.

The microscopic biopsy images in our BreakHis dataset were collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). The images are provided in their raw PNG (Portable Network Graphic) format, without normalization or color standardization and are all the same size (700x460 pixels, 3-channel RGB, 8-bit depth per channel). The dataset is divided into two main groups: benign tumors and malignant tumors. A lesion is referred to as histologically benign when it does not match any criteria of malignancy. Malignant tumors are cancerous lesions that can invade and destroy adjacent structures (locally invasive) and spread to distant sites (metastasize) to cause death. The samples present in this dataset were collected by SOB (Surgical Open Biopsy) method, also called partial mastectomy or excisional biopsy. This type of procedure removes a large tissue sample and is done in a hospital with general anesthesia.

The benign and malignant groups are further divided into sub-groups describing the specific kind of anomaly. For benign lesions, the anomalies present are fibroadenoma, Phyllodes tumor and tubular adenoma. For the malignant lesions, the anomalies present are ductal carcinoma, lobular carcinoma, mucinous carcinoma and papillary carcinoma. In our experiments, we will only consider images at the 400X magnification level, where we count a total of 1,606 samples. Out of that total, 374 samples are benign and 1,232 are malignant. We can see that this is a greatly imbalanced dataset in favor of malignant tumors. Table 2.2 below shows how the samples are distributed within each group.

Table 2.2 Benign and Malignant Tumor Distribution in the Breast Cancer Biopsy Images

Benign Cases	Malignant Cases
Fibroadenoma (129 samples)	Ductal Carcinoma (788 samples)
Phyllodes Tumor (115 samples)	Lobular Carcinoma (137 samples)
Tubular Adenoma (130 samples)	Mucinous Carcinoma (169 samples)
	Papillary Carcinoma (138 samples)

2.1.1 Breast Cancer Mammogram Dataset Exploration

The 3103 full mammogram images included in the classification experiments are not all the same size. They will need to be resampled to make one uniform set. The original DICOM images are very large, some more than 25 MB in size. When converted to JPG, the images are much smaller (less than 1 MB at most, depending on the final resolution.) The images are also a mixture of left and right breasts. The angle and zoom level of the x-ray image varies from image to image. Many images also have markings on them, like letters, numbers and other unidentified obstructions. Figure 2.1 below illustrates the variation in the images.

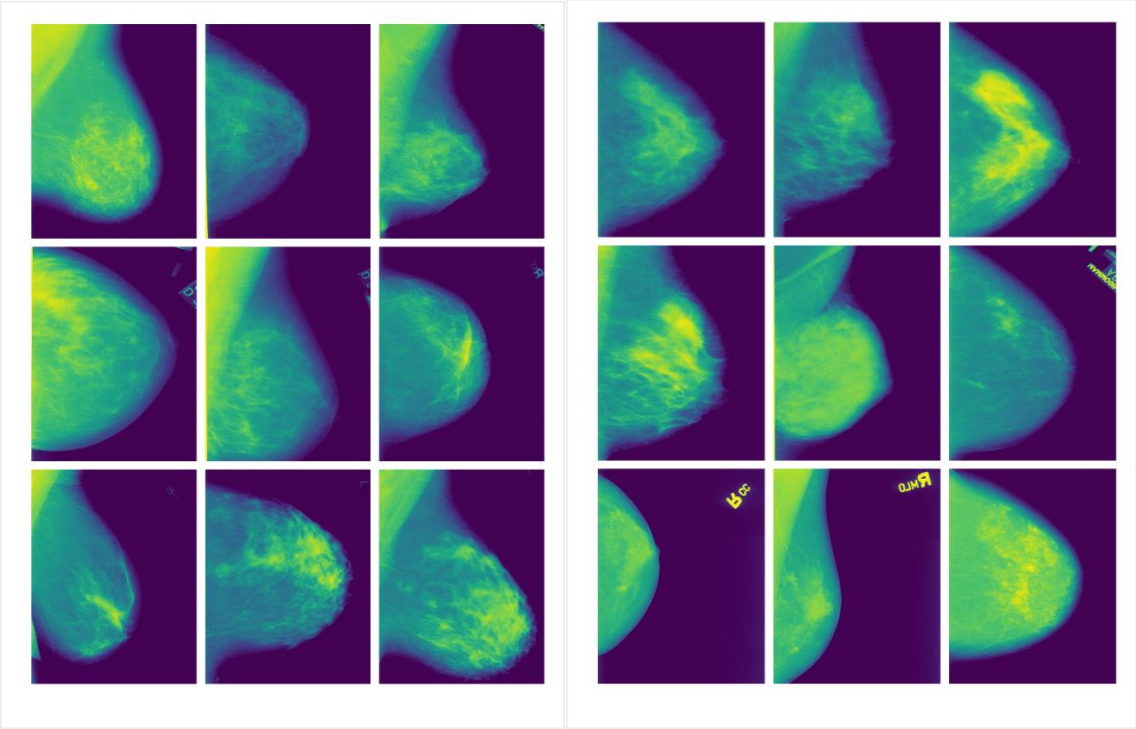


Figure 2.1 An illustration of the breast cancer images and their variations. Images shown have been resampled to the same aspect ratio and same size.

2.1.2 Breast Cancer Tissue Biopsy Dataset Exploration

The BreakHis images are relatively large in their native format. Each image is about about 500 KB on disk and a uniform size of 700x460 pixels, which would generate a feature vector of 315,000 features. To make these images more manageable, we downsized them to 350x230 pixels. This will require a lot less memory and less compute resources. Additionally, we note that the dataset contains about 6 times more ductal carcinoma images than all other classes (average of 136 samples per class). We will use augmentation (adding more samples by rotating and flipping the original images) to balance the samples, and we will test with both non-augmented and augmented datasets to evaluate and compare classifier performance.

2.2 Classification Methods

In this section, we will look at classifying both breast cancer image sets first as two classes (**cancer [class 1]** or **no cancer [class 0]**), then later consider the specific anomalies in each dataset and label the images in a multi-class configuration. For the breast mammogram images, we will label the images in the following four classes:

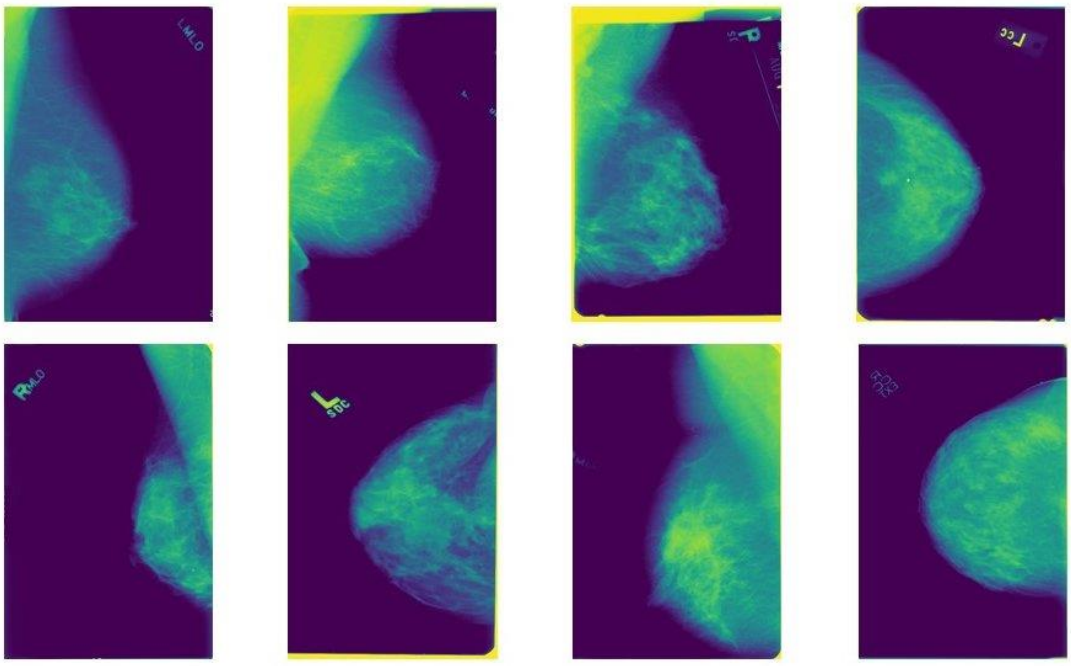
a. **Calcification, Benign** [class 0, tissue calcification, not cancerous]

b. **Calcification, Malignant** [class 1, tissue calcification, cancerous]

c. **Mass, Benign** [class 2, tissue mass, not cancerous]

d. **Mass, Malignant** [class 3, tissue mass, cancerous]

Figure 2.2 shows a sample of the images in each of the four classes used. In the two class machine learning experiments, mass and calcification are grouped in one category, and classified as cancerous or non-cancerous.



(a) Calcification, Benign, class 0 (b) Calcification, Malignant, class 1 (c) Mass, Benign, class 2 (d) Mass, Malignant, class 3

Figure 2.2 Sample mammogram images showing the 4 anomalies being classified. (a) Calcification, Benign [class 0: tissue calcification, not cancerous], (b) Calcification, Malignant [class 1: tissue calcification, cancerous], (c) Mass, Benign [class 2: tissue mass, not cancerous], (d) Mass, Malignant [class 3: tissue mass, cancerous].

For the breast cancer tissue biopsy images, we will label the images in the following seven classes:

a. **Ductal Carcinoma** [class 0, malignant]

b. **Lobular Carcinoma** [class 1, malignant]

c. **Mucinous Carcinoma** [class 2, malignant]

d. **Papillary Carcinoma** [class 3, malignant]

e. **Fibroadenoma** [class 4, benign]

f. **Phyllodes Tumor** [class 5, benign]

g. **Tubular Adenoma** [class 6, benign]

Figure 2.3 shows a sample of the images in each of the seven classes we defined. In the two class machine learning experiments, images are classified as cancerous or non-cancerous.

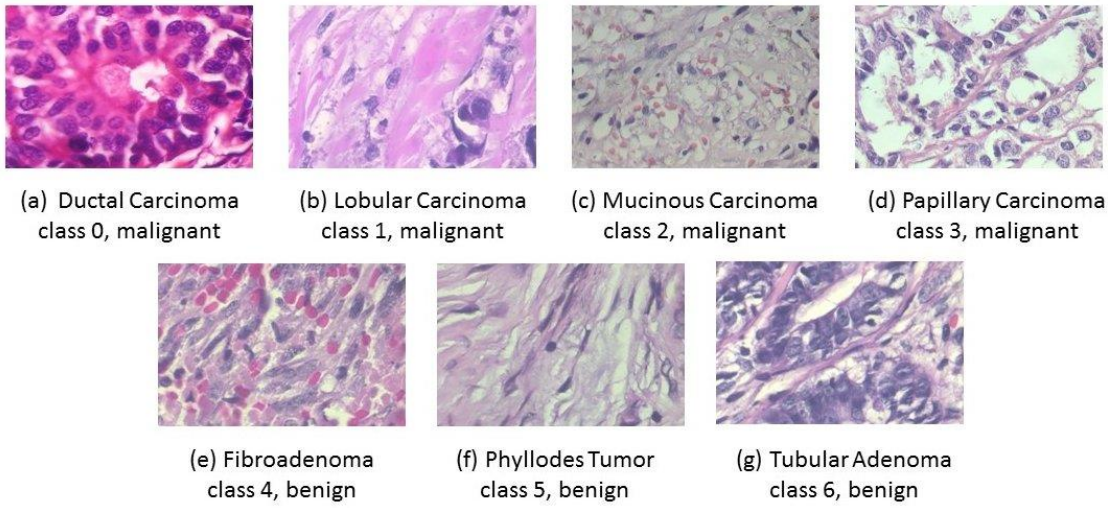


Figure 2.3 Sample biopsy images showing the 7 anomalies being classified. (a) Ductal Carcinoma [class 0, malignant], (b) Lobular Carcinoma [class 1, malignant], (c) Mucinous Carcinoma [class 2, malignant], (d) Papillary Carcinoma [class 3, malignant], (e) Fibroadenoma [class 4, benign], (f) Phyllodes Tumor [class 5, benign], (g) Tubular Adenoma [class 6, benign].

In the multi-class experiments, our classifiers will attempt to detect each of the individual anomaly as described above. Three methods are used for classifying both datasets: Random Forest (RF), Support Vector Machine (SVM) and Convolutional Neural Networks (CNN). In addition, the breast cancer biopsy images are classified using Logistic Regression (LR).

2.2.1 Breast Cancer Mammogram Image Pre-Processing and Preparation

Before our image data can be analyzed, it has to be converted to a format that is acceptable to the classifier functions. As mentioned previously, the raw breast cancer images are not uniform in size or aspect ratio. A decision was made to test with several

image sizes and aspect ratios. For some input datasets, the images were kept in the original 1-channel DICOM format; for others, the images were converted to 3-channel color JPG images. The following image sizes were used (in pixels): 750x500, 256x256, 224x224, 130x80, 96x96.

Once the image parameters are set, the next step is to convert our image data to classifier friendly NumPy arrays (n-dimensional array structure in Python). Each image is flattened out and converted to a feature vector. Looking at the input image resolution, it becomes clear that the resulting vectors will have high dimensionality. For example, a 750x500 DICOM image generates a vector with 375,000 features (which is the total number of pixels in a single image). A 256x256 JPG image generates a vector with 196,608 features (a JPG image has 3 channels). For a DICOM image, the value of each dimension is equal to the HU value of the corresponding pixel. For a JPG image, the value of each dimension is equal to the RGB value of the corresponding pixel. The vectors can then be normalized if needed or required by the classification method.

In order to train and evaluate our classifier, a label vector that stores the true class of each data sample (e.g., [0,1] for two classes, [0,1,2,3] for four classes) is created and used as a true label input along with the sample data. We note that 1,385 samples are labeled as cancerous, and 1,718 samples as non-cancerous, for a total of 3103 samples. When considering the four classes, the samples are distributed as follows:

- Class 0: Calcification, Benign; 885 samples.
- Class 1: Calcification, Malignant; 626 samples.
- Class 2: Mass, Benign; 833 samples.
- Class 3: Mass, Malignant; 759 samples.

2.2.2 Breast Cancer Tissue Biopsy Image Pre-Processing and Preparation

The breast biopsy images will be kept in their original 3-channel PNG format, and since they are already uniform in size, the images are simply downsized and stored in NumPy arrays. Each image is flattened out and converted to a feature vector. At 350x230 pixels, the feature vector will have 80,500 features, reducing memory requirements and training times. The vectors are then normalized if needed or required by the classification method.

In order to train and evaluate our classifier, a label vector that stores the true class of each data sample (e.g., [0,1] for two classes, [0,1,2,3,4,5,6] for seven classes) is created and used as a true label input along with the sample data. We note that for the non-augmented set, 1,232 samples are labeled as cancerous, and 374 samples as non-cancerous for a total of 1,606 samples. Given that the data is imbalanced in favor of one of the cancer types, we augment the data to create a more balanced set. Augmentation is done only 1-fold on the most common class (ductal carcinoma) and 6-fold on all other classes. A combination of horizontal and vertical flip, rotation (180 degrees of rotation is used to maintain the shape of the image) and shifting is performed on the resized images to create more samples. Figure 2.4 shows the augmented versions of a biopsy sample.

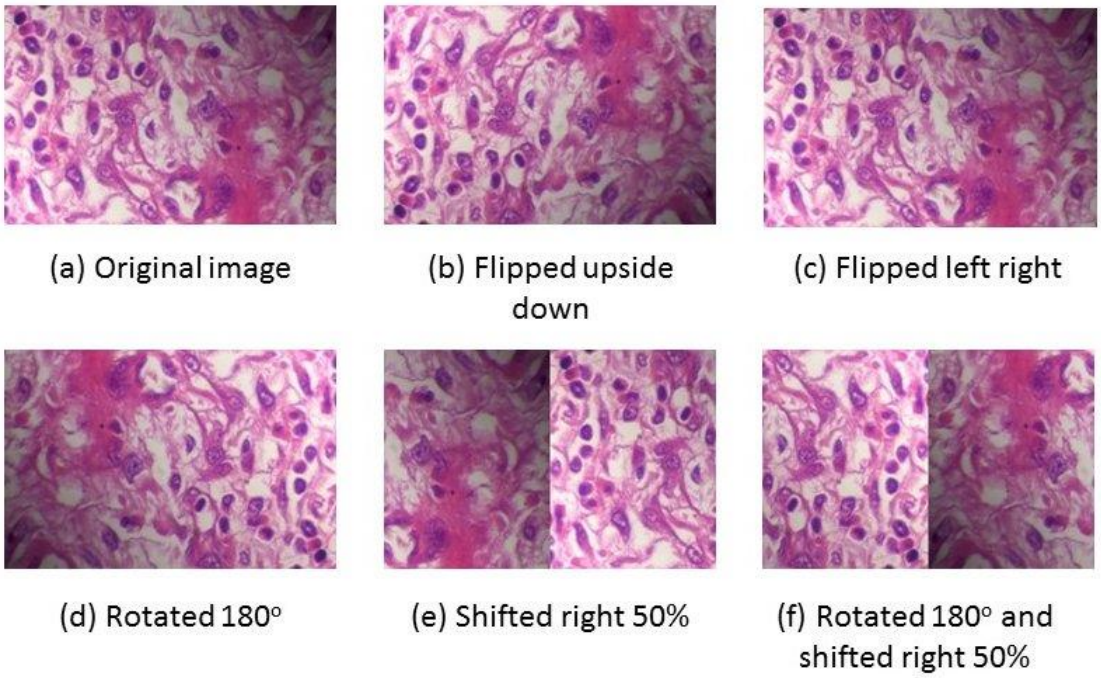


Figure 2.4 Sample biopsy image with various augmentation transformations. (a) Original resized image, (b) Flipped upside down, (c) Flipped left/right, (d) Rotated 180°, (e) Shifted right 50%, (f) Rotated 180° and Shifted right 50%.

Table 2.3 shows the number of breast cancer biopsy images by class in the original non-augmented dataset and the augmented dataset (1,606 vs. 8,120 samples), which is more balanced.

Table 2.3 Tumor Class Distribution in the Original and Augmented Biopsy Datasets

	Original Dataset	**Augmented Dataset**
Fibroadenoma	129 samples	1,032 samples
Phyllodes Tumor	115 samples	920 samples
Tubular Adenoma	130 samples	1,040 samples
Ductal Carcinoma	788 samples	1,576 samples
Lobular Carcinoma	137 samples	1,096 samples
Mucinous Carcinoma	169 samples	1,352 samples
Papillary Carcinoma	138 samples	1,104 samples

2.2.3 Support Vector Machine, Logistic Regression and Random Forest

In this series of experiments, we perform our classification in Python using the Scikit-Learn library which includes Support Vector Machine (SVM), Logistic Regression and Random Forest classifier implementations. In a binary class system, the SVM looks for a line or a hyperplane (in high dimensional space) that separates the two classes. The line or hyperplane is defined by a vector *W*. Figure 2.5 illustrates a binary classification problem which uses a non-linear kernel that maps the data into an alternate feature space where the two-class samples are linearly separable. In our case, we test both SVC (Support Vector Classifier) and Linear SVC, which uses a linear kernel, but also has more flexibility in the choice of penalties and loss functions and usually scales better to large numbers of samples. For SVC, we use the Sigmoid function for the non-linear kernel, and we set the tolerance for our stopping criteria at 10^{-6}.

