UCLA
UCLA Electronic Theses and Dissertations

Title

Improving Acute Ischemic Stroke Clinical and Imaging Outcome Classification using Machine Learning and Deep Learning Methods

Permalink

https://escholarship.org/uc/item/4673d3qd

Author

Ho, King Chung

Publication Date

2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Improving Acute Ischemic Stroke Clinical and Imaging Outcome

Classification using Machine Learning and Deep Learning Methods

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioengineering

by

King Chung Ho

2019

ABSTRACT OF THE DISSERTATION

Improving Acute Ischemic Stroke Clinical and Imaging Outcome

Classification using Machine Learning and Deep Learning Methods

by

King Chung Ho

Doctor of Philosophy in Bioengineering

University of California, Los Angeles, 2019

Professor Corey Wells Arnold, Co-Chair

Professor Denise R Aberle, Co-Chair

Stroke is the fifth leading cause of death in the United States, with approximately 795,000 new cases each year. The goal of stroke treatment is to rescue salvageable tissue by reperfusion therapy. Clinical trials have shown that intravenous tissue plasminogen activator (IV tPA) and clot retrieval devices are effective treatments for recanalizing occluded blood vessels. However, determining an optimal stroke treatment plan is not a straightforward decision because it involves different factors, such as patient risk of hemorrhage and penumbra size. The relationships between these factors and patient outcomes are not clearly understood. This dissertation attempts to overcome these challenges by developing machine learning and deep learning models for acute ischemic stroke clinical and imaging outcome classification. A novel deep learning model was first proposed using source perfusion imaging to predict voxel-wise tissue outcome. The model architecture is designed to include contralateral patches to improve the feature learning process. Second, an end-to-end machine learning approach was developed to classify stroke onset time, which is a major clinical variable in selecting patients for IV tPA treatments. The approach combines baseline descriptive features and deep features to improve stroke onset time classification using machine learning

ii

models. Third, a bi-input convolutional neural network was developed for perfusion parameter estimation. This model lays a foundation to estimate perfusion parameters using pattern recognition techniques. Finally, a machine learning model trained with a balanced data set was developed for acute stroke patient outcome prediction. Rigorous experiments and results have shown the effectiveness of these proposed methods. This dissertation describes methods that lead to better understanding of stroke imaging, which lays the foundation to offer decision-making guidance for clinicians providing acute stroke intervention treatments.

The dissertation of King Chung Ho is approved.

Suzie May El-Saden

Ricky Kiyotaka Taira

Daniel Ennis

Alex Anh-Tuan Bui

William F Speier

Denise R Aberle, Committee Co-Chair

Corey Wells Arnold, Committee Co-Chair

University of California, Los Angeles

2019

Dedicated to my beloved family, friends, and the medical & engineering community.

Don't just aspire to make a living; aspire to make a difference.

— DENZEL WASHINGTON

It is not the things we do in life that we regret on our death bed; it is the things we do not… Find your passion, and in my experience, no matter what you do at work or what you do in official settings, that passion will be grounded in people. It will be grounded in the relationships you have with people and what they think of you when your time comes.

— RANDY PAUSCH

To the world you may be one person, but to one person you may be the world.

— THEODOR SEUSS GEISEL

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Shiwen Shen, Nicholas Matiasz, Mary Zide, Edgar Rios, Maurine Tong, Anna Wu, Kyle Singleton, Sagari Grandhi, and Daniel Johnson.

I am very grateful to my parents, Kwing Yam Ho and Chui Lin Yan, who have sacrificed selflessly to support my education. I am also very thankful to my brother, King Ning Ho, who shouldered the responsibility to take care of our parents while I pursued my studies abroad.

I would like to acknowledge others who have helped and inspired me, whether it is directly, through their friendship, or indirectly through their work. These people are Harsha Kittur, Janay Kong, Mrinal Rath, Clara Chan, Mayuri Kerr, Mayank Jog, James Che, Xianjie Chen, Samatha Ma, Andrew Li, Wai Yan Tang, Karrie Sio, Sarah Lau, Ivan Cheng, Hayden Chiu, Victor Yeung, Joseph Lo, Jonathan Ho, Vanessa AuYeung, Bobby Lee, Felix Liang, Janice Kuang, Christie Fung, Savio Kwan, Nikki Lui, Cheuk Yu, Clarrisa Chan, Deep Ng, Alice Cheung, Claudia Ng, Mandy Leung, Jaco Wong, Osanna Tam, Fiko Fan, Stanley Fong, Alice Fong, Chester Hung, Max Ng. I would also like to acknowledge the organizations that I participated in during my graduate study. These organizations include IReach Fellowship (CBMC), Joseph Fellowship (SJCAC), YF (CBC), CF (FCBC) and CCF (UCLA).

I want to express my sincere gratitude to Fion Kung, Auntie Bonnie Yip and Uncle Chris Yip's family. They were the people whom I spent a lot of time with throughout my Ph.D study and they had showered me with tremendous love and support throughout my graduate study.

I would like to thank all the people who have contributed to the following article contents, which are included in this dissertation:

Chapter 3 contains materials published in "Ho KC, El-Saden S, Scalzo F, Bui AAT, Arnold CW. Predicting Acute Ischemic Stroke Tissue Fate using Deep Learning on Source Perfusion MRI. International Stroke Conference (ISC) 2016", and "Ho KC, Sarma KV, Speier W, El-Saden S, Arnold CW. Predicting Ischemic Stroke Tissue Fate Using a Deep Convolutional Neural Network on Source MR Perfusion Imaging. Journal of Medical Imaging, 2019, in submission."

Chapter 4 contains materials published in "Ho KC, Speier W, Zhang H, El-Saden S, Arnold CW. Classifying Acute Ischemic Stroke Onset Time using Deep Imaging Features. In AMIA Annual Symposium Proceedings (Vol. 2017)", and "Ho KC, Speier W, El-Saden S, Arnold CW. A Machine Learning Approach for Classifying Ischemic Stroke Onset Time from Imaging. IEEE Transactions on Medical Imaging, 2018, in submission."

Chapter 5 contains materials published in "Ho KC, Scalzo F, Sarma KV, El-Saden S, Arnold CW. A Temporal Deep Learning Approach for MR Perfusion Parameter Estimation in Stroke. 23rd International Conference on Pattern Recognition. Cancun; 2016", and "Ho KC, Scalzo F, Sarma KV, El-Saden S, Bui AAT, Arnold CW. A Novel Bi-Input Convolutional Neural Network for Deconvolution-Free Estimation of Stroke MR Perfusion Parameters. 2016 Annual Meeting of the Radiological Society of North America. Chicago, IL; 2016."

Contains 6 contains materials published in "Ho KC, Speier W, El-Saden S, Liebeskind DS, Saver JL, Bui AAT, Arnold CW. Predicting Discharge Mortality after Acute Ischemic Stroke Using Balanced Data. In AMIA Annual Symposium Proceedings (Vol. 2014, p. 1787)."

One last thing: I really like a sketch of an umbrella. This is not only because the sketch looks like a "J" (which is the first character of my English nickname), but also because it symbolizes that I can hold an umbrella to cover others on the rainy days. When I applied for graduate schools, my goal was to make our society a better place. I hope that all I have done in this dissertation can

make an impact in the medical community, and all I will do beyond this can continue to satisfy this desire, and not just to make a living.

VITA

2011	Bachelor of Science in Bioengineering, UC Berkeley
2012	Software Development Engineering Intern (summer), Amazon.com
2013	Master of Science in Bioengineering, UCLA
2013-2018	Graduate Student Researcher in Bioengineering, UCLA
2016	Machine Learning Engineering Intern (summer), IBM
2017	Data Scientist Intern (summer), GE
2018	Data Scientist Intern (summer), SurveyMonkey
2019 (expected)	Doctor of Philosophy in Bioengineering (Medical Informatics), UCLA

PUBLICATIONS

Ho KC, Speier W, et al. Predicting Discharge Mortality after Acute Ischemic Stroke Using Balanced Data. In AMIA Annual Symposium Proceedings (Vol. 2014, p. 1787). Finalist on best student paper award.

Ho KC, El-Saden S, et al. Predicting Acute Ischemic Stroke Tissue Fate using Deep Learning on Source Perfusion MRI. International Stroke Conference (ISC) 2016. Poster presentation

Ho KC, Scalzo F, et al. A Novel Bi-Input Convolutional Neural Network for Deconvolution-Free Estimation of Stroke MR Perfusion Parameters. 2016 Annual Meeting of the Radiological Society of North America. Chicago, IL; 2016. Oral presentation

Ho KC, Scalzo F, et al. A Temporal Deep Learning Approach for MR Perfusion Parameter Estimation in Stroke. 23rd International Conference on Pattern Recognition. Cancun; 2016. Oral presentation. Finalist on IBM best student paper award.

Li J, Sarma KV, **Ho KC**, et al. A Multi-scale U-Net for Semantic Segmentation of Histological Images from Radical Prostatectomies. In AMIA Annual Symposium Proceedings (Vol. 2017).

Ho KC, Speier W, et al. Classifying Acute Ischemic Stroke Onset Time using Deep Imaging Features. In AMIA Annual Symposium Proceedings (Vol. 2017). Best student paper award.

Li W, Li J, Sarma KV, **Ho KC**, et al. Path R-CNN for Prostate Cancer Diagnosis and Gleason Grading of Histological Images. IEEE transactions on medical imaging, 2018.

Li J, Speier W, **Ho KC**, et al. An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies. Computerized Medical Imaging and Graphics 69 (2018): 125-133.

Ho KC, Sarma KV, et al. Predicting Ischemic Stroke Tissue Fate Using a Deep Convolutional Neural Network on Source MR Perfusion Imaging. Journal of Medical Imaging, 2019, in submission.

Ho KC, Speier W, et al. A Machine Learning Approach for Classifying Ischemic Stroke Onset Time from Imaging. IEEE transactions on medical imaging, 2018, in submission.

CHAPTER 1

Introduction

1.1 Motivation

Stroke is the primary cause of long-term disability and the fifth leading cause of death in the United States, with approximately 795,000 Americans experiencing a new or recurrent stroke each year [1]. Acute stroke treatments focus on restoring blood flow to hypoperfused regions to minimize infarction (i.e., tissue death). Intravenous (IV) tissue plasminogen activator (tPA) remains the dominant thrombolytic treatment for acute stroke, with a strict time usage guideline (no more than 4.5 hours from witnessed stroke symptom onset, i.e., time-since-stroke (TSS) < 4.5hrs) due to the increased risk of hemorrhage when administered beyond that time interval [2]. More recently, mechanical thrombectomy (clot retrieval) [3]–[5] has been developed as an alternative or adjunct therapy to IV tPA, in which a device is used to navigate through the arterial tree under fluoroscopic guidance to physically retrieve blood clots. Yet, the optimal treatment time window of the clot retrieval devices remains uncertain [6], and overall stroke outcomes (e.g., mortality) after treatments have to be further determined [7].

Effort has been placed in identifying risk factors and creating models predictive of stroke and its outcomes [8]–[10]. Yet, the existing stroke outcome prediction model performance is often biased because of the problem of between-class imbalance in medical data, which is common in binary prediction tasks when one class has substantially more samples than the other class [11]–[13]. Also, models with the greatest ability to predict an individual's functional outcome after stroke are based primarily on pre-stroke variables (e.g., age) and functional status at presentation.

These variables shed little light on the pathophysiological processes supporting their predictions and offer minimal information on how to tailor individual treatment.

Stroke imaging analysis represents a key area for further investigation. Magnetic resonance (MR) images are often obtained, including diffusion-weighted images (DWIs), apparent diffusion coefficient (ADC) maps, and perfusion-weighted images (PWIs). Early work using manual and threshold-based analysis on these MR images has demonstrated the utility in salvageable tissue calculation [14]–[16]; recent work has shown that imaging features can be used with clinical features to improve outcome prediction [17], [18]; and image preprocessing methods [19], [20] and diagnostic tools [21] are readily accessible. Currently employed methods for stroke image analysis typically make voxel-based predictions on tissue outcome prediction (i.e., tissue fate) [22], [23]. These studies apply machine learning techniques using shallow features extracted from images to predict infarcts. These hand-crafted features are derived from the perfusion parameter maps (e.g., cerebral blood volume map), which describe the brain tissue characteristics. While the machine learning techniques have proven to be beneficial, the performance of tissue outcome prediction is still limited and recent studies suggest that this is due to drawbacks of using hand-crafted features in the classification models [24]–[29]. Another stroke imaging research is time since stroke (TSS) classification. An imaging feature (a mismatch pattern) is derived from the magnetic resonance (MR) image. The mismatch volume is calculated to estimate the patient stroke onset time [30]–[33], which determines the eligibility of IV tPA (TSS < 4.5hrs). While this is the current state-of-the-art method for determining eligibility for thrombolytic therapy in cases of unknown TSS, computing mismatch using MR imaging remains a difficult task and the performance is moderate [34], [35].

This research focuses on two areas to address some of the aforementioned challenges: 1) tissue outcome classification; and 2) time since stroke classification. The ensuing two sections detail the motivation for each task.
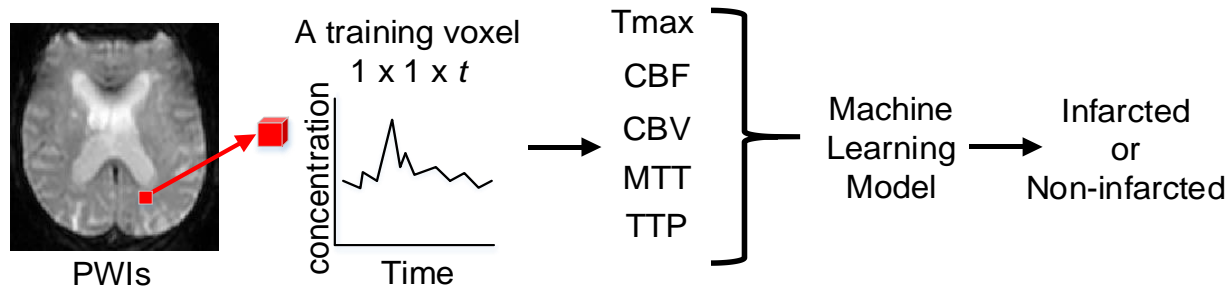
Figure 1.1: An overview of voxel-wise tissue outcome prediction from source perfusion-weighted images (PWIs). Hand-crafted features are derived from a voxel concentration time curve. These features are time-to-maximum (Tmax), cerebral blood flow (CBF), cerebral blood volume (CBV), mean transit time (MTT), and time-to-peak (TTP).

1.1.1 Tissue outcome classification

Predicting ischemic tissue outcome is an important task for better stroke evaluation and treatment planning. Knowing the potential tissue outcome can provide important information to clinicians about the relative value of interventions. For example, if the volume of predicted infarct tissue is large given any treatment variable in the model, a clinician may reconsider the utility of a treatment such as thrombolysis or clot retrieval, either of which can increase the chance of hemorrhage and may worsen the clinical outcome. In addition, predicting stroke tissue outcome helps generate new knowledge that may be useful in patient selection for clinical trials [6].

Researchers have attempted to predict the final infarct volume from the pre-treatment MR images using machine learning techniques [22], [36], [45]–[47], [37]–[44]. These models rely on hand-crafted features that are derived from the perfusion parameter maps for prediction (Figure 1.1). While these techniques have proven to be useful, recent concerns have been raised about the use of these parameters [24], [25], due to drawbacks that have been discussed in several studies [26]–[29]. One such drawback is the sensitivity of PWI to vascular delays and dispersion effects caused by physiologic changes, such as heart rate and cardiac output, that can substantially change
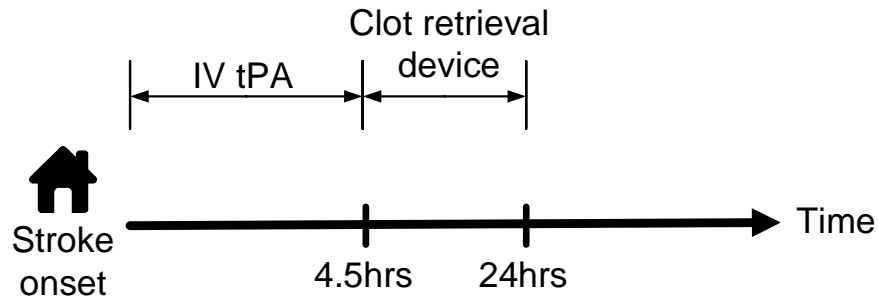
Figure 1.2: Treatment guideline for acute ischemic stroke patient.

the perfusion image parameters [26]. Another drawback is the fact that choosing the appropriate arterial input function (AIF) from PWIs, which describes the contrast input to the vasculature over time, is a challenging and generally subjective task due to the need to account for the partial volume effect. This can lead to variability in blood flow measurements caused by varying delays and/or dispersion based on different AIF choices [27]. Although deconvolution by singular value decomposition (SVD) can address this problem, studies have found that the deconvolution process can introduce distortions that influence the measurement of perfusion parameters [28], and the decoupling of delay may negatively impact infarct prediction [29]. All of these factors have contributed to the imperfect prediction of tissue outcome by current methods. Thus, it is highly desirable to have a novel machine learning model, where no hand-crafted features are needed, to achieve optimal performance on tissue outcome prediction.

1.1.2 Time since stroke classification

Figure 1.2 shows the current stroke treatment guideline based on time since stroke onset. IV tPA remains the dominant thrombolytic treatment for acute ischemic stroke, with a strict time usage guideline (no more than 4.5 hours from witnessed stroke symptom onset) due to the increased risk of hemorrhage when administered beyond that time interval. About 30% of the population cannot receive IV tPA because of unknown TSS (e.g., wake-up strokes or unwitnessed

4

strokes) [48]. Therefore, determining stroke onset time independent of patient history is an important task for better stroke evaluation and stroke treatment decision-making.

The current state-of-the-art and experimental methods that can provide clinicians with insight into stroke onset time are mostly based on observable mismatch patterns between MR DWI and fluid attenuated inversion recovery (FLAIR) imaging, known as "DWI-FLAIR mismatch" [30]–[33]. Most of the previous studies [31]–[33] reported that the mismatch method could only achieve a medium performance: a specificity of 0.60 to 0.80 with a sensitivity of 0.50 to 0.60 and a negative predictive value (NPV) of 0.20 to 0.50. Computing mismatch using MR imaging is a difficult task that requires extensive training and for which clinician inter-observer agreement has been found to be only moderate [34], [35]. Another research work has also suggested that the mismatch method may be too stringent and simple, and therefore miss individuals who could benefit from thrombolytic therapy [49]. Thus, it is desirable to develop a classification pipeline, which includes novel feature generation and advanced machine learning model classification, to classify stroke onset time.

1.2 Contributions

Recent work has shown that deep learning techniques outperform many state-of-the-art classification tasks [50], [51]. The deep-layer-learning approach is effective in discovering hidden hierarchical non-linear representations that are powerful for prediction [52]. To address the issues described in Section 1.1, this work presents novel machine learning and deep learning methods to improve acute ischemic stroke tissue outcome prediction and provide a state-of-the-art solution for TSS classification. Two specific aims are defined:

1. *To develop a deep learning algorithm for acute stroke clinical MR imaging sequences that allows for more accurate prediction of stroke tissue outcome.* A novel deep convolutional neural network (CNN) architecture was proposed to utilize the patch of interest and its contralateral patch to improve voxel-wise tissue outcome prediction. The deep learning model

automatically learns hierarchical representations from the stroke images and makes predictions of tissue outcome without using hand-crafted perfusion parameters. This method eliminates the use of the traditional deconvolution method which may cause distortions in residue function. The contralateral patch input also acts as a matched control (i.e., a patch for comparison) for the patch of interest which makes tissue outcome prediction possible without the information of the arterial input function. The results show that the proposed deep learning model outperforms existing tissue fate models.

2. *To develop a machine learning framework that utilizes unsupervised feature learning of stacked autoencoders on quantitative image representations to classify acute stroke patient onset time (TSS).* The framework is one of the earliest approaches to apply machine learning in TSS classification. It includes developing a set of hundreds of imaging features from stroke images and training machine learning models for TSS classification. A deep learning algorithm based on an autoencoder architecture is developed to extract latent representative imaging features from stroke images. This work shows that deep learning features improve TSS classification significantly and the best model outperforms the current state-of-the-art thresholding approach (DWI-FLAIR mismatch method). This framework serves as the foundation to move towards advanced imaging analysis for TSS classification, which ultimately impacts the treatment guidance for IV tPA.

Towards Aim 1, a novel deep CNN architecture was developed for voxel-wise tissue outcome prediction using source perfusion images. The proposed deep CNN had a new layer to learn pairs of unit voxel-wise temporal filters. These unit temporal filters help paring feature learning from modified training data, in which each input data was a pair consisting of the patch of interest and its contralateral patch. The model was evaluated and compared with the existing tissue fate models (generalized linear regression model, support vector machine, and spectral regression kernel discriminant analysis). While the deep CNN tissue outcome prediction model was being built, a bi-input convolutional neural network (bi-CNN) was also developed to investigate if a machine

learning approach can be used to estimate perfusion parameter values. The proposed bi-CNN had two separated convolution chains that were used as deconvolution operations for the input (the patch of interest and the arterial input function). The results show that the bi-CNN can be used as an alternative to traditional deconvolution methods for perfusion parameter map generation. Additional analyses were performed to correlate imaging and clinical features to stroke outcome (i.e., mortality) with an imbalanced data set.

For Aim 2, a machine learning framework was proposed for TSS classification. Notably, this framework is the first complete machine learning work on addressing issues on the simple thresholding approach (DWI-FLAIR mismatch). The developed framework comprises of four major steps: 1) defining the region of interest (ROI) in the MR images for feature generation; 2) generating a set of imaging features from the ROI; 3) creating additional imaging features using the proposed deep autoencoder architecture from the source perfusion-weighted images; and 4) training five machine learning models with imaging features for TSS classification. The trained models were then compared with the DWI-FLAIR mismatch method.

Collectively, the design and implementation of these aims results in imaging features and outcome classification models that present deeper understanding of acute ischemic stroke and provide insights for better acute ischemic stroke treatment guidance.

1.3   Organization of the Dissertation

The dissertation is organized as follows. Chapter 2 provides background on acute ischemic stroke and its treatment, stroke patient outcome prediction, stroke imaging, deep learning, current limitations, and literature review. Chapter 3 describes Aim 1 work on tissue outcome prediction using the proposed deep learning architecture with contralateral inputs. A machine learning framework is introduced in Chapter 4, addressing Aim 2 for advanced imaging analysis in TSS classification. Chapter 5 describes the work we have done in perfusion parameter estimation using bi-input convolutional neural network and Chapter 6 describes the work we have done on acute

stroke patient outcome prediction. Chapter 7 concludes by summarizing this work results, discussing the limitations of this work, and suggesting future directions.

CHAPTER 2

Background

2.1 Stroke Overview

There are two types of strokes: ischemic strokes (87%) and hemorrhagic strokes (13%) [1]. Ischemic stroke occurs when arteries to brain tissues are narrowed or blocked by one or multiple occlusions, leading to severely reduced blood flow (ischemia) and tissue death. Hemorrhagic stroke occurs when a weakened blood vessel ruptures, causing bleeding inside the brain. In this dissertation, acute ischemic stroke is the primary research focus.

Figure 2.1 shows an overview of the ischemic stroke regions caused by a blood clot. The infarct core is defined as a dead tissue region with insufficient blood supply and irreversible damage. The
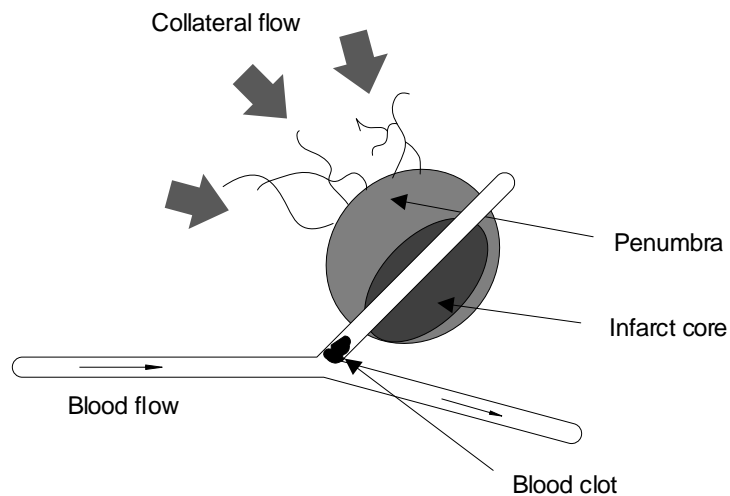
Figure 2.1: A schematic of ischemic stroke regions. The damage to the infarct core is irreversible, whereas the penumbra can be possibly saved by immediate treatment. The collateral flow provides microvascular perfusion to the ischemic regions, maintaining the salvageable tissues.

penumbra is a salvageable tissue region and has hypoperfused tissue that is at risk of dying, unless blood flow is restored. Penumbra tissue can recover from damage and therefore is often the target for interventional treatments. Because penumbra size determines how much tissue can be saved in treatments, research is conducted actively to determine the factors that influence penumbra size. Factors including reperfusion success, clinical variables (e.g., age), and vascular risk factors (e.g., hypertension) are being investigated [53].