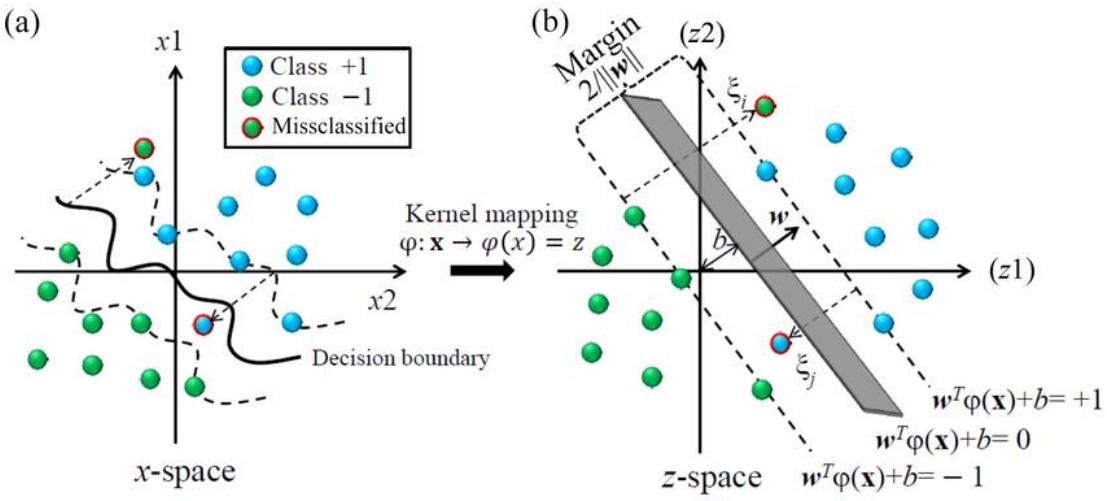


Figure 2.5 Graphical representation of the support vector machine classifier with a non-linear kernel, (a) complex binary pattern classification problem in input space, and (b) non-linear mapping into high-dimensional feature space where a linearly separable data classification takes place.

Source: https://www.researchgate.net/figure/Graphical-presentation-of-the-support-vector-machine-classifier-with-a-non-linear-kernel_fig1_299529384; available via CC BY 3.0 license (https://creativecommons.org/licenses/by/3.0/) (accessed in April 2019)

Logistic Regression is a statistical method for predicting the value of a dependent variable which can only have one of two possible values (0 or 1, true or false, yes or no). For example, it can be utilized to find the probability of success or failure of an event. The formula is obtained by applying the Sigmoid function to the Linear Regression model. Figure 2.6 below illustrates how Logistic Regression can be used to solve binary classification problems.

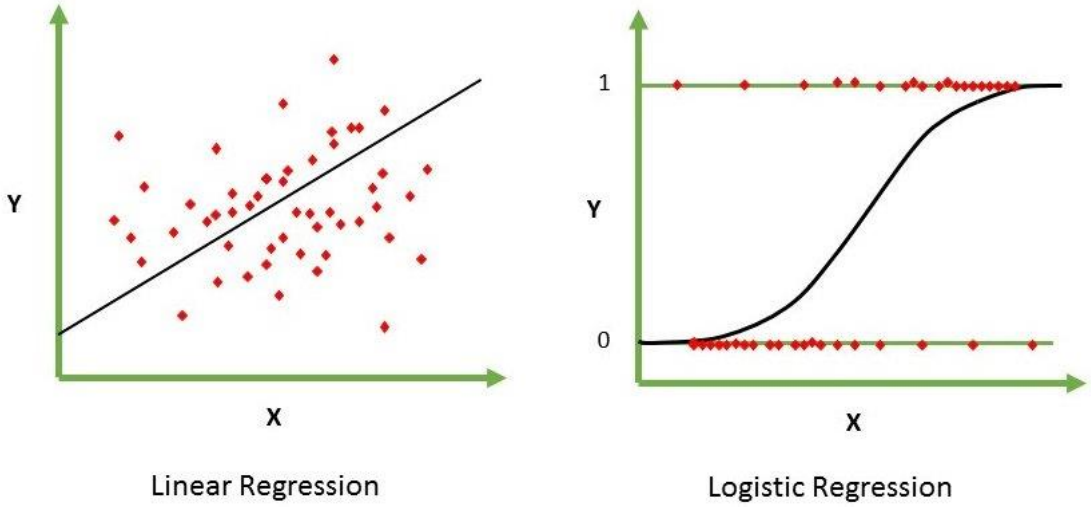


Figure 2.6 Comparison between Linear Regression and Logistic Regression. While Linear Regression generates continuous values, Logistic Regression gives only one of two possible values for the target variable.

Source: Tech Differences. https://techdifferences.com/difference-between-linear-and-logistic-regression.html. (accessed in April 2019)

Random Forest uses an ensemble of decision trees to separate the data and identify the various classes. In our implementation, we use 2000 estimator trees, we place no restriction on tree depth, and we use "*gini*" (Gini impurity; see Appendix B on page 70 for background) as the function to measure the quality of a split (at the decision tree level). We will use the well-known IRIS dataset example to illustrate how the Random Forest algorithm identifies the different flower classes within the set. Figure

2.7 below provides a graphical representation of the decision-making process.

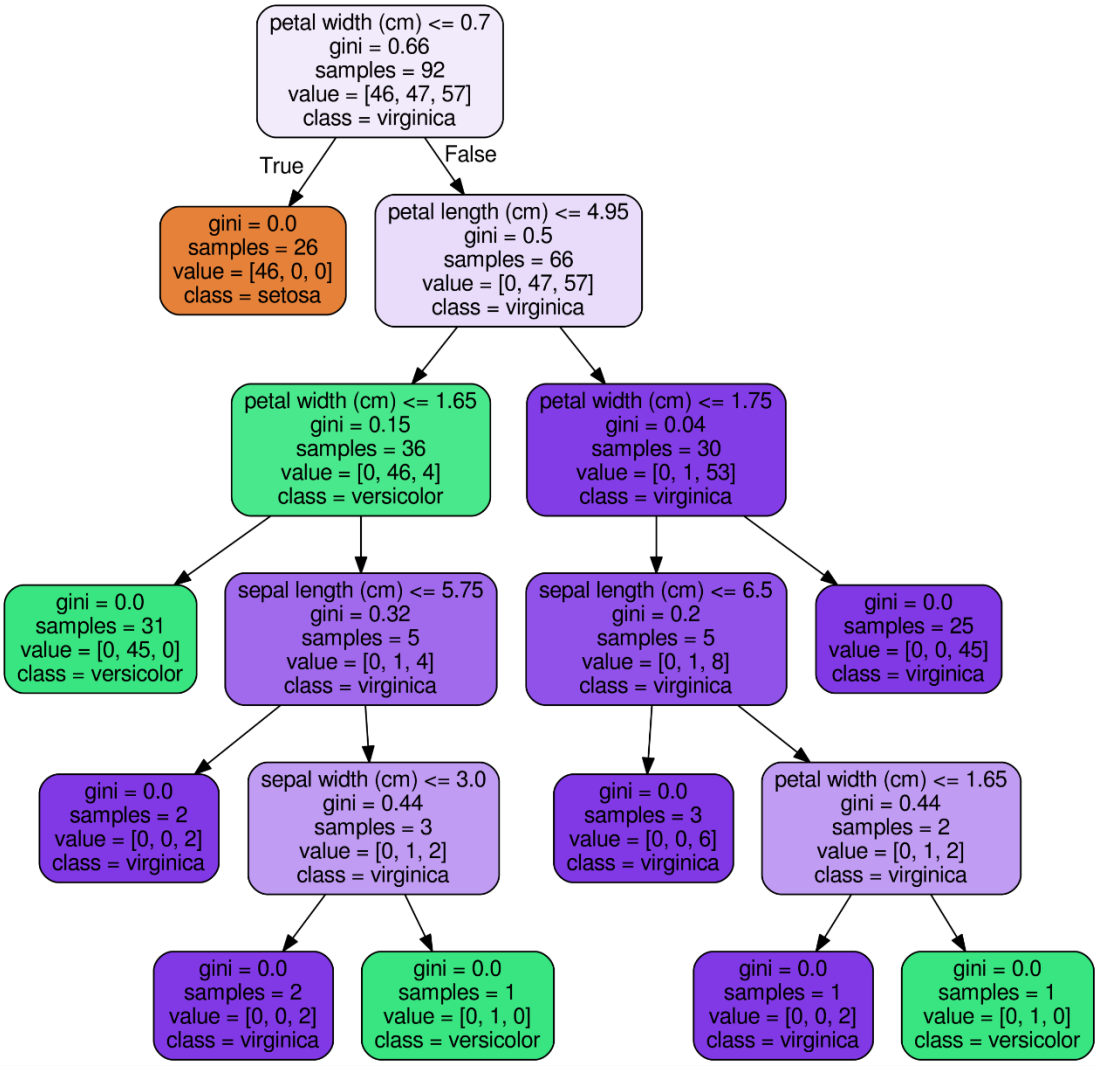


Figure 2.7 Graphical illustration of Random forest decision tress using the IRIS dataset. The box color indicates the type of flower detected.

Source: Will Koehrsen, Data Scientist at Cortex Intel, Data Science Communicator. https://towardsdatascience.com/how-to-visualize-a-decision-tree-from-a-random-forest-in-python-using-scikit-learn-38ad2d75f21c. (accessed in April 2019)

Considering that our NumPy arrays have very high dimensions, we will use the feature selection method to find the top and most relevant features before the data is classified. Feature selection allows us to discard irrelevant information and greatly reduces training times. In our experiment, we will use the Pearson Correlation

Coefficient to select the top 10,000 features. For a pair of random variables (X,Y), the Pearson correlation is given by:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

where ***cov(X,Y)*** is the covariance of random variables ***X*** and ***Y***, $\boldsymbol{\sigma_X}$ and $\boldsymbol{\sigma_Y}$ are the standard deviations of ***X*** and ***Y*** respectively. In our experiment, we will calculate the correlation between the target variable (true label) and each dimension in the feature vector.

2.2.4 Convolutional Neural Networks

Convolutional Neural Networks are typically composed of alternating convolution and pooling layers followed by a final flattened layer. A convolution layer is specified by a filter size and the number of filters in the layer. Each convolution layer performs a moving dot product against pixels given by a fixed filter of a predetermined size. The dot product is made non-linear by passing the output to an activation function such as a sigmoid or ReLU (Rectified Linear Unit) function. Three convolutional neural networks are used to classify the breast cancer images. A simple 3-layer CNN, the popular VGG-16 network, and the recently developed Random Depth-wise Convolutional Neural Network (RDCNN). This new network attempts to learn a feature space with random depth-wise convolutions on which a linear support vector machine or stochastic gradient descent is then applied. Figures 2.8, 2.9 and 2.10 below illustrate the three networks and give a brief description of each.

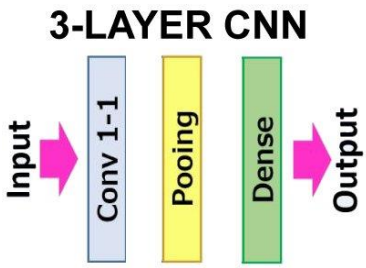


Figure 2.8 Flow diagram of a simple 3-layer convolutional neural network. The network consists of one convolutional layer (with ReLU activation), one pooling layer, and one fully connected dense layer (with Softmax activation).

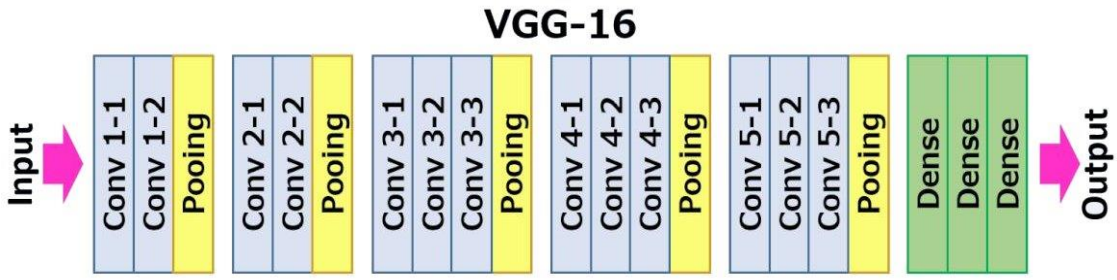


Figure 2.9 Flow diagram of the VGG-16 convolutional neural network. It contains 13 convolutional layers (with ReLU activation), 5 pooling layers and 3 fully connected layers (with ReLU and Softmax activation). The model was proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition."

Source: Muneeb ul Hassan. https://neurohive.io/en/popular-networks/vgg16/ (accessed in April 2019)

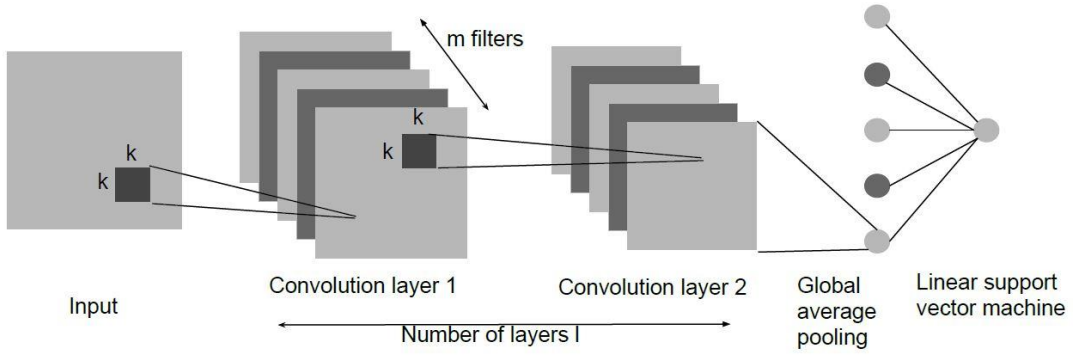


Figure 2.10 A random depth-wise convolutional neural network with two layers. Diagram shows filter size of k, and m = 5 filters in each layer.

Source: Yunzhe Xue and Usman Roshan. http://scinapse.io/papers/2808187103 (accessed in April 2019)

Our CNN experiments are conducted in Python using the Keras library implementations for the simple 3-layer CNN and VGG-16. The loss function used is Categorical Cross-entropy (since it generated the highest accuracy after many trials), and the SGD (**S**tochastic **G**radient **D**escent) optimizer is selected, as it generated the best results with a learning rate of 10^{-4}. For the 3-layer CNN, the following parameters are used: number of filters = 4; kernel dimension = 16x16 pixels; average pooling = 64x64 pixels. The networks are trained using an increasing number of epochs, starting with 10 epochs, up to 50 epochs, in steps of 10. For VGG-16, the following parameters were used: number of filters = 64, 128, 256, 512; kernel dimension = 3x3 pixels; max pooling = 2x2 pixels.

For RDCNN, a TensorFlow implementation is used (faster than Keras). We experimented with the following parameters: models: STL10, Cifar10; number of features: 2500, 10000; iterations: 5000; structures: 7 layers, 25 layers. In addition to using the full x-ray images, we also used cropped tissue images by removing most of the irrelevant background. This reduces the image size and makes the images more uniform.

2.3 Classification Results

We begin with the breast cancer x-ray image classification. For both 2-class and 4-class datasets, the Random Forest methods did best, followed by SVM. The 3-layer CNN, VGG-16 and RDCNN did not do as well. VGG-16 was the worst performer on the breast cancer images. Detailed results for the most significant experiments are given in sections 2.3.1 and 2.3.2 below. (CNN-3 generated all zero output; results not included in table 2.4). A possible explanation for why deep learning methods performed poorly on this dataset is given in Appendix A starting on page 68.

2.3.1 Two-class Breast Cancer Mammogram Image Classification Results

Random Forest (with features selection) did quite well on the large DICOM images (750x500 pixels), with a validation accuracy of almost 68%. Other experiments with smaller images using the same method did not perform as well. The Random Forest results were also checked per class, and those accuracy results were comparable. The SVM methods (both SVC and Linear SVC implementations in Scikit-Learn) did not do as well, but the validation accuracy for Linear SVC came close to 51%, and it was confirmed that the SVM classifier was able to detect the classes (accuracy per class comparable to the overall accuracy). VGG-16 did better than the SVM classifiers but could not beat Random Forest. Several image sizes were attempted, but the result was about the same. Table 2.4 below outlines the most notable results from the dozens of experiments that were run on this dataset.

Table 2.4 Two-Class Mammogram Image Classification Results

Input Data	Classifier	Training Accuracy	Validation Accuracy	Training Time [hours]
750x500; DICOM	SVC	52.32%	43.79%	0.5
750x500; DICOM	Linear SVC	56.97%	50.79%	0.5
750x500; DICOM	**Random Forest**	**98.20%**	**67.80%**	**1**
750x500; DICOM	VGG16	56.20%	55.00%	5
224x224; DICOM	VGG16	54.59%	53.67%	4
230x146; DICOM	VGG16	57.65%	55.71%	4

2.3.2 Four-class Breast Cancer Mammogram Image Classification Results

In the four-class experiments, Random Forest did better than all other methods, but not as well as it did on the two-class datasets. Several experiments were conducted using RDCNN, but the results were not as good as Random Forest. Cropping the images and using more learned features (10K vs. 2.5K) helped boost RDCNN's performance, but not significantly (increased only by 1.13%, see table 2.6). It remained lower than that of the Random Forest method. The 3-layer CNN came in third in performance, with validation accuracies hovering around 30%. Even when data augmentation was used with CNN-3 (by rotating and flipping the images), performance improved only marginally. Once again, VGG-16 did very poorly, and virtually generated a zero output for all classes.

One of the biggest challenges of machine learning in general is training time. We note that conventional methods do not take as long to train as convolutional neural networks, which can have very long training times. While conventional methods typically require an hour or two to train, CNN's can take many hours, and even days when very large datasets are trained on deep networks with a large number of layers. Not counting the feature learning (which can take more than 10 hours), RDCNN has the fastest training times, often measured in minutes. Tables 2.5 and 2.6 below display the most notable results obtained.