An additional important factor is time. "Time is brain" [54] is a term emphasizing the importance of time in treating acute ischemic stroke patients. For every minute a stroke is left untreated, an estimated 1.9 million neurons are destroyed, resulting in a larger infarct core and smaller penumbra. Early clinical trials on mechanical thrombectomy often failed because the door-to-imaging time was poorly controlled [7], [55], [56]. However, later clinical trials demonstrated one of the important factors to successful stroke treatments is time [57]–[61]. As motivated by "time is brain", computer tomography perfusion (CTP) imaging is becoming more popular over magnetic resonance perfusion (MRP) imaging in pre-treatment diagnosis. CTP requires an averagely shorter imaging time (about 78 minutes) than MRP and it generates similar amount of information to physicians for treatment decision [62]. However, CTP often requires a higher radiation dose to generate detailed anatomic and physiologic images for analysis, limiting its clinical usage [63]. Active research is being conducted in low-dose CTP and commercial tools are being developed for CTP imaging in clinical practice [64].

Another factor that affects the penumbra and infarct core size is the degree of collateral flow, which is defined as the alternate circulation around a blocked artery via nearby minor vessels [65]. Despite the occlusion in the blood vessel, ischemic tissue can still receive sufficient cerebral blood flow via a subsidiary network of vascular channels. With a good degree of collateral flow, a large volume of penumbra may survive before treatment. Good collateral circulation often correlates well with smaller infarct cores, better reperfusion, and good clinical outcomes [66], [67]. A recent

finding also suggests that collateral flow may be more important than time to determine the volume of infarct core within 60 minutes of stroke onset [68], [69].

2.2 Treatments

The current US Food and Drug Administration (FDA) approved treatments for ischemic stroke are tissue plasminogen activator (tPA) and clot retrieval devices. tPA is often administered intravenously, with a cutoff time at three hours within stroke onset for all patients or an extended cutoff time at 4.5 hours within stroke onset for a group of eligible patients [70]. Patients may receive clot retrieval devices up to six hours within stroke onset [71]. In 2018, FDA approved the use of a particular clot retrieval device (i.e., Trevo) to treat certain stroke patients up to 24 hours after stroke onset [72]. The following sections describes the overview of each treatment.

2.2.1 Intravenous and intra-arterial tPA

Tissue plasminogen activator is a protein that breaks down blood clots to restore blood flow. tPA can be administrated intravenously (IV) or intra-arterially (IA), which is known as IV tPA and IA tPA respectively. IV tPA remains effective within 4.5 hours of stroke onset, yet it dramatically increases the chance of parenchymal hematoma (intracerebral bleeding) beyond this time point [73]. Therefore, the clinical guidelines suggest the use of IV tPA only up to 4.5 hours within stroke onset. IA tPA may improve patient outcomes between 4.5 hours and 6 hours of stroke onset, according to two clinical trials, the Prolyse in Acute Cerebral Thromboembolism II (PROACT II) [74] and the Middle Cerebral Artery Embolism Local Fibrinolytic Intervention Trial (MELT) [75]. However, there is still insufficient clinical data to conclude the optimal treatment time window for IA tPA and no corresponding clinical guidelines are established [76].

Figure 2.2: A Trevo revascularization device. The image is obtained from stryker.com [79].

2.2.2   Clot retrieval devices

All clot retrieval devices adopt a catheter-like design to physically remove clots. MERCI retriever was the first clot retrieval device released in 2004 [77], followed by Penumbra Systems (2009) [78], Trevo retriever (2012) [3], and Solitaire revascularization device (2012) [4]. In 2013, three clinical trials reported that the mechanical thrombectomy is no more effective than IV tPA [7], [55], [56]. Two years later, five clinical trials reported that clot retrieval devices are indeed highly beneficial as compared to using IV tPA alone with better imaging time and newer devices in patients with large vessel occlusion [57]–[61]. In 2016, two clot retrieval devices (Trevo and Solitaire) were approved by FDA. An example of Trevo device is shown on Figure 2.2 [79].

2.2.3   Revascularization

The goal of stroke treatments is to revascularize blocked blood vessels and save penumbral tissue. A roadmap was published in 2013 and suggested that revascularization is a combination of three different mechanisms: recanalization, reperfusion, and collateralization [80]. In the past, researchers often evaluated treatment effectiveness based on whether recanalization was achieved [81]. Yet recent work suggests that reperfusion is as important as recanalization [82]. Recanalization and reperfusion are strongly associated but not always equivalent in ischemic stroke [83]. It is therefore important to apply treatment that can reperfuse ischemic tissues in a short amount of time when collateral flow is poor [84].

Pre-FLAIR	Pre-DWI	Pre-ADC	Post-FLAIR
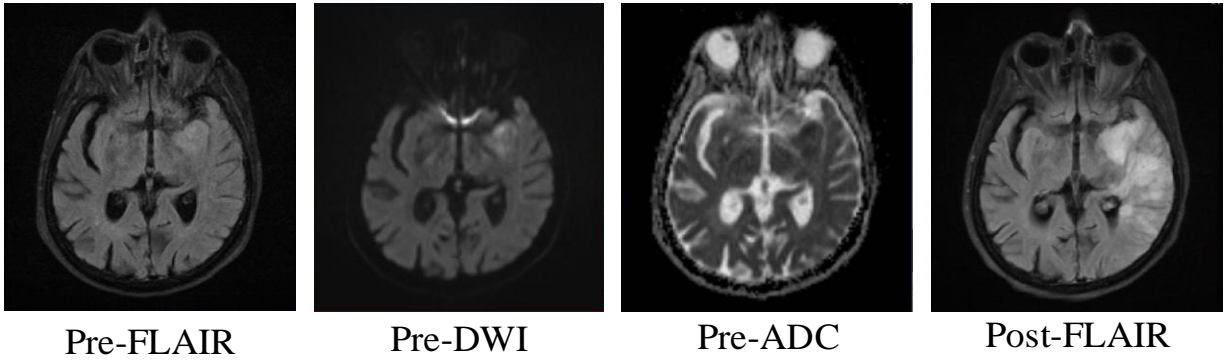
Figure 2.3: Imaging sequences for a patient with MCA (middle cerebral artery) stroke. Images include a pre-treatment FLAIR image (pre-FLAIR), pre-treatment DWI (pre-DWI), pre-treatment ADC map (pre-ADC), and a post-treatment FLAIR image (post-FLAIR).

2.3   Stroke Imaging

CT imaging and MR imaging are commonly used in clinical acute ischemic stroke diagnosis to identify infarct and penumbra. CT imaging provides excellent evidence on intracerebral/subarachnoid hemorrhage in a short amount of time while MR imaging provides a more precise measurement and details in ischemic and penumbral regions identification [85]. In this dissertation, MR imaging is the primary research interest.

The following sections will describe several major types of pre-treatment and post-treatment MR images (Figure 2.3). They are: (1) diffusion-weighted imaging (DWI)/apparent diffusion coefficient (ADC) map, (2) T2-weighted fluid-attenuated inversion recovery (FLAIR) imaging, and (3) perfusion-weighted imaging (PWI).

2.3.1   Diffusion-weighted imaging and apparent diffusion coefficient map

DWI measures the random Brownian motion of water molecules within a voxel of tissue and it is used to quickly locate the ischemic parenchymal injury within the first few minutes after stroke onset. When acute ischemic stroke occurs, the random motion of water molecules decreases in the ischemic region because of swelling and cytotoxic edema. The decrease manifests as an increase

in the signal intensity on diffusion weighted images (DWIs). To entangle diffusion and relaxation effects on image contrast, an ADC map is often generated from multiple diffusion-weighted imaging sequences acquired from different b-values (the degree of diffusion weighting applied) [86], [87]. DWIs and ADC maps are used together to identify the area of ischemic core; the infarct core appears bright (hyperintense) in DWI and appears dark (hypointense) on ADC maps [26]. Yet, a reversal of DWI-ADC ischemic core may occur under certain circumstances, including patients who are treated within 4.5 hours of stroke onset [88], immediate reperfusion after a tPA treatment [89], and patients who have a small infarct core [90].

2.3.2 FLAIR

FLAIR images are generated from MR imaging with long repetition time (TR) and time to echo (TE) [91]. They suppress cerebrospinal fluid (CSF) effects on the image, so as to bring out hyperintense lesions. Typically, there are no signs of ischemic region on FLAIR images within six hours of stroke onset, and areas of hyperintensity slowly evolve thereafter. Therefore, pre-treatment FLAIR images are often a good imaging source to identify past patient lesions, and to estimate the onset time of a new stroke by measuring the mismatch of hyperintensities in DWIs [92]. Post-treatment FLAIR images are, in contrast, often used to measure the final infarct volume, which can be assessed as early as the subacute stage (3-6 days after treatment) [93].

2.3.3 Perfusion-weighted imaging

PWI generates spatio-temporal imaging data (4-dimensional) that shows the flow of a gadolinium-based contrast bolus into and out of the brain over time. The perfusion-weighted images (PWIs) contain concentration time curves (CTCs) for each brain voxel that describe the flow of the contrast (i.e., signal intensity change) over time. Model-based perfusion parameters such as cerebral blood volume (CBV), cerebral blood flow (CBF), time-to-peak (TTP), time-to-maximum (Tmax), and mean transit time (MTT) can be derived from PWIs and used to estimate the volume of

14
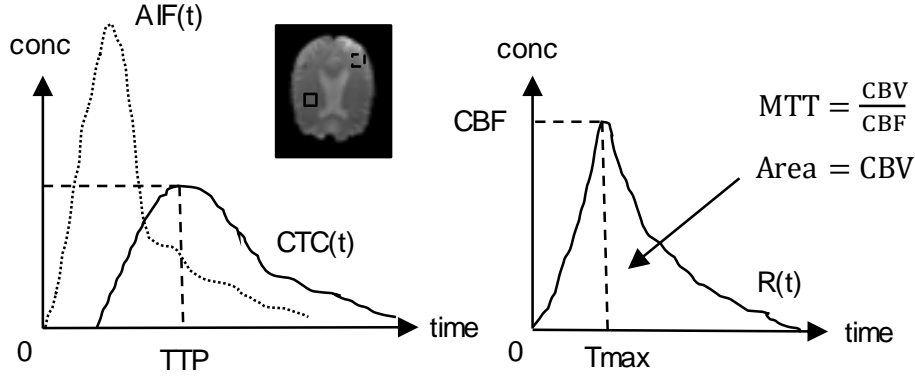
Figure 2.4: An illustration of the tissue concentration time curve (CTC), arterial input function (AIF), and residue function (R) corresponding to a voxel. The CTC and AIF are observable from the raw perfusion images, whereas R is obtained via deconvolution of CTC and AIF. The perfusion parameters are defined in the CTC(t) and R(t). CBV, CBF, Tmax, and MTT are normally defined in the R(t); the TTP is defined in the CTC(t).

salvageable brain tissue. To be able to compute these physiological parameters (CBF, CBV, TTP, Tmax, and MTT) from perfusion imaging, the CTCs at any voxel in the collected raw data must be processed. Because the injected contrast bolus is not a perfect impulse and varies across acquisitions, it is generally assumed that it follows the Indicator-Dilution Theory [94], i.e., the observed CTC of a voxel is the convolution of the arterial input function, *AIF(t)*, with a residue function, *R(t)*, scaled by CBF:

$$CTC(t) = \text{CBF} \cdot \big(AIF(t) * R(t)\big), \tag{2.1}$$

where *R(t)* represents the fraction of observed contrast remaining in the vasculature (within a voxel) at a certain time *t*, and *AIF(t)* describes the contrast input to the vasculature (within a voxel) at a certain time *t* [27]. The AIF can be defined locally or globally. The global AIF is defined from a large artery, such as middle cerebral artery (MCA), and it is used as the global input function for every voxel in a PWI. The local AIF is defined from the local small arteries and every voxel has its own unique local AIF. The global AIF is typically used in perfusion parameter estimation

15

because oftentimes local AIFs suffer from partial volume effect [27]. Figure 2.4 shows an illustration of a CTC and an AIF.

2.3.3.1 Deconvolution

The common standard technique to obtain *R(t)* is based on the deconvolution of CTC with AIF using singular value decomposition (SVD). The concentration time curve (CTC) from equation (2.1) can be first expressed in an integral form [95]:

$$CTC(t) = \text{CBF} \int_0^t AIF(t)R(t-\tau)d\tau, \tag{2.2}$$

where both *CTC(t)* and *AIF(t)* can be observed from the raw perfusion weighted images (PWIs). To obtain *R(t)* by SVD, equation (2.2) is first discretized to:

$$ctc(t_j) = \Delta t \cdot \text{CBF} \cdot \sum_{i=0}^{j} AIF(t_i) \cdot R(t_j - t_i), \tag{2.3}$$

where Δt is the sampling frequency. Equation (2.3) is then formulated as an inverse matrix problem:

$$\begin{bmatrix} ctc(t_0) \\ ctc(t_1) \\ \vdots \\ ctc(t_{N-1}) \end{bmatrix} = \text{CBF} \cdot \Delta t \begin{bmatrix} AIF(t_0) & 0 & \cdots & 0 \\ AIF(t_1) & AIF(t_0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ AIF(t_{N-1}) & AIF(t_{N-2}) & \cdots & AIF(t_0) \end{bmatrix}$$

$$\times \begin{bmatrix} R(t_0) \\ R(t_1) \\ \vdots \\ R(t_{N-1}) \end{bmatrix}, \tag{2.4}$$

$$\boldsymbol{c} = \boldsymbol{A} \cdot \boldsymbol{b}, \tag{2.5}$$

where $\boldsymbol{c}$ represents the *CTC(t)*, $\boldsymbol{A}$ represents the *AIF(t)*, and $\boldsymbol{b}$ represents the *R(t)* (constants are not shown for simplification). Using SVD, we can decompose $\boldsymbol{A}$:

$$A = U \cdot S \cdot V^T, \tag{2.6}$$

$$A^{-1} = V \cdot W \cdot U^T, \tag{2.7}$$

where S is a non-negative square diagonal matrix, and U and V are orthogonal matrices. W is equal to the inverse of S (i.e., S^{-1}). Then, b, which is also equivalent to $R(t)$, can be obtained as following:

$$b = V \cdot W \cdot U^T \cdot c. \tag{2.8}$$

2.3.3.2 Perfusion parameter maps

Perfusion parameters (CBF, CBV, TTP, Tmax, and MTT) are defined from $CTC(t)$ and $R(t)$. CBV describes the total volume of flowing blood in a given volume of a voxel and CBF describes the rate of blood delivery to the brain tissue within a volume of a voxel. CBV and CBF are used to derive MTT, which represents the average time it takes the contrast to travel through the tissue voxel. TTP is the time required for the CTC to reach its maximum, which approximates the time needed for the bolus to arrive at the voxel with delay caused by brain vessel narrowing or obstruction. Tmax is the time point where the contrast residue function reaches its maximum, which approximates the true time needed for the bolus to arrive at the voxel. Tmax>6s is often used as a cutoff time point to capture both the dead tissue core and the salvageable tissue that can possibly be saved by intervention aimed at restoring blood flow [96]. The mathematical expressions of these parameters are listed in the following:

$$\text{CBV} = \int_0^\infty R(t)dt, \tag{2.9}$$

$$\text{CBF} = \max(R(t)), \tag{2.10}$$

$$\text{MTT} = \frac{\text{CBV}}{\text{CBF}}, \tag{2.11}$$

$$\text{TTP} = \arg\max_t(CTC(t)), \tag{2.12}$$

$$\text{Tmax} = \arg\max_t(R(t)), \tag{2.13}$$

2.3.3.3 PWI-DWI mismatch

Perfusion parameter thresholds are defined to identify tissue regions that are dead, hypoperfused, or hyperperfused [97]–[99]. The most frequently used threshold is Tmax>6s, which captures hypoperfused regions that can possibly be saved by intervention aimed at restoring blood flow [96]. A PWI lesion (as defined by Tmax>6s) is often larger than a DWI lesion. The volume difference between lesions defines the PWI-DWI mismatch [16], which represents the volume of penumbra (salvageable tissue) that can be possibly saved by treatments.

2.4   Classification and predictive modeling in stroke

This section is divided into four parts to cover the background of classification and predictive modeling work in stroke outcome and imaging analysis: 2.4.1. Stroke Clinical Outcome Prediction describes the work on acute stroke patient mortality prediction; 2.4.2. Stroke Tissue Outcome Prediction describes the work on stroke tissue fate prediction after treatments are applied; 2.4.3. Perfusion Parameter Estimation describes the current techniques on generating perfusion parameter maps; 2.4.4 Time Since Stroke Classification describes the state-of-the-art imaging methods for classifying unknown acute stroke onset time.

2.4.1   Stroke clinical outcome prediction

Significant effort has been placed in creating models predictive of stroke and its outcomes since 1997. In early work, researchers were interested mostly in multivariate logistic regression techniques. Stineman *et al.* [100] created a logistic regression to predict patient functional recovery

status after rehabilitation and to provide a better understanding of discharge patient proportion. Counsell *et al.* [8] developed a six simple variable (SSV) multivariate logistic regression model to predict 30-day mortality using a dataset of 530 patients. The model was validated using two external independent cohorts of stroke patients with an area under the receiver operating curve (AUROC) of 0.840. Nam *et al.* [101] used univariate logistic regression analysis and found that percent improvement at 24 hours is an important factor for long-term outcome prediction.

While modeling techniques were the primary interest, effort has also been placed in identifying predictive clinical variables for acute stroke patient outcome. Weimar *et al.* [102] and König *et al.* [103] developed a multivariate logistic regression model and found out both the age and pre-treatment National Institutes of Health Stroke Scale (pre-NIHSS) are strong predictors for mortality. Later, Smith *et al.* [104], Yoo *et al.* [18], and Fonarow *et al.* [105] also showed that pre-NIHSS score provides substantial incremental information for mortality prediction.

Unlike most models that were built to predict mortality probability using logistic regression, Saposnik *et al.* [106] developed a coefficient-based scoring method, "iScore", to calculate patient mortality risk score based on information routinely available in hospital, such as demographics, clinical presentation (diplopia, dysarthria, aphasia, etc.), and patient comorbidities. Additional scoring models are under development, including the DRAGON score [107], and the HIAT2 score [108]. Table 2.1 summarizes the existing prediction and scoring models for stroke patient outcome prediction.

While significant work has been done in predicting ischemic stroke patient outcomes and identifying clinical predictors, there has been little work regarding the problem of between-class imbalance, which occurs frequently in binary prediction tasks when one class (the majority) is more common than the other class (the minority) [13]. Additionally, limited work has been done to investigate the effect of clinical variables and treatment variables in stroke patient outcome prediction. Therefore, one of the contributions in this dissertation is to address the imbalance

Table 2.1: Summary of existing models on stroke patient outcome prediction

Author	Outcome	Model	Variable in the best model	Performance
Stineman *et al.* (1997) [100]	Functional independence at discharge	Logistic regression	Age, and 9 other variables	N/A*
Counsel *et al.* (2002) [8]	30-day mortality	Forward stepwise multiple logistic regression	Age, and 5 other variables	AUROC: 0.840
Weimar *et al.* (2004) [102]	100-day mortality	Binomial logistic regression	Age, pre-NIHSS	Accuracy: 0.747
König et al. (2008) [103]	90-day mortality	Logistic regression with intercepts adjustment	Age, pre-NIHSS	Accuracy: 0.748
Nam *et al.* (2009) [101]	90-day mRS[¥] after IV-tPA	Logistic regression	Smoking, and 3 other variables	N/A
Smith *et al.* (2010) [104]	In-hospital mortality	Logistic regression	Age, and 13 other variables	AUROC: 0.850
Yoo *et al.* (2010) [18]	90-day mRS	Logistic regression	Pre-NIHSS[†], acute DWI and MTT lesion volumes	N/A
Saposnik et al. (2011) [106]	30- day mortality	iScore	Age, and 11 other variables.	AUROC: 0.850
Fonarow *et al.* (2012) [105]	30-day mortality	Logistic regression	Pre-NIHSS[†], and 16 other variables	AUROC: 0.820
Strbian et al. (2012) [107]	90-day mRS	DRAGON score	Age, and 4 other variables	AUROC: 0.800
Sarraj et al. (2013) [108]	Discharge mRS after IA-tPA	HIAT2 score	Age, and 3 other variables	N/A

[*]N/A = Not available

[¥]mRS = modified Rankin Score; scale: 0 (no stroke symptoms) – 6 (dead)

[†]Pre-NIHSS = pre-treatment National Institutes of Health Stroke Scale; scale: 0 (no stroke symptoms) – 42 (severe stroke)

problem in acute ischemic stroke dataset and investigating the correlation of clinical variables in patient outcome prediction. The details of this work are described in Chapter 6.

2.4.2 Stroke tissue outcome prediction

Models have been developed to predict tissue outcome, and estimate the growth of infarcts in order to provide more information for clinicians to make treatment decisions. One of the earliest models is the MR tissue signature model developed by Welch *et al.* [41], which utilized the ADC and T2 images to identify reversible and irreversible volumes in ischemic brain regions. Another early model is the generalized linear model (GLM) developed by Wu *et al.* [36]. This model used patients' DWIs and perfusion parameters (e.g., CBF) to predict tissue outcome at a voxel level. The result shows that using both PWIs and DWIs provided better performance in prediction compared to using DWIs alone. Later, Wu *et al.* [37] applied the same model to investigate the effectiveness of tPA on animal data. Nguyen *et al.* [38] further improved the basic GLM by introducing a correlation term that integrated spatial correlation information of voxels.

In addition to linear classifiers, non-linear classifiers have been explored. Bagher-Ebadian *et al.* [39] developed a four-layer artificial neural network (ANN) to predict the final extent of the three month post-stroke T_2-lesion in stroke patients using T_1-weighted, T_2-weighted, diffusion-weighted, and proton density-weighted images. Huang *et al.* [40] developed an ANN using voxel-wise neighboring information to predict stroke tissue outcome of rats. The model was trained to use a spatial infarction incidence map and nearest-neighborhood information (eight neighboring voxels for a 2D patch; 26 neighboring voxels for a 3D patch). Scalzo *et al.* [22] exploited spectral regression kernel discriminant analysis (SR-KDA) [109] to predict voxel infarction using voxel time-to-maximum (T_{max}) and ADC values. SR-KDA is an effective algorithm to project high-dimensional non-linear distributed data into a low-dimensional space, enabling efficient non-linear dimension reduction. The result shows that learning non-linear functions of the neighboring relationships of a patch is important for classification.

Clustering-based methods have also been investigated. Carano *et al.* [110] proposed a multispectral image analysis of four MR parameters (ADC value, T_2, M_0, CBF) to predict ischemic

tissue fate on animal data. The analysis included the use of two unsupervised (K-means, fuzzy C-means) and two supervised (multivariate Gaussian, K-nearest neighbor) clustering techniques. Rose *et al.* [111] developed a parametric Gaussian mixture model to predict infarcted tissue. To further improve the performance, Rose *et al.* [112] later developed a model based on expectation maximization and K-means clustering algorithms to account for the time-delay problem in perfusion parameters. Shen *et al.* [113] developed an iterative-self-organizing-data-analysis-algorithm (ISODATA) to forecast rat tissue outcome longitudinally on a pixel-by-pixel basis. The result shows that the model performed the best with ADC and CBF maps.

Other methods have included region-growing approaches [114] and instance-based learning [115]. While most of the existing models incorporated features from multiple imaging sequences, only two studies have investigated the use of voxel neighboring information to improve performance [22], [40]. Surprisingly, there has only been one model [37] that studied the effect of treatment (IV tPA) on tissue outcome prediction. In addition to that, all models have utilized hand-crafted features extracted from perfusion parameter maps (e.g., CBV), which may be problematic because of the vascular delays caused by physiologic changes [26] and signal distortions caused by the deconvolution process [28]. This motivates part of this dissertation work in developing a novel deep learning approach to automatically learn predictive imaging features for voxel-wise tissue outcome prediction. The details of this work are described in Chapter 3.

2.4.3   Perfusion parameter estimation

Perfusion parameter maps are important sources of imaging features to predict stroke patient clinical (mortality) and imaging (tissue fate) outcomes. Singular value decomposition (SVD) is the most common deconvolution technique to calculate the perfusion parameter maps. While SVD is simple and quick, it has limitations on time delays and therefore variations of SVD have been proposed. Delayed-corrected SVD (dSVD) [116] was developed to perform deconvolution while doing delay correction for contrast delay. Another delay-insensitive method is the block-circulant

SVD (bSVD) [95], which employs a block-circulant decomposition matrix to remove the causality assumption built into standard SVD. Oscillation-index SVD (oSVD) [95] is a variant of bSVD, which used an oscillation index in an iterative process of repetitive bSVD deconvolution to identify the best residue function. Because the acquired concentration curves are generally very noisy, SVD techniques may produce residue functions that are not physiologically plausible and subject to distortions that can underestimate the perfusion parameters [28], which may negatively impact tissue outcome prediction [29]. Other less common techniques are being developed to improve perfusion parameter estimation, including a Bayesian approach [117], Gaussian process deconvolution [118], and Tikhonov regularization [119].