Table 2.5 Four-Class Mammogram Image Classification Results: RF, VGG16, CNN3

Input Data	Classifier	Training Accuracy	Validation Accuracy	Training Time [hours]
130x80; DICOM	VGG16	25.00%	27.33%	5
130x80; DICOM	CNN3	29.70%	30.11%	4
130x80; DICOM	CNN3 12,412 samples	30.14%	30.84%	4
224x224; DICOM; [cropped]	RDCNN	64.54%	37.68%	1.24
750x500; DICOM	**Random Forest (FS)**	**97.92%**	**45.66%**	**0.16**
750x500; DICOM	Random Forest (no FS)	97.92%	44.69%	0.27

Table 2.6 Four-Class Mammogram X-Ray Image Classification Results: RDCNN

Input Data	Classifier	Training Accuracy	Validation Accuracy	Training Time [hours]
256x256; JPG	STL10 Model, 2500 features	98.19%	32.00%	0.35
96x96; JPG	cifar10 Model, 2500 features	98.19%	31.72%	0.36
224x224; DICOM	STL10 Model, 2500 features	77.24%	36.55%	0.53
224x224; DICOM; [cropped]	STL10 Model, 10K features	64.78%	36.71%	1.32
224x224; DICOM; [cropped]	**STL10 Model, 10K features, 5K iterations**	**64.54%**	**37.68%**	**1.24**
256x256; DICOM; [full images]	STL10 Model, 10K feat., 5K iterations, 7-layer structure	61.56%	34.14%	0.02
256x256; DICOM; [full images]	STL10 Model, 10K features, 5K iterations, 25-layer structure	61.56%	34.14%	0.02

Next, we present the breast cancer biopsy image classification results. For both 2-class and 7-class datasets, the RDCNN neural network did best, followed by VGG-16, then Random Forest, Logistic Regression and SVM. Detailed results for the most significant experiments are given in sections 2.3.3 and 2.3.4 below.

2.3.3 Two-class Breast Cancer Biopsy Image Classification Results

RDCNN performed very well in the two-class experiments, reaching a validation accuracy of over 92%. VGG did relatively well at about 80% accuracy. Random Forest, SVM and Logistic Regression came in last at around 75%. The biopsy images are clearly a lot easier to classify than the breast x-rays. Surprisingly, SVM and Logistic Regression did worse on the augmented data, but all other classifiers did better with 8,120 total samples. Training accuracies were very high, but the validation accuracies fell short for SMV and Logistic Regression. RDCNN did even better with the augmented dataset, with the accuracy reaching 98%. Although training times for RDCNN were in the order of minutes, it took about 20 hours to generate the new features and then train the Linear SVC classifier. Tables 2.7 and 2.8 below outlines the most notable results from the many experiments that were run on this dataset. A possible explanation for why linear classifications methods (SVM and Logistic Regression) performed worse on the two-class augmented dataset is given in Appendix A starting on page 68.

Table 2.7 Two-Class Breast Biopsy Image Classification Results – No Augmentation

Input Data	Classifier	Training Accuracy	Validation Accuracy	Training Time [hours]
230x350; PNG	SVM	100%	75.64%	0.08
230x350; PNG	Logistic Regression	100%	75.43%	0.13
230x350; PNG	Random Forest	100%	74.44%	0.17
230x350; PNG	VGG16	89.41%	80.49%	1
230x350; PNG	**RDCNN**	**100%**	**92.55%**	**0.04**

Table 2.8 Two-Class Breast Biopsy Image Classification Results – Augmented Data

Input Data	Classifier	Training Accuracy	Validation Accuracy	Training Time [hours]
230x350; PNG	SVM	100%	70.83% ↓	1.5
230x350; PNG	Logistic Regression	100%	71.53% ↓	0.3
230x350; PNG	Random Forest	100%	78.76% ↑	2
230x350; PNG	VGG16	82.81%	81.78% ↑	0.5
230x350; PNG	**RDCNN**	**100%**	**98.77% ↑**	**0.12**

2.3.4 Seven-class Breast Cancer Biopsy Image Classification Results

In the seven-class experiments, it is clear that the classes are harder to detect. The accuracies obtained were below 50%, with the exception of Random Forest which yielded a validation accuracy of about 53%, and RDCNN which reached a 67% accuracy on the original dataset (1,606 samples). SVM and Logistic Regression did worse on the augmented dataset, but the rest of the classifiers did considerably better. Random Forest went from 53% to 63%. VGG-16 went from about 46% to about 71%. RDCNN accuracy increased from 67% to about 95%. New feature generation took about 25 hours.

Again, we note the very long training times for the deep learning methods (VGG16 and RDCNN if we include the feature generation time). Tables 2.9 and 2.10 below display the most notable results obtained. A possible explanation for why linear classifications methods (SVM and Logistic Regression) performed worse on the seven-class augmented dataset is given in Appendix A starting on page 69.

Table 2.9 Seven-Class Biopsy Image Classification Results – No Augmentation

Input Data	Classifier	Training Accuracy	Validation Accuracy	Training Time [hours]
230x350; PNG	SVM	98.60%	43.79%	0.13
230x350; PNG	Logistic Regression	98.60%	43.79%	0.25
230x350; PNG	Random Forest	98.60%	53.42%	0.13
230x350; PNG	VGG16	98.60%	52.48%	0.5
230x350; PNG	**RDCNN**	**98.60%**	**67.08%**	**0.28**

Table 2.10 Seven-Class Biopsy Image Classification Results – Augmented Data

Input Data	Classifier	Training Accuracy	Validation Accuracy	Training Time [hours]
230x350; PNG	SVM	99.49%	35.45% ↓	4
230x350; PNG	Logistic Regression	99.49%	35.63% ↓	3
230x350; PNG	Random Forest	99.49%	63.05% ↑	0.5
230x350; PNG	VGG16	99.14%	70.61% ↑	7.5
230x350; PNG	**RDCNN**	**99.49%**	**95.38% ↑**	**1**

CHAPTER 3

BRAIN LESION MRI IMAGE CLASSIFICATION

3.1 Description of Datasets Used

Our OASIS dataset is a subset of the free and publicly available set provided by the Open Access Series of Imaging Studies (OASIS) Brains Project, which makes neuroimaging datasets freely available to the scientific community. “By compiling and freely distributing this multi-modal data, the hope is to facilitate future discoveries in basic and clinical neuroscience,” say the authors. The description adds that it “consists of a cross-sectional collection of 416 subjects aged 18 to 96. For each subject, 3 or 4 individual T1-weighted brain MRI scans obtained in single scan sessions are included. The subjects are all right-handed and include both men and women.” (OASIS: D. Marcus, et al.). The OASIS samples used in our experiment have no lesion masks included.

Our MCW dataset is from the Medical College of Wisconsin (MCW) in Milwaukee, Wisconsin. We obtained a subset of the data as part of a collaboration research project between NJIT and Rutgers University. The authors describe the 45 participants as “patients with focal encephalomalacia from chronic left hemisphere stroke (males and females). Participants had to have at least minimal ability to read aloud, defined as greater or equal to 10% accuracy in single word reading, but were otherwise included regardless of behavioral profile. All participants were at least 180 days post-stroke, native English speakers, and right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971).” (Binder et al., Brain, 2016).

3.1.1 OASIS Dataset Exploration

Our OASIS subset contains 34 subjects of lesion-free brain MRI scans. The 3D scans are all the same size: 256x256x128 pixels, which is a volume of 128 2D images (or slices), each slice being 256x256 pixels. Additionally, all OASIS images are oriented the same way, have the skull intact, as well as non-brain matter tissue. These are the raw images produced by the scanner. A few sample slices from the same subject are shown in Figure 3.1 below for illustration purposes.

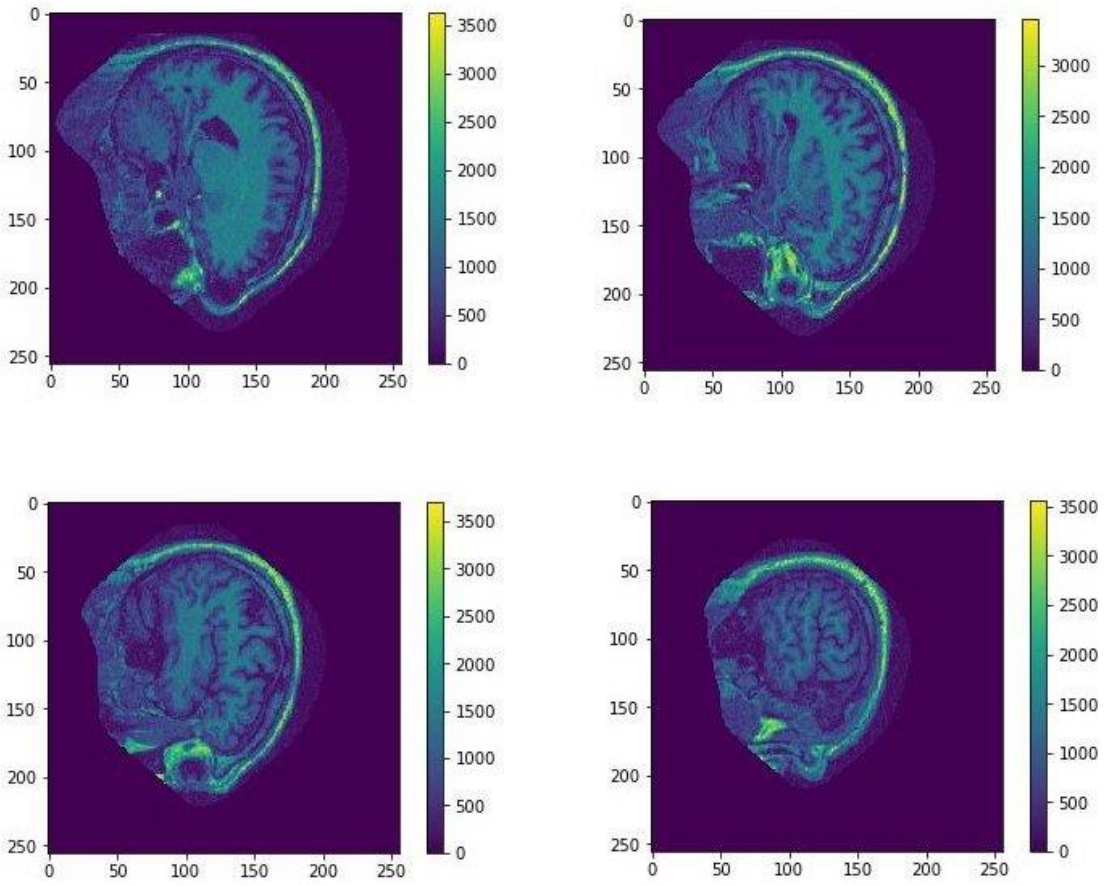


Figure 3.1 Sample lesion-free brain MRI slices from the OASIS dataset.

3.1.2 MCW Dataset Exploration

The MCW subset included in our study has 20 subjects. The 3D scans are all the same size: 256x256x136 pixels, which is a volume of 136 2D images (or slices), each slice being 256x256 pixels. Additionally, all MCW images are oriented the same way, appear to have the skull intact, as well as non-brain matter tissue. We note that these images are not in the same reference space as the OASIS images. We will discuss this in more detail in the next section when we describe the classification experiments which will use a combined dataset containing both OASIS and MCW samples. For each patient, we will use two images to construct our dataset. An anatomical scan (T1 weighted image), and the corresponding 3D lesion mask, which is traced by a human specialist. Both images are used for preparing the image data for the classifier and labeling it. Figure 3.2 shows two sample MCW MRI slices from the same subject and their corresponding true lesion masks. We note that the MCW images are not oriented the same way as the OASIS images and will need to be re-oriented (or re-aligned) before they can be combined in one dataset to be fed to the classifier.

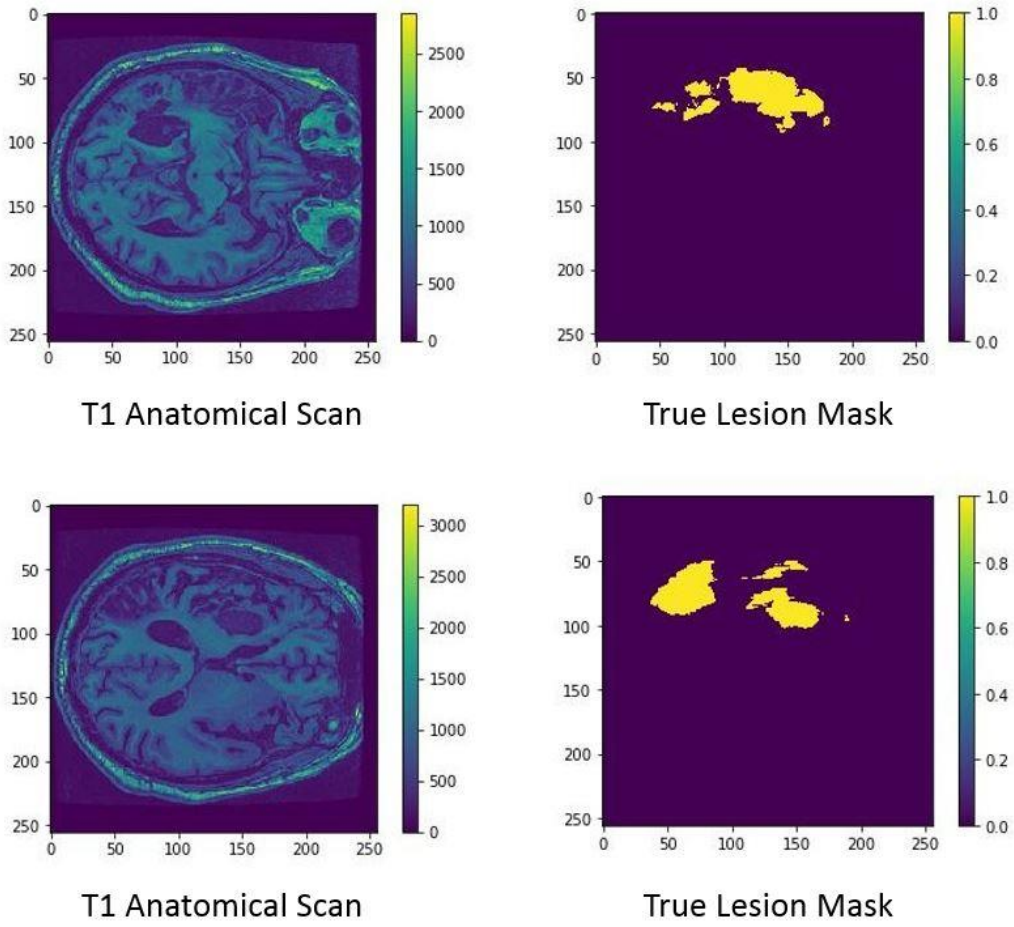


Figure 3.2 Sample brain MRI slices and their lesion masks from the MCW dataset.

Overall, the MCW dataset appears to have large lesions. These are chronic lesions, which means that the scans were taken a long time (about 6 months) after the stroke occurred. The graph in figure 3.3 below shows the lesion volume distribution for all samples used in this study.

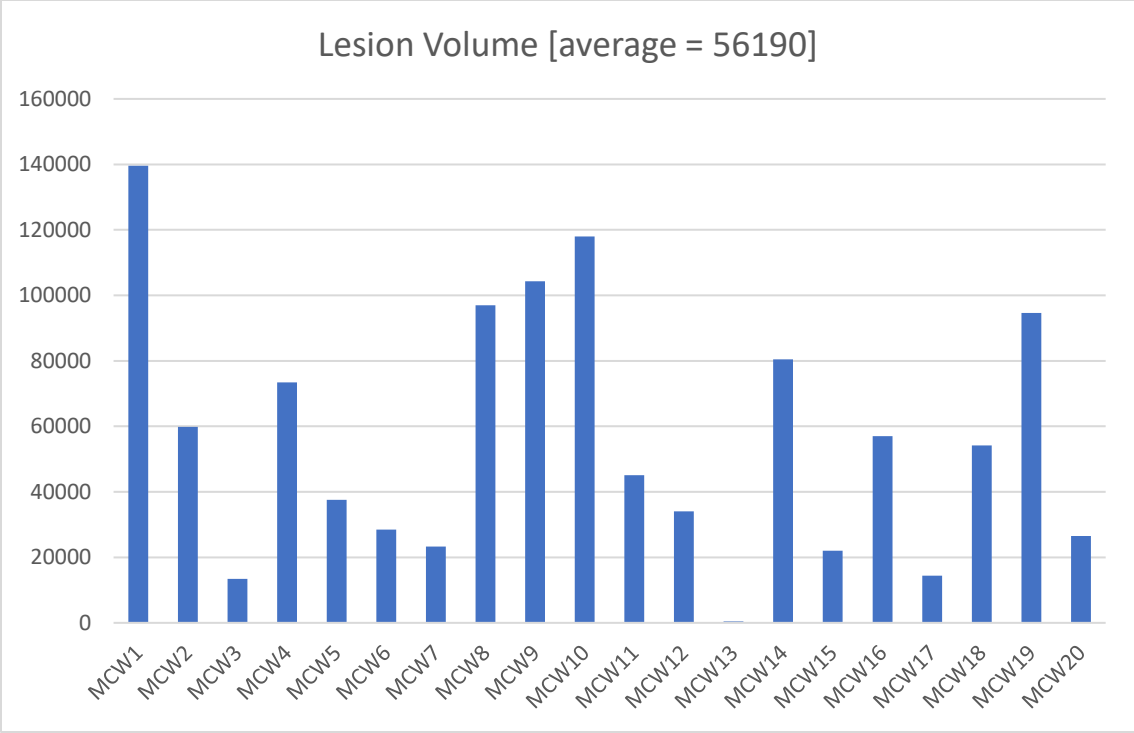


Figure 3.3 Lesion volume distribution (in voxels) of all MCW MRI scans included in this study. The lesions are chronic, and therefore appear to be relatively large.

3.2 Classification Methods

In this section, we will look at classifying the brain MRI images automatically (lesion or no lesion) using three methods: Random Forest (RF), Support Vector Machine (SVM) and Convolutional Neural Networks (CNN). The machine learning models used were trained using a combined dataset containing both healthy and unhealthy scans (80% of the total samples) and tested on the remaining 20% of the data.

3.2.1 Data Pre-Processing and Preparation

In the experiments that follow, we create our classification datasets based on individual 2D slices taken from the 3D MRI scans, as well as patches taken from the 2D slices. To perform two-class classification (lesion/no lesion), we combine both OASIS and MCW

images into one dataset. The lesion-free OASIS images are used as controls and the MCW images as cases. Before the images can be combined, some pre-processing is required. The images have to be skull stripped, all non-brain matter removed, and they need to be re-aligned to the same template space. As explained previously, the raw images are not oriented the same way, nor would the 2D slices be the same size if simply rotated. This is an important step to be able to create a uniform set that can be used as a classifier input. The template used to warp the raw images will generate scans of shape 193x229x193, which can be viewed as a volume of 229 slices, each slice being 193x193 pixels.

The next step is to carefully choose a lower boundary and upper boundary for the slices to be included in our dataset. It turns out that the top and bottom 15% of the slices can be discarded, as they do not contain any significant brain tissue, with most of those slices containing no tissue at all. The last step in the image preparation stage is to trim the slices to reduce the background area, the overall data size, and therefore reduce training times. The final image size in the dataset is 145x170 pixels. Once the image shape is finalized, we convert our image data to classifier friendly NumPy arrays. Our combined image sets are split into training and test datasets. When the OASIS and MCW images are warped (aligned) to the same template space, they are stored as NIfTI (**N**euroimaging **I**n**f**ormatics **T**echnology **I**nitiative) volumes, which like DICOM images, are 1-channel (HU value based) color images. A 145x170 NIfTI image generates a vector with 24,650 features (equal to the total number of pixels in a single image). Figure 3.4 shows what the final (re-aligned) OASIS and MCW images look like.