While a number of techniques have been developed, limited work has been done using a deconvolution-free machine learning approach to approximate perfusion parameters. Perfusion parameter estimation is an important step in classification and predictive modeling in stroke because most models rely on perfusion parameters as input features. This motivates part of this dissertation work in developing an alternative solution to estimate perfusion parameters based on recognizing patterns from input imaging data using deep learning. The details of this work are described in Chapter 5.

2.4.4 Time since stroke classification

Time since stroke (TSS) symptom onset is a key factor in selecting tPA candidates, in which guidelines support its administration within a maximum of 4.5 hours from TSS. Yet up to 30% of acute stroke patients have unknown TSS [48]. Typically, the ischemic stroke tissue is nearly immediately visible in DWI at stroke onset whereas it takes 3-4 hours for the ischemic tissue to appear in FLAIR [30]–[33]. Studies therefore utilize this biological phenomenon to investigate the use of a simple imaging feature, a mismatch pattern between magnetic resonance (MR) diffusion weighted imaging (DWI) and fluid attenuated inversion recovery (FLAIR) imaging, to estimate TSS. The mismatch pattern is known as "DWI-FLAIR mismatch," which is defined as the presence

DWI   FLAIR   DWI   FLAIR

DWI-FLAIR mismatch   No DWI-FLAIR mismatch
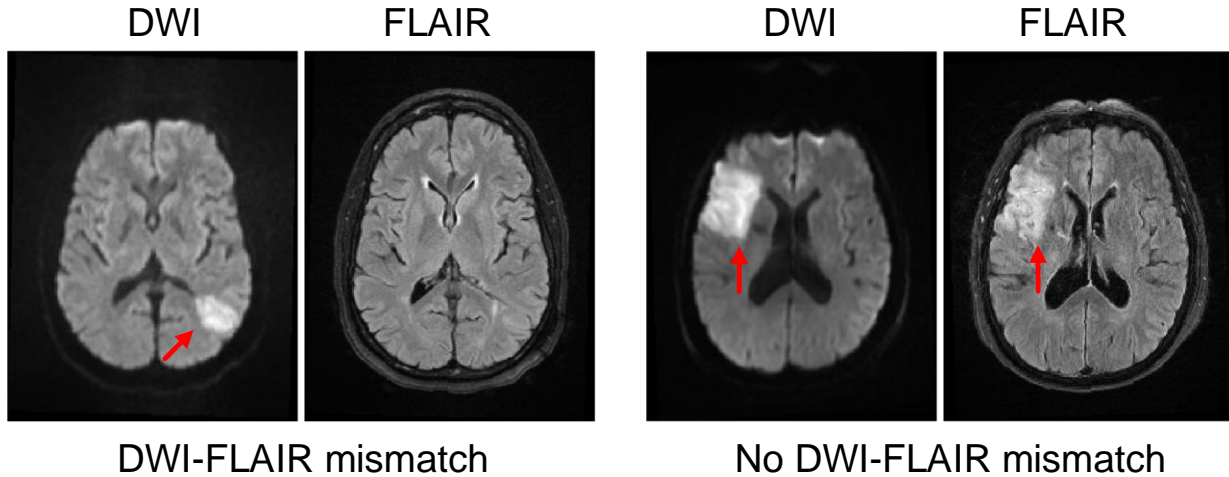
Figure 2.5: An example of DWI-FLAIR mismatch and no DWI-FLAIR mismatch. LEFT: presence of DWI-FLAIR mismatch (TSS = 1hr); RIGHT: absence of DWI-FLAIR mismatch (TSS = 8hrs). Hyperintensities are indicated by the red arrows.

of visible acute ischemic lesions on DWI with no traceable hyperintensity in the corresponding region on FLAIR imaging. Figure 2.5 shows an example of DWI-FLAIR mismatch and no DWI-FLAIR mismatch. In the left example of Figure 2.5, the clinical TSS is one hour and there is clearly a mismatch between the DWI and FLAIR. In contrast, both DWI and FLAIR have hyperintensities in the right example of Figure 2.5, in which the clinical TSS is eight hours.

The DWI-FLAIR mismatch method has been investigated in multiple studies to classify TSS<3hrs and TSS<4.5hrs. The work of using DWI-FLAIR mismatch was first introduced by Thomalla *et al.* [30], in which they used the mismatch pattern to identify stroke patients with less than 3-hour stroke onset. The method achieved a high specificity of 0.93 and a high positive predictive value (PPV) of 0.94, with a low sensitivity of 0.48 and a low negative predictive value (NPV) of 0.43. Aoki *et al.* [92] and Petkova *et al.* [120] followed the same method and applied it to their datasets. Both achieved a high sensitivity (0.83 and 0.90 respectively) and a high specificity (0.71 and 0.93 respectively), but Aoki *et al.* reported a moderate PPV of 0.64. Later, interest has been shifted to classifying TSS<4.5hrs because this may allow more acute stroke patients to be treated. Ebinger *et al.* [32] developed a mismatch model and it achieved a specificity of 0.79 and

Table 2.2: Summary of studies using DWI-FLAIR mismatch on TSS classification

Study	Label	Sensitivity	Specificity	Precision	NPV	AUROC
Thomalla *et al.*(2009) [30]	<3hrs	0.48	0.93	0.94	0.43	N/A
Aoki *et al.* (2010) [92]	<3hrs	0.83	0.71	0.64	0.87	N/A
Petkova *et al.*(2010) [120]	<3hrs	0.9	0.93	N/A	N/A	N/A
Ebinger *et al.*(2010) [32]	<4.5hrs	0.46	0.79	N/A	N/A	N/A
Thomalla *et al.* (2011) [31]	<4.5hrs	0.62	0.78	0.83	0.54	N/A
Emeriau *et al.* (2013) [33]	<4.5hrs	0.55	0.60	0.88	0.19	0.58

[*]N/A = Not available

a sensitivity of 0.46. Later, a large multicenter study was done by Thomalla *et al.* [31] to assess the ability of DWI-FLAIR mismatch. The mismatch method achieved a specificity of 0.78 and a PPV of 0.83, with a sensitivity of 0.62 and an NPV of 0.54. The study interobserver agreement of acute ischemic lesion visibility on FLAIR imaging was moderate (kappa = 0.569). Emeriau *et al.* [33] also investigated the use of mismatch pattern and the model achieved a high PPV of 0.88, but a low NPV of 0.19 with a moderate sensitivity of 0.55 and a moderate specificity of 0.60. The AUROC of using mismatch patterns in the identification of TSS was 0.58. Table 2.2 summarizes the existing studies on TSS classification.

There are also large multicenter clinical trials to further investigate the use of magnetic resonance DWI-FLAIR mismatch in guiding stroke treatment decisions, including the WAKE-UP trial in the European Union [121], the THAWS in the Japan [122], and the MR WITNESS trial in the United States [123]. A recent study has also shown that lesion water uptake obtained from CT images may estimate TSS more accurately than MRI [124], yet CT research on TSS classification remains limited. While much work has been done in estimating TSS, the developed methods are mostly limited by using a simple mismatch feature. This motivates part of this dissertation work

in developing a novel classification framework that is based on machine learning and deep learning techniques to classify TSS. The details of this work are described in Chapter 4.

2.5 Deep Learning

Conventional machine learning techniques are often limited in classification because they heavily rely on careful engineering and domain expertise to design feature extractors that can transform raw data (e.g., image) into a suitable feature representation (e.g., descriptive statistics of an image), which a classifier may use to classify patterns [52]. Deep learning methods are representation learning methods, which allow a machine to be fed with raw data and features are then automatically discovered for classification or prediction. Deep learning methods often consist of many layers of information-processing stages, and this hierarchical architecture is exploited for feature representation learning and pattern recognition [125]. A single building block of deep learning methods is a neuron, and a deep learning method often consists of many neurons in many layers, in which each neuron performs non-linear transformation to the input and generates an output for the next neuron. The simplest form of a deep learning model is a multi-layer neural network. Variants of neural network methods have been proposed in both supervised and unsupervised manners, including convolutional neural networks [126], stacked autoencoders [127], recurrent neural network [128], U-Nets [129], generative adversarial networks [130], and deep Boltzmann machines [131]. The first method receives the most attention in recent computer vision research [52] and is the primary focus in this dissertation. In the subsequent sections, the background of multi-layer neural networks is first introduced, followed by the background of convolutional neural networks and its association with multi-layer neural networks. The training details of neural networks are then described, with a literature review of deep learning applications in natural image and medical image analyses.

2.5.1 Multi-layer neural networks

A single layer neural network is called perceptron, which maps its input vector to an output value via a linear or non-linear mathematical transformation. A multi-layer neural network (NN) consists of multiple hidden layers with many neurons. In an NN, a neuron is always connected to the next layer neurons via a weight vector with linear or non-linear transformation. The final output value of an NN is used for classification based on a decision function. One example of a decision function can be:

$$f(x) = \begin{cases} 1 \text{ if sigmoid}(\boldsymbol{W} \cdot \boldsymbol{X} + b) > \boldsymbol{0} \\ 0 \text{ otherwise} \end{cases}, \tag{2.14}$$

where $\text{sigmoid}(\cdot)$ is a non-linear sigmoid transformation [132], $\boldsymbol{W}$ is the weight vector, $\boldsymbol{X}$ is the input vector, and b is the bias. A shallow NN is a model that contains only one or zero hidden layers; a deep NN is a model that contains multiple hidden layers with many neurons. While a deep NN is often needed for classifying highly non-linear data, training a deep NN can be inefficient because of a large number of trainable parameters. This limitation motivates the creation of convolutional neural networks (CNNs) [126], which have fewer trainable parameters while achieving superior performance in many challenging tasks.

2.5.2 Convolutional neural networks

Convolutional neural networks (CNNs) are invented to mimic human eye operations in observing an object and sending signals to the brain via receptive fields [133]. Therefore, CNNs have unique architectures that allow much easier training and better generalizability than NNs [52]. CNNs reduce learning computation requirements and overfitting by sharing weights within layers. They also possess a translation invariance characteristic that make them robust to translation artifacts in images. A deep CNN typically consists of four standard layer types: a convolutional layer, pooling layer, fully-connected layer, and output classification layer. Figure
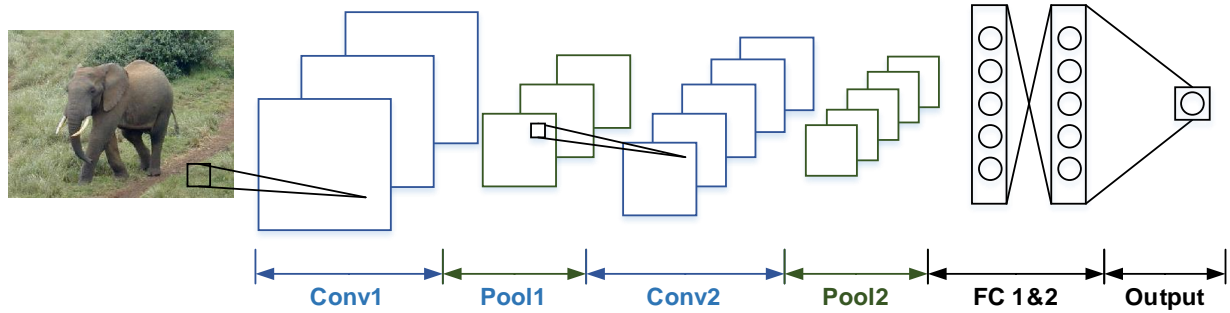
Figure 2.6: An example of a convolutional neural network. Conv: a convolutional layer; Pool: a pooling layer; FC: a fully-connected layer. The image is obtained from Caltech-101 dataset [242].

2.6 shows an example of a CNN with two layers of each type and an output layer. A pooling layer often follows a convolutional layer; fully-connected layers are usually the last few layers before the output layer. A non-linear activation function is often added between layers to introduce non-linearity to the network (not shown in Figure 2.6). Key concepts are as follows:

- **Weights**. They are adjustable parameters (continuous real values) that are learned during model training. Every weight value represents the numerical influence of one neuron on another in a connection. In a typical deep learning model, there may be a hundreds of millions of adjustable weights [52], which requires many labeled data to learn.

- **Convolutional layer.** Every local region of an input, also known as a local receptive field [134], is connected to neurons in a subsequent layer. A local receptive field is associated with a weight matrix, which is a feature filter to detect input patterns using convolution operations. However, it is impossible to have weight matrices for all local regions in a convolutional layer because this might result in many tunable weights. Therefore, a weight-sharing strategy is often applied, i.e., a fixed number of weight matrices are learned and shared among all the local regions. Through using these weight matrices, elementary features, such as oriented edges and endpoints, may be captured from an input. The outputs of a convolutional layer are therefore also called feature maps. Feature filter size, stride value, and zero-padding value are three important hyperparameters in a convolutional layer [135]. Feature filter size controls a receptive field size,

28
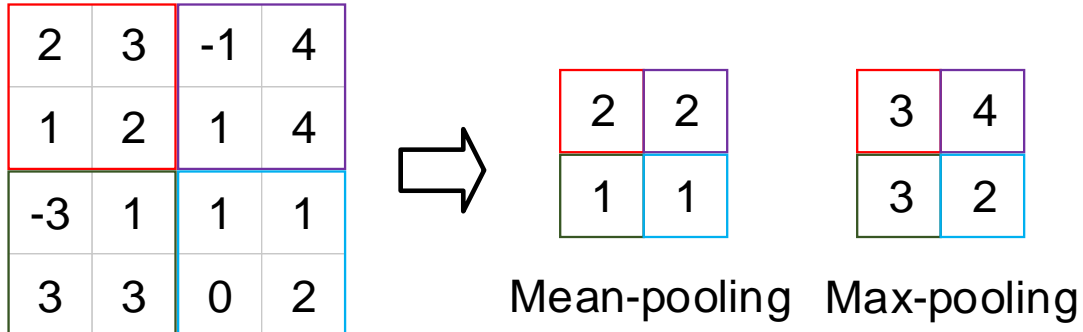
Figure 2.7: An example of a mean-pooling and a max-pooling on a 4×4 input.

which defines the area of a local region for consideration. Stride and zero-padding values control how much spacing and area are needed for a convolution to be performed.

- **Pooling layer.** Local characteristics of input values are summarized via pooling operations (mean-pooling or max-pooling). Pooling reduces input spatial size and limits the number of adjustable weights, and therefore controls overfitting. The pooling layer produces summary statistics for the next layer that are robust to local transformations. In mean-pooling, averages of local regions are calculated. In contrast to mean-pooling, max-pooling only captures maximum values of local regions and passes them to the next layer. Figure 2.7 shows an example of mean-pooling and max-pooling on a 4×4 input.

- **Fully-connected Layer.** Every fully-connected layer neuron connects to every next-layer neuron via a weight vector. Because every neuron is connected to the next layer, there are often thousands of trainable weights and therefore only two to three fully-connected layers are used in a typical CNN. Fully-connected layers are usually used before an output layer.

- **Output Classification Layer.** Classification is performed in this layer. It usually comes with a softmax classifier [136], which is a generalized form of logistic regression. Softmax classifier is trained in a supervised manner and its cost function is defined as:

$$J(\boldsymbol{W}, b; \boldsymbol{X}, y) = -\frac{1}{m}\left(\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\} log \frac{e^{w_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{w_l^T x^{(i)}}}\right), \tag{2.15}$$

29
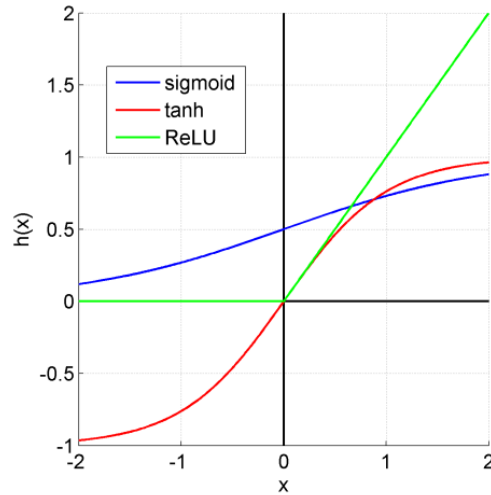
Figure 2.8: Three typical activation functions in deep learning. Sigmoid: sigmoid function; tanh = hyperbolic tangent function; ReLU = rectified linear unit

where m is the number of training data, k is the number of unique labels, $\boldsymbol{W}$ is the weight vector, b is the bias constant, $\boldsymbol{X}$ is the input vector, and y is the label. A L1- or L2- regularization term may be added to regularize the learning of the parameters and avoid overfitting [137].

- **Activation Function.** A non-linear activation function is usually added between fully-connected layers and after the convolutional/pooling layers. Widely accepted activation functions are sigmoid and hyperbolic tangent functions. The major difference between a sigmoid function and a hyperbolic tangent function is the upper and lower bounds: the former has a range of [0, 1], whereas the latter has a range of [-1, 1]. Rectified linear unit (ReLU) is another common activation function, which improves network sparsity learning and enables faster and more stable learning [132]. Figure 2.8 plots the activation functions and their mathematical expressions are as follows:

$$\text{sigmoid function} = \text{sigmoid}(x) = \frac{1}{1 + \exp(-x)}, \tag{2.16}$$

$$\text{hyperbolic tangent function} = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{2.17}$$

$$\text{ReLU} = \text{rectifier}(x) = \max(0, x), \tag{2.18}$$

30
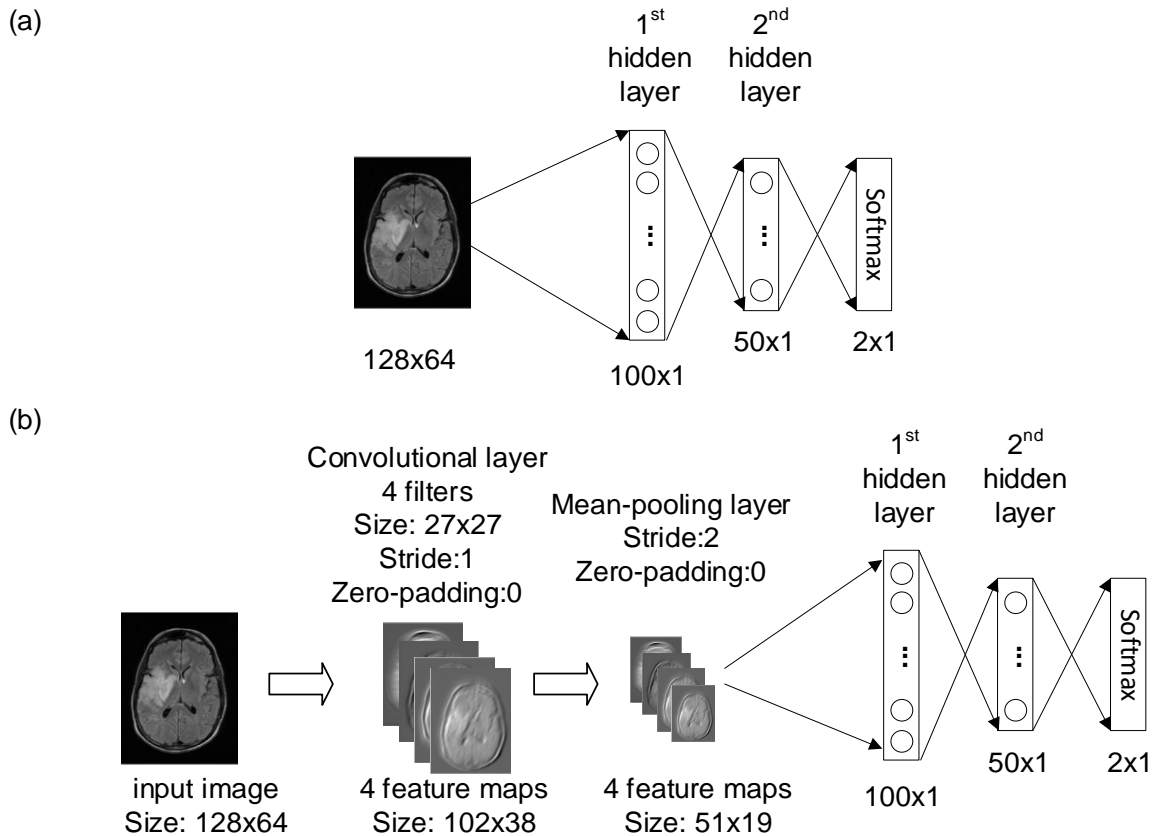
Figure 2.9: An example of an NN and a CNN. (a) NN; (b) CNN. The number of adjustable weights in the CNN (104,916) is approximately 8 times fewer than the NN (824,300).

- **Additional Components.** Dropout [135] is an operation added to a fully-connected layer, in which every neuron is randomly turned off based on a probability, p, during training. Dropout prevents complex co-adaptions among feature filters and helps preventing overfitting. Batch normalization [138] is another technique to prevent overfitting and speedup training, in which outputs of a layer are normalized (i.e., reduce covariance shift) before passing to the next layer.

2.5.3 Neural networks versus convolutional neural networks

The architecture of CNNs allows translation invariance and feature filter learning for pattern recognition. Additionally, CNNs typically have fewer weights than NNs, enabling faster training and avoiding overfitting. Figure 2.9 shows an example of an NN and a CNN to illustrate how

31

efficient a CNN can be. The NN has two fully-connected layers and an output layer, whereas the CNN has a convolutional layer and pooling layer, followed by two fully-connected layers and an output layer. The convolutional layer has four feature filters, each one with a filter size of 27×27, stride of 1, and zero-padding of 0. Assuming there are no bias constants in both networks, the total number of weights in the NN is: $128 \times 64 \times 100 + 100 \times 50 + 50 \times 2 = 824{,}300$. In contrast, the total number of weights in the CNN is: $27 \times 27 \times 4 + 51 \times 19 \times 100 + 100 \times 50 + 50 \times 2 = 104{,}916$, which is approximately eight times fewer than the NN.

2.5.4 Training a deep network

Training a deep network refers to learning a set of weights that minimize a loss function while maximizing an evaluation metric (e.g., accuracy). Training a deep network can be supervised, where both input and output is required, or unsupervised, where only input is required, depending on the deep network architecture [139]. For example, an autoencoder is often trained in an unsupervised manner while a CNN is often trained in a supervised manner. In general, deep network training is performed by iteratively minimizing a loss function while adjusting the network weights. Several components are required in training: a loss function, gradient descent, and backpropagation. Key concepts include the following:

• **Loss function.** This is a function that is optimized during training. It is important to define a suitable loss function for a training task because it determines how fast a network converges to an optimal point. An example of a simple loss function is the mean square loss between the input and the output, which is often used as a reconstruction loss in an autoencoder [127] :

$$J(\boldsymbol{W}, b; \boldsymbol{X}, y) = \frac{1}{2} \left\| y - h_{\boldsymbol{W}, b}(x) \right\|^{2}, \tag{2.19}$$

where $\boldsymbol{W}$ is the weight vector, b is the bias constant, $\boldsymbol{X}$ is the input vector, y is the label, and $h_{\boldsymbol{W}, b}(x)$ is the network output. Another example of typical loss function is softmax loss for multi-class classification (Equation 2.15).

- **Gradient descent.** This is an optimization algorithm that is commonly used in deep network training [140], [141]. General gradient descent requires a complete forward pass of all inputs in order to compute the overall loss and update the network weights. Variants of gradient descent are proposed: stochastic gradient descent [142] and mini-batch gradient descent. A weight update is performed for each training example (instead of all training examples) in stochastic gradient descent, which eliminates redundant computations [142]. Yet, per-example update has high variance that causes the objective loss function to fluctuate heavily. Therefore, mini-batch gradient descent is often used instead, in which the loss of a mini batch of training data (e.g., n=20) is computed in every update. The weights are then updated by gradients (which are calculated from the loss), namely:

$$W = W - \alpha \frac{\partial}{\partial W} J(W, b), \qquad (2.20)$$

where W is the weight vector, $J(W, b)$ is the loss, the derivative is the gradient, and α is the learning rate. The learning rate is a user-defined hyperparameter that determines how much a weight is updated. Because it is a sensitive hyperparameter, it often requires careful manual tuning. Alternative gradient ascent algorithms have been proposed to adaptively adjust the learning rate during training and these algorithms show better performance than general gradient descent. These algorithms include AdaGrad [143] , Adadelta [144], and Adam [145].

- **Backpropagation.** This refers to the movement of the loss (calculated from a loss function) from the output layer to the input layer, in which the weight gradients are calculated, and weights are updated. One important trick of backpropagation for gradient calculation is the "chain rule trick" [126], which enables gradient computations in all the hidden layers.

- **CPU and GPU training.** Training a deep network requires high computation power due to many adjustable weights. To address this challenge, many tools exploit hardware features such as multi-core computer processing units (CPUs) (10-20) and many-core graphical processing units (GPUs) (1000+) to shorten the training and inference time [146]. For example, a high-end GPU

(e.g., Nvidia Tesla K40c) can achieve about 10x improvement in training speed than a high-end CPU (e.g., Intel Xeon E5-2697 v2).

2.5.5 Deep learning in natural and medical images analysis

Many state-of-the-art classification records have been achieved by deep learning models [52]. The success of these models is heavily based on their powerful capability to automatically learn to detect patterns from input images [126], [147]. These data-driven patterns are learned by hierarchical convolutional feature filters, which have been shown to be effective in detecting local characteristics that improve classification [148].