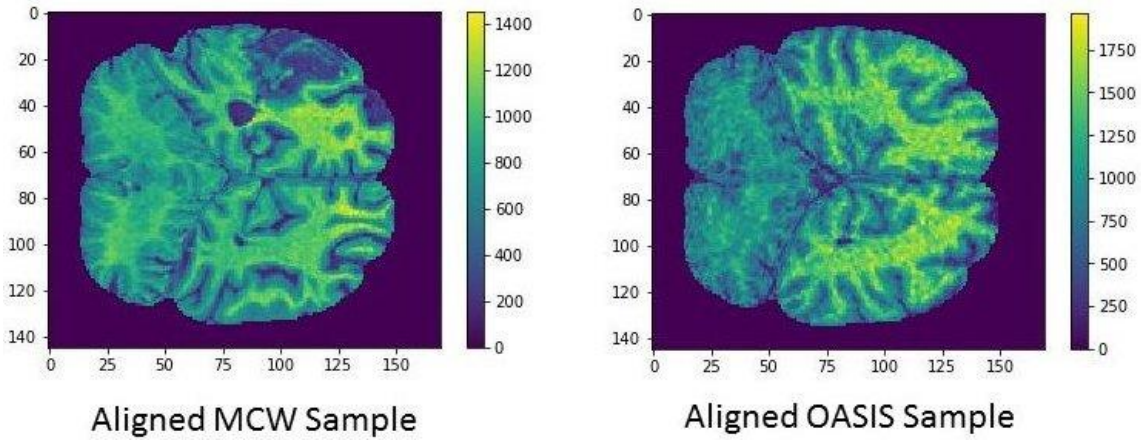


Figure 3.4 Sample brain MRI slices that have been aligned to the same template space. (samples shown are from the MCW and OASIS datasets.)

3.2.2 Full Image Datasets

For the full image datasets, we use individual scan slices from both MCW and OASIS datasets as our samples. All samples obtained from the MCW dataset are given the label (1), meaning that the lesion is present, while samples obtained from the lesion-free OASIS images are given the label (0). Looping through both sets of scans yields a dataset of 8,480 slices. 80% are used for training and 20% for testing. The sample labels break down as follows:

- Out of 1,760 test samples, 640 are labeled as having lesions.
- Out of 6,720 train samples, 2560 are labeled as having lesions.

3.2.3 Patch Datasets

For the patch dataset, we use only four random MCW scan images. The idea is to create a dataset that contains both lesion and lesion-free patches (obtained from slices), which can be used as cases and controls respectively. Each slice is cut in half. Only the upper part (or left side of the brain) is retained. The lower part, which has no damaged tissue,

is discarded. The proposed method is to split the slices of the left side of the brain into patches, and then compare them to their corresponding lesion mask patches. The brain patches are then classified based on the content of their masks. If the mask patch contains damage tracing, the MRI patch is classified as (1). If not, it is classified as (0). The dimensions of the half slices are 128x256. Multiples of 2 are used in creating the patches to make sure the whole image is accounted for. After experimenting with several patch sizes, we determined that 64x64 is the most optimal for this image size, which yields 8 patches per half slice. Figure 3.5 shows 8 patches that come from the same slice and their corresponding masks.

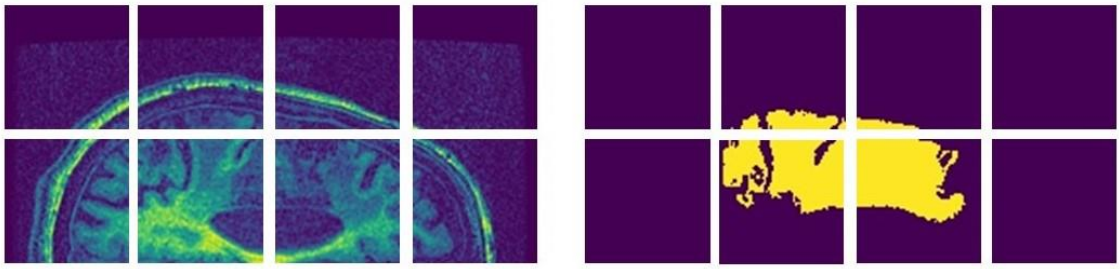


Figure 3.5 Sample brain MRI patches obtained from a single MCW 2D image. The eight patches and their lesion masks shown represent the left side of a single MRI slice.

Similar to the slice datasets, the patch dataset is split into training and test subsets, with 80% for training and 20% for test. Here is how it breaks down:

- Train Data [80%]: 16 MRI scans, 6808 samples, each sample is a 64x64 image
- Test Data [20%]: 4 MRI scans, 2008 samples, each sample is a 64x64 image

3.2.4 Random Forest, Support Vector Machine and Logistic Regression

The classification experiments on the MCW and OASIS datasets are performed in Python using the Scikit-Learn library which includes Random Forest, Support Vector

Machine (SVM) and Logistic Regression classifier implementations described in Section **2.2.2**. For SVM, we use the Sigmoid function for the non-linear kernel, and we set the tolerance for our stopping criteria at 10^{-6}. In our Random Forest experiments, we use 2000 estimator trees, we place no restriction on tree depth, and we use "*gini*" (Gini impurity) as the function to measure the quality of a split (at the decision tree level). For Logistic Regression, we set the maximum iterations at 1000, and use 0.1 for the C parameter.

3.2.5 Convolutional Neural Networks

The Python Keras based VGG-16 network described in section **2.2.3** is used to classify the full MCW/OASIS brain MRI images. The Binary Cross-entropy loss function is used, and the SGD (**S**tochastic **G**radient **D**escent) optimizer is selected again, as it generated the best results on the MRI images with a learning rate of 10^{-4}. For the 3-layer CNN, the following optimized parameters (through experiments) were used: number of filters = 4; kernel dimension = 16x16 pixels; average pooling = 64x64 pixels. The networks are trained using an increasing number of epochs, starting with 10 epochs, up to 50 epochs, in steps of 10. For VGG-16, the following optimized parameters were used: number of filters = 64, 128, 256, 512; kernel dimension = 3x3 pixels; max pooling = 2x2 pixels.

3.3 Classification Results

The flattened vectors have only 24,650 features, which is not a very high number. No feature selection was performed on the data. We start with the full image dataset results. VGG-16 did best, even after just a few epochs of training, followed by Random Forest

then Logistic Regression. SVM did not do as well but was not too far behind. Although the deep learning method produced the highest accuracy, its training time was at least 3 times more than Random Forest, and 6 times more than the other methods. Detailed results for the most notable experiments are given in tables 3.1 below.

Table 3.1 Full Image Brain MRI Classification Results

Input Data	Classifier	CV / Training Accuracy	Validation Accuracy	Training Time [hours]
145x170; NIfTI	SVM	88.69%	87.57%	0.5
145x170; NIfTI	Logistic Regression	89.37%	88.13%	0.5
145x170; NIfTI	Random Forest	91.06%	88.75%	1
145x170; NIfTI	**VGG16**	**98.68%**	**94.63%**	**3**

For the patch image dataset, no feature selection is performed with only 4,096 features. The accuracies obtained were not as good as those for the full image datasets, but the performance order was the same with VGG-16 yielding the best performance and SVM coming in last. Detailed results for the most notable experiments are given in tables 3.2 below.

Table 3.2 Patch Image Brain MRI Classification Results

Input Data	Classifier	CV / Training Accuracy	Validation Accuracy	Training Time [hours]
64x64; NIfTI	SVM	67.00%	61.09%	1.5
64x64; NIfTI	Logistic Regression	67.48%	64.34%	0.5
64x64; NIfTI	Random Forest	79.20%	68.13%	1
64x64; NIfTI	**VGG16**	**82.74%**	**72.05%**	**3**

In addition to classifying the patches (lesion / no lesion), the classification accuracy per slice is also calculated by getting the ratio of patches classified correctly to the total number of patches per slice. For example, consider the following example:

Slice true label vector: [0, 1, 1, 0, 0, 1, 0, 0] (0: no lesion in patch, 1: lesion in patch)

Slice predicted label vector: [0, 1, 1, 1, 0, 1, 0, 0]. In this case, the accuracy for the slice would be 7/8, as 7 out of the 8 patches were classified correctly. The last step is to average all slice accuracies per patient. The results are shown in table 3.3 below.

Table 3.3 Average Brain MRI Scan Classification Results Based on Patch Accuracies

Patient ID	Number of Patches	Averaged Accuracy
1	696	68.53%
2	**344**	**75.58%**
3	544	56.25%
4	424	59.43%

Lastly, we calculate the **DICE** similarity coefficient, **Precision** and **Recall** for every MRI scan in the validation set. Results are shown in table 3.5. These metrics are defined by:

$$DICE = \frac{2 * True\ Positive}{2 * True\ Positive + False\ Positive + False\ Negative} \tag{3.1}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3.2}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3.3}$$

A true predicted class (positive or negative) is one that was predicted correctly. Similarly, classes predicted incorrectly are called false. The confusion matrix in table 3.4 below tabulates these quantities and clarifies what they represent.

Table 3.4 Confusion Matrix Showing Relationships Between True and False Predictions

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative	False Positive
Actual Positive	False Negative	True Positive

Table 3.5 Patch Image Brain MRI Classification Results – Dice, Precision and Recall

Patient ID	DICE Coefficient	Precision	Recall
1	54.09%	60.56%	48.86%
2	**59.29%**	**77.78%**	**47.91%**
3	43.14%	34.20%	48.41%
4	36.37%	44.21%	30.88%

CHAPTER 4

BRAIN MRI LESION DETECTION

4.1 Description of Datasets Used

Our Kessler dataset is provided by the Kessler Foundation as part of a joint research project between NJIT and Rutgers University. The samples are brain MRI scans with lesion, and according to the authors, "data were collected from patients (males and females) who were undergoing inpatient rehabilitation. Inclusion criteria: right-handedness prior to stroke, English as first language, left-hemisphere stroke within five weeks of study enrollment, and ability to carry out the experimental tasks. Exclusion criteria: contraindication to MRI (claustrophobia, pregnancy, extreme obesity, inability to lie flat, implanted ferromagnetic devices), uncorrectable hearing or vision difficulties, dementia, head trauma, tumor, multiple infarcts, severe psychiatric illness, and pre-stroke diagnosis of a reading or learning disability." (Boukrina et al., Frontiers in Human Neuroscience, 2015).

Our ATLAS (Anatomical Tracings of Lesions After Stroke) Release 1.1 is an open-source dataset consisting of 220 T1-weighted MRIs with manually segmented diverse lesions and metadata. The authors state that "the goal of ATLAS is to provide the research community with a standardized training and testing dataset for lesion segmentation algorithms on T1-weighted MRIs. These images were collected from research groups in the ENIGMA Stroke Recovery Working Group consortium." (Sook-Lei Liew, et al.) The images are anatomical MRIs of individuals after stroke and were

collected primarily for research purposes and are not representative of the overall general stroke population.

<u>Note</u>: the MCW dataset described in Chapter 3 is used again in the experiments included in this chapter.

4.1.1 Kessler Dataset Exploration

Our Kessler brain MRI subset contains a total of 28 subjects. The 3D scans are all the same size: 161x191x151 pixels, which is a volume of 151 2D images (or slices), each slice being 161x191 pixels. Additionally, all images are aligned to the standard MNI-152 (**M**ontreal **N**eurological **I**nstitute) template, have the skull stripped, as well as non-brain matter tissue removed. Figure 4.1 shows two samples of the Kessler MRI slices, and their corresponding lesion masks.

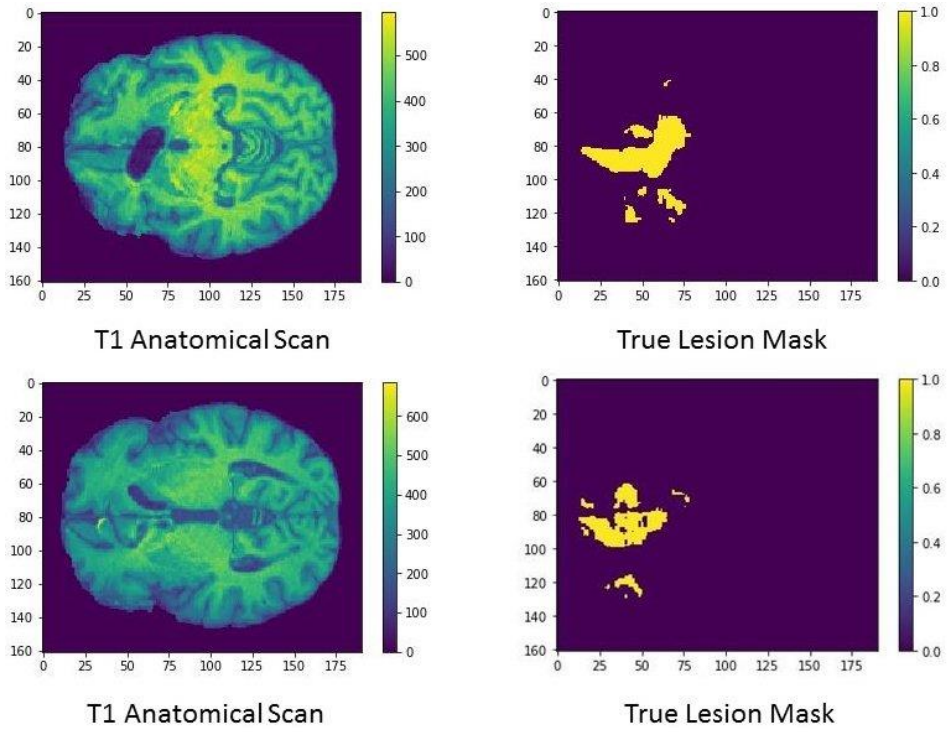


Figure 4.1 Sample brain MRI slices and their lesion masks from the Kessler dataset.

Generally, the Kessler dataset appears to have smaller lesions than the MCW set. These are subacute lesions, which means that the scans were taken a short time after the stroke occurred. Below is a graph that shows the lesion volume distribution for all samples used in this study (figure 4.2).

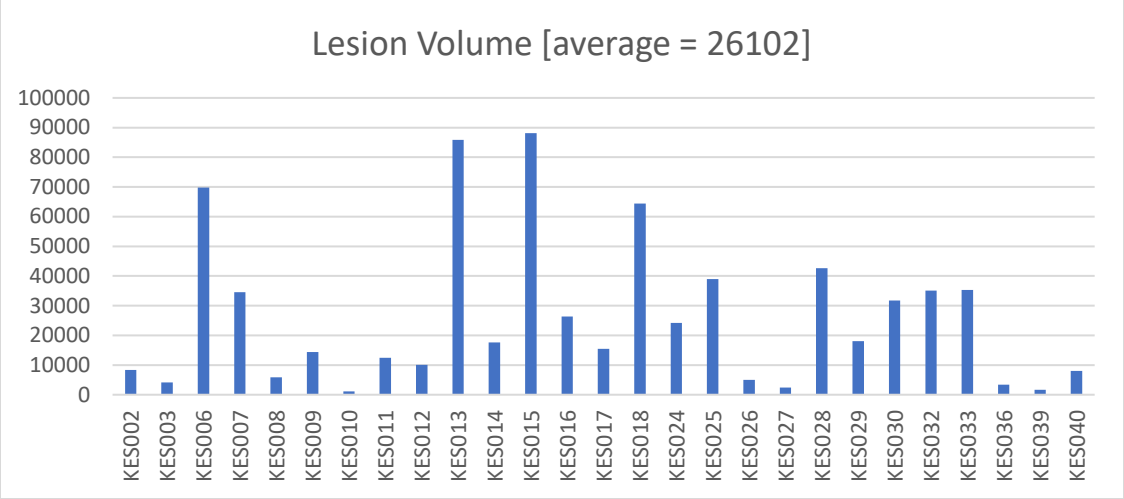


Figure 4.2 Lesion volume distribution (in voxels) of all 28 Kessler MRI scans included in this study. The lesions are smaller than those found in the MCW dataset.

4.1.2 ATLAS Dataset Exploration

Our ATLAS brain MRI subset consists of 54 subjects. These subjects were selected from a pool of 220 scans based on certain conditions required by our study. We only included T1-weighted scans (some T2-weighted scans were provided in the dataset). We also only include scans that have a single lesion (cortical or sub-cortical but not both) in the left hemisphere, and no lesion anywhere else in the brain. It turns out that most neuropsychological studies exclude patients with more than one stroke (indicated by a lesion), likely because multiple strokes can have super-additive effects (Dr. William Graves, Rutgers University – Newark).

The 3D scans are all the same size: 193x229x193 pixels, which is a volume of

193 2D images (or slices), each slice being 193x229 pixels. Additionally, all images are oriented the same way, have the skull intact, as well as non-brain matter tissue. Figure 4.3 shows two samples of the ATLAS MRI slices, and their corresponding lesion masks.

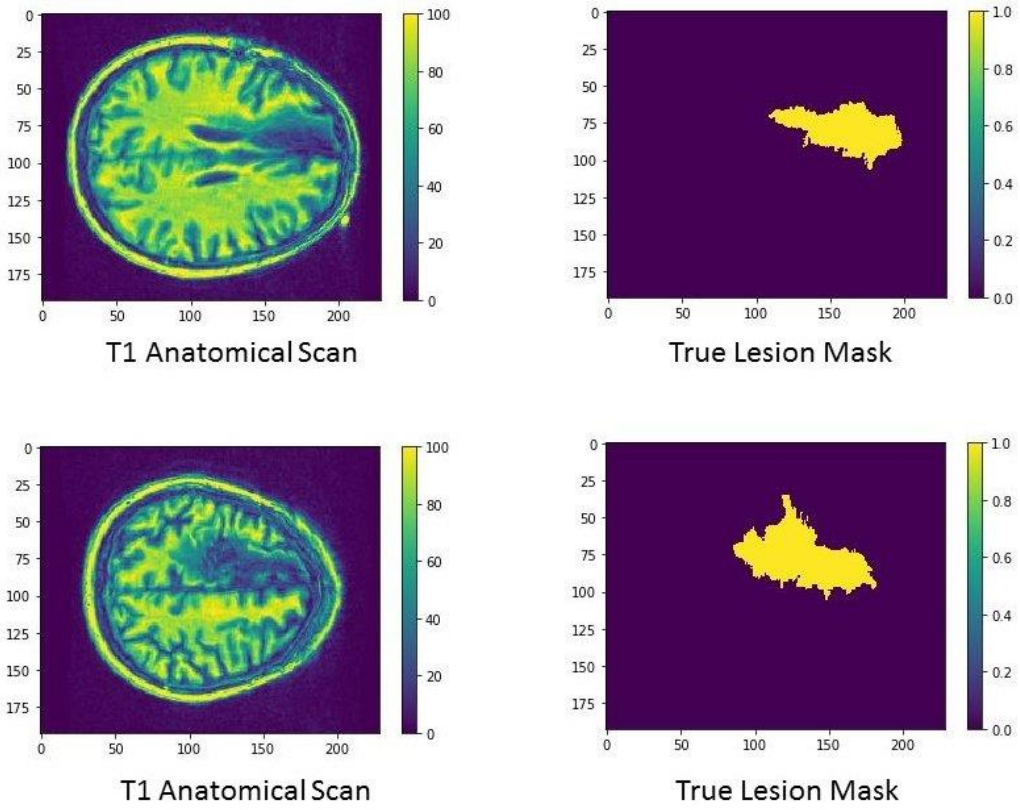


Figure 4.3 Sample brain MRI slices and their lesions mask from the ATLAS dataset.

Overall, the ATLAS dataset lesions appear to be smaller than those in the MCW set. Below is a graph that shows the lesion volume distribution for all samples used in this study.

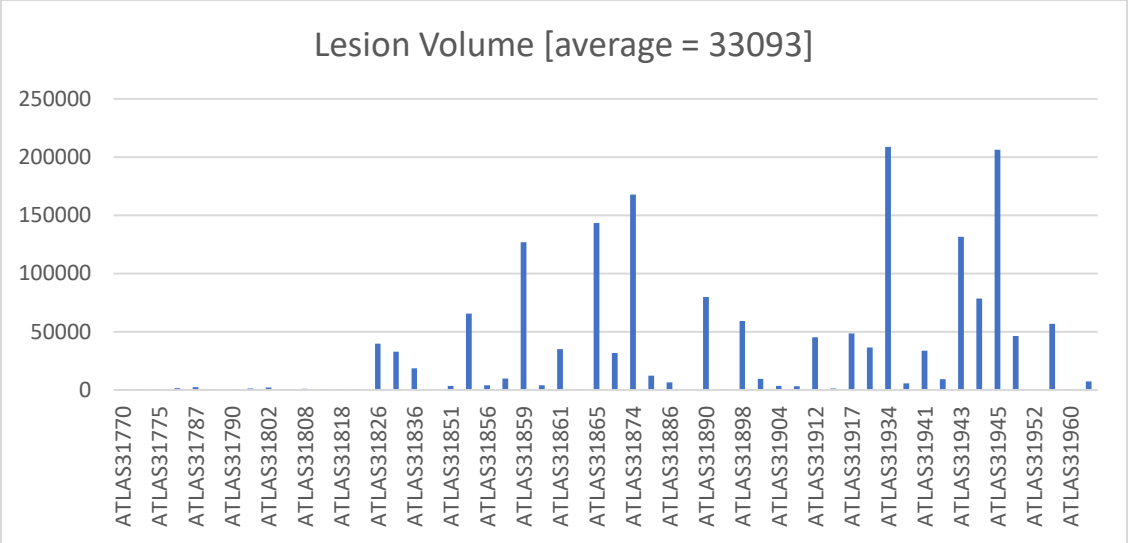


Figure 4.4 Lesion volume distribution (in voxels) of all 54 ATLAS MRI scans included in this study. The lesions are generally smaller than those found in the MCW dataset.

4.2 Lesion Detection Methods

In this experiment, we use pre-trained models to perform automatic segmentation of brain MRI scans with lesions that have been manually traced by a radiologist. The automatic segmentation generates a predicted mask which is compared to the true mask using the Dice Similarity Coefficient (DSC). Once we obtain the DSC values, we plot them against the lesion volumes for all subjects to determine if there is any correlation between the prediction quality and the lesion volume. Lastly, the results of the automatic segmentation methods are compared to a multi-modal CNN developed in a related research project. Note: the data preparation described in the next section is not relevant to the multi-modal CNN method for which data pre-processing and experiments were conducted in a separate study.

4.2.1 Data Pre-Processing and Preparation

The datasets used in the segmentation experiments are not uniform or standardized in any way. Some are raw images, some are skull stripped, and they are not all in the same warp space. After several segmentation attempts, it was determined that our chosen methods perform best when the input MRI scans have the skull intact since they perform skull stripping as part of the segmentation process, and optimal segmentation results are obtained when the MRI scans are aligned to the standard template (MNI-152). Experiments are conducted on both raw and aligned images. The outcomes are discussed in detail in the results section.

To perform the alignment and brain matter segmentation we use a suite of tools referred to as AFNI (**A**nalysis of **F**unctional **N**euro**I**mages), which are C programs originally developed at the Medical College of Wisconsin to work with neuroimaging and are currently maintained by the National Institute of Mental Health. Figure 4.5 shows a sample ATLAS image from a raw scan, and what is looks like after alignment.

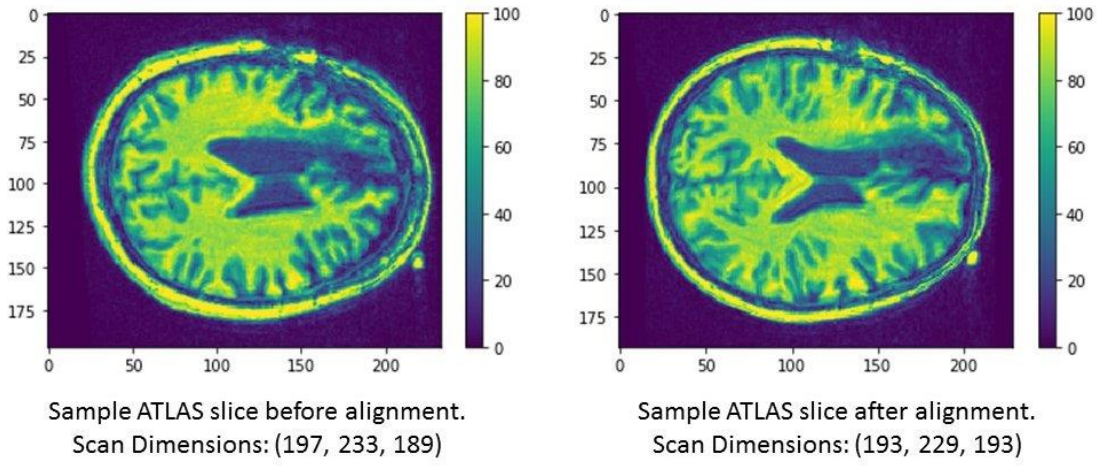


Figure 4.5 Sample ATLAS brain MRI slice before and after alignment. The aligned image is properly centered and is resampled to the dimensions of a pre-selected template.

Unlike the previous classification methods used in this study, the lesion segmentation methods discussed in this chapter take the whole 3D scan as input, are somewhat interactive, and process (test) one scan at a time.

4.2.2 Lesion Gaussian Naïve Bayes (Lesion GNB)

When the arteries to the brain become narrow or get blocked, the blood flow is greatly reduced causing what is called ischemia. The voxel-based Gaussian Naïve Bayes method proposes automatic segmentation of ischemic stroke lesions in individual T1-weighted MRI scans and is distributed as a MATLAB toolbox developed by Dr. Joseph Griffis from Washington University in St. Louis. The pre-processing of the data includes probabilistic tissue segmentation and image algebra to create feature maps encoding information about missing and abnormal tissue. Leave-one-case-out training and cross-validation is used to obtain out-of-sample predictions for each of 30 cases with left hemisphere stroke lesions. This GNB toolbox uses the SPM12 MATLAB software package, which has been designed for the analysis of brain imaging data sequences. The flow chart in figure 4.6 explains the training and testing procedures of the Lesion GNB method, which output a predicted lesion mask that can be compared to the manually traced mask.

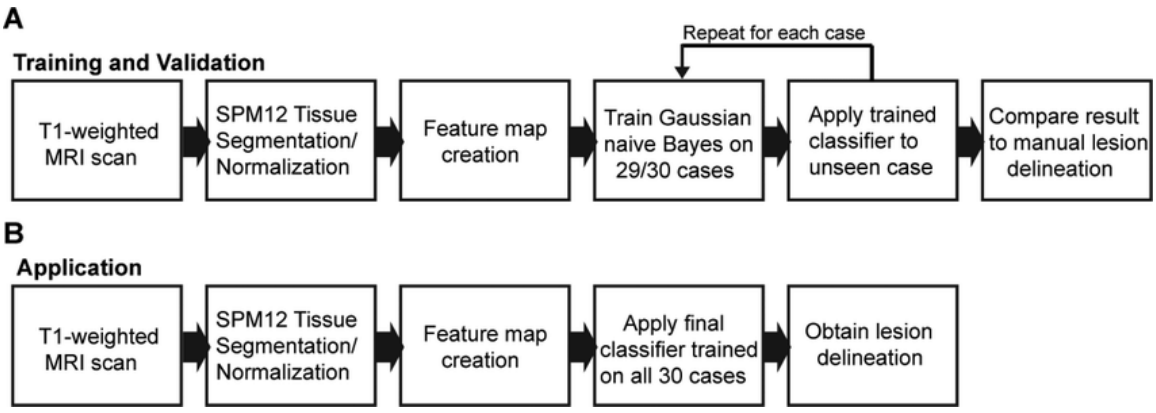


Figure 4.6 Training and testing procedures in the Lesion GNB segmentation method. Diagram (A) shows the training phase. Diagram (B) explains how the trained model is applied on a new, unseen T1-weighed MRI scan.

Source: Dr. Joseph Griffis. https://www.researchgate.net/figure/fig10_282019589 (accessed in April 2019)

Running the trained Lesion GNB model on the unseen MRI scan generates a predicted lesion mask. Before it can be compared to the true mask, the author recommends running a smoothing step and a thresholding step, which is similar to what is normally done on manually delineated lesions. Figure 4.7 shows that post-processing improves the prediction results.

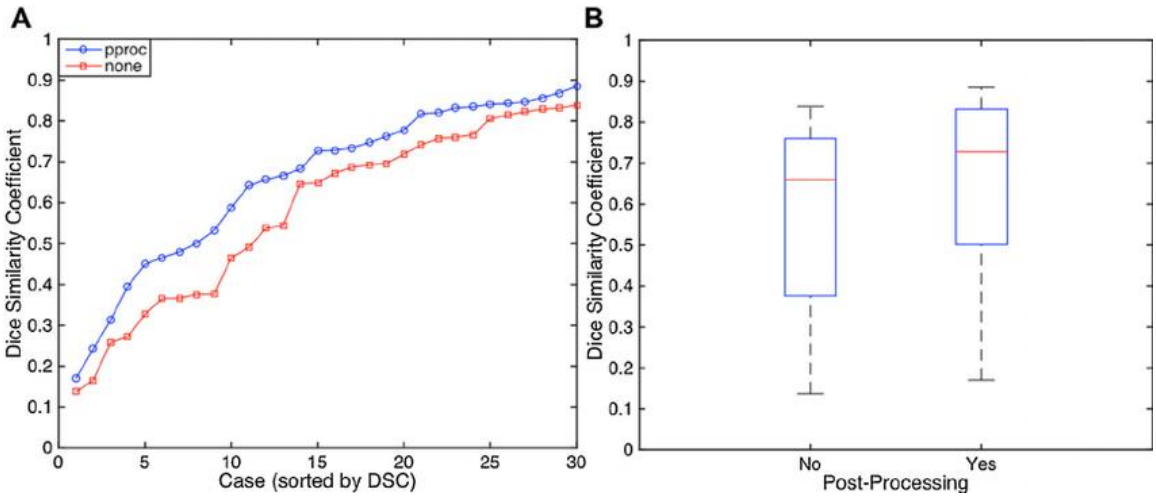


Figure 4.7 Dice Similarity Coefficients (DSCs) for 30 predicted lesions. (A) Plots of Dice Similarity Coefficients (DSCs) for all 30 predicted lesion delineations without (red) and with (blue) post-processing. (B) Boxplots of DSCs for all 30 predicted lesion delineations without and with post-processing.

Source: Dr. Joseph Griffis. https://www.researchgate.net/figure/A-Plots-of-dice-similarity-coefficients-DSCs-for-all-30-predicted-lesion-delineations_fig4_282019589 (accessed in April 2019)

4.2.3 Lesion Identification with Neighborhood Data Analysis (LINDA)

The authors describe LINDA (Lesion Identification with Neighborhood Data Analysis) as "an automated (Random Forest based) segmentation algorithm capable of learning the relationship between existing manual segmentations and a single T1-weighted MRI." LINDA was developed in R programming language and uses the ANTsR (**A**dvanced **N**ormalization **T**ool**s** for **R**) library.

The algorithm was trained on a dataset of 60 left hemispheric chronic stroke patients and was tested with k-fold and leave-one-out procedures. The method was also tested on a dataset of 45 patients, achieving high accuracy rates and confirming its cross-study value according to the authors.

The authors explain the training algorithm as follows: "a series of Random Forest (RF) models are trained at different image resolutions, starting at low resolution and

ending at high resolution. At each level, a matrix containing data from all subjects is used to train the RF model. Each row of the matrix contains information about a single voxel of a single subject and includes values from neighboring voxels on all features as columns. Thus, the model is trained to classify voxels based not only on the value of the voxel itself but also on its neighbors. The status of the voxel (e.g., 1=healthy, 2=lesion) is used as ground truth outcome to train the RFs. Once training is performed at the coarsest resolution level, it is immediately applied to the same subjects in order to obtain a set of additional features consisting of posterior probability maps (i.e., posterior probability of healthy tissue, posterior probability of lesion). These new features are passed to the next (usually finer scale) resolution step together with the existing features, and a new RF model is trained at this resolution. Then, a new set of posterior probabilities is obtained and passed at the successive resolution step. This procedure is repeated hierarchically up to the highest resolution, and RF models are produced at all resolutions (i.e., three RF models for three resolution steps)." (Dorian Pustina, et al.) Figure 4.8 depicts the LINDA workflow, as well as the Random Forest algorithm.

The trained model can be applied to segment new cases. To predict the lesion map in new MRI scans, the algorithm follows the same hierarchical steps, but uses the trained RF models to create the posterior probabilities at each resolution step and predict the unknown outcome/label. Probabilities at the highest resolution are converted into a discrete class map, meaning that each voxel is given the class with the higher probability. (i.e., 60% healthy and 40% lesioned is classified as healthy).

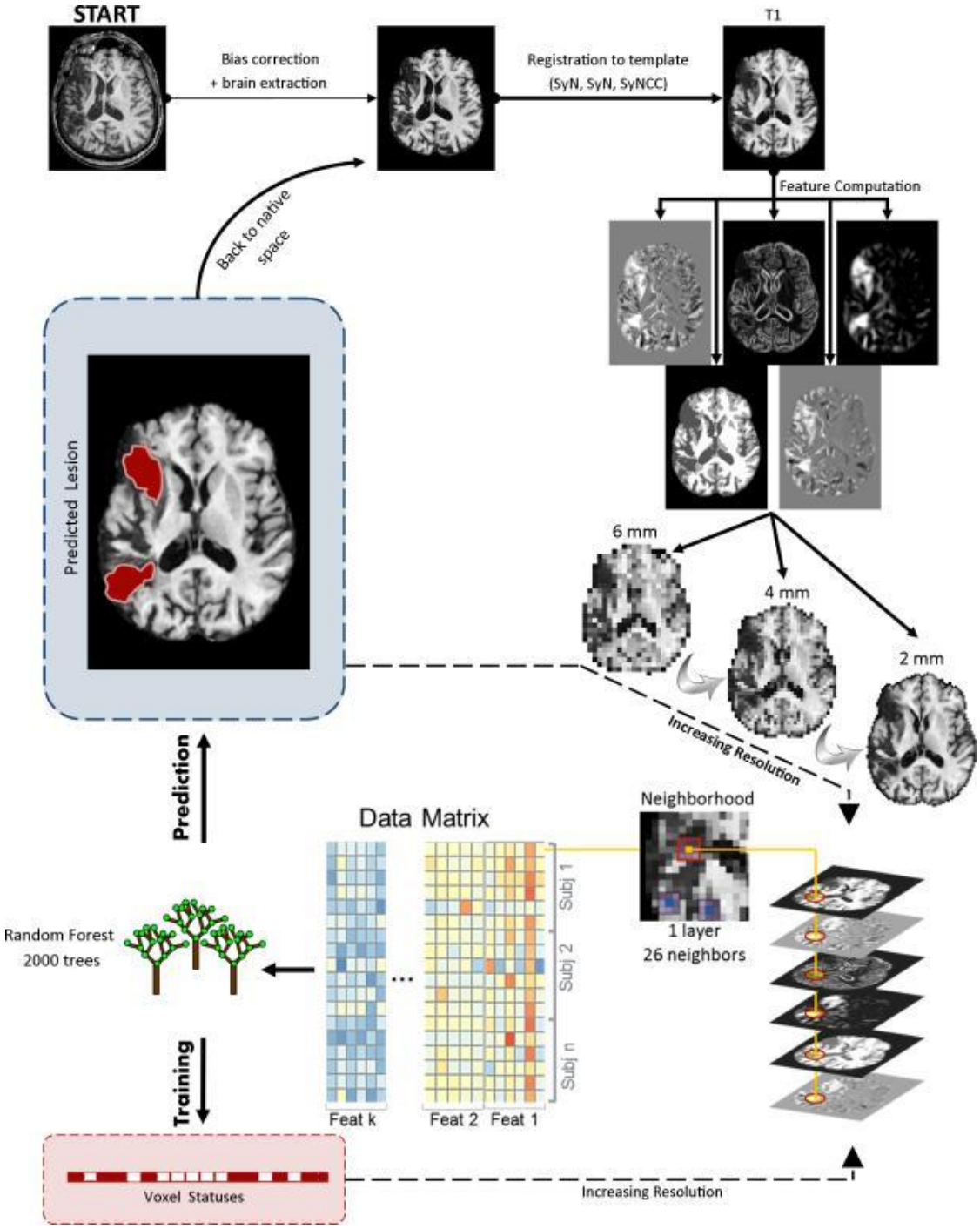


Figure 4.8 Depiction of the LINDA workflow. The multi-resolution voxel neighborhood random forest algorithm is displayed in the lower part of the image.

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4783237/figure/F1/ (accessed in April 2019)

4.2.4 Multi-Modal Convolutional Neural Network (MMCNN)

The Multi-Modal multi-path Convolutional Neural Network (MMCNN) was recently developed in a related joint Deep Learning research project between NJIT and Rutgers University. The system has nine end-to-end U-Nets that take as input 2D slices and examines all three planes (taken from the 3D MRI scan) with three different normalizations. Outputs from the nine paths are collected and the 2D slices are arranged into a 3D volume that is fed into a 3D convolutional neural network for a final prediction. An overview of the MMCNN architecture is shown in figure 4.9.

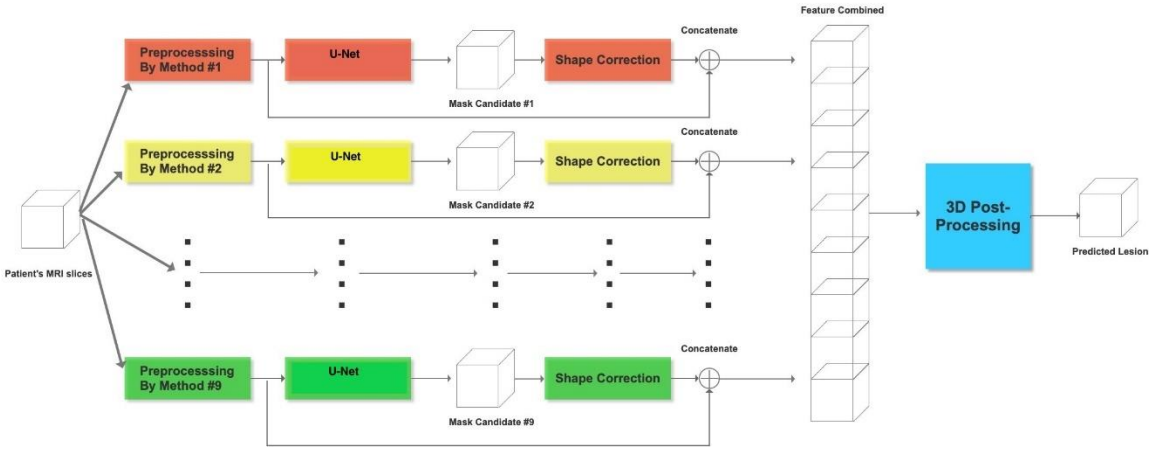


Figure 4.9 Overview of entire 9-path U-Net based MMCNN architecture.