One example is the ImageNet competition [50], in which participating teams are ranked based on the performance of their classifiers on classifying 1,000 different image categories. The winning teams from the recent years' competition all built their classifiers based on deep learning techniques, and teams had built deep learning classifiers with new architectures and building blocks that can even achieve error rates lower than the human error. For example, Ioffe *et al.* [138] proposed the batch normalization method that can improve deep model performance by reducing covariance shifts of training data. He *et al.* [149] proposed a new deep model architecture that is designed to learn residual functions with reference to a layer input, which makes the learning easier to optimize. Hu *et al.* [150] developed a new building block, the squeeze-and-excitation block, that can adaptively recalibrate channel-wise feature responses (i.e., channel feature importance) to significantly improve the model performance in ImageNet competition. All the techniques are generalizable to different network architectures and have been adapted widely in many imaging applications. The applications of deep learning techniques are not limited to two-dimensional imaging data, but also include higher dimensional data, such as video classification [151]–[154]. Instead of using two-dimensional spatial feature filters, spatio-temporal filters are learned during training, and they are used to extract meaningful patterns from input videos for classification.

In the medical domain, researchers have recognized the tremendous ability of deep learning techniques in image and video analysis and have begun to apply these techniques in medical image challenges. In segmentation tasks, Davy *et al.* [155] developed a multi-scale CNN approach in a cascade architecture that exploits both local features as well as more global contextual features to perform brain tumor segmentation. The model achieved state-of-the-art performance while being over 30 times faster than a single-scale CNN. Ronneberger *et al.* [129] alternatively proposed a new CNN architecture, U-Net, for favors precise localization using symmetric expanding paths to improve cell segmentation. This architecture has been adapted in many applications, including high-resolution histological segmentation [156]. Recently, Li *et al.* [157] have developed a region-based convolution neural network that utilizes R-CNN [158] to achieve an epithelial cell segmentation with an accuracy of 99.1%.

In medical prediction and classification tasks, Suk *et al.* [159] used a Deep Boltzmann Machine to predict Alzheimer's Disease and its prodromal stage by learning joint imaging features from MRI and positron emission tomography. Shin *et al.* [160] tested the application of unsupervised deep autoencoders on organ (e.g., liver, kidney, spleen) identification on MR images. Roth *et al.* [161] proposed a classification method that exploits random aggregation of deep CNN outputs from rotated image patches to predict bone lesions. Ertosun *et al.* [162] trained a deep CNN for automated classification of gliomas grading using digital pathology images. All of these proposed deep learning models have shown superior performance over the existing methods.

Notably, several of these models are inspired by the clinical understanding of the diseases, such as the multi-scale CNN approach for local and global feature extraction in tumor segmentation [155]. Chapters 3 and 4 detail the deep network architecture designs on stroke tissue outcome prediction and stroke onset time classification, both of which are driven by stroke pathophysiology.

CHAPTER 3

Stroke Tissue Outcome Prediction Using Deep Learning

3.1 Overview

Treatment guidelines have been established to limit the morbidity and mortality associated with acute stroke [6], [76], [163]. Researchers have actively pursued new methods and information to assist neuroradiologists in stroke treatment guidance. One area of research is to predict stroke infarct volume from magnetic resonance perfusion weighted imaging (PWI), which could provide helpful information to clinicians in deciding how aggressively to treat acute stroke patients [22], [36], [113]–[115], [37]–[41], [110]–[112]. While these techniques have proven to be useful, there are two major limitations. First, the proposed models were mostly built using hand-crafted features (e.g., perfusion parameters) derived from perfusion images, and performance may be limited due to feature simplicity. Alternatively, deep learning techniques have not been fully explored in tissue outcome prediction, yet they have drawn tremendous interest in other areas of research because of their ability to learn data-driven features for state-of-the-art performance [164]. Second, most of the existing models rely on estimated model-based perfusion parameters (e.g., cerebral blood flow (CBF)) to predict tissue infarction. Recently, concerns have been raised about the use of these parameters [24], [25], due to drawbacks discussed in several studies, including the sensitivity of perfusion parameters to vascular delays [26], and their robustness to distortions from deconvolution process [28].

In this chapter, we first present a deep convolutional neural network (CNN) that predicts voxel-wise stroke tissue outcome using source perfusion weights images (PWIs). This model consists of several novel designs to improve the prediction performance: 1) patch sampling to incorporate

36

neighboring voxel information, 2) the use of matched controls to improve feature learning, and 3) a new layer to learn paired unit filters for temporal feature learning. The proposed model is compared to several existing tissue outcome models and baseline deep CNNs. Second, I describe analyses to predict final infarct volume using only clinical variables and investigate whether treatment variables can improve imaging outcome prediction. In the remainder of this chapter, Section 3.2 details the data and preprocessing steps used in this study. Section 3.3 describes the proposed CNN architecture. Section 3.4 describes the experimental setup for voxel-wise infarct prediction using imaging data and the validation results. Section 3.5 details the experimental setup and the results on using only clinical and treatment variables to predict final infarct volume. Section 3.6 discusses several interesting observations and the limitations. The content of this chapter have partly been published in [165].

3.2   Data Cohort and Data Processing

A total of 444 patients were identified from UCLA Research Electronic Data Capture (REDCap) database [166]. These patient MR images were retrieved and examined from the UCLA picture archiving and communication system (PACS) between December 2005 and December 2015 with the approval from institutional review board, IRB#11-000728. These patients have had been treated with none, one, or more of the following treatments: intra-arterial tissue plasminogen activation (IA tPA), intravenous tissue plasminogen activator (IV tPA), or clot retrieval devices. Two sets of data were collected: the first set was for the study in which source PWIs were used for the final infarct prediction; the second set was for the study in which clinical and treatment variables were used for the final infarct volume prediction.

3.2.1   Voxel-wise tissue outcome prediction using source PWI

The patient inclusion criteria were all patients with: 1) acute ischemic stroke due to middle cerebral artery (MCA) occlusion; 2) MR imaging performed before treatment and 3-7 days after

37

Table 3.1: Patient cohort characteristics for final infarct prediction using only source PWI.

	Patients (n = 48)
Demographics	
Age	65.4 ± 17.3
Gender	20 males
Clinical Presentation	
Time since stroke	206 ± 122 mins
NIHSS[†]	15.1 ± 7.7
Treatments (received)*	
IV tPA	24
IA tPA	6
Clot retrieval devices	29
Treatment Evaluation	
TICI score[¥]	0 (4), 1 (0), 2a (13), 2b (8), 3 (1), N/A (22)
AOL score[Ω]	0 (3), 1(0), 2(4), 3(16), N/A (25)
Outcome	
Discharge mRS[‡]	3.44 ± 1.47
Discharge lesion size	65.9 ± 81.7 cm^3

[†]NIHSS = NIH Stroke Scale International; scale: 0 (no stroke symptoms) - 42 (severe stroke)

*A patient could receive more than one treatment in a visit.

[¥]TICI = Thrombolysis in Cerebral Infarction; scale: 0 (no reperfusion), 1, 2a, 2b, 3 (full reperfusion), N/A (missing). Available only for patients with clot-retrieval devices.

[Ω]AOL = Arterial Occlusive Lesion; scale: 0 (complete occlusion) – 3 (complete recanalization), N/A (missing). Available only for patients with clot-retrieval devices.

[‡]discharge mRS = discharge modified Rankin Scale; scale: 0 (no stroke symptoms) – 6 (dead)

treatments (e.g., clot retrieval, tPA); and 3) absence of hemorrhage. A total of 48 patients satisfied these inclusion criteria and were used in the study. The patient characteristics are summarized in Table 3.1.

Final infarct volumes were semi-automatically determined and measured on the 3-7 days post-treatment (post-FLAIR) images [93] by an expert neuroradiologist (Dr. S. El-Saden) using Medical

Table 3.2: A list of clinical and treatment variables for final infarct prediction.

Clinical Variables	Treatment Variables
Gender, hypertension, diabetes, hyperlipidemia, atrial Fib, myocardial Infarction, congestive heart failure, NIHSS, age, time since stroke	IV tPA, clot retrieval device, IV tPA & clot retrieval device

Image Processing, Analysis, and Visualization (MIPAV) software [167]. Pre-treatment FLAIR images (pre-FLAIR) were used to identify pre-existent lesions that were not related to the current stroke; these lesions were not labeled as part of the final infarct volumes.

All patients underwent MR imaging using a 1.5 or 3 Tesla echo planar MR scanner (Siemens Medical Systems); scanning was performed with 12-channel head coils. The PWIs were acquired using a repetition time (TR) range of 1,490 to 2,890 ms and an echo time (TE) range of 23 to 50 ms. The pixel dimension of the PWIs varied from $1.00 \times 1.00 \times 5.00$ to $2.00 \times 2.00 \times 7.00$ mm. The pre-FLAIR images were acquired using a TR range of 8,000 to 10,000 ms and a TE range of 82 to 123 ms. The pixel dimension of the pre-FLAIR images varied from $0.45 \times 0.45 \times 3.00$ to $0.94 \times 0.94 \times 7.00$ mm. The post-FLAIR images were acquired using a TR range of 8,000 to 10,000 ms and a TE range of 82 to 134 ms. The pixel dimension of the post-FLAIR images varied from $0.43 \times 0.43 \times 3.00$ to $0.94 \times 0.94 \times 7.00$ mm. The perfusion parameter maps of time-to-maximum (Tmax), cerebral blood flow (CBF), time-to-peak (TTP), cerebral blood volume (CBV), and mean transit time (MTT) were calculated using block-circulant singular value decomposition (bSVD) as provided by the sparse perfusion deconvolution toolbox [168]. The z-dimension is ignored in patch data generation due to slice thickness. Therefore, the notation of the size of a "voxel" is simplified to 1 x 1 only (x-, and y- axes).

3.2.2 Final infarct volume prediction using clinical and treatment variables

The patient inclusion criteria were all patients with: 1) acute ischemic stroke due to MCA occlusion; 2) FLAIR imaging performed 3-7 days after treatment; 3) absence of hemorrhage; and 4) all clinical and treatment variables. A total of 82 patients satisfied these inclusion criteria and were used in the study. The clinical and treatment variables are summarized in Table 3.2. Similarly, the ground truth final infarct volumes were determined and measured on the 3-7 days post-treatment (post-FLAIR) images by an expert neuroradiologist (Dr. S. El-Saden).

3.2.3 Image preprocessing

Image preprocessing steps were performed on MR images before they were used. First, each voxel in pre-PWIs and post-FLAIR should align to the same anatomical location in pre-FLAIR. Therefore, intra-patient registration of axial pre-PWIs, pre-/post-FLAIR images, and post-FLAIR images were performed with a six-degree of freedom rigid transformation using FLIRT [169]. Then, Gaussian filters with a size of 2.35 mm full width at half maximum (FWHM) were applied to remove spatial noise for all pre-PWIs [170], followed by interpolation to ensure all pre-PWIs had the same unit increment in the time dimension. Multi-atlas skull-stripping algorithm [19] was used to remove the skulls and brain ventricle voxels were excluded in training. Each source signal of pre-PWIs, *S(t)*, was then converted to a tissue concentration time curve, *CTC(t)*, based on the baseline signal *S(0)*, *TE*, and β [171]:

$$CTC(t) = -\beta \, log \left(\frac{S(t)}{S(0)} \right) / TE, \qquad (3.1)$$

where *S(0)* is the average of the first five values of the signal curve, TE is the echo time of the MR image, and β is a scaling factor provided by Mouridsen *et al.* [171], with a value of 2000. All CTCs were standardized to zero-mean and unit-variance globally on a voxel-by-voxel basis [172].
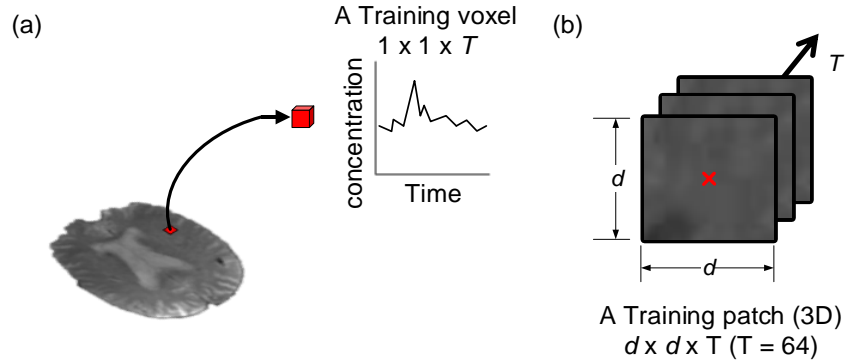
Figure 3.1: Training data generation. (a) A training voxel with a dimension of 1 x 1 x T, containing the concentration change along time t (concentration time curve). (b) A training patch with a dimension of d x d x T ($T = 64$). The center of the patch (red) is the voxel of interest. A patch is associated with the outcome value (0 or 1) of the voxel of interest.

3.3 Methods for Voxel-wise Tissue Outcome Prediction

This section details the methods for voxel-wise tissue outcome prediction. The task was to predict the final outcome of every brain voxel (i.e., infarcted or non-infarcted) given its concentration time curve (CTC) in PWIs. The ground truth binary mask was derived from the post-FLAIR images, where the positive class was infarcted and the negative class was non-infarcted. The proposed deep CNNs learned features directly from the CTCs to predict voxel-wise outcomes, which were different to previous approaches that often rely on perfusion parameters [22], [36], [113]–[115], [37]–[41], [110]–[112]. The proposed CNNs did not depend on the AIFs, which describes the overall contrast input to the vasculature within a voxel.

3.3.1 Patching sampling

One straightforward approach to generate training data from a perfusion signal associated with a given voxel is to only use a voxel's concentration time curve (Figure 3.1a). For example, a single training sample has a dimension of 1 x 1 x T, in which T is the total time length of a pre-PWI.

However, this approach is sensitive to noise (e.g., small motion artifacts can cause apparent concentration changes in a voxel) and it incorrectly assumes voxels are independent. Recent work [22] revealed that incorporating neighboring voxel information improved classification performance. Therefore, training data was sampled as a three-dimensional (3D) "patch" sequence instead of a single sequence; each training example had a size of *d* x *d* x *T*, where *d* is the width/height of the patch and the center of the patch is the voxel of interest (Figure3.1b). Deep CNNs then learn filters to extract spatio-temporal features from the patch to predict the outcome of the central voxel. The optimal value of *d* (*d*=25, as described in the next section) is determined in a nested 10-fold cross-validation.

3.3.2 Deep CNN architecture for voxel-wise tissue outcome prediction

3.3.2.1 Baseline deep CNNs

Typical deep CNNs consist of multiple convolutional, pooling, non-linear, fully-connected layers, and a softmax classifier. The convolutional layers produce feature maps (usually > 10) from the input through automatically learned feature filters (weight matrices). The pooling layers pool values in a local region together using pooling operators to make the network more translation invariant. The non-linear rectified linear layer [173] (ReLU) is often added in between the convolutional and pool layers to enable the network to learn non-linear features. After multiple stacking of convolutional-ReLU-pooling layers, several fully-connected layers are added to generate summarized features, which are the inputs to the softmax classifier for classification. Given a deep CNN with *L* layers and *N* data samples (*X*, *y*), the parameters ($\boldsymbol{\theta}$) of a softmax classifier in binary form (i.e., logistic classifier) are obtained by minimizing the cost function:

$$\text{J}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log h_{\boldsymbol{\theta}}(\boldsymbol{X}_n^L) + (1-y_n)\log\left(1 - h_{\boldsymbol{\theta}}(\boldsymbol{X}_n^L)\right)\right] + \gamma f(\boldsymbol{\theta}), \qquad (3.2)$$

$$h_{\boldsymbol{\theta}}(\boldsymbol{X}_n^L) = \frac{1}{1+\exp(-\boldsymbol{\theta}\boldsymbol{X}_n^L)}, \qquad (3.3)$$

42
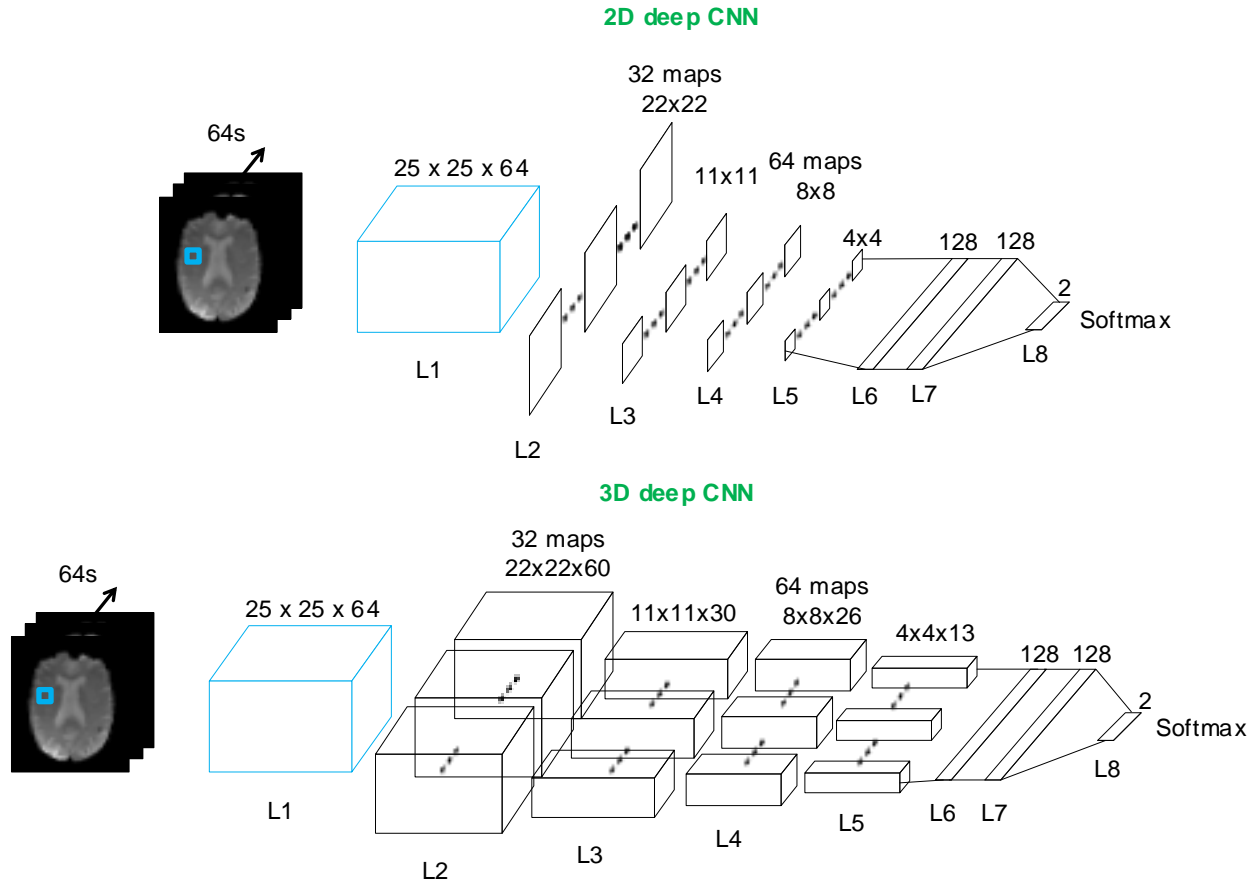
Figure 3.2: An eight-layer 2D deep CNN and an eight-layer 3D deep CNN for tissue outcome prediction. Input patch regions (width x height x time) are convolved in the convolutional layers, where different number of feature maps are learned (32 & 64). After two series of convolutional-ReLU-pooling layers, overall representations are learned in the two fully-connected layers, which form the input to a softmax classifier for binary classification (infarct or not-infarct). Input layer: L1; 2D/3D convolutional layer with ReLU: L2, L4; max-pooling layer: L3, L5; fully-connected layer: L6, L7; output softmax layer: L8.

where γ is a parameter for controlling the L_2 weight decay term ($f(\boldsymbol{\theta})$), $h_{\boldsymbol{\theta}}(\boldsymbol{X}_n^L)$ is the sigmoid function, and $\boldsymbol{X}_n^L$ is the output of the fully-connected layer before the softmax classifier. The weights in a deep CNN layer ($\boldsymbol{W}^{(l)}$) are updated via gradient descent and backpropagation [126].

There are two basic approaches to train deep CNNs with the 3D training data for voxel-wise tissue outcome prediction (Figure 3.2). In the first approach, the temporal dimension of the training data is treated as if it were a color channel. This therefore allows the use of a standard 2D deep CNN architectures that are typically applied to images [135]. The filter learning of a convolutional layer in a 2D deep CNN is two-dimensional. In the second approach, a 3D deep CNN architecture [152] can be used, which is composed of 3D filters instead of 2D filters in convolutional layers. Spatio-temporal features can be extracted progressively along the network and they are more descriptive for small changes in both the spatial and temporal dimensions.

3.3.3 Baseline CNN limitations

The motivation for using deep 2D and 3D CNNs is their strong ability to learn data-driven filters to obtain complex features that are predictive of infarction. However, when we trained these standard architectures using the perfusion image training patches, we implicitly assumed that every training patch was sampled from a distribution generated by the same global AIF, which describes the unique pattern of flow of the contrast agent traveling within the cerebrovascular, and which also reflects the effects of both of the administration method as well as the cardiac function and vasculature between the intravenous administration site and the brain [27]. This assumption may not hold across patients due to a variety of factors, such as a patient's unique cerebrovascular architecture. We can also define a lot of local AIFs [27], in which each is based on the closest artery to the voxel of interest. Yet this approach is difficult to use because the PWIs are low-resolution and it is challenging to find suitable arterial voxels throughout the brain.

Figure 3.3 illustrates an example of the tissue concentration time curves (CTCs) of a non-infarcted voxel and infarcted voxel in two different patients with their corresponding AIFs. A non-infarcted voxel always has a CTC with an earlier and higher peak (solid line) relative to the CTC of the infarcted voxel (dotted line). However, when we compare curves across patients, the CTC of the non-infarcted voxel of patient #2 is delayed and lower than the CTC of the non-infarcted
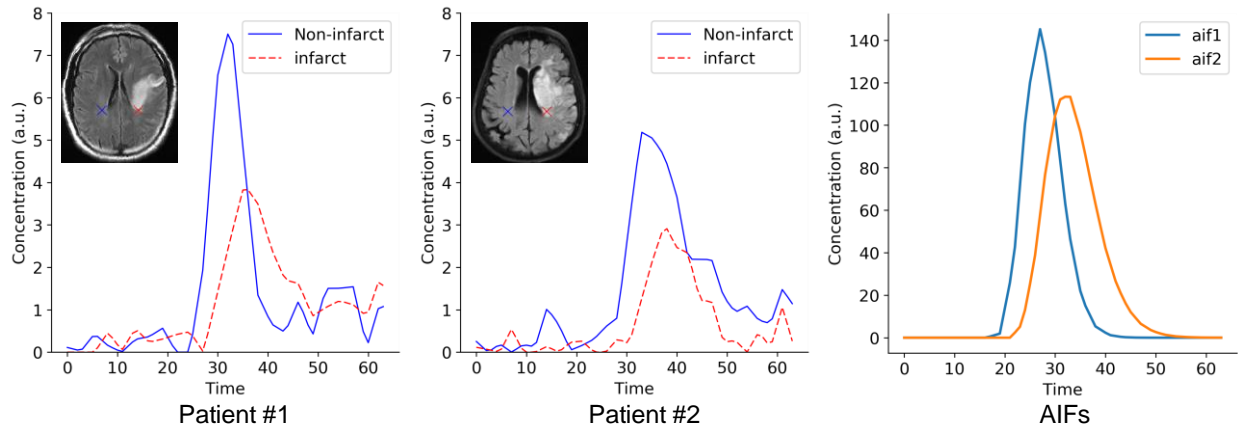
Figure 3.3: An illustration of CTCs of a non-infarcted voxel and infarcted voxel in two different patients (#1 and #2). The left plot shows the corresponding AIFs of patients #1 and #2. Patient #1 has an AIF with earlier and higher peak than patient #2.

voxel of patient #1. These differences are due to their unique AIFs, which are not easy to incorporate into the training, and this makes the learning of representative features difficult. The learned feature filters from these 2D and 3D deep CNN architectures are therefore likely to be limited to only detect features within a patch signal (e.g., peak maximum value) and do not account for the difference in patient AIFs.

3.3.4 The proposed CNN architecture

To overcome abovementioned limitations, the network must be capable of learning features that are independent of confounding patient-specific variables, such as AIFs, and that are predictive of tissue outcome. Therefore, a novel deep CNN model was proposed to improve feature learning. The model included a new form of input patches, in which a patch of interest is paired with its contralateral patch (as a matched control [174]). Additionally, a new architecture for the convolutional layer was proposed to enable explicit learning of unit temporal filters (i.e., a set of filters that have a size of 1 x 1 x t) for the new form of input patches.

3.3.4.1   Matched controls for input patches

The features learned on an input patch of interest are implicitly affected by the AIF. In contrast, features that are derived from the comparisons between the input patch of interest and a matched control might be independent of the confounding variables [174]. We exploited the natural symmetry of the brain to create a matched control for each patch, which is the patch contralateral to the patch of interest ("contralateral patch"). Each original training patch was paired with its contralateral patch, and it was then used to train the CNNs.