Source: Yunzhe Xue, Fadi G. Farhat, Olga Boukrina, A. M. Barrett, Jeffrey R. Binder, Usman W. Roshan, William W. Graves, "A multi-path 2.5-dimensional convolutional neural network system for segmenting stroke lesions in brain MRI images", Preprint submitted to Neuroimage Clinical, March 31, 2019.

The model was trained and tested on the same datasets used in this chapter: Kessler, ATLAS and MCW. To demonstrate cross-study validity, the network was also trained on a combined Kessler and MCW dataset and tested on ATLAS. Results are discussed in detail in section 4.3.

4.3 Lesion Detection Results

Both Lesion GNB and LINDA methods proved to be sensitive to lesion size in the brain. The bigger the lesion, the better the Lesion GNB and LINDA prediction masks are. We provide correlation coefficients (see page 63 for definition) and p-values (based on t-distribution test that measures the difference in the means between lesion volumes and dice values; p-value < 0.05 is desirable). Results were very poor for small lesions. Lesion GNB performed better than LINDA on the ATLAS and Kessler datasets. For the MCW set, LINDA yielded a higher accuracy. The MMCNN performed better than both LINDA and Lesion GNB, and did significantly better on smaller lesions, making it less sensitive to lesion size, and therefore more useful when segmenting lesions that are hard to detect using other methods.

4.3.1 Lesion Segmentation Predictions using Lesion GNB & LINDA

We start with the ATLAS dataset. The MRI scans used were aligned to warp space with the skull intact, as required by both the Lesion GNB and LINDA methods. Lesions in this set of scans are relatively small with an average lesion volume of 33,093 voxels. For Lesion GNB, the average Dice Similarity Coefficient (DSC) was 0.32. It was almost the same with the LINDA method with 0.30. Figures 4.10 and 4.11 show a graph of the DSC results for both methods. We note that for some subjects, there was no lesion detection at all. All subjects have been sorted by increasing lesion volume. We can clearly see how the models perform better for subjects with large lesion volumes.

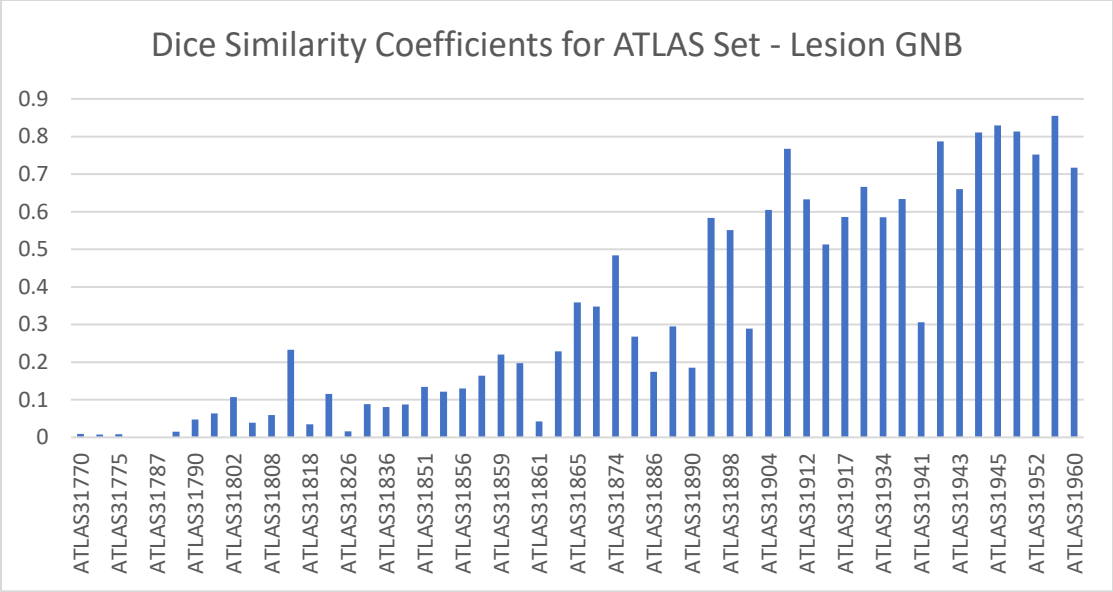


Figure 4.10 Dice Similarity Coefficient (DSC) graph for ATLAS data samples (Lesion GNB method). The average DSC accuracy for this dataset is 0.32 (average lesion volume is 33,093). Subjects are sorted by increasing lesion volume.

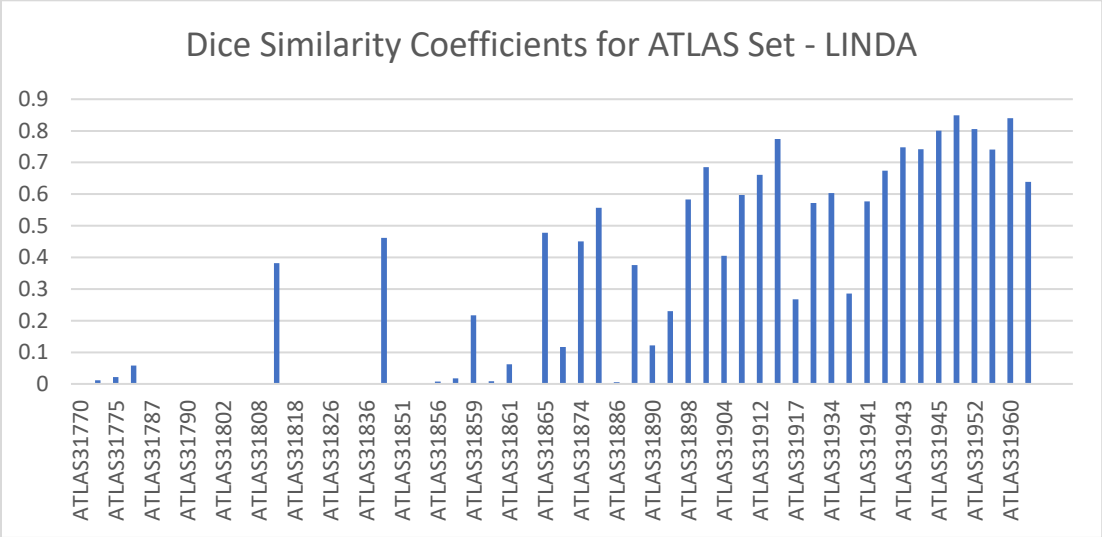


Figure 4.11 Dice Similarity Coefficient (DSC) graph for ATLAS data samples (LINDA method). The average DSC accuracy for this dataset is 0.30 (average lesion volume is 33,093). Subjects are sorted by increasing lesion volume.

Figures 4.12 and 4.13 show the strong correlations (high correlation coefficients and small p-values) between true lesion volume and the DSC value for the ATLAS

dataset. Lesion mask predictions for MRI scans with larger lesions tend to be much better. The graphs below illustrate how both the Lesion GNB and LINDA methods generated very poor or no predictions at all for subjects with small lesions.

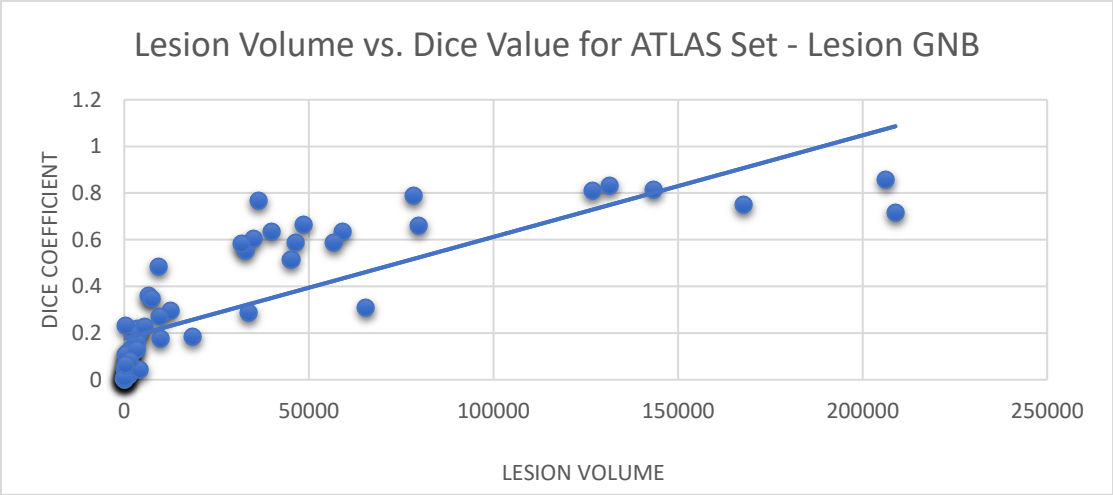


Figure 4.12 A plot of lesion volume versus Dice value for the ATLAS dataset (Lesion GNB method). Correlation Coefficient = 0.804; p-value(t-test) = 0.0000130234.

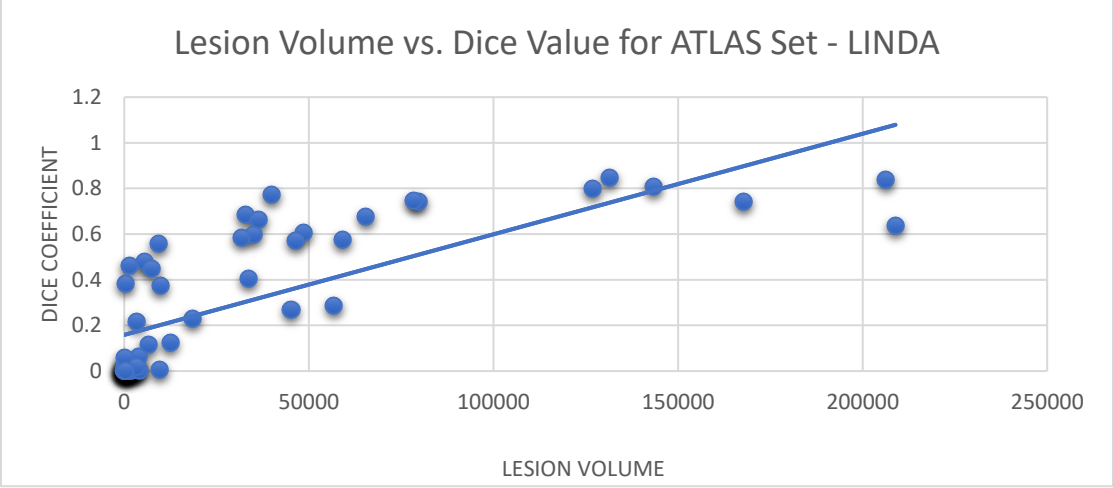


Figure 4.13 A plot of lesion volume versus Dice value for the ATLAS dataset (LINDA method). Correlation Coefficient = 0.745; p-value(t-test) = 0.0000130233.

The Kessler scans we used had the skull already stripped, which was not ideal for our methods. Samples with the skull intact were not available. We note that LINDA did a

lot worse than Lesion GNB on this dataset which has relatively small lesions with an average volume of 26,101 voxels. The average DSC we obtained is 0.25 for Lesion GNB, and a mere 0.02 for LINDA. Again, we observe that the models perform better with large lesion volumes. Results are shown in figures 4.14 and 4.15.

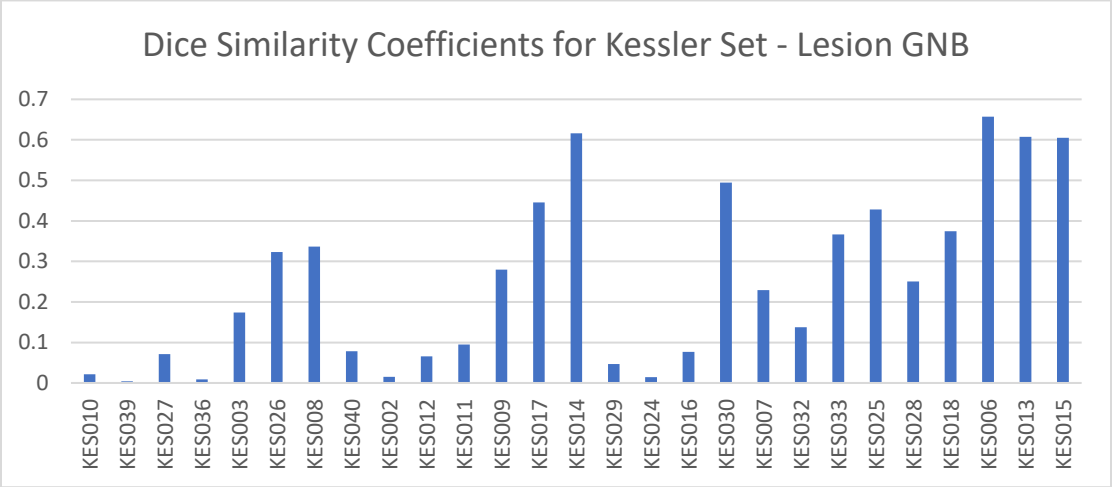


Figure 4.14 Dice Similarity Coefficient (DSC) graph for Kessler data samples (Lesion GNB method). The average DSC accuracy for this dataset is 0.25 (average lesion volume is 26,102). Subjects are sorted by increasing lesion volume.

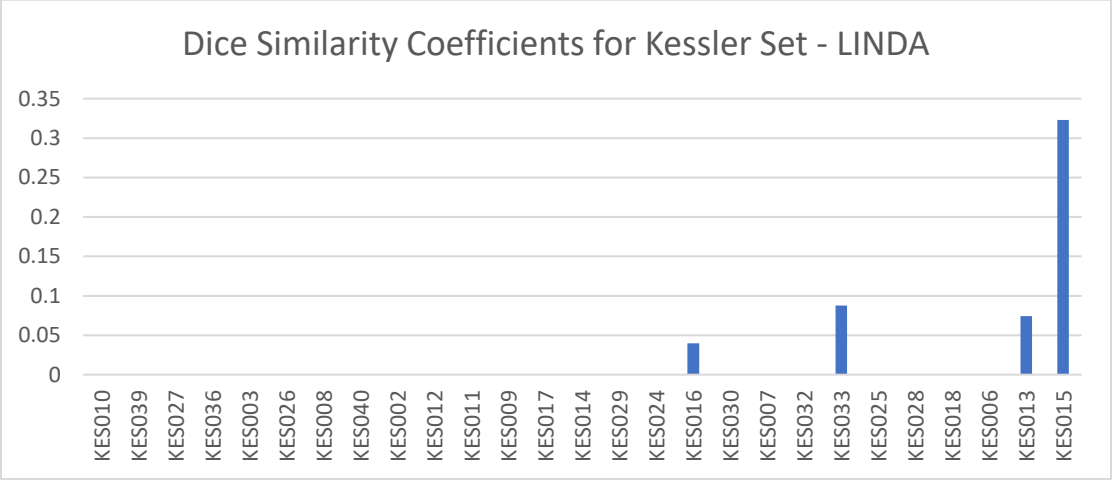


Figure 4.15 Dice Similarity Coefficient (DSC) graph for Kessler data samples (LINDA method). The average DSC accuracy for this dataset is 0.02 (average lesion volume is 26,102). Subjects are sorted by increasing lesion volume.

The strong/moderate correlations (high correlation coefficients and small p-values) between true lesion volume and the DSC value for the Kessler dataset is shown in figures 4.16 and 4.17. Lesion mask predictions for MRI scans with larger lesions tend to be much better. Overall, Lesion GNB performed better on this dataset.

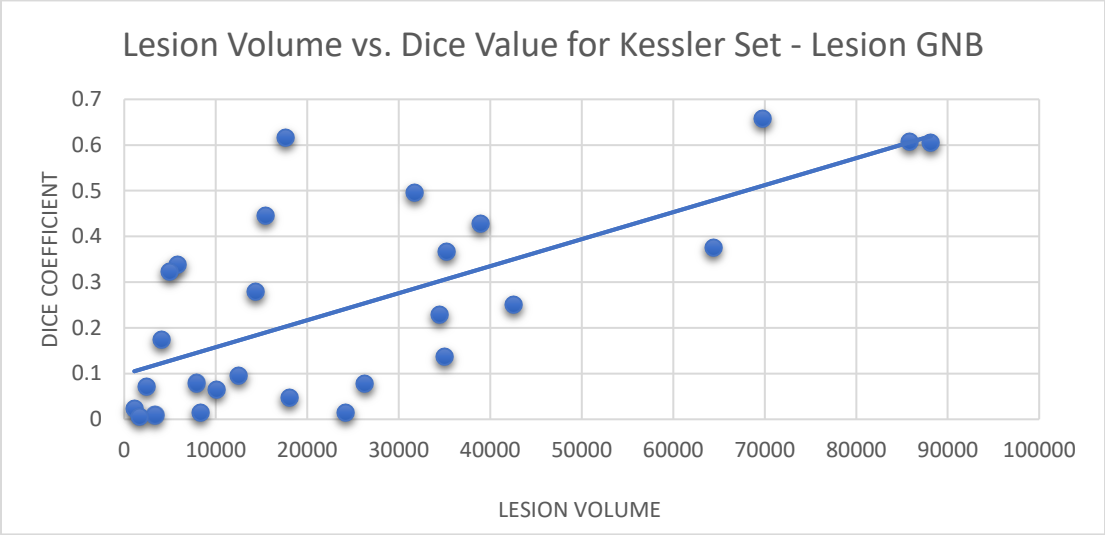


Figure 4.16 A plot of lesion volume versus Dice value for the Kessler dataset (Lesion GNB method). Correlation Coefficient = 0.690; p-value(t-test) = 0.0000064261.

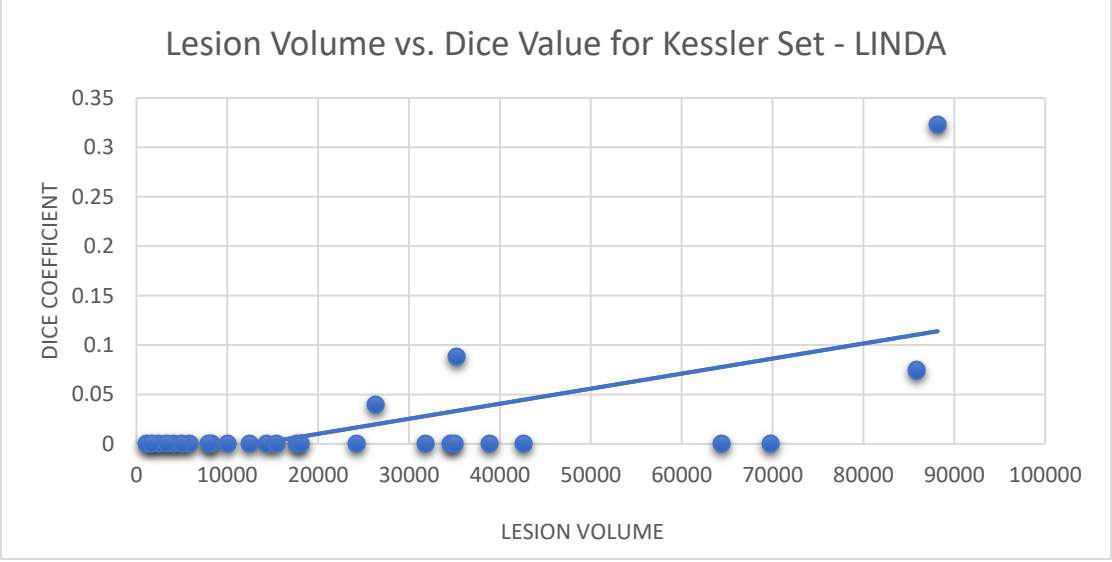


Figure 4.17 A plot of lesion volume versus Dice value for the Kessler dataset (LINDA method). Correlation Coefficient = 0.595; p-value(t-test) = 0.0000064256.

For the MCW dataset, the segmentation methods were run on samples with the skull intact. The average lesion volume on MCW is relatively high at 56,190 (68,078 on raw images), and so the average DSC obtained was 0.48 for Lesion GNB and 0.54 for LINDA. Figures 4.18 and 4.19 display the results obtained from both methods.

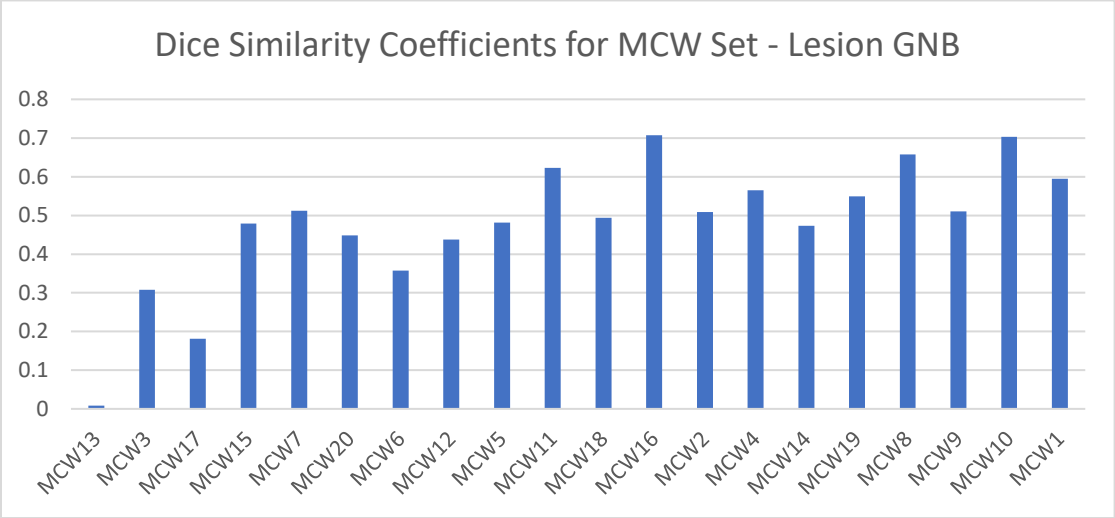


Figure 4.18 Dice Similarity Coefficient (DSC) graph for MCW data samples (Lesion GNB method). The average DSC accuracy for this dataset is 0.48 (average lesion volume is 56,190). Subjects are sorted by increasing lesion volume.

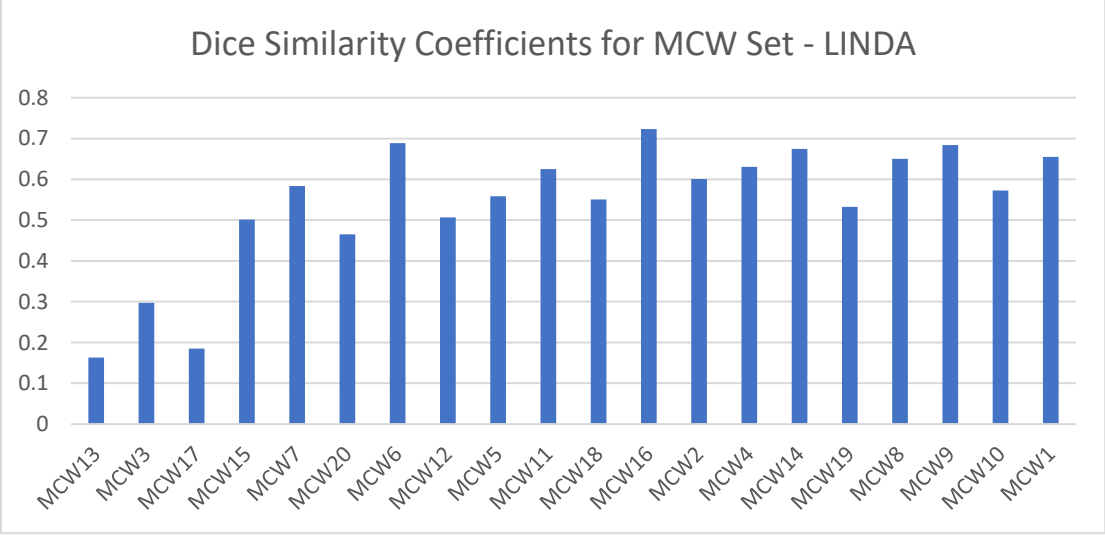


Figure 4.19 Dice Similarity Coefficient (DSC) graph for MCW data samples (LINDA method). The average DSC accuracy for this dataset is 0.54 (average lesion volume is 56,190). Subjects are sorted by increasing lesion volume.

Figures 4.20 and 4.21 show the strong correlations (high correlation coefficients and small p-values) between true lesion volume and the DSC value for the MCW dataset. Predictions for these MRI scans were generally much better that those obtained for ATLAS and Kessler. The graphs below illustrate how both the Lesion GNB and LINDA methods generated valid mask predictions for all subjects in the set.

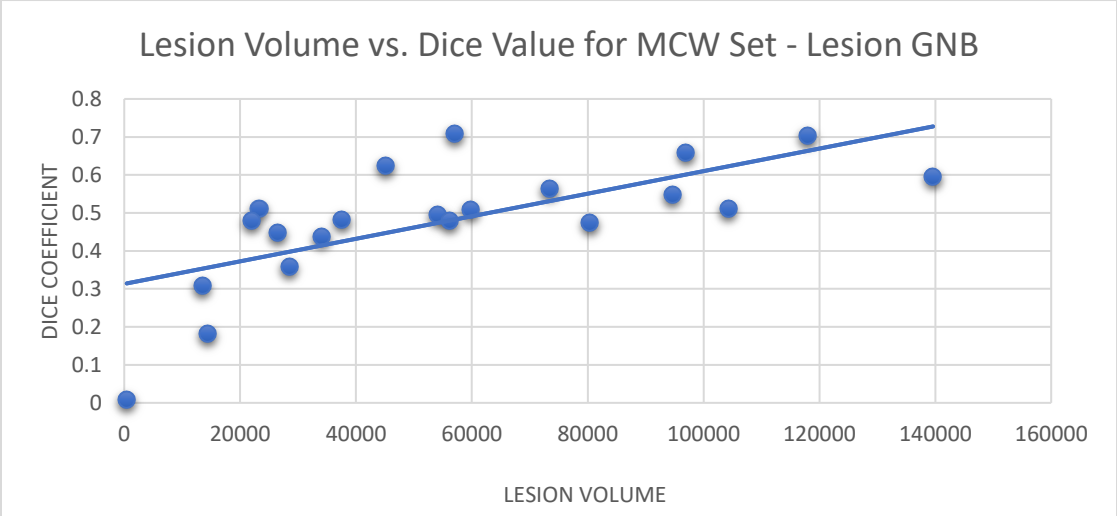


Figure 4.20 A plot of lesion volume versus Dice value for the MCW dataset (Lesion GNB method). Correlation Coefficient = 0.684; p-value(t-test) = 0.0000016864.

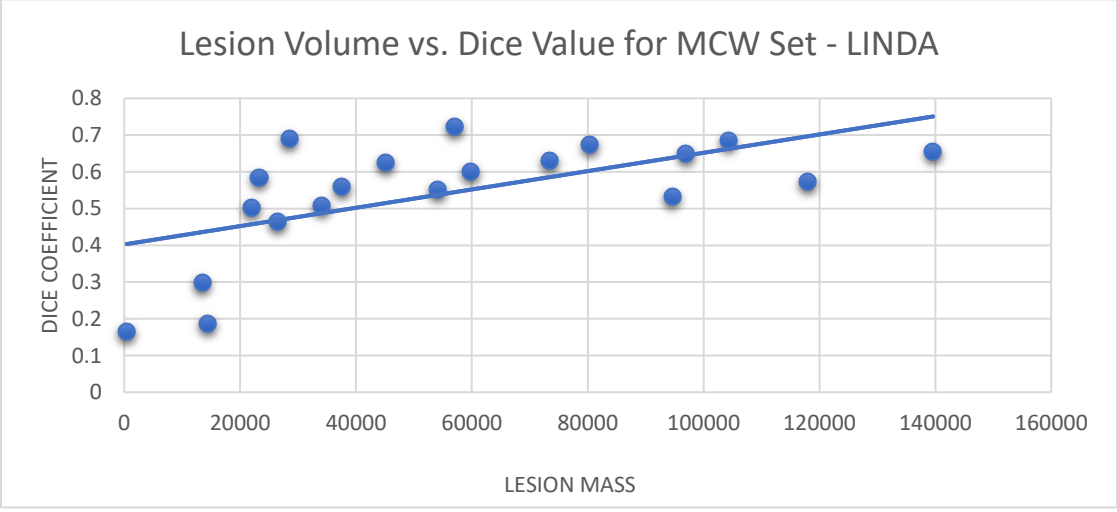


Figure 4.21 A plot of lesion volume versus Dice value for the Kessler dataset (LINDA method). Correlation Coefficient = 0.611; p-value(t-test) = 0.0000016864.

4.3.2 Lesion Segmentation Predictions using MMCNN

In this section, we compare the results we obtained using Lesion GNB and LINDA with the MMCNN Deep Learning results from a related study. In addition to our methods, MMCNN was compared to U-Net, UResNet and DeepMedic (Yunzhe Xue, et al., Neuroimage Clinical, 2019). The network was trained on a combined dataset of Kessler and MCW images and was tested on the ATLAS dataset. This is particularly challenging since most existing systems perform poorly in cross-study testing. MMCNN performed better than all other methods with a mean DSC value of 0.54. Table 4.1 shows all Dice values obtained in the tests on the ATLAS dataset.

Table 4.1 Mean Dice Coefficients of All Models Tested on the ATLAS Dataset

Method	**MMCNN**	UNet	UResNet	DeepMedic	LINDA	GNB
Mean Dice	**0.54**	0.47	0.45	0.47	0.30	0.32

Figure 4.22 provides more details and comparisons from the same experiment. The Raincloud plots show the distribution of Dice coefficients across all test images as well as the five summary values: median (middle horizontal line), Q3 (upper horizontal line), Q1 (lower horizontal line), minimum (lowermost bar), and maximum (uppermost bar). All models except for LINDA and GNB are trained on KES+MCW. Deep learning methods which are based on convolutional neural networks look promising and are already surpassing other methods that are based on conventional machine learning algorithms like Random Forest or SVM.

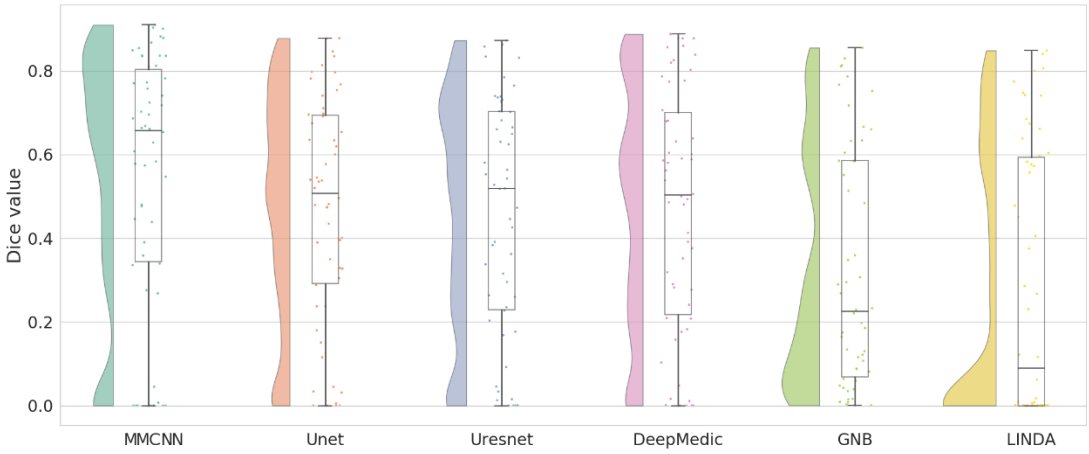


Figure 4.22 Raincloud plots of Dice coefficient values of all models (trained on Kessler+MCW and tested on ATLAS.)

Source: Yunzhe Xue, Fadi G. Farhat, Olga Boukrina, A. M. Barrett, Jeffrey R. Binder, Usman W. Roshan, William W. Graves, "A multi-path 2.5-dimensional convolutional neural network system for segmenting stroke lesions in brain MRI images", Preprint submitted to Neuroimage Clinical, March 31, 2019.

The correlation coefficient calculated for lesion volumes and dice values in this chapter is given by:

$$\rho = \left(\frac{1}{n-1}\right) \Sigma \left(\frac{x-\mu_x}{\sigma_x}\right) * \left(\frac{y-\mu_y}{\sigma_y}\right) \tag{4.1}$$

where $\boldsymbol{n}$ is the number of data points, $\boldsymbol{x}$ is our first variable (lesion volume), $\boldsymbol{y}$ is our second variable (DSC values), $\boldsymbol{\mu_x}$, $\boldsymbol{\mu_y}$ are the means of variables $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively and $\boldsymbol{\sigma_x}$, $\boldsymbol{\sigma_y}$ are the standard deviations of variables $\boldsymbol{x}$ and $\boldsymbol{y}$.

When describing the correlation coefficient values, we will use the following naming convention: 0-0.19 will be considered very weak, 0.2-0.39 weak, 0.40-0.59 moderate, 0.6-0.79 as strong and 0.8-1 as very strong correlation.

CHAPTER 5

SUMMARY AND CONCLUSIONS

5.1 Work Summary

In this work, we've presented comparisons of several machine learning and deep learning methods to solve medical imaging problems using two types of images: 2D breast cancer images (x-ray and biopsy) and 3D MRI brain scans. For image classification, we used conventional machine learning methods like Random Forest, Logistic Regression and Support Vector Machine, as well as deep learning methods like a simple 3-layer CNN, and deeper convolution neural networks like VGG16 and RDCNN. For lesion detection, existing Random Forest and Gaussian Naïve Bayes methods (LINDA and Lesion GNB) were tested and compared to the newly developed deep learning system, MMCNN (tested against UNet, UResNet and DeepMedic in a related study).

On the breast cancer x-ray image data, Random Forest beat all other classification methods, including the CNN's. SVM also performed better than neural networks on this dataset. The great variation in the images and the abundance of zero-pixel background in the images made it especially hard for the computer vision methods to detect the anomalies in the otherwise very similar breast tissue images. Further pre-processing and possibly zoomed in views of the x-ray images may improve neural network performance.

On the breast cancer biopsy image data, deep learning methods, especially RDCNN, beat all other classifiers. While augmenting the data actually degraded the

performance of SVM and Logistic Regression, it helped produce better Random Forest results, and greatly improved the performance of deep learning methods, in particular that of RDCNN, with a validation accuracy reaching 98% on the two-class datasets.

On the combined MCW/OASIS MRI brain scan image and patch datasets, VGG16 outperformed all other classification methods followed by Random Forest, then Logistic Regression and SVM. The neural network achieved an average accuracy of more than 94% in the full image classification experiments, and around a 72% accuracy on the patch dataset. The actual brain tissue was more uniformly distributed across the entire image than in the case of the breast x-rays. The images were also aligned to the standard MNI-152 template. These factors gave the neural network the advantage and allowed it to detect the classes more easily, especially for the full image dataset.

For lesion detection, images that are pre-processed and properly aligned yielded the best performance in all systems that were tested in this study. Also, lesion size in brain scans played a key role. The larger the lesions, the easier they are to detect. The MCW dataset produced the best segmentation results, in part due to the large lesions in its subject scans. The ATLAS and Kessler datasets did not do as well, partially due to their smaller lesion size. Neural network-based systems as well as the LINDA and Lesion GNB methods performed best on subjects with large lesions. Smaller lesions were difficult to detect, and MMCNN was the least sensitive system to lesion size (Yunzhe Xue, et al., Neuroimage Clinical, 2019).

5.2 Contribution and Limitations

The findings in this work demonstrate the promise of automatic classification and

segmentation systems in aiding with the diagnosis and treatment of disease. While none of the solutions presented can replace human experts, they can surely help with time consuming tasks that involve performing tedious and repetitive tasks on large amounts of data. For breast cancer images for example, the automatic classification methods can be used to quickly classify a new test subject and with high accuracy. For brain lesion detection problems, new MRI images can be automatically segmented, then given to a human expert for further assessment and review. The same processes can be easily adapted to other medical imaging applications and other types of medical images.

As explained throughout this paper, all methods described are supervised which rely primarily on and are limited by the accuracy of human expertise. Another limitation to be mindful of is the absence of a unified method for data preparation and pre-processing. While critical in any Machine Learning or Deep Learning system, data preparation varies greatly for medical images based on application objectives, specific data type and format, as well as methods used. We've also shown that the success of any system depends highly on the subject data, and while Deep Learning techniques seem to offer the best performance in general, non-linear classifiers such as Random Forest can still play an important role in the design of computer aided diagnosis systems.

The most serious limitation of the automatic systems presented is probably the absence of an end to end integrated solution that can solve problems with minimal human intervention. Current systems require trial and error at every step, from data preparation, to system training and testing, all the way to system deployment and implementation. Although these systems can be a great aid for experts, getting them to perform efficiently and adequately still requires a lot of time, close supervision and plenty of resources.

5.3 Future Work

The research presented in this work can be thought of as a sampling of select Machine Learning and Deep Learning methods that are currently available, and is by no means an exhaustive comparative study, and much work remains to be done. More types of data (in terms of image formats and human anatomy) are needed to evaluate the application of these methods more broadly. Other Deep Learning networks (Like U-Net, ResNet and DeepMedic) should be tested on medical imaging problems. More cross study experiments are also needed to ensure the portability of the presented solutions. Unsupervised learning methods need to be explored further. This may allow us to surpass human expertise which currently limits the performance of supervised methods.