3.3.4.2   A proposed architecture for learning paired unit temporal filters

To learn feature filters that extract relationships between a pair of input patches (i.e., the patch of interest and its contralateral patch), a new convolutional layer with paired convolutions was proposed. In this layer, pairs of 3D convolutional filters are learned simultaneously. Each "filter" in the new convolutional layer consists of two 3D filters (w_1, w_2), each of which convolved local regions on the patch of interest and the contralateral patch respectively. An output location ($O_{i,j}$) on a feature map produced from the new convolution layer is defined to be the sum of the convolved values of the same local regions on the patch of interest (i_{poi}) and the contralateral patch (i_{cp}):

$$O_{i,j} = i_{poi} \otimes w_{poi} + i_{cp} \otimes w_{cp}, \tag{3.4}$$

The advantage of this formulation is that all pairs of filters (w_{poi}, w_{cp}) are learned simultaneously, without modification to our loss backpropagation and weight updating methodology.

Figure 3.4 shows the proposed CNN architecture. It is designed to first favor the learning of temporal features between patches, followed by learning the spatial features. The network is made to first *explicitly* learn n paired unit voxel-wise temporal filters in the new convolutional layer (L1.5). More specifically, each filter in a pair had a size of 1 x 1 x t (where t is the time dimension);
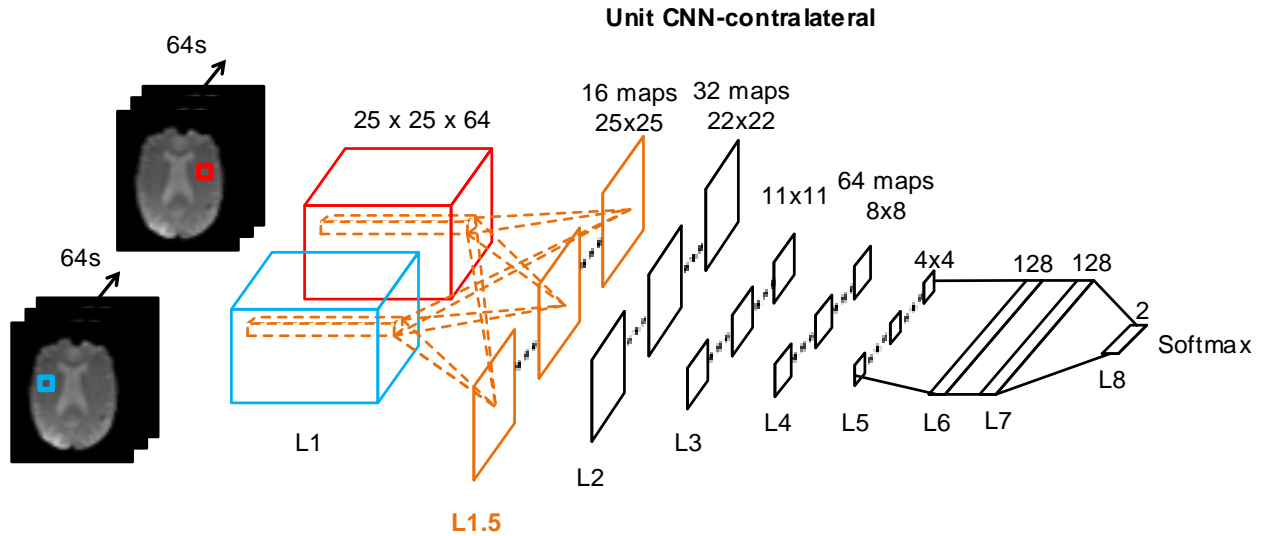
Figure 3.4: The proposed eight-layer deep CNN (Unit CNN-contralateral) with the new architecture of the convolutional layer and unit temporal filter learning (orange) for tissue outcome prediction (only the interconnections within first layer are shown). An input consists of a pair of 25 x 25 voxel patches (the patch of interest, red, and its contralateral patch, blue). Pairs of unit temporal filters (L1) are learned simultaneously, which feed into the first convolutional layer (L1.5) and the remaining layers in the network (same as 2D CNN).

these filters capture the local voxel-wise temporal characteristics. Then, normal 2D convolution and pooling are performed to extract spatial features (similar to a 2D deep CNN), which are fed to the fully-connected layers with a softmax classifier to predict voxel-wise tissue outcome. The new proposed deep CNN is denoted as "Unit CNN-contralateral".

3.4 Evaluations and Results for Voxel-wise Tissue Outcome Prediction

3.4.1 CNN configurations and implementations details

The architecture details of the Unit CNN-contralateral, baseline 2D deep CNN and baseline 3D deep CNN are described in Figure 3.5. In summary, all three deep CNNs consisted of two sequences of convolutional-nonlinear-pooling layers, followed by two fully-connected layers. The
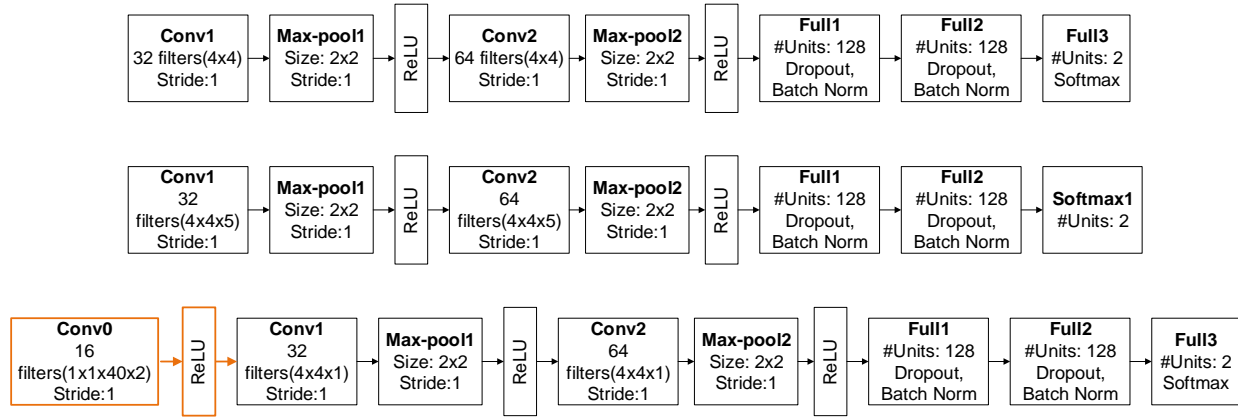
Figure 3.5: The deep CNNs for predicting voxel-wise tissue outcome with the hyperparameters setup. Top: 2D deep CNN. Middle: 3D deep CNN. Bottom: the proposed deep CNN (denoted as "Unit CNN-contralateral"). These deep CNNs learn feature filters to generate 128 complex hierarchical features in the last fully-connected layer, which are then used by the softmax classifier to predict outcome. Abbreviations: conv (convolutional layer), max-pool (max-pooling layer), full (fully-connected layer), batch norm (batch normalization), and ReLU (rectifier linear units).

two fully-connected layers were incorporated with dropout [175] and batch normalization [138] to reduce overfitting of the data and address the issue of internal covariate shift. There were 128 features in the last fully-connected layer, which were the inputs to a softmax classifier for voxel outcome prediction. The Unit CNN-contralateral had an additional first layer (the new convolutional layer) which learns paired unit temporal filters. This layer (orange boxes) was inserted into the 2D deep CNN architecture. For these deep CNNs, the number of filter maps, and the parameters (e.g., size and stride) of the filters and the max-pooling were based on published architectures [135], [152].

The deep CNNs were trained with batch gradient descent (batch size: 50) and backpropagation. A momentum of 0.9 and a learning rate of 0.05 were used. A heuristic was repeated three times to improve the learning of weights [135], where the learning rate was divided by 10 when the validation error rate stopped improving with the current learning rate. An early-stopping strategy was also applied to improve the learning of deep CNNs weights and prevent overfitting. In this

strategy, the training was terminated if the performance failed to improve in five consecutive epochs (maximum number of training epochs: 40). The CNNs were implemented in Torch7 [176], and the training was done on two NVIDIA Titan X GPUs and an NVIDIA Tesla K40 GPU.

3.4.2 Baseline tissue outcome model comparison

In addition to the baseline 2D and 3D deep CNN models, we compared the proposed deep CNN with published state-of-the-art tissue fate models (GLM [36], SR-KDA [22]) and a SVM designed for large-scale classification [177]. Briefly, SR-KDA is a patch-based model for which Tmax patches [22] were generated for model training. GLM is a single-voxel-based model in which several image parameters (ADC, pre-flair, CBF, CBV, and MTT) are the model inputs. SVMs were trained using the same features as the GLM.

To investigate the importance of using contralateral patches, the proposed Unit CNN architecture was trained with two additional types of training data: 1) the patch of interest with a patch randomly selected from the brain (Unit CNN-random), and 2) the patch of interest with a copy of itself (Unit CNN-duplicate). Additionally, the tissue perfusion status may be different in gray matter and white matter. This may affect the Unit CNN model performance when trained with random patches sampled from both types of tissue. To evaluate this, two additional Unit CNNs were trained using: 1) random patches from only the white matter (Unit CNN-random-white), and 2) random patches from only the gray matter (Unit CNN-random-gray).

The performance of the deep CNNs were compared using 10-fold patient-based cross-validation, with a nested validation in each cross-validation fold to identify the optimal cutoff point that optimizes the Youden Index [178] for receiver operating characteristic (ROC) curve. A total of 1,000 patches were randomly selected without replacement from the set of infarcted voxels, and the set of non-infarcted voxels respectively, generating a randomized, stratified, and balanced training dataset with a total of 96,000 patches from 48 patients to avoid biased training [179].

49
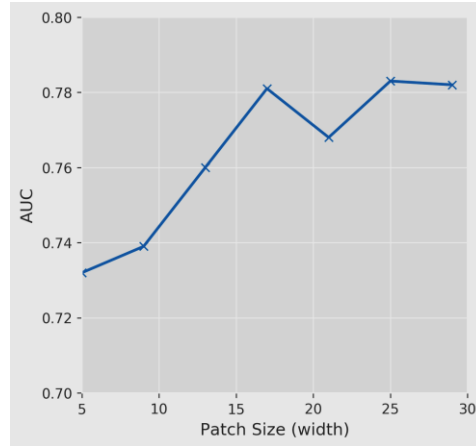
Figure 3.6: The AUROCs of the deep 2D CNN with different training patch sizes in the nested 10-fold cross-validation. The deep 2D CNN achieved the best AUROC (0.783) at patch size = 25. Therefore, a training patch size of 25 x 25 x 64 was used for all subsequent deep CNN models.

3.4.3 Evaluation metrics

Area under the ROC curve (AUROC) was used to evaluate model performance, which is the probability that a model ranks a random positive example highly than a random negative example [180]. The accuracy and overlap coefficient were also calculated [181]. Ten-fold cross-validation was used to the compute all the evaluation metrics for the whole brain data set to which the held-out validation set belonged. To determine if the performance of the models significantly differed, we used Hanley and McNeil significant test [182] to compare the model AUROCs, and used two-tailed Wilcoxon signed-rank test [183] to compare the overlap coefficient and the accuracy of the models.

3.4.4 Results

3.4.4.1 Optimal input patch size

Different values of patch size (*d*) were experimented to identify the optimal number of neighboring information to be incorporated. Figure 3.6 shows the AUROCs with respect to
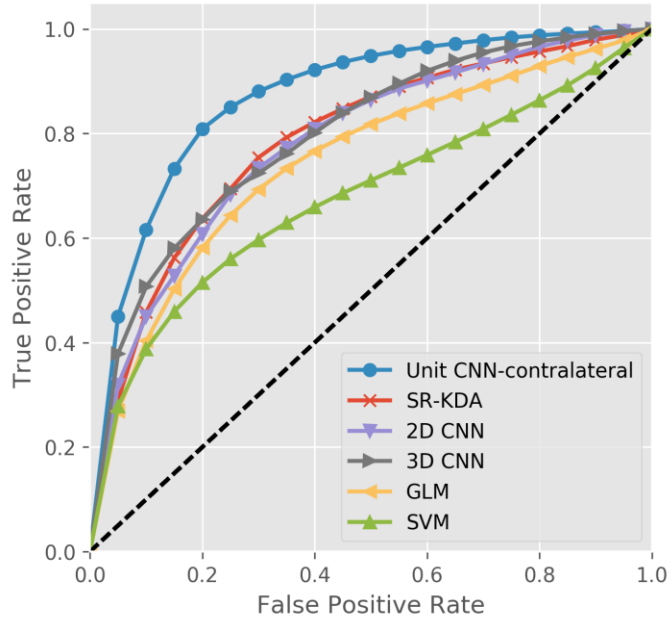
Figure 3.7: Ten-fold cross-validation AUROC curves for Unit CNN-contralateral, Unit CNN-random, Unit CNN-duplicate, and baseline models.

different value of *d*. The optimal value of *d* is 25 in the baseline 2D CNN architecture. Therefore, a training patch size of 25 x 25 x 64 was used for all the deep CNN model training and testing.

3.4.4.2 Baseline models versus Unit CNN-contralateral

Figure 3.7 shows the ROCs of the deep CNNs and the baseline tissue outcome models (GLM, SR-KDA, standard SVM) on predicting tissue outcome, and Table 3.3 shows the average accuracy, overlap coefficient, and AUROC of each classifier calculated from the 10-fold patient-based cross-validation.

Among all the baseline tissue outcome models, SR-KDA achieved the best AUROC, and overlap coefficient. Compared to the best baseline results (SR-KDA), deep CNNs achieve better performance. The 3D deep CNN is slightly better than 2D deep CNN, yet these two deep CNNs did not offer significant performance improvement of AUROC compared to SR-KDA. Among all deep CNNs, the Unit CNN-contralateral achieved the best performance in all evaluation metrics.

51

Table 3.3: Ten-fold cross-validation result using CNNs and baseline models.

Model	Accuracy	Overlap	AUROC
GLM	0.751*	0.628*	0.746*
SVM	0.724*	0.633*	0.691*
SR-KDA	**0.784**	**0.679***	**0.788***
2D CNN	0.746	0.728*	0.783*
3D CNN	0.790	0.717*	0.799*
Unit CNN-contralateral	**0.818**	**0.811**	**0.871**

*Statistically Significant result (p-value<0.05) against the Unit CNN-contralateral model

It achieved the highest AUROC and overlap coefficient of 0.871 and 0.811 respectively. The significance test results show that the proposed deep CNN achieved significantly better AUROC than all other deep CNN classifiers. Overall, the AUROC and overlap coefficient indicates that the proposed deep CNN performed significantly better than SR-KDA, 0.871 vs 0.788 (p-value = 0.003) and 0.811 vs 0.679 (p-value = 0.0001). These results indicate that the proposed deep CNN is the best model for predicting tissue outcome.

3.4.4.3   Unit CNNs with different types of patches

To investigate the significance of using the contralateral patch as part of the input, the proposed Unit CNN architecture was trained with four additional types of patches. The results are shown in Table 3.4. The performance of the deep CNNs significantly dropped when trained with random or duplicate paired patches instead of contralateral patches. The Unit CNN trained with duplicate patches performed worse than other deep CNNs; the Unit CNN trained with random patches achieved only slightly higher performance than a 3D deep CNN (Table 3.3). Unit CNN-random-white, Unit CNN-random-gray, and Unit CNN-random all had similar accuracies, overlap scores, and AUROCs. The results suggest the deep CNN model can capture hidden features from tissue outcome prediction that are robust to tissue types and using the contralateral patch provides

Table 3.4: Ten-fold cross-validation result using CNNs with different patches.

	Accuracy	Overlap	AUROC
Unit CNN-random-white	0.784	0.764*	0.803*
Unit CNN-random-gray	0.781	0.774*	0.805*
Unit CNN-random	0.780	0.753*	0.805*
Unit CNN-duplicate	0.770	0.698*	0.757*
Unit CNN-contralateral	**0.818**	**0.811**	**0.871**

*Statistically Significant result (p-value<0.05) against the Unit CNN-contralateral model

additional information that allows the deep CNN to learn useful comparison filters rather than merely signal filters to extract features for tissue outcome prediction.

3.4.4.4   Visualization of convolutional filters in the Unit CNNs

Figure 3.8 shows the seven of the 16 learned pairing filters of the proposed deep CNNs (trained with contralateral patches, random patches, and duplicate patches respectively). Each pair of filters consists of two 1 x 1 x t ($t = 64$) 3D filters, and each 3D filter is composed of 64 unit filters with a size of 1 x 1. Instead of capturing spatial features (e.g., edges) as typical 2D filters do, these pairs of 3D filters capture the relationships between two input signals in the time dimension. When looking at the pairing filters of the deep CNN trained with duplicate patches, one may observe that filters are similar within a pairing filter. In contrast, the variability of the pairing filters of the deep CNN trained with contralateral patches is higher: some pairing filters appear to be the "opposite" of each other, while others appear to be similar. For example, in filter pair number 7, the first filter (top) can detect later signals (indicated as black first and then white along time) whereas the second filter detects early signals (indicated as white first and then black along time). Compared to these pairing filters, those paired filters trained by the deep CNN with random patches were inconsistent and more random in appearance; those paired filters trained by the deep CNN with duplicate
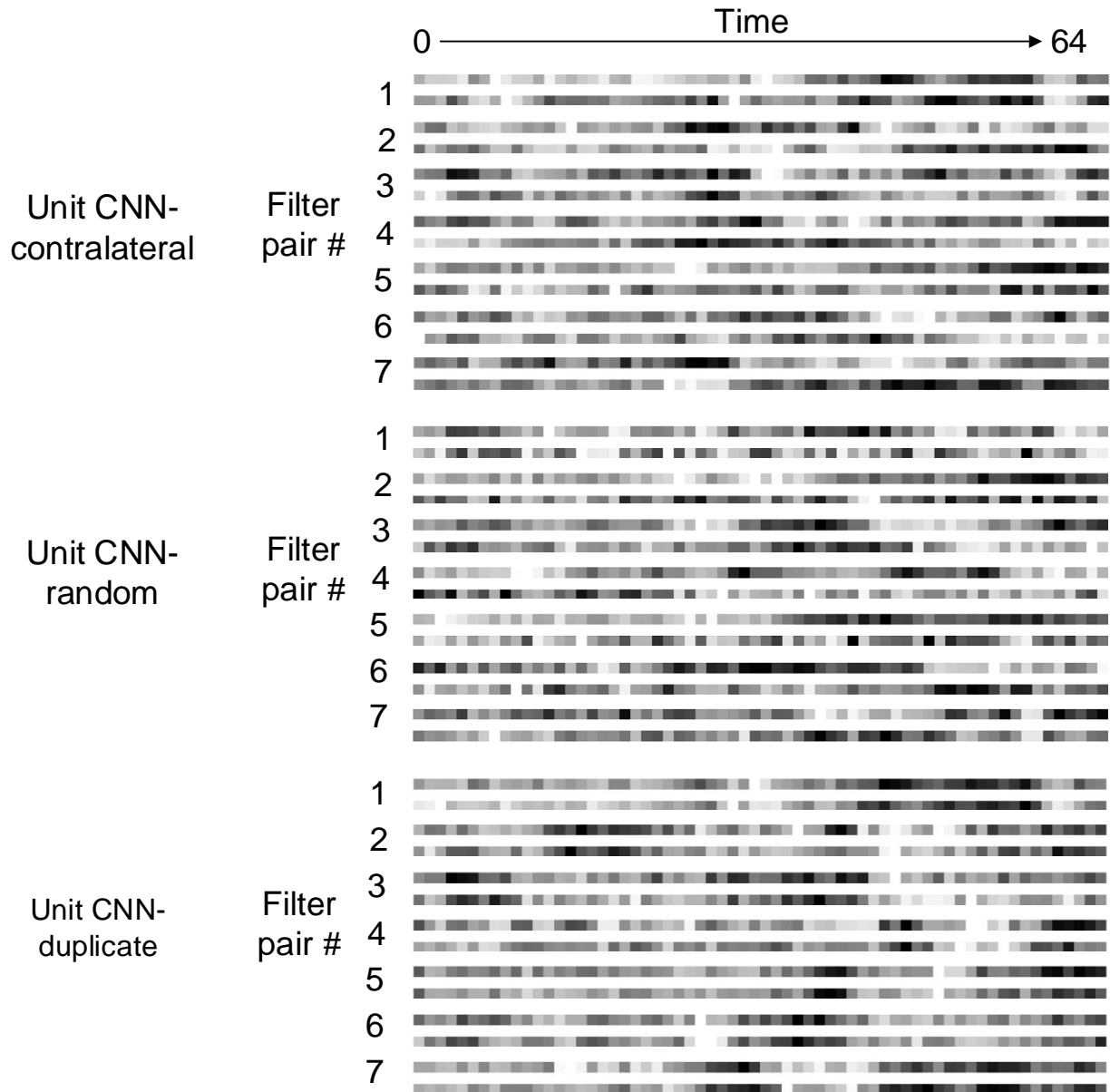
Figure 3.8: Seven of the 16 pairs of 3D filters learned in the first layer of the proposed Unit CNNs trained with contralateral patches, random patches, and duplicate patches respectively. Each 3D filter is composed of 64 unit filters (along time) with a size of 1 x 1; therefore, each 3D filter has a size of 1 x 1 x t ($t = 64$).

patches looked similar. This result shows that training with contralateral patches results in a distinct class of filters that are associated with better model performance.
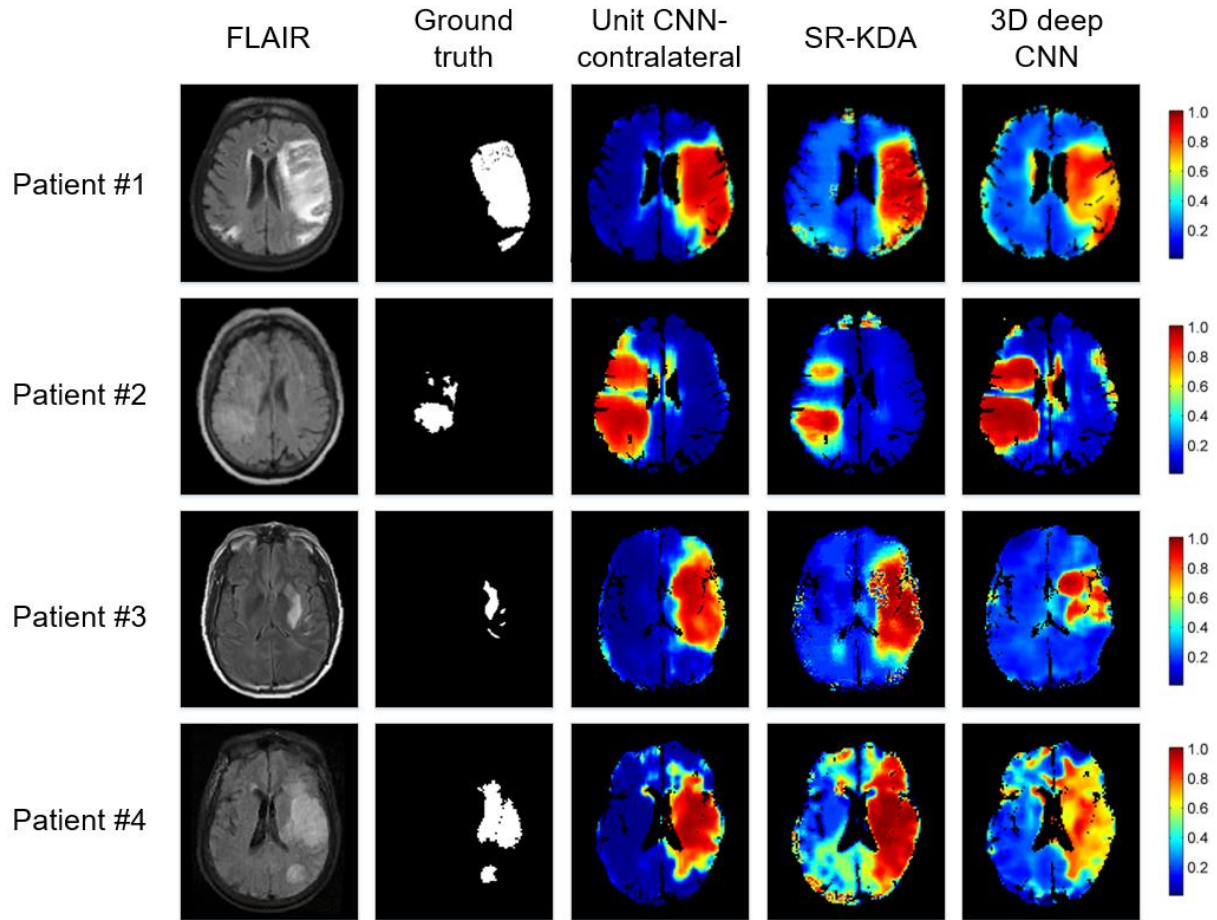
Figure 3.9: Tissue outcome prediction of Unit CNN-contralateral, SR-KDA, and 3D deep CNN.

3.4.4.5 Examples of prediction

Figure 3.9 shows the examples of tissue outcome predictions. For patient #1, Unit CNN-contralateral and SR-KDA showed better visual inspection than 3D deep CNN, i.e., 3D deep CNN predicted high probability (red color) only in certain parts of the final infarct regions, whereas Unit CNN-contralateral and SR-KDA both predicted most of the final infarct regions. For patient #2, all models predicted the correct locations of the final infarct regions. Unit CNN-contralateral and 3D deep CNN both predicted larger final infarct volumes than SR-KDA. The volume of prediction with the highest probability (>0.90) of Unit CNN-contralateral matches well with the ground truth. For patient #3, SR-KDA and Unit CNN-contralateral predicted larger final infarct volume with

Table 3.5: Classifier AUROCs on different final infarct volume cutoff using different variables.

Variable Type	Final infarct volume cutoff (cm³)								
	20	40	60	80	100	120	140	160	180
Clinical only	0.648	0.507	0.647	0.607	0.613	**0.766**	0.619	0.619	0.499
Treatment only	0.529	**0.566**	0.561	0.521	0.324	0.259	0.265	0.265	0.365
Clinical + treatment	0.632	0.507	0.660	0.634	0.589	**0.772**	0.633	0.633	0.460

high probabilities whereas 3D deep CNN predicted only partial volume. For patient #4, the regions with high probability of SR-KDA and 3D deep CNN were larger and smaller respectively, whereas the region of high probability of Unit CNN-contralateral matched well with the ground truth.

3.5 Evaluations and Results for Tissue Outcome Volume Prediction

3.5.1 Evaluations

In addition to using only images (i.e., PWIs) to predict voxel-wise tissue outcome, studies [17], [18], [184] have shown interests in determining whether clinical and treatment variables are predictive of final tissue outcome volume. Therefore, in this study, machine learning models were trained to predict the final tissue outcome volume using clinical and treatment variables.

Random forest (RF) classifiers [185] were trained to perform binary final infarct volume classification at different final infarct volume cutoff. Three groups of classifiers were compared: (1) trained with only clinical variables, (2) trained with only treatment variables, and (3) trained with both clinical and treatment variables. AUROCs of these classifiers were calculated and the nested five-fold cross-validation was used to evaluate the model performance.

3.5.2 Final tissue volume prediction

Table 3.5 summarizes the RF performance on different final infarct volume cutoff values trained with three groups of variables. The best classifier using only the clinical variables achieved

an AUROC of 0.766 with a final infarct volume cutoff of 120 cm^3, whereas the best classifier using only the treatment variables achieved an AUROC of 0.566 with a final infarct volume cutoff of 40 cm^3. Among all the classifiers, the classifier trained with both clinical and treatment variables achieved the best performance at a final infarct volume cutoff of 120 cm^3, with an AUROC of 0.772. However, there was not much AUROC improvement as compared to the classifier using only the clinical variable (AUROC = 0.766). These results show that clinical variables are predictive of the final infarct volume, and the addition of treatment variables did not provide enough information to greatly improve the classifiers' performance. We also notice that some AUROCs are lower than 0.5 (e.g., treatment only variables and cutoff > 100 cm^3). This may be due to data imbalance problem when one class is more common than another class in training dataset, which can make the trained classifiers biased in classification when a small dataset is used.

3.6 Discussion

In this work, I proposed a new architecture for the convolutional layer (Unit CNN), which learned pairs of unit temporal filters simultaneously from the patch of interest and its contralateral patch. This new layer was inserted as the first layer to the standard 2D deep CNN and allowed the deep CNN to derive paired filters to obtain useful temporal correlations between inputs in the first layers. This enables the learned filters in the subsequent convolutional layers no longer detected the spatio-temporal features of a single input. Instead, these filters were learned to detect spatial features of the *differences* between two input patches (i.e., the patch of interest and its contralateral patch). Such differences were further expanded and stacked through multiple layers in the deep CNN and finally became the 128 features that were used to train a softmax classifier. The incorporation of the new convolutional layer changed the nature of the learned features in Unit CNNs, which ultimately led to features that achieved the best performance. Another contribution of this work is that the type of input patch does matter. Unit CNN achieved the best performance when it was trained with the contralateral patches. One interesting observation is that the deep

CNNs trained with random patches has higher AUROC (not statistically significant) than the deep CNNs trained with duplicate patches. This illustrates that additional random information may boost model performance. However, this boosting is not as good as using contralateral patches. Overall, the superior performance of the deep learning algorithm reinforces the findings of Christensen *et al.* [29] and Willats *et al.* [186] that summary parameters calculated without an AIF from source perfusion images contain enough information to determine tissue outcome.

In addition to the imaging results, the study demonstrates that clinical variables are predictive of the final infarct volume. Surprisingly, the random forest models did not improve with the addition of treatment variables. This result contradicts some studies that treatments can improve patient outcomes [59], [187]. We suspect that this may be due to the small data set (n=82) and the relationship of outcome and treatment variables are not fully captured.

This study does have some limitations. The training and evaluation data sets are relatively small and are not amenable to subgroup analysis. The data sets are small because often patients are discharged before three days after stroke onset in UCLA, and it is challenging to collect their post-FLAIR images. Another limitation is that the proposed deep CNNs only used the standard CNN architecture [188] as the baseline structure. There are new CNN building blocks that can achieve better performance, such as the ResNet [149] and the DenseNet [189]. A possible improvement is to incorporate these new building blocks in the existing architectures to generate better composite feature representations. More details about future works will be presented in Chapter 7.

CHAPTER 4

Stroke Onset Time Classification using Deep Learning

4.1 Overview

Intravenous tissue plasminogen activator (IV tPA) and clot retrieval devices are the two currently approved treatments by the National Institutes of Health for acute ischemic stroke patients. While the guidelines of using clot retrieval devices are still changing, the administration of the IV tPA has a clear cut-off time, i.e., within a maximum of 4.5 hours from time since stroke (TSS). Unfortunately, as many as 30% of acute stroke patients have unknown symptom onset times (e.g., wake up or unwitnessed strokes) [190]–[192]. These patients have historically been excluded from thrombolytic and endovascular treatments because of unknown TSS, even though their actual TSS may fall within the medically appropriate window for treatments [193]. This population may be possibly saved by methods that can determine the TSS without clinical history or witnessed events. Studies [30]–[33], [92], [120] have attempted to use an imaging feature, diffusion weighted imaging and fluid attenuated inversion recovery imaging mismatch (DWI-FLAIR mismatch), to approximate TSS. Yet these methods often achieved a moderate performance: a negative predictive value of 0.20 to 0.50, with a specificity of 0.60 to 0.80 and a sensitivity of 0.50 to 0.60. The major criticisms of these methods are the simplicity of the mismatch feature and the non-standardized mismatch identification guidelines [34], [35], [49].

This chapter details a novel machine learning approach for better TSS classification. The approach mainly consists of three steps: 1) defining and generating a set of imaging features from MR images and perfusion parameter maps; 2) proposing a novel deep autoencoder (AE) approach to learn deep features from four-dimensional perfusion images; and 3) training and comparing

several machine learning models using the generated features for TSS classification. The evaluation results demonstrate that the proposed machine learning approach outperforms the DWI-FLAIR mismatch method. In the remainder of this chapter, Section 4.2 describes the dataset and the image preprocessing steps used for this study. Section 4.3 describes the proposed machine learning approach. Section 4.4 summarizes the evaluation methods and Section 4.5 reports the results. Section 4.6 discusses the limitations and future works. Some of the work presented in this chapter have been published in [194].

4.2 Data Cohort and Image Preprocessing

4.2.1 Patient cohort and imaging data

The patient characteristics are summarized in Table 4.1. Under institutional review board (UCLA IRB#18-000329) approval, a total of 181 patient MR images (taken between December 2011 and December 2017) were retrieved and examined from the UCLA picture archiving and communication system (PACS). The inclusion criteria were all patients with: 1) acute ischemic stroke due to middle cerebral artery (MCA) occlusion; 2) a recorded time of observed stroke symptom onset; 3) a recorded time of initial pretreatment imaging; and 4) a complete MR imaging sequence set (PWI, FLAIR, DWI, and ADC). After applying the inclusion criteria, a total of 131 patients were retrieved for the analysis (85 positive class; 46 negative class). This cohort subset was used to build the models for TSS classification. Patients' TSS was calculated by subtracting the time at which the stroke symptoms were first observed from the time at which the first imaging was obtained. Patient TSS was binarized based on an existing DWI-FLAIR TSS classification task [33]: positive (<4.5hrs) and negative (≥4.5hrs).

All patients underwent MRI using a 1.5 or 3 Tesla echo planar MR imaging scanner (Siemens Medical Systems); scanning was performed with 12-channel head coils. The PWIs were acquired using a repetition time (TR) range of 1,490 to 2,640 ms and an echo time (TE) range of 23 to 50

Table 4.1: Patient cohort characteristics for TSS classification.

	Patients (n = 131)
Demographics	
Age	72.9±13.9
Gender	59 males
Clinical Presentation	
Time since stroke (continuous)	256±247 minutes
NIHSS[†]	10.1±7.87
Atrial fibrillation	37
Hypertension	87
Stroke location (hemisphere)	
Left	65
Right	66
Classification Label	
Time since stroke (binary)	<4.5hrs (85); ≥4.5hrs (46)

[†] NIHSS = NIH Stroke Scale International; scale: 0 (no stroke symptoms) - 42 (severe stroke)

ms. The pixel dimension of the PWIs varied from $1.72 \times 1.72 \times 6.00$ to $2.61 \times 2.61 \times 6.00$ mm. The FLAIR images were acquired using a TR range of 8,000 to 9,000 ms and a TE range of 88 to 134 ms. The pixel dimension of the FLAIR images varied from $0.688 \times 0.688 \times 6.00$ to $0.938 \times 0.938 \times 6.50$ mm. The DWI/ADC images were acquired using a TR range of 4,000 to 9,000 ms and a TE range of 78 to 122 ms. The pixel dimension of the DWI images varied from $0.859 \times 0.859 \times 6.00$ to $1.85 \times 1.85 \times 6.50$ mm. The z-dimension is ignored in patch data generation due to slice thickness. Therefore, the notation of the size of a "voxel" is simplified to two-dimensional only (x-, and y- axes).

4.2.2 Image preprocessing

The image preprocessing steps of TSS classification follow the image preprocessing steps of tissue outcome prediction (Chapter 3) closely. MR image intra-patient registration, noise removal, skull stripping, tissue type mask identification, perfusion parameter maps and arterial input functions (AIF) generation, and image standardization were performed. Please see Section 3.2.3 for details.

4.3 Methods

Inspired by research in other medical domains (e.g., lung nodule detection [195]), in which hundreds of hand-crafted imaging features were defined and used for classification, a machine learning approach was proposed that involves imaging features derived from MR images and perfusion parameter maps to classify TSS. The machine approach consists of multiple components: 1) defining regions of interest (ROIs); 2) building deep learning models to learn deep feature maps from PWIs; 3) generating a set of baseline imaging and deep learning features from ROIs; and 4) training machine learning models using the generated features for TSS classification. Figure 4.1 shows the overview of the TSS classification. The details for the approach are described as the following subsections.

4.3.1 Imaging feature generation

PWIs are spatio-temporal imaging data (4-D) that show the flow of a gadolinium-based contrast bolus into and out of the brain over time. They contain concentration time curves (CTCs) for each brain voxel, that describe the flow of the contrast (i.e. signal intensity change) over time. Perfusion parameter maps can be derived from PWIs and describe tissue perfusion characteristics. The details of PWIs and the definition of different perfusion parameter maps are described in Chapter 2.3.3.
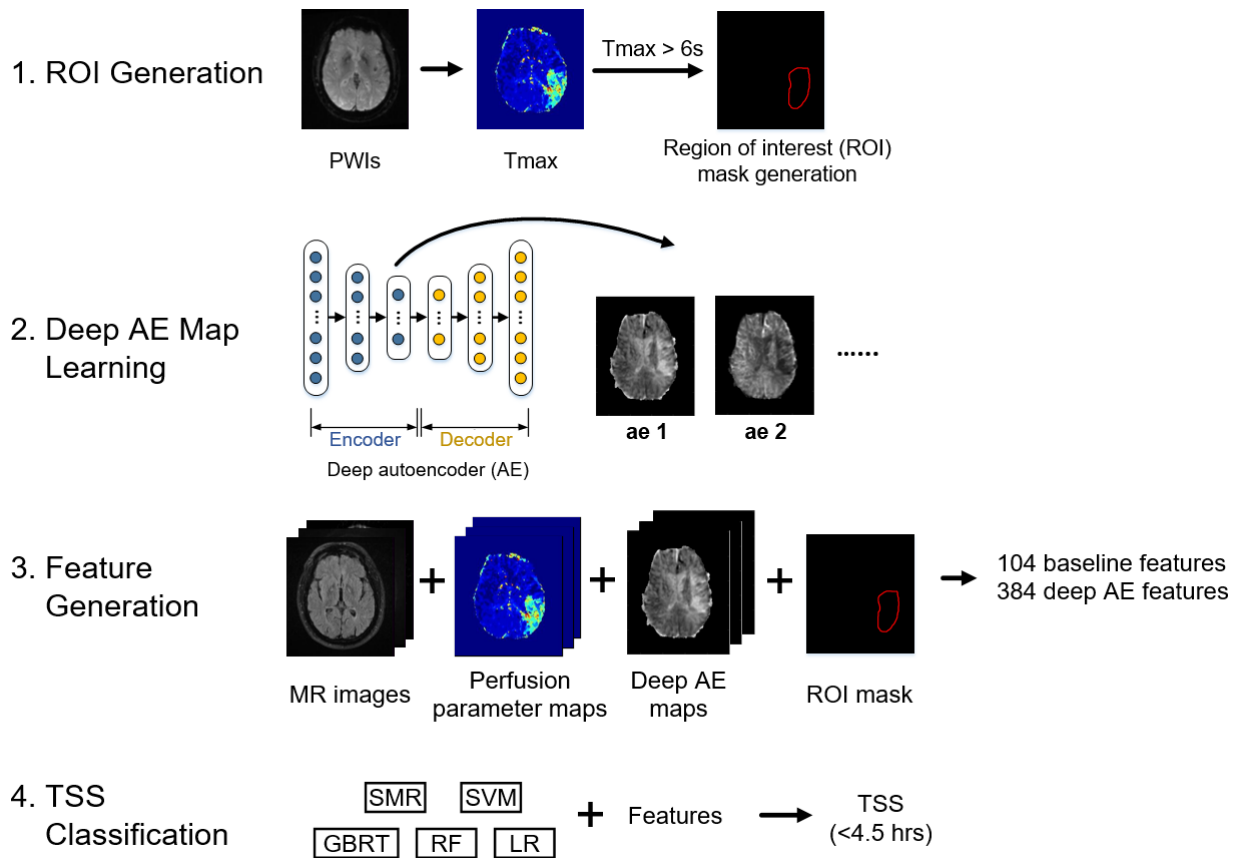
Figure 4.1: The overview of the proposed machine learning approach for TSS classification. The classification involved four steps: (1) ROI generation, (2) deep AE map learning, (3) feature generation, and (4) TSS classification. A set of 104 baseline imaging features and 384 deep AE features were generated and used for TSS classification.

In this study, two ways to generate the imaging features for TSS classification were proposed and compared. The first way was to generate imaging features from MR images and perfusion parameter maps, in which descriptive statistical features (e.g., mean) and morphological features were defined. The second way was to generate imaging features directly from 4-D PWIs using deep learning. A deep learning approach was proposed to learn hidden representations of every CTC within the PWIs using a deep autoencoder. Then, these hidden representations were aggregated into new feature maps, which we hypothesized can indicate hidden characteristics of

stroke tissue and can improve TSS classification. The descriptive statistical features could then be generated from these new feature maps for TSS classification. The imaging feature generation involves three parts: (1) region of interest generation, (2) baseline imaging feature generation, and (3) deep imaging feature generation.

4.3.1.1 Region of interest generation

Stroke often occurs in only one cerebral hemisphere and generating imaging features based on entire brain MR images may be less descriptive to the stroke pathophysiology and less predictive of TSS. Therefore, ROIs were defined by a time-to-maximum cut-off (Tmax>6s), which captures both the dead tissue core and the salvageable tissue [96]. The largest connected region in which Tmax>6s on the stroke hemisphere was used as the ROI mask.

4.3.1.2 Baseline imaging feature generation

There are two major types of baseline imaging features: descriptive statistics and morphological features. These baseline imaging features were generated from MR images (DWI, ADC, and FLAIR), the perfusion parameter maps, and Tmax>6s ROI masks. The perfusion parameter maps included the cerebral blood volume (CBV), cerebral blood flow (CBF), mean transit time (MTT), and time-to-peak (TTP). Descriptive statistics included the maximum, minimum, median, mean, standard deviation, and variance of the intensity/parameter value within the ROI. The ratio between a value of interest and the corresponding value on the contralateral side of the brain has been shown to be predictive in stroke tissue outcome prediction (Ch. 3), and therefore this type of relative features (e.g., relative maximum) was included as part of the descriptive statistics. Relative features of ADC-to-FLAIR and DWI-to-FLAIR were also included, as inspired by the DWI-FLAIR mismatch method. This resulted in a set of 96 baseline descriptive features. Morphological features [195] were calculated using the ROI mask, including area, volume, circularity, and sphericity. Two shape features [196] were included: the ratio between the volume of the ROI and its bounding box

Table 4.2: Imaging features for TSS classification.

Type	Features	Sources
Descriptive Statistics (n=96 for baseline features) (n=384 for deep features)	(Relative[†]) maximum, (relative) minimum, (relative) median, (relative) mean, (relative) standard deviation, (relative) variance	DWI, ADC, FLAIR, CBV, CBF, MTT, TTP, deep feature maps
Morphological Features (n=8 for baseline features)	Area, volume, circularity, sphericity, the ratio between the volume of the ROI and the bounding box, the ratio between the lesion surface area and the lesion volume, maximum diameter, minimum diameter	Tmax > 6s ROI mask

[†]Relative = the ratio between the value of interest and the value in its contralateral side

(BE), and the ratio between the lesion surface area and the lesion volume (SV). The maximum and minimum diameter of the ROI mask were also included. This resulted in a set of eight baseline morphological features.

In total, a set of 104 baseline imaging features were generated. All the features were standardized independently to zero mean with a standard deviation of one. The baseline imaging features are summarized in Table 4.2.

4.3.1.3   Deep imaging feature generation

PWIs are high-dimensional data, in which a typical convolutional neural network (CNN) [197] may not be capable of extracting representative features from the small stroke imaging dataset (n=131), even with a large amount of data augmentation and a well-tuned deep architecture. Therefore, we proposed to use deep autoencoders (deep AEs) [127] coupled with novel input approach to generate imaging features for TSS classification. Figure 4.2 shows the overview of the feature generation using the deep AE approach.

65
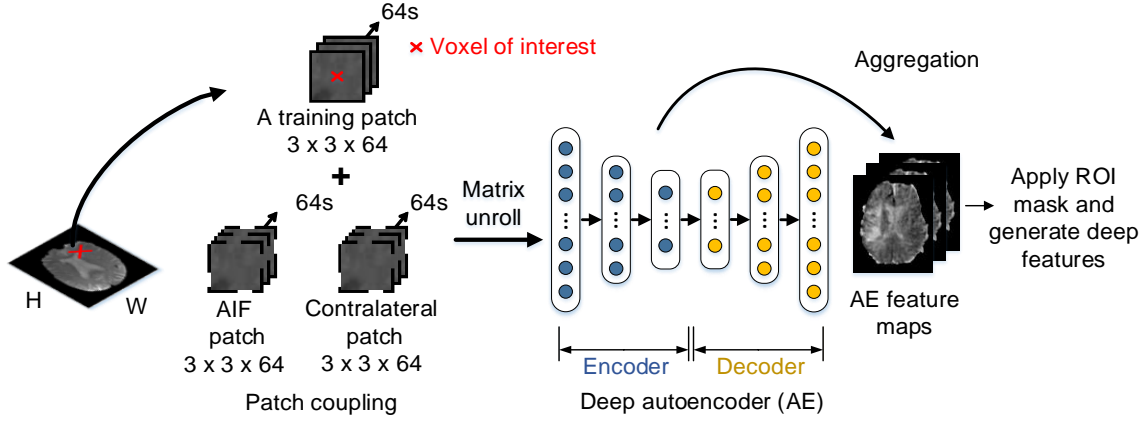
Figure 4.2: Deep AE feature generation. Training patches (with a size of 3 x 3 x 64) were randomly generated from PWIs. Each patch was coupled with an extra patch (AIF only, contralateral only, or AIF+contralateral) and the combined matrix was unrolled into a 1D vector that would be fed into the deep network. The proposed deep AE consisted of an encoder and decoder. The encoder outputs were aggregated into the final deep AE feature maps, in which a ROI mask (Tmax>6s) was applied to them and generate the imaging features (descriptive statistics). Note that the input z-dimension is not included.

Each PWI voxel CTC at location *ijz*, with a size of 1 x *t* (*t* = time for perfusion imaging), was transformed by the deep AE into *K* new feature representations that could represent complex voxel perfusion characteristics. The deep AE consisted of an encoder and a decoder. The encoder consisted of two components: 1) an input layer; and 2) fully-connected layers. The encoder was connected to the decoder, which followed reversely the same layer patterns of the encoder. The encoder output (i.e., the middle layer output of the deep AE) was the set of *K* new feature representations. Each new feature representation of all CTCs, f_{ijz}^k, was aggregated to form a new feature map, known as "AE feature map" (F^k):

$$F^k = \{f_{ijz}^k\}, \forall ijz \in I \,, \tag{4.1}$$

where *I* was the set of pixels in a PWI. In total, there were *K* new AE feature maps for a PWI (*K* is equivalent to the encoder dimension in an optimal AE model). New AE deep imaging features

66

(descriptive statistics) were then generated from the AE features maps following the same procedure as used in the baseline feature generation. In total, a set of 384 deep imaging features were generated.

The advantage of using the proposed deep AE is that it is trained via an unsupervised learning procedure, in which thousands of training data could be sample from the PWIs. The network was optimized to obtain weights, Θ, that minimized the binary cross-entropy loss between the input, I, and the reconstructed output, $\hat{I}(\Theta)$, across the samples with size n [176]:

$$\underset{\Theta}{\mathrm{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left[(I_i * \log(\hat{I}(\Theta)) + (1 - I_i) * log(1 - \hat{I}(\Theta)) \right], \qquad (4.2)$$

4.3.1.4 Input training patch coupling and feature generation

As Chapter 3 suggests, regional information corresponding to a voxel's surroundings can improve classification in the MR images. Therefore, a small region (8 neighboring voxels) was included in each training voxel, leading to a size of 3 x 3 x t patch (width x height x time; the z-dimension is omitted; $t = 64$), where the center of the patch is the voxel of interest for the deep AE feature learning. Additionally, the patch coupling strategies in voxel-wise stroke classification indicate that reference patches may improve the learning of hidden features. This motivates the proposal of three patch coupling ways in TSS classification: (1) training patch with global AIF patch; (2) training patch with its corresponding contralateral patch; and (3) training patch with both the AIF patch and the contralateral patch.

In the experiments, each training patch (with the coupled patch(es)) was unrolled from a size of 3 x 3 x t x p ($p = 2$ or $p = 3$, depending on the number of coupled patches) into a 1-D vector. The 1-D data were used to train the deep AEs, which consisted of the fully-connected layers. Three different deep AEs were optimized for the three patch coupling methods. In total, 131,000 training patches were generated by random and balanced sampling from all patient PWIs.

4.3.2   Classifying TSS using machine learning model

Five machine learning models were trained using the generated features for binary TSS classification (TSS<4.5hrs or TSS≥4.5hrs). These classifiers included logistic regression (LR), random forest (RF), gradient boosted regression tree (GBRT), support vector machine (SVM), and stepwise multilinear regression (SMR). Briefly, LR is a probabilistic classification model in which binary label probabilities are found by fitting a logistic function of feature values [198]. RF is an ensemble learning method in which a multitude of decision trees are randomly constructed and the classification is based on the mode of the classes output by individual trees [185]. GBRT is an ensemble learning method similar to RF, in which a multitude of decision trees are randomly generated, yet these trees are added to the model in a stage-wise fashion based on their contribution to the objective function optimization [199]. SVM is a supervised learning classification algorithm that constructs a hyperplane (or set of hyperplanes) in a higher dimensional space for classification [200]. SMR is a stepwise method for adding and removing features from a multilinear model based on their statistical significance (e.g., F-statistics) to improve model performance [201].

In addition to the five machine learning models, four end-to-end convolutional neural networks (CNNs) were trained for TSS classification. The first three CNN architectures were adapted from the popular ImageNet architectures: Alexnet [135], VGG-16 [202], and ResNet-18 [149]. The fourth CNN architecture was a shallow-Alexnet, which had only one convolution-pooling-rectified linear unit (ReLU) layer sequence and two fully-connected layers, to investigate if a smaller end-to-end CNN is preferable for TSS classification. The inputs to the CNNs were the four-dimensional ROIs (width × height × depth × number of images) extracted from the stacked MR images (DWI, ADC, and FLAIR) and the perfusion parameter maps. The convolutional filters of each CNN architecture were modified from two dimensions to three dimensions to accommodate the change of the input data size.