Once we reach acceptable levels of accuracy in classification and segmentation, the scope of automation can be expanded to include the interpretation and usage of the results to take the next action. For breast cancer, this can be automatic recommendation of treatment plans for example. For brain MRI, this can be automatically correlating predicted lesion findings with impact on cognition and/or behavior, then recommending treatment or appropriate care.

5.4 Final Thought

Current classification and anomaly detection systems in medical imaging applications continue to improve and are quickly approaching the performance of human experts. With the help of new and innovative methods, automated systems may even outperform humans and become reliably more accurate. Expert human resources will be more empowered and will be able to focus more on other critical aspects of medical care.

APPENDIX A

BREAST CANCER IMAGES AND CLASSIFIER PERFORMANCE

A.1 Deep Learning Performance on the Breast Cancer X-Ray Dataset

The breast cancer x-ray images proved to be a challenge for the Deep Learning methods that we tested on this dataset, namely CNN-3, VGG16 and RDCNN. Surprisingly, Support Vector Machine (SVM) methods competed with the convolutional neural networks (CNN's), while Random Forest (RF) performed best. Actually, when looking at the prediction accuracy per class (especially the case or "1" class), both SVM and RF did better than CNN's whose predictions were mostly "0", with only a small percentage correctly predicted as "1". Two factors may have contributed to these results: the image content, as well as how each classifier interprets and processes input data.

The breast cancer x-rays all contain the image of the same object, albeit in varied sizes, shapes and orientations. The images are taken at the macroscopic level, and therefore do not show the differences between cancerous and non-cancerous tissue clearly. The images also contain a large area of zero-value pixel background, which normally does not provide any information to a classifier, especially a computer vision method such as a convolutional neural network. In contrast, the breast biopsy images are more easily distinguishable, even to the naked eye. Also, being at the microscopic level, the biopsy images show distinct objects of different shapes and even colors in some cases, which computer vision methods are very good at detecting.

Both RF and SVM classifiers treat their input as a single feature vector. These classifiers do not recognize images. In fact, when images are fed to a conventional

classifier, they are flattened to one dimension first, and if they have multiple channels, each channel is flattened as a vector, and all channels are concatenated into one feature vector as well. Similarly, a 3D image has to be flattened to one vector before it is fed to an SVM or RF classifier. An image essentially becomes a sequence of numbers, with each number being in a specific column (also called feature or dimension) as far as the classifier is concerned. When learning the correlation between the class (target variable) and the features, a linear classifier (such as SVM) considers all features at once, which means all pixel values for a particular image enter in the calculation (such as a dot product for example) at the same time. Using all features could have helped the classifier detect the subtle differences in the breast tissue pixel values in this case.

We should point out that although RF uses the entire feature vector to classify a sample, it uses decision trees over a large number of iterations to predict the class. Decision trees consider one or more features at a time to cluster samples in similar groups, but eventually test all features and generate one predicted label for each sample.

Computer vision methods work differently. To begin with, a CNN takes the actual image (2D or 3D) as input. Rather than looking at it as a flat feature vector, a CNN uses a kernel (typically 2D) to interpret the local contents of an image patch and tries to find objects that are similar in images that are in the same class. It should be obvious by inspection that the breast cancer x-ray images do not contain distinct objects that a CNN might correlate with a class, but rather one object, which is the breast itself. The difficulty is that the same object appears in all classes, which could explain the poor performance of the neural networks on these images.

A.2 Linear Classifier Performance on the Breast Cancer Biopsy Dataset

In general, data augmentation improves classifier performance. It is an established fact that classifiers learn better with more data. When it comes to image data and linear classifiers, this does not appear to be the case. Let's look at what is involved when augmenting image data in terms of flattened feature vectors.

Image data is typically augmented by flipping and rotating the original sample images. This is good for a CNN as augmentation creates more instances of the same object in the image. The object may be flipped or rotated, it may even be in a different part of the image, but it is still the same object, and a CNN is very good at detecting objects, even if they are relocated within the image. For SVM, this can be detrimental and can degrade performance. When an image is flipped or rotated, the location of each pixel in the modified image changes, and when the augmented image is flattened into a feature vector, the features of the original vector are now in different columns, and for SVM, the location of a feature in the vector matters.

Consider a vector V_O (representing the original image) and its flipped equivalent V_F. If V_O contains the following features: [f1, f2, f3, f4, f5, f6, f7, f8], then V_F might look like this: [f5, f6, f7, f8, f1, f2, f3, f4]. In a classification scenario, these two vectors are given the same label of course, since they are in the same class, but as far as SMV, the correlation between features and label is now different. This could explain why both SVM and Logistic Regression did worse with the augmented breast biopsy images. This hypothesis can be tested using simple (and small) text examples that can be manually observed and analysis.

APPENDIX B

CLASSIFIER BACKGROUND INFORMATION

B.1 Gini Impurity Index in the Random Forest Algorithm

Random Forest is based on the CART (**C**lassification **A**nd **R**egression **T**rees) algorithm, which refers to decision trees used for classification (and regression), as illustrated in figure 2.7 on page 18. A decision tree uses stumps (a line in a two-dimensional space) to split samples that may not be linearly separable. The graph in figure B.1 shows a two-dimensional dataset that contains two classes: circles (C_1) and squares (C_2). The first split "S_1" splits the data at a random value of f_1 (the first feature or dimension) and creates two child nodes (subsets of the whole dataset, which is the parent node). The second split "S_2" splits the data again at a random value of f_2 and creates four child notes (by splitting their parent nodes). After the two splits are made, all data points are classified, albeit with a small margin of error (samples shown in red on the graph).

Since a split can introduce classification errors, the goodness of the split has to be measured to determine if it is the best split or if a better split (with less error) can be found. The Gini Impurity Index is one way to measure the quality of a split. Next, we will define the Gini Impurity Index or "***gini***" and use a simple one-dimensional dataset example to explain how it is calculated.

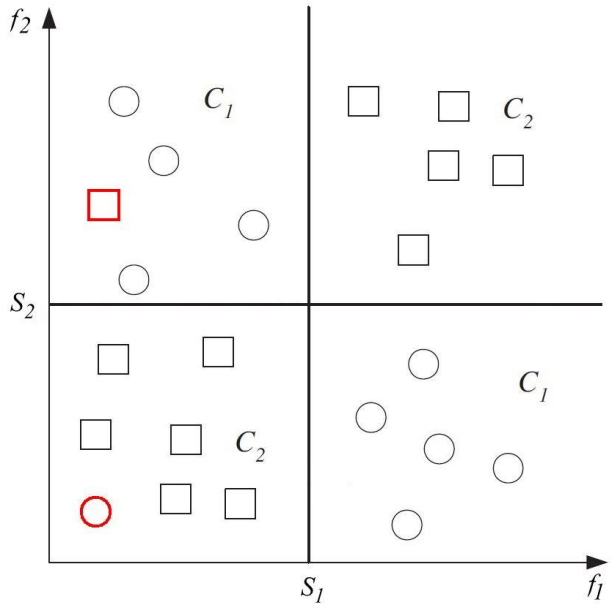


Figure B.1 Illustration of a classification decision tree with two stumps. Although splits S_1 and S_2 (created by the stumps) introduce some impurity (or misclassification), they classify the data with high accuracy.

In a binary classification system (classes 0 and 1), we define $\boldsymbol{Z_i}$ as the number of class 0 instances in child node *i*, $\boldsymbol{C_i}$ as the total number of instances in child node *i*, and $\boldsymbol{N}$ as the number of instances in the parent node. The objective is to find a split that minimizes the Gini Impurity Index which is given by:

$$gini = \sum \frac{Zi}{N} * \left(1 - \frac{Zi}{Ci}\right) \tag{B.1}$$

Which is the weighted sum of impurities in all child nodes created by the split.

Consider the one-dimensional dataset: [0 0 0 1 1 1 1 0]. We define a split after the first data point to create two nodes. The dataset will be as follows after the split: [0 | 0 0 1 1 1 1 0]. Next, we will calculate the *gini* for this split using $N = 8$, $Z_1 = 1$, $C_1 = 1$, $Z_2 = 3$, $C_2 = 7$.

gini = 1/8 * (1 – (1/1)) + 3/8 * (1 – (3/7)) = 0.2143.

To check if we can find another split that would give a smaller impurity index, we will define a split after the fourth data point this time. The dataset will now be as follows: [0 0 0 1 | 1 1 1 0]. We repeat the *gini* calculation we did before, but this time using $N = 8$, $Z_1 = 3$, $C_1 = 4$, $Z_2 = 1$, $C_2 = 4$.

gini = 3/8 * (1 – (3/4)) + 1/8 * (1 – (1/4)) = 0.1875.

The second split is clearly better than the first. When $gini = 0$ for a node, it means that all instances in the node are of the same class, and we say that the node is pure. In practice, the minimum *gini* is found by trying all possible splits and calculating the *gini* for each split across all child nodes.

REFERENCES

Rachmadi, M., Valdés-Hernández, M., Agan, M., Komura, T. "Deep Learning vs. Conventional Machine Learning: Pilot Study of WMH Segmentation in Brain MRI with Absence or Mild Vascular Pathology", Journal of Imaging, Volume: 3, Number: 4, Pages: 482–493, 2017, https://www.mdpi.com/2313-433X/3/4/66, (accessed in April 2019)

Zhou, L., Wenxiang D., Xiaodong W. "Robust Image Segmentation Quality Assessment without Ground Truth", CoRR, Volume: abs/1903.08773, Number: 1, Pages: 1, 2019, https://arxiv.org/abs/1903.08773v1, (accessed in April 2019)

Chen, X., Chen K., Ender. "Unsupervised Detection of Lesions in Brain MRI Using Constrained Adversarial Auto-Encoders", Volume: 1, Number: 1, Pages: 1, 2018, https://arxiv.org/abs/1903.08773, (accessed in April 2019)

Sawyer Lee, R., Gimenez, F., Hoogi, A., Rubin, D. "Curated Breast Imaging Subset of DDSM", The Cancer Imaging Archive, Volume: 1, Number: 1, Pages: 1, 2016, http://dx.doi.org/10.7937/K9/TCIA.2016.7O02S9CY, (accessed in April 2019)

Sawyer Lee, R., Gimenez, F., Hoogi, A., Miyake, K., Gorovoy, M., Rubin, D. "A Curated Mammography Data Set for Use in Computer-Aided Detection and Diagnosis Research", Scientific Data, Volume: 4, Number: 170177, Pages: 1, 2017, https://www.nature.com/articles/sdata2017177, (accessed in April 2019)

Clark K., Vendt B., Smith K., Freymann J., Kirby J., Koppel P., Moore S., Phillips S., Maffitt D., Pringle M., Tarbox L., Prior F. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository", Journal of Digital Imaging, Volume: 26, Number: 6, Pages: 1045-1057, 2013, https://link.springer.com/article/10.1007%2Fs10278-013-9622-7, (accessed in April 2019)

Spanhol, F., Oliveira, L., Petitjean, C., Heutte, L. "A Dataset for Breast Cancer Histopathological Image Classification", IEEE Transactions on Biomedical Engineering (TBME), Volume: 63, Number: 7, Pages: 1455-1462, 2016, http://www.inf.ufpr.br/lesoliveira/download/TBME-00608-2015-R2-preprint.pdf, (accessed in April 2019)

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., Whyntie, T., Nachev, P., Modat, M., Barratt, D., Ourselin, S., Cardoso, M., Vercauteren, T. "NiftyNet: A Deep-Learning Platform for Medical Imaging, Computer Methods and Programs in Biomedicine", ScienceDirect, Volume: 158, Number: 1, Pages: 113-122, 2018, https://doi.org/10.1016/j.cmpb.2018.01.025, (accessed in April 2019)

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso M., Vercauteren T. "On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task", Springer, Cham., New York, NY, Volume: 10265, Number: 1, Pages: 348-360, 2017, http://doi.org/10.1007/978-3-319-59050-9_28, (accessed in April 2019)

Finlayson, S., Chung, H., Kohane, I., Beam, A. “Adversarial Attacks Against Medical Deep Learning Systems” , CoRR, Volume: abs/1804.05296, Number: 1, Pages: 1, 2018, https://arxiv.org/abs/1804.05296, (accessed in April 2019)

Teramoto, A., Tsukamoto, T., Kiriyama, Y., Fujita, H. “Automated Classification of Lung Cancer Types from Cytological Images Using Deep Convolutional Neural Networks”, BioMed Research International, Volume: 2017, Number: 4067832, Pages: 6, 2017, https://doi.org/10.1155/2017/4067832, (accessed in April 2019)

Xue, Y., Roshan, U. “Random Depthwise Signed Convolutional Neural Networks", Scinapse,Volume: 1, Number: 1, Pages: 1, 2018, http://scinapse.io/papers/2808187103, (accessed in April 2019)

Al-Nahid, A., Mehrabi, M., Kong, Y. “Histopathological Breast Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering”, BioMed Research International, Volume: 2018, Number: 2362108, Pages: 20, 2018, https://doi.org/10.1155/2018/2362108, (accessed in April 2019)

Mohsen, H., El-Dahshan, A., El-Horbaty, M., Salem, A. “Classification Using Deep Learning Neural Networks for Brain Tumors”, NeuroImage, Volume: 1, Number: 1, Pages: 1, 2017, http://www.sciencedirect.com/science/article/pii/S2314728817300636, (accessed in April 2019)

Preston, D. “Magnetic Resonance Imaging (MRI) of the Brain and Spine: Basics”, Neuroanatomy Review, Volume: 1, Number: 1, Pages: 1, 2006, http://hcasemed.case.edu/clerkships/neurology/web%20neurorad/mri%20basics.htm, (accessed in April 2019)

Fotenos, A., Snyder, A., Girton, L., Morris, J., Buckner, R. "Normative Estimates of Cross-Sectional and Longitudinal Brain Volume Decline in Aging and AD", Neurology, Volume: 64, Number: 6, Pages: 1032-1039, 2005, https://doi.org/10.1212/01.WNL.0000154530.72969.11, (accessed in April 2019)

Buckner, R., Head, D., Parker, J., Fotenos, A., Marcus, D., Morris, J., Snyder, A. "A Unified Approach for Morphometric and Functional Data Analysis in Young, Old, and Demented Adults Using Automated Atlas-Based Head Size Normalization: Reliability and Validation Against Manual Measurement of Total Intracranial Volume", NeuroImage, Volume: 23, Number: 2, Pages: 724-738, 2004, https://doi.org/10.1016/j.neuroimage.2004.06.018, (accessed in April 2019)

Zhang, Y., Brady, M., Smith, S. "Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation Maximization Algorithm", IEEE Transactions on Medical Imaging, Volume: 20, Number: 1, Pages: 45-57, 2001, https://doi.org/10.1109/42.906424, (accessed in April 2019)

Rubin, E., Storandt, M., Miller, J., Kinscherf, D., Grant, E., Morris, J., Berg, L. "A Prospective Study of Cognitive Function and Onset of Dementia in Cognitively Healthy Elders", Archives of Neurology, Volume: 55, Number: 3, Pages: 395-401, 1998, https://www.ncbi.nlm.nih.gov/pubmed/9520014, (accessed in April 2019)

Morris, J. "The Clinical Dementia Rating (CDR): Current Version and Scoring Rules", Neurology, Volume: 43, Number: 11, Pages: 2412-2414, 1993, https://www.ncbi.nlm.nih.gov/pubmed/8232972, (accessed in April 2019)

Boukrina, O., Barrett, A., Alexander E., Yao, B., Graves W. "Neurally Dissociable Cognitive Components of Reading deficits in Subacute Stroke", Frontiers in Human Neuroscience, Volume: 9, Number: 1, Pages: 298, 2015, https://www.frontiersin.org/article/10.3389/fnhum.2015.00298, (accessed in April 2019)

Liew, S. "The Anatomical Tracings of Lesions after Stroke (ATLAS) Dataset - Release 1.1", Ann Arbor, MI: Inter-University Consortium for Political and Social Research, Volume: 1, Number: 1, Pages: 1, 2017, https://doi.org/10.3886/ICPSR36684.v1, (accessed in April 2019)

Cox, R. "AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages", Computers and Biomedical Research, Volume: 29, Number: 3, Pages: 162-173, 1996, https://www.ncbi.nlm.nih.gov/pubmed/8812068, (accessed in April 2019)

Cox, R., Hyde, J. "Software Tools for Analysis and Visualization of FMRI Data", NMR in Biomedicine, Volume: 10, Number: 4-5, Pages: 171-178, 1997, https://www.ncbi.nlm.nih.gov/pubmed/9430344, (accessed in April 2019)

Gold, S., Christian, B., Arndt, S., Zeien, G., Cizadlo, T., Johnson, D., Flaum, M., Andreasen, N. "Functional MRI Statistical Software Packages: A Comparative Analysis", Human Brain Mapping, Volume: 6, Number: 2, Pages: 73-84, 1998, https://www.ncbi.nlm.nih.gov/pubmed/9673664, (accessed in April 2019)

Griffis, J. et al. "Lesion GNB Toolbox for SPM 12", ResearchGate, Volume: 1, Number: 1, Pages: 1, 2016, https://www.researchgate.net/project/Chronic-lesion-identification-in-T1-weighted-MRI, (accessed in April 2019)

Pustina, D., et al. "Automated Segmentation of Chronic Stroke Lesions Using LINDA: Lesion Identification with Neighborhood Data Analysis", Human Brain Mapping, Volume: 37, Number: 4, Pages: 1405-1421, 2016, https://doi.org/10.1002/hbm.23110, (accessed in April 2019)

Xue, Y., Farhat, F., Boukrina, O., Barrett, A., Binder, J., Roshan, U., Graves, W. "A Multi-Path 2.5-Dimensional Convolutional Neural Network System for Segmenting Stroke Lesions in Brain MRI Images", Preprint submitted to Neuroimage Clinical, Volume: 1, Number: 1, Pages: 18, 2019

Various. "Symptoms and Diagnosis of Breast Cancer", Breastcancer.org, Ardmore, PA, Volume: 1, Number: 1, Pages: 1, 2019, https://www.breastcancer.org/symptoms, (accessed in April 2019)

Various. "Breast Cancer Facts", National Breast Cancer Foundation Inc., Frisco, TX, Volume: 1, Number: 1, Pages: 1, 2019, https://www.nationalbreastcancer.org/breast-cancer-facts, (accessed in April 2019)

Howard, J. "Now Anyone Can Train ImageNet in 18 minutes", Fast.Ai, Volume: 1, Number: 1, Pages: 1, 2018, https://www.fast.ai/2018/08/10/fastai-diu-imagenet/, (accessed in April 2019)

Binder, J., Pillay, S., Humphries, C., Gross, W., Graves, W., Book, D. "Surface Errors Without Semantic Impairment in Acquired Dyslexia: A Voxel-Based Lesion–Symptom Mapping Study", Brain, Volume: 139, Number: 5, Pages: 1517–1526, 2016, https://doi.org/10.1093/brain/aww029, (accessed in April 2019)