4.3.3 Experimental setup

4.3.3.1 Deep learning model configurations and implementations details

The deep AE and the CNNs were optimized using Adam, which computes adaptive learning rates during training and has demonstrated superior performance over other optimization methods [145]. An early-stopping strategy was applied to improve the learning of the model weights and prevent overfitting, where the training would be terminated if the performance did not improve over five consecutive epochs (maximum number of training epochs: 50). The models were implemented in Torch7 [176] and Pytorch [203], and the training was done on two NVIDIA Titan X GPUs and an NVIDIA Tesla K40 GPU. Ten-fold patient-based cross-validation was performed to determine the optimal deep AE architectures, including the number of encoder hidden layers (from 1-3) and the number of hidden units (factor of 4, 8, 16, 32).

4.3.3.2 Machine learning model training

The LR, RF, and SVM were developed using the Python Scikit-learn library [204]. The SMR and GBRT were developed using MATLAB and the XGBoost library [205], respectively. Different model hyperparameters (e.g., a LR's hyperparameter, C) contribute differently to the classification and different machine learning methods may not perform equally on the same feature set. Therefore, nested 10-fold cross-validations were used to evaluate all five classifier [206]. Briefly, an outer 10-fold cross-validation was performed to obtain the overall classifier performance. Within each outer fold (in which a validation fold was held out), an inner 10-fold cross-validation was performed to determine the optimal model hyperparameters using the training data (i.e., the nine out of ten folds), and then the model was trained with the optimal hyperparameters and applied to the validation fold.

The optimal hyperparameters for LR (C, L2 regularization), RF (maximum tree depth, 100 random trees), GBRT (gamma, maximum tree depth, 100 random trees), SVM (C, radial basis

function (rbf) kernel), and SMR (p-enter=0.05, p-remove=0.35, maximum iteration = 60) classifiers were determined using the inner cross-validation in the nested 10-fold cross-validation [206]. The range of tested C values for LR was [0.00001, 1]. The range of tested maximum tree depths for RF was [3, 12]. The range of tested gamma values and maximum tree depths for GBRT were [0.01, 1] and [3, 12], respectively. The range of tested C values for SVM was [0.00001, 1].

4.3.3.3 DWI-FLAIR mismatch

The Medical Image Processing, Analysis, and Visualization (MIPAV) software [167] was used by an expert neuroradiologist (Dr. S. El-Saden) to determine the presence of DWI-FLAIR mismatch, following the published protocol [33]. The presence of a DWI-FLAIR mismatch was labeled 1; the absence of a DWI-FLAIR mismatch was labeled 0.

4.4 Evaluations

Several evaluations were performed to compare the machine learning models. These evaluations include ROI sensitivity analysis, feature correlation analysis, TSS subgroup classification analysis, and performance evaluation metrics. The details are described as follow.

4.4.1 ROI sensitivity analysis

The generation of ROIs (Tmax>6s) could have significant impact on the TSS classification. To investigate the effect of the ROIs on the classification, two additional Tmax cutoff values were evaluated [96], [99]. One is Tmax>4s, which is a softer cutoff value that may include normal brain tissue; one is Tmax>8s, which is a stricter cutoff value that captures only the severe hypoperfused stroke region. The same experimental procedures were used to extract the imaging features from the ROIs generated by the two new cutoff values and these features were used to train the machine learning classifiers. The same evaluation metrics were applied to these classifiers.

4.4.2   Feature correlation analysis

Recently, deep learning has been criticized as a "black-box" approach [207] that yields state-of-the-art performance, yet the classification mechanism is unclear. A question one may ask is the correlation of the new deep features to the baseline imaging features. To understand what the deep features represented, an approach was proposed based on a correlation analysis. First, the correlations between the deep AE features and the baseline imaging features were calculated. Then, the most correlated deep AE feature was identified for each baseline imaging feature. For each identified deep AE feature, the top five correlated baseline imaging features were obtained. All correlations were calculated using Pearson correlation [208].

4.4.3   TSS subgroup classification analysis

Changes in MR image acquisition parameters (e.g., field strength) over time may impact the classifier performance because the study data set was created from the patient imaging exams obtained from 2011 to 2017 [209]. Two image-related variations, magnetic field strength and year of imaging acquisition, were investigated on the TSS classification. For the field strength, two-fold cross-validation was used to evaluate the classifiers on TSS classification, i.e., trained on a data subset with one field strength (e.g., 1.5T) and evaluated on a data subset with another field strength (e.g., 3T), and vice versa. For the year of imaging, the classifiers were first trained with the data collected from 2011-2014 and then evaluated with the data collected from 2015-2017. This evaluation was meant to explore whether the model still performed well on the newer data when training on the older data.

4.4.4   Metrics

The AUROCs were computed for all five classifiers. To determine if the performance of the TSS classification models significantly differed from each other, the Hanley and McNeil

Table 4.3: The optimal deep AE architectures for different coupling patch.

Coupling Patch Type	Optimal AE Architecture (# of hidden units/layer)	Optimal MSE (Average Deep AE MSE)
AIF patch only	1152-192-32-32-192-1152	0.606 (1.54)
Contralateral patch only	1152-288-32-32-288-1152	1.16 (1.95)
AIF + Contralateral patch	1728-288-32-32-288-1728	1.06 (4.49)

significance test [182], [210] was used to compare the model AUROCs. Sensitivity, specificity, F1-score, positive predictive value (PPV), and negative predictive value (NPV) were calculated for the DWI-FLAIR mismatch method. Given the DWI-FLAIR mismatch method specificity, the performance (sensitivity, F1-score, PPV, and NPV) was calculated for the machine learning classifiers and compared against the DWI-FLAIR mismatch method.

4.5 Results

4.5.1 TSS classification

Table 4.3 shows the optimal AE model architectures for three types of coupling patches as described in Section 4.3.1.4. All three optimal AE architectures achieved a mean square error (MSE) of at least 40% smaller than the average MSE of all of the trained AEs. All three optimal AE architectures had 32 hidden units (AE1 to AE32) in the middle layer (i.e., 32 deep feature maps). These optimal AE models were used to generate deep feature maps from patient PWIs for feature extraction.

The classifiers were trained with three different groups of features (Table 4.2): (1) the baseline imaging (BI) features (96 descriptive statistics and 8 morphological features); (2) the deep AE features (384 descriptive statistics); (3) the baseline and deep AE features. The AUROCs of the classifiers are depicted in Table 4.4.

Table 4.4: The AUROCs of classifiers on TSS classification.

Model	Ho, *et al*. (2017) [26]	No AE BI[†]	AIF coupling patch only		Contralateral coupling patch only		AIF + contralateral patch	
			AE[¥]	BI+AE [Ω]	AE	BI+AE	AE	BI+AE
LR	0.574	0.618	0.650	0.658	0.647	0.676	0.710	**0.765***
RF	0.624	0.640	0.650	0.669	0.662	0.682	0.592	**0.690**
GBRT	0.567	0.608	0.590	0.570	**0.676**	0.674	0.612	0.670
SVM	0.669	0.636	0.477	0.736	0.605	0.666	0.600	**0.746***
SMR	0.683	0.661	0.574	0.707	0.650	0.677	0.705	**0.730**

Bold indicated the highest AUROC for a given classifier

*statistically significant result (p-value<0.05) against model with BI features only.

[†]BI = Models were trained with the baseline imaging features (94 descriptive statistics and 8 morphological features)

[¥]AE = Models were trained with the 384 descriptive features generated from deep AE feature maps)

[Ω]BI + AE = Models were trained with the baseline and deep AE features

With the baseline imaging features alone or the deep AE features alone, all classifiers (LR, RF, GBRT, SVM, and SMR) achieved an AUROC of at least 0.6 on TSS classification, showing that the MR imaging features were predictive of TSS. With the combination of baseline imaging features and deep features, all classifiers (except the GBRT trained with AIF coupling patch) showed improvement in AUROC, demonstrating that the proposed deep AEs extracted hidden features in PWIs which could improve TSS classification. Among all the patch coupling methods, deep features generated from the AIF + contralateral coupling method improved TSS classification in most of the classifiers, e.g., LR has the best AUROC with the AIF + contralateral patches (0.765 vs. 0.658 vs. 0.676). Both LR and SVM had significantly better AUROCs (p-value=0.003 and p-value=0.024 respectively) with the features from the AIF + contralateral coupling than with the features from only the baseline imaging. Comparing to the method published by Ho *et al.* in 2017,

Table 4.5: The AUROCS of CNNs on TSS classification.

	Alexnet	VGG-16	ResNet-18	Shallow-Alexnet
AUROC	0.525	0.548	0.557	0.575
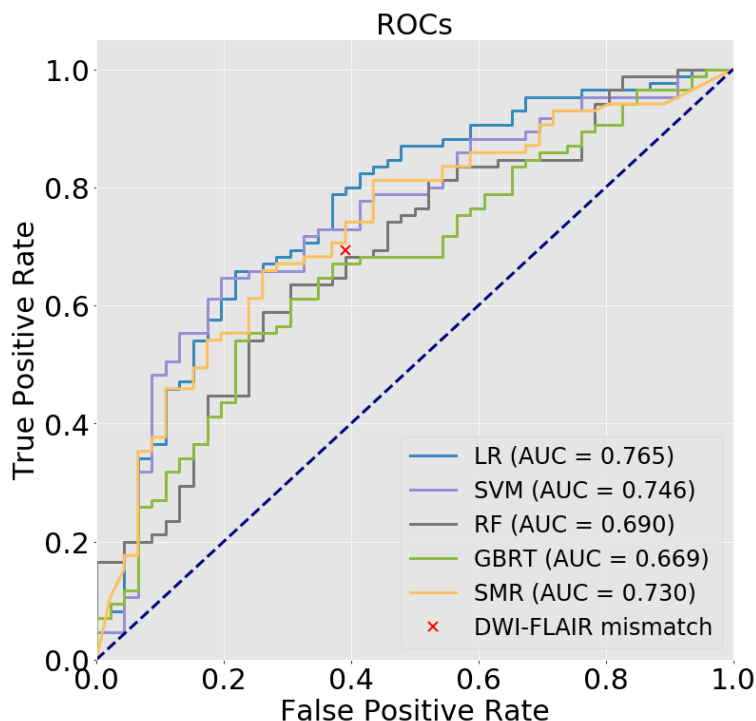
Figure 4.3: The ROCs of different classifiers trained with both the baseline imaging features and the deep features (generated from the deep AE with the AIF + contralateral coupling patch). The red cross indicated the neuroradiologist classification using the DWI-FLAIR mismatch method.

all classifiers (AIF + contralateral patches) performed better using the current method. Compared to the best proposed model (LR), the best end-to-end CNN had a lower AUROC (0.575 vs. 0.765, p-value = 0.0001; Table 4.5). The low performance may be due to the limited training data (n=131) and a large number of trainable weights (>100,000). Transfer learning [211] and semi-supervised learning [156] may be explored to improve the CNN performance.

Figure 4.3 shows the neuroradiologist performance using the DWI-FLAIR mismatch method, and the ROCs of the classifiers trained with the baseline features and the deep features (generated

from the deep AE with the AIF + contralateral coupling patch). Among all the classifiers, the LR trained with baseline imaging features and the deep features performed the best, with an AUROC of 0.765. Comparing to the mismatch method, three classifiers (LR, SMR, and SVM) achieved higher sensitivity (while having the same specificity) with the addition of the deep features, demonstrating the ability of using imaging features with machine learning models to classify TSS. Overall, LR achieved higher sensitivity (0.788 vs 0.694), F1-score (0.788 vs 0.728), NPV (0.609 vs 0.519), and PPV (0.788 vs 0.766) while maintaining same specificity (0.609) as the DWI-FLAIR mismatch method. Therefore, LR with the baseline imaging features and the deep AE features was determined to be the most suitable classifier for the TSS classification.

4.5.2 Example of classification

Figure 4.4 shows the TSS classification example of the optimal LR classifier, trained with both the baseline and deep AE features (generated from the AIF + contralateral coupling patch). When mismatch or absence of mismatch were clear between DWI and FLAIR, the classifier was able to classify correctly (patient #1 and patient #2). In cases where the mismatch was not clear, the classifier was able to correctly classify some cases (patient #3), but occasionally resulted in misclassification (patient #4).

4.5.3 ROI sensitivity analysis

Table 4.6 shows the TSS classification results on the sensitivity analysis. The deep AE features were still able to improve the performance of almost every classifier with two additional thresholds (Tmax>4s and Tmax>8s). The only exception was SMR with Tmax>8s, but the difference was not statistically significant (p-value=0.089). These results show that the deep AE feature generation is robust despite the change of the ROIs. Among all the cutoff values, Tmax>6s remained the threshold that provided the ROIs with the best performance in all classifiers with baseline and deep
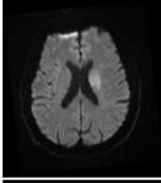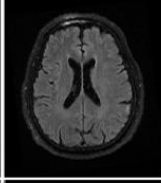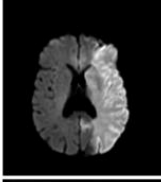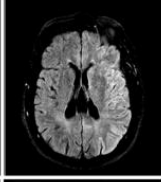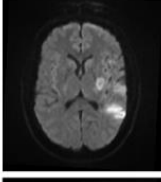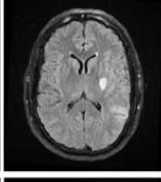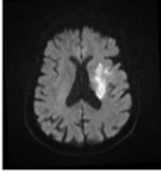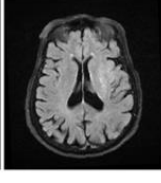
Patient	DWI	FLAIR	Ground Truth TSS	DWI-FLAIR Mismatch	Model Classification
1			<4.5hrs	1 (<4.5hrs)	<4.5hrs
2			≥4.5hrs	0 (≥4.5hrs)	≥4.5hrs
3			≥4.5hrs	1 (<4.5hrs)	≥4.5hrs
4			≥4.5hrs	1 (<4.5hrs)	<4.5hrs

Figure 4.4: Examples of TSS classification of the optimal LR classifier trained with both the baseline and deep AE features. Patient #1 and patient #2 were correctly classified and showed clear mismatch and absence of mismatch respectively between DWI and FLAIR. In patient #3, the classifier still classified correctly even though the mismatch between DWI and FLAIR was less obvious. Patient #4 was misclassified. In patient #4, there was a visible mismatch between DWI and FLAIR images, but clinical history determined TSS to be >4.5hrs.

AE features. This verifies that Tmax>6s is the optimal threshold to define hypoperfused regions for TSS classification.

4.5.4 Feature correlation analysis

Table 4.7 shows the top-10 correlated imaging features to the TSS label and Table 4.8 shows several examples of the deep AE feature correlation to the baseline imaging features. Four out of

Table 4.6: Classifier AUROCs (with AIF + Contralateral patch) in ROI sensitivity analysis.

Classifier	Tmax>4s		Tmax>6s[¥]		Tmax>8s	
	BI[†]	BI+AE[Ω]	BI	BI+AE	BI	BI+AE
LR	0.520	0.690*	0.618	0.765*	0.622	0.651
RF	0.667	0.678	0.640	0.690	0.610	0.666
GBRT	0.607	0.650	0.608	0.670	0.618	0.644
SVM	0.479	0.649*	0.636	0.746*	0.624	0.683
SMR	0.494	0.591	0.661	0.730	0.696	0.624

* statistically significant result (p-value<0.05) against model with BI features only
[†] BI = Models were trained with the baseline imaging features (94 descriptive statistics and 8 morphological features)
[Ω] BI + AE = Models were trained with the baseline and deep AE features
[¥] The result is obtained from Table 4.4

Table 4.7 Top-10 correlated (Pearson correlation) baseline + deep features to the TSS label.

Features	$|\rho|$
AE25* relative maximum	0.307
DWI mean	0.286
AE25 relative variance	0.267
FLAIR relative mean	0.261
DWI variance	0.255
AE23 minimum	0.246
DWI relative minimum	0.241
AE8 mean	0.239
FLAIR relative minimum	0.236
ADC-FLAIR maximum	0.236

[¥] AE25 = Deep AE feature map #25. This applies to other abbreviations (e.g., AE23)

the top ten correlated features were the deep AE features. It is interesting to observe that the relationships can be grouped into certain categories. For example, the AE8 feature correlated well

Table 4.8: Feature correlation between the deep AE features and the baseline imaging features.

Rank	AE8 relative minimum	AE7 Relative max	AE16 variance	AE23 Relative variance
1	TTP relative minimum	Area	ADC variance	DWI relative maximum
2	TTP minimum	Maximum diameter	ACD-FLAIR relative mean	DWI variance
3	MTT relative minimum	Volume	ADC-FLAIR relative variance	FLAIR relative maximum
4	TTP maximum	Minimum diameter	MTT variance	DWI-FLAIR relative variance
5	DWI minimum	TTP minimum	ADC mean	SV
Interpretation	Time-related	Morphology-related	ADC-related	DWI-related

with the time-related baseline imaging features (TTP and MTT), whereas the AE7 feature correlated well with the morphological baseline imaging features (e.g., area). Some deep AE feature (e.g., AE16) correlated well to an image type (e.g., ADC). The correlation analysis demonstrates that the deep AE features capture a variety of complex representations (i.e., shape, morphology).

4.5.5   TSS subgroup classification analysis

Table 4.9 summarized the classifier performance in the TSS subgroup classification analysis. Three out of five classifiers showed improvement with the addition of deep AE features on the field strength subgroup analysis, and four out of five classifiers showed improvements on the year of imaging subgroup analysis. Overall, the results still align with the general finding, i.e., the proposed method generated deep features that could improve the TSS classification.

Table 4.9: The AUROCs of classifiers in subgroup analysis.

Classifier	Field Strength			Year of Imaging		
	BI[†]	AE[¥]	BI+AE[Ω]	BI	AE	BI+AE
LR	0.637	0.673	**0.751**	0.554	0.660	**0.648**
RF	**0.620**	0.610	0.606	0.664	0.713	**0.740**
GBRT	0.603	0.631	**0.624**	0.692	0.664	**0.700**
SVM	0.605	0.496	**0.728**	0.577	0.596	**0.673**
SMR	**0.625**	0.608	0.603	0.538	**0.787**	0.488

BOLD indicated higher AUROC of model with BI+AE features against model with BI features only

[†]BI = Models were trained with the baseline imaging features (94 descriptive statistics and 8 morphological features)

[¥]AE = Models were trained with the 384 descriptive statistics generated from the deep AE feature maps)

[Ω]BI + AE = Models were trained with the baseline and deep AE features

4.6 Discussion

The DWI-FLAIR mismatch method is the current state-of-the-art method that can provide clinicians with insight into stroke onset time based on observable mismatch patterns between DWI and FLAIR. Yet this method suffers from its simplicity, i.e., the mismatch pattern between DWI and FLAIR may not capture all patients in whom TSS<4.5hrs [49], which can lead to a misclassification. Additionally, convolutional neural networks are often the first choices of model to generate hierarchical features for classification [135], yet they are not applicable in this study because of a limited number of patients (n=131) and high dimensional input data (4-D PWIs). This hypothesis was verified by the poor performance of standard CNNs in TSS classification (Table 4.5). In this work, we proposed an approach for TSS classification that included machine learning model training, and novel imaging feature generation from MR images and perfusion parameter maps using deep learning. With only the proposed baseline imaging features, the best classifier

(SMR) can achieve an AUROC of 0.661 on TSS classification (Table 4.4). This indicates that the machine learning models capture signal changes from MR images and perfusion parameter maps that are predictive of TSS. One possible signal is the change of the perfusion parameter value (e.g., CBV) over time within the ischemic stroke regions, previously demonstrated in animal studies [212], [213]. This also shows that the enriched baseline imaging feature set improves the TSS classification, in which previous work [194] showed a limited performance (AUROC<0.700) with a single mean intensity value feature.

Another interesting observation is that the deep AE feature maps generated from the AIF + contralateral coupling input are more predictive than the deep AE feature maps generated from either the AIF coupling input or the contralateral coupling input. This supports the hypothesis that the AIF patch provides the base for the initial bolus setting (e.g., how fast the bolus is injected) whereas the contralateral patch provides a matched control for the healthy brain concentration time curve. Adding the deep imaging features (from the AIF + contralateral coupling) could also improve the classifiers' performance, e.g., the best classifier (LR) was improved by at least an AUROC of 0.1, and the correlation shows that 4 out of 10 top-10 correlated are the deep AE features. These observations suggest that the deep AE features are important for improving the TSS classification.

This study does have some limitations. First, the machine learning models were trained and validated on only the MR images because MR images are commonly retrieved in the clinical routine of UCLA hospital. Yet, CT perfusion is becoming more widely used [214], and a study has shown that CT perfusion may help classifying TSS [215]. Second, clinical variables (e.g., age) were not investigated in the classification, which may further improve TSS classification because it is likely that they correlate with TSS. A possible improvement is to develop a multimodal framework [51] that uses both the imaging and clinical features to do TSS classification. More details about TSS classification and future works will be presented in Chapter 7.

CHAPTER 5

Estimating Perfusion Parameter Maps Using Deep Learning

5.1 Overview

Magnetic resonance (MR) and computed tomography (CT) perfusion-weighted imaging generate four-dimensional spatial-temporal images that are often acquired during acute stroke evaluation. Perfusion parameter maps (e.g., CBF and Tmax) are derived from these high dimensional images for quantification of cerebral perfusion in the clinical diagnosis of acute stroke. These perfusion parameter maps are widely used. In the simplest application, a thresholding approach is applied to these maps for comparison with other MR images to identify volumes of salvageable tissue [16], [96]. In advanced analyses, the perfusion parameter maps are used in machine learning models to predict stroke tissue outcomes (Chapter 3) and estimate stroke onset time (Chapter 4).

Methods have been developed to generate perfusion parameter maps from source perfusion-weighted imaging. These methods are mostly based on the theory of singular value deconvolution (SVD), including delayed-corrected SVD (dSVD) [116], block-circulant SVD (bSVD) [95], and oscillation-index SVD (oSVD) [95]. Alternative approaches have been explored, including a Bayesian approach [117], and Tikhonov regularization [119]. Because the acquired concentration curves are generally noisy, some of these methods may produce residue functions that are not physiologically plausible and subject to distortions that can underestimate the perfusion parameters. Therefore, ongoing studies directly use source perfusion images instead for stroke image analysis, such as tissue outcome prediction in Chapter 3.

This chapter presents a novel approach to estimate perfusion parameters based on recognizing patterns from data using deep learning. We developed a deconvolution-free deep model to estimate perfusion parameters and the evaluation results demonstrate that the proposed approach is an alternative method that has the potential to improve the current quantitative analysis of perfusion images. In the remainder of this chapter, Section 5.2 describes the dataset and the data preprocessing steps. Section 5.3 describes the proposed deep learning model architecture for perfusion parameter estimation. Section 5.4 summarizes the experimental results. Section 5.5 discusses the limitations and future works. The work presented in this chapter has been published in [216], [217].

5.2 Dataset and Data Preprocessing

A set of 11 acute ischemic stroke patient MR perfusion data was collected retrospectively from UCLA picture archiving and communication system (PACS). The sparse perfusion deconvolution toolbox [168] and the ASIST-Japan perfusion mismatch analyzer [218] were used to generate the ground truth perfusion parameter maps and AIFs respectively. The perfusion parameter maps include the cerebral blood volume (CBV), cerebral blood flow (CBF), mean transit time (MTT), and time-to-maximum (Tmax). All the perfusion images were interpolated and truncated to have a consistent 70 second time interval. The ranges of CBV, CBF, MTT, and Tmax values are between 0-201 ml/100g, 0-1600 ml/100g/min, 0-25 s, and 0-69 s (Tmax was clipped at 11s because there were too few examples beyond this value) respectively. All perfusion parameter values were partitioned into ten equal-sized bins and training samples were drawn equally from each bin to avoid imbalance in the training dataset [11]. The number of samples for the CBV, CBF, MTT, and Tmax values were 91,950; 97,110; 87,080; and 74,850 respectively.

5.3 Methods

The estimation task was to estimate the perfusion parameters for a voxel, given its concentration time curve (CTC) and arterial input function (AIF). The pattern recognition method

for perfusion parameter estimation was a bi-input convolutional neural network (bi-CNN), which took the two inputs (CTC, AIF) and generated an estimated perfusion value for a voxel. The overall estimation task is defined as:

$$v = f(AIF, CTC), \tag{5.1}$$

where v is the estimated value, and $f(\cdot)$ is the bi-CNN with trained weights. Separate bi-CNNs were trained to estimate four perfusion parameter (CBV, CBF, MTT, and Tmax). Each bi-CNN was trained to learn important features from the input data to make an approximation. The details for bi-CNN models to complete the estimation task are described in the following subsections.

5.3.1 Defining training data

Each training example (voxel) consists of a pair: the CTC and its AIF. A CTC or AIF was a one-dimensional vector with a size of T, where T is the number of time points (in this study, $T = 70$). A small region was included in each training voxel because a previous work suggests that regional information corresponding to a voxel's surroundings can improve prediction in MR images [22]. This resulted in a patch size of 3 x 3 x T patch (width x height x time; the z-dimension is omitted) for every training example, where the center of the patch was the voxel of interest.

5.3.2 Bi-CNNs for perfusion parameter estimation

The proposed bi-CNN model was inspired by a denoising architecture [219]. A simple signal with artifacts model can be defined as:

$$y = x * k, \tag{5.2}$$

where y is the observed 1D signal (instead of a 2D image), x is the original artifact-free signal, and k is the convolution kernel that is caused by the artifacts. When a Fourier transform operator, $F(\cdot)$, is applied with Tikhonov regularizer, x can be re-expressed as:
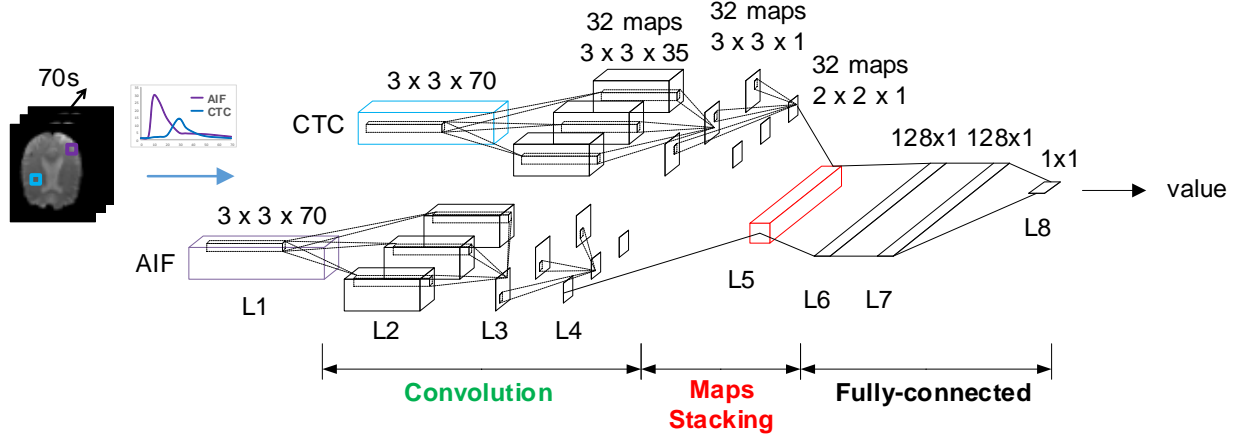
Figure 5.1: The proposed architecture of a bi-CNN for perfusion parameter estimation. It consists of three components: (1) convolution, (2) maps stacking, and (3) fully-connected. Feature maps are first learned separately for a CTC and its AIF in the convolution chains which follow the denoising architecture [219]. The feature maps are then stacked together in the maps stacking component, followed by two fully-connected layers for parameter estimation.

$$x = F^{-1}\left(\frac{1}{F(k)}\left\{\frac{|F(k)|^2}{|F(k)|^2 + \frac{1}{SNR}}\right\}\right) * y = k^* * y, \tag{5.3}$$

where SNR is the signal to noise ratio and k^* is the pseudo inverse kernel. The new representation of x can be further expanded into a matrix representation by the kernel separability theorem, where k^* is decomposed into $k^* = \boldsymbol{U} \cdot \boldsymbol{S} \cdot \boldsymbol{V}^T$. This leads to a new representation of x:

$$x = k^* * y = \sum_j s_j \cdot u_j * (v_j^T * y), \tag{5.4}$$

where u_j and v_j are the jth columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively, and s_j is the jth singular value. This new expression shows that the original artifact-free signal, x, can be obtained via the weighted sum of separable 1D filters [219]. This motivates the design of a bi-CNN (Figure 5.1), which consisted of three components: (1) convolution, (2) maps stacking, and (3) fully-connected.

In the convolution, a CTC and its AIF were convolved independently via multiple convolutional layers (i.e., two convolution chains) to remove artifacts (e.g., noise, distortion) that were often seen in the input perfusion signals, which was important to identify fine-grained features from CTC and AIF signals that help estimation. Following the denoising architecture [219], two separate 1D convolutions were first performed (L1 to L2 and L2 to L3), with a filter size of 1 x 1 x 36 and 1 x 1 x 35 respectively. Then, a convolutional layer (L3 to L4) was added after the denoising architecture to learn filters for detecting the spatial contributions of neighboring voxels. In total, 32 maps were learned (with zero-padding and a stride of one) in every convolutional layer. The output feature maps of the convolution chains were stacked together in the maps stacking layer (L5), resulting in a matrix with a size of 64 x 2 x 2 x 1. The matrix was connected to two fully-connected layers where hierarchical features were learned to correlate the AIF and CTC derived features. A rectified linear unit (ReLU) layer was attached to every convolutional and fully-connected layer. The output (L8) was the estimated parameter value.

The training optimization of the network was to obtain network weights, Θ, that minimized the mean squared loss between the true value, V, and the estimated value, $\hat{V}(\Theta)$, across the samples with size n:

$$\arg min_{\Theta} \, loss = \frac{1}{n} \sum_{i=1}^{n} (V_i - \hat{V}_i(\Theta))^2, \qquad (5.5)$$

5.3.3   Model architecture for each perfusion parameter

The proposed architecture (Figure 5.1) worked well for CBV, MTT, and Tmax estimation. For CBF, it is a physiological parameter representing the maximum cerebral blood flow. Mathematically, a CBF can be defined as the maximum value of a residue function:

$$CBF = max(R(t)), \qquad (5.6)$$

where *R(t)* is the residue function of a CTC. This motivates a change in the original architecture, in which a max-pooling layer (with a max operator) was added to help identifying maximum values. The new max-pooling layer was inserted into L3 to replace the second convolutional layer in each convolutional chain for bi-CNNs of CBF. The size of the max-pooling layer was set to 1 x 1 x 35 to maintain the size consistency across the rest of the network.

5.4 Evaluation and Results

5.4.1 Bi-CNN configuration and implementation

There are two changes that are important to optimize the performance of the model, which are different from standard CNN configurations [188]. First, dropout was not included in the fully-connected layers because it decreased performance during validation. This may be due to the nature of the problem of parameter estimation (i.e., estimating a continuous value versus predicting a categorical label), where every output unit may contribute (to some degree) to the estimated value. Second, the training losses can easily explode when a learning rate is too high, especially for perfusion parameters with high maximum values (e.g., max(CBF) = 1600). Therefore, the initial learning rates are different for different parameter estimations, i.e., they are 0.0005, 0.00005, 0.005, 0.005 for CBV, CBF, MTT, and Tmax respectively, with a learning rate decay of 1e-8, 1e-9, 1e-7, 1e-7 respectively.

The bi-CNNs were trained with batch gradient descent (batch size: 50; epochs: 10) and backpropagation. A momentum of 0.9 was used. A heuristic was applied to improve the learning of deep CNN weights [188], where the learning rate was divided by 10 when the validation error rate stopped improving with the current learning rate. This heuristic was repeated three times. The deep CNNs were implemented in Torch7, and all the experiments were performed on a NVIDIA Tesla K40 GPU.

5.4.2 Evaluations

There were two evaluations. First, models were evaluated using leave-one-patient-out cross-validation, in which a validation was performed on a held-out patient and the model training was done using the remaining patients. The average root-mean-square error (ARMSE) of a validation was calculated using following definition:

$$ARMSE = \frac{1}{n_T} \sum_{j=1}^{n_T} \sqrt{\frac{1}{s_j} \sum_{i=1}^{s_j} (V_{i,j} - \hat{V}_{i,j})^2} \, , \tag{5.7}$$

where n_T is the total number of patients, V is the ground truth value from bSVD, $\hat{V}$ is the estimated value from bi-CNN, and s_j is the number of samples. Second, published CBF and Tmax thresholds [96], [220] were used to define the salvageable tissue masks generated from the bi-CNN and the ground truth perfusion maps. The similarity between these masks (the ground truth salvageable tissue mask, A, and the estimated mask, B) was calculated using the Dice coefficient [221]:

$$\text{Dice}(A, B) = 2 \frac{|A \cap B|}{|A| + |B|}, \tag{1}$$

A value of 0 indicates no overlap, and a value of 1 indicates perfect similarity. A good overlap between masks is considered to have occurred when the Dice coefficient is larger than 0.7.

5.4.3 Results

5.4.3.1 Perfusion parameter estimation using Bi-CNNs

The bi-CNNs achieved an ARMSE of 4.80 ml/100g, 27.4 ml/100g/min, 1.18 s, 1.33 s for CBV, CBF, MTT, and Tmax respectively, which are equivalent to 2.39%, 1.71%, 4.72%, and 1.19% of the individual perfusion parameter's maximum value. The small ARMSE results show that the bi-CNNs are capable of learning important patterns from the inputs to approximate perfusion
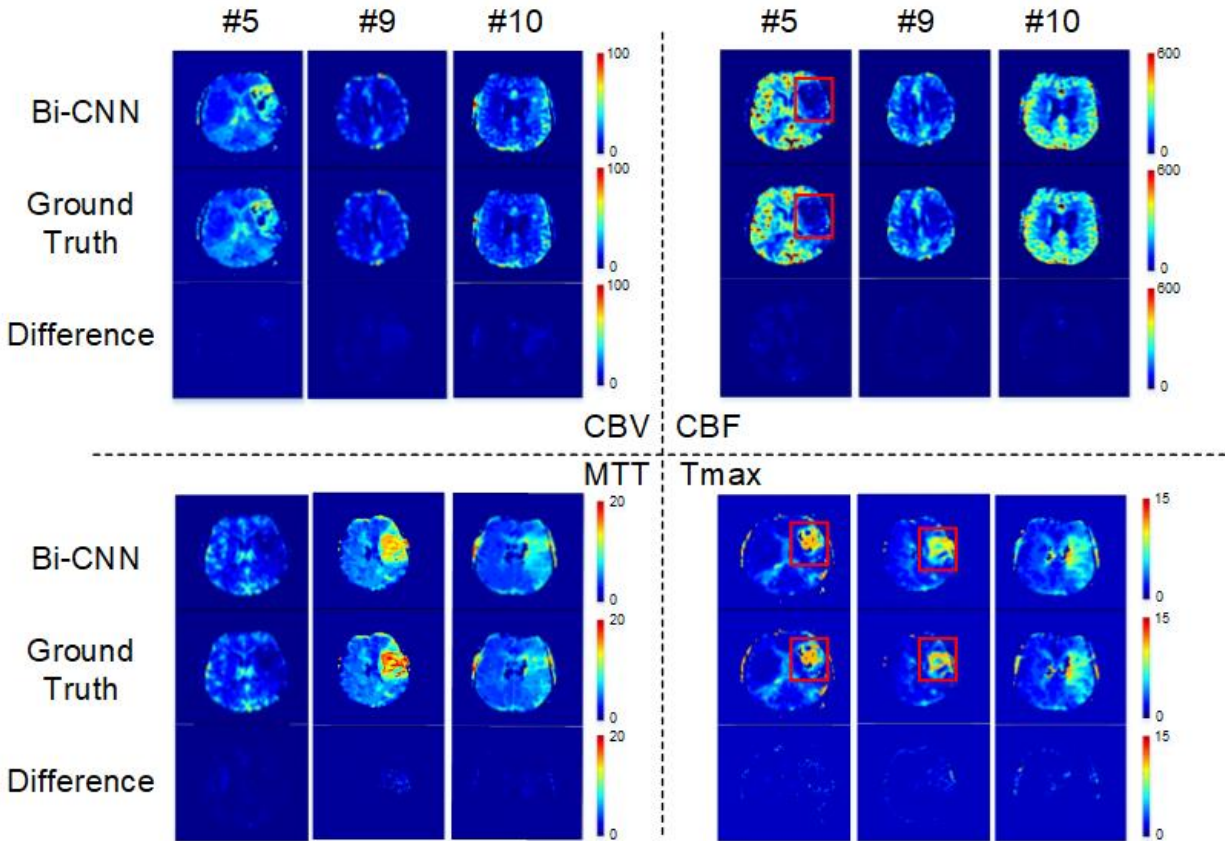
Figure 5.2: Examples of estimated perfusion maps (CBV, CBF, MTT, Tmax) generated by the bi-CNNs. Top row: bi-CNN maps; middle row: ground truth; bottom row: the difference between the bi-CNN maps and the ground truth. The estimated perfusion maps show good alignment with the ground truth. Perfusion abnormalities (i.e., hypoperfusion) can be visually detected in some estimated perfusion maps (red boxes).

parameters from CTCs and AIFs without using standard deconvolution. Figure 5.2 shows several examples of estimated perfusion maps. We observed that all the estimated perfusion maps (CBV, CBF, MTT, and Tmax) showed good alignment with the ground truth and we could even identify hypoperfusion (i.e., less blood flow) in some of the estimated maps (red boxes). The differences between the estimated maps and the ground truth were minimal.

To verify the usability of the estimated perfusion maps, the salvageable tissue masks (Figure 5.3) were generated from a CBF cutoff of 50.2 ml/100g/min [220] and a Tmax cutoff of 4s [96]
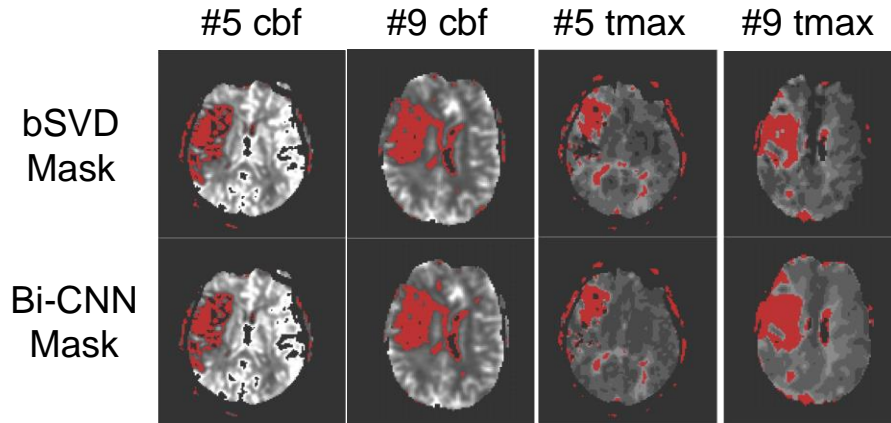
Figure 5.3: Salvageable tissue masks (red) defined on the CBF and Tmax maps generated bSVD and bi-CNNs. The bi-CNN generated salvageable tissue masks have good alignments with the bSVD generated masks. Note that the difference in contrast grayscale scale is caused by different range of perfusion parameter values.

using the ground truth (from bSVD) and the estimated perfusion maps (from bi-CNN). The average Dice coefficients for the CBF and Tmax masks were 0.830 ± 0.109 and 0.811 ± 0.071 respectively, showing good overlap between the ground truth masks and the estimated masks. This result shows that the bi-CNN can generate useful masks for salvageable tissue approximation.

5.4.3.2   Feature filters in Bi-CNNs

Feature filters were learned in the convolution chains during training. Figure 5.4 shows some examples of learned convolutional filters from the first layer of the CTC convolution chain. Each row represents a 1 x 1 x 36 temporal filter and each column is a unit filter at a time point. As can be seen, these filters capture high signals (white) and low signals (black) at different time points, which helps the fine-grained temporal feature detections from the source signals. For example, feature filter 2 captured high signal mostly after 20s. In contrast, feature filter 4 captured high signal mostly before 15s. These feature filters are important to capture input signal patterns for perfusion parameter estimation.
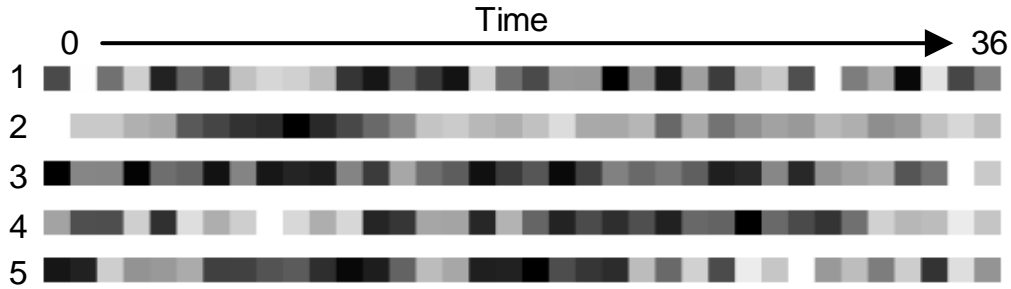
Figure 5.4: Examples of learned temporal filters (1 x 1 x 36) in the first layer of the CTC convolution chain. Each row represents a temporal filter; each column represents a unit filter at a time point. These filters capture signal changes along the time dimension for parameter estimation.

5.5 Discussion

The cross-validation results show that the patch-based bi-CNN model can estimate four perfusion parameters in stroke patients without using a standard deconvolution method (e.g., SVD) and the estimated perfusion parameter maps were numerically close to the ground truth maps generated from bSVD. Additionally, the binary mask results show the utility of the bi-CNN perfusion parameter maps in salvageable tissue identification.

One limitation of this study is that the bi-CNNs were not evaluated using digital phantoms [218], which is a more accurate source of ground truth. The results show that it is feasible to apply a pattern recognition method (i.e., bi-CNN) to estimate perfusion parameter instead of using the standard deconvolution, which is a merit of this study. However, it would be more promising if the proposed models is evaluated using digital phantoms or even be compared against *in vivo* data [117]. Another limitation of the proposed method is that it generally requires more computational time than standard deconvolution (~5x slower) to generate an estimated perfusion map. This may limit the application of the proposed method in clinical practice because a fast imaging analysis is often required for acute stroke evaluation [222]. Therefore, future work will include evaluating the

proposed bi-CNNs using digital phantom, and exploring batch and multi-GPU processing to shorten the map generation time. More details about future works will be presented in Chapter 7.

CHAPTER 6

Acute Stroke Patient Clinical Outcome Predication

6.1 Overview

In recent decades, predictive models have been developed to predict acute stroke patient outcomes, such as patient mortality. Yet, these models often neglect the problem of between-class imbalance in binary classification tasks, which is common in medical datasets because oftentimes one class is highly dominant over another class [13]. Training classification models using an imbalanced dataset without proper handling may decrease model performance and lead to biased prediction [11]. This chapter details the comparison between six machine learning models for predicting stroke patient mortality at discharge. It includes methods for data balancing, which previous work has been focused on using only logistic regression models to predict stroke outcomes with imbalanced datasets [8], [100]–[104], [223], [224].

The remainder of this chapter is organized as follows. Section 6.2 describes the dataset used for this study. Section 6.3 details the methods with data balancing techniques for stroke outcome prediction, which includes a systematic approach to identify the optimal feature set. Section 6.4 describes evaluation methods and results for the machine learning model predictions. Section 6.5 concludes with a discussion of the strengths and limitations of the analysis. This chapter is based on the content of a published work [179].

6.2 Dataset

The machine learning models for stroke patient outcome prediction were validated using data retrieved from UCLA REDCap database [166]. A total of 778 patients were identified from this
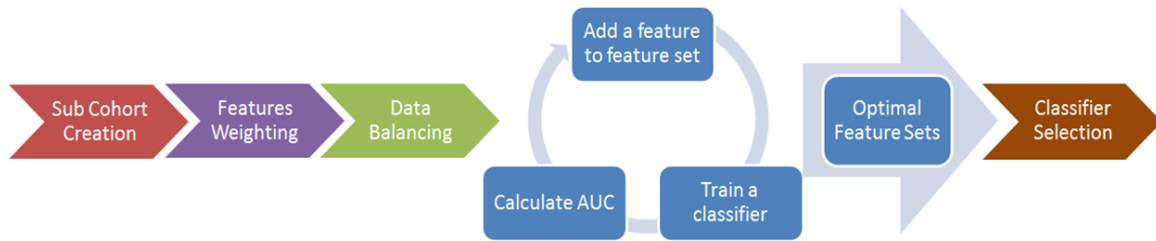
Figure 6.1: The proposed prediction framework for acute ischemic stroke patient mortality prediction at discharge. A sub-cohort was first created. Then, the relevance between features and the classes were weighted by chi-square statistics, and a balanced dataset was created. The fourth step was an iterative process in which the highest weighted feature was first used to build classifiers and AUROC was calculated. Features were added to the training feature set sequentially in the order of weighting. After several iterations, optimal feature sets for all classifiers were obtained. Finally, six classifier performances were compared, and the best classifier was identified.

study cohort, each with more than 500 features, including demographic information, laboratory results, and medications. The patient inclusion criteria were all patients with: 1) acute ischemic stroke; 2) treatments received solely at UCLA; 3) hospital stays less than 20 days (patients who stay longer are more likely to have other conditions in addition to stroke; 4) all features values after using feature inclusion criteria; and 5) discharge modified Rankin Scale (discharge mRS). The feature inclusion criterion was only features that were available in over 90% of patient cases after the patients' inclusion criteria. The prediction task was to predict patient mortality at discharge, and therefore discharge mRS was collapsed to two groups: alive (0-5) and dead (6). After applying the inclusion criteria, a total of 190 patients with 26 features were retrieved (156 alive, 34 dead). Alive patients were labeled as class-1 and dead patients were labeled as class-2.

6.3 Methods

The prediction framework is shown in Figure 6.1. It mainly consists of three steps: 1) addressing the imbalance problem using a sampling method, called the Synthetic Minority Over-

sampling Technique (SMOTE) [225]; 2) developing a systematic approach to identify the optimal feature set for every machine learning method by sequentially adding informative features until the performance no longer increased; and 3) training and comparing six machine learning methods for balanced and imbalanced datasets and determining the best model for predicting mortality at discharge. The details for each step are described in the following subsections.

6.3.1 Balancing data

The cohort subset was imbalanced (156 alive vs. 34 dead), in which the majority class (alive) was about five times more than the minority class (dead). In this case, trained models may predict all patients as alive to achieve high accuracy, but with low precision and recall [11]. Most machine learning algorithms do not deal with imbalanced datasets during training. In some cases, the minority are scattered in the feature space and the decision boundary is too specific. Therefore, we proposed to address this issue by using SMOTE, which is a sampling technique combining under-sampling of the majority class with over-sampling of the minority class [225].

SMOTE algorithm has two parts. In the first part, the minority class is over-sampled by taking each minority class sample and introducing new synthetic samples joining any or all the *k* minority class nearest neighbors, in which the closest neighbors are identified by Euclidean distance. Neighbors from the *k* nearest neighbors are randomly and uniformly chosen depending upon the number of samples required. Synthetic samples are generated using the following equation:

$$F_{new} = F_{orig} + rand(0,1) \times (F_{near} - F_{orig}), \qquad (6.1)$$

where F_{orig} is the normalized feature vector of a minority sample, F_{near} is a randomly selected normalized nearest neighbor, and F_{new} is the new synthetic sample. This approach essentially creates a random point along the line segment between two specific features and effectively forces the decision region of the minority class to become more general. Equation (6.1) is suitable for samples with only continuous features. If a sample also has binary features, the new value of each

binary feature is obtained by the majority vote (0 or 1) of all the neighbors. In the second part of algorithm, the majority class is under-sampled, in which random majority samples are neglected. However, this was not performed because the dataset was small, and all data should be considered.

In the over-sampling step, five nearest neighbors (k) were used and one of them was randomly selected to generate a synthetic sample. The synthetic step was repeated until the number of minority was equivalent to the number of majority. The SMOTE algorithm was implemented in MATLAB and the new balanced dataset is denoted as SMOTE-dataset.

6.3.2 Feature selection

It is time-consuming to collect a comprehensive set of features for every patient, and not all the features are relevant. In addition, using all the features to construct a classifier may lead to decreased performance due to over-fitting, especially on small, imbalanced datasets, which are common in stroke. Therefore, a systematic approach was proposed to select a minimum feature set to mitigate these modeling challenges, which could also be externally validated more easily.

The systematic approach consisted of several steps. First, chi-square tests were used to weight the association between features (categorical variables/binned continuous variables) and discharge mRS in the original dataset. Then, the feature with the highest association weight was used to construct a single-variable classifier. The area under the receiver operating curve (AUROC) of each classifier was computed. The feature with the second highest association weight was then added, and the AUROC was re-calculated. This process of adding features was repeated until an optimal feature set for a machine learning method was obtained. The optimal feature set was defined as the first point at which adding any additional features did not increase the performance.

After identifying the optimal feature set, the database was revisited to retrieve an additional set of patients who were not used in training and had all the optimal features. This set of patients was used as an independent test dataset.

6.3.3 Discharge mRS prediction

Early research has emphasized using logistic regression models [24] to predict stroke outcome [7-10]. None have compared the performance of different machine learning methods in stroke outcome prediction, particularly with a balanced dataset. Therefore, five common machine learning methods were investigated. They were Naïve Bayes (NB) [226], Support Vector Machine (SVM) [200], Decision Tree (DT) [226], Random Forests (RF) [185], and Logistic Regression (LR) [198]. In addition, Gumus *et. al.* [227] have shown the effectiveness of using PCA to extract principal components for SVM classification. Therefore, the performance of a combined method, principal component analysis followed by support vector machine (PCA+SVM), was also investigated. All models were compared using AUROC and the best one was chosen for stroke patient mortality prediction at discharge. Models were fitted and compared using RapidMiner (RM) [228]. After identifying the best model and the number of top features, the database was revisited to identify patients that were not included in the initial filtering due to missing data, but had all the top features. These patients were then used as an independent test dataset to test the classifier. A Mann-Whitney U test and a weighted Wilcoxon signed-rank test were used to verify if the best classifier performance was better than random and was statistically significant.

6.4 Results

6.4.1 SMOTE sampling

There were 190 patients in the data cohort with 156 class-1 patients (alive) and 34 class-2 patients (dead). The balanced SMOTE-dataset had 156 class-1 patients and 156 class-2 patients. Table 6.1 shows the feature distribution of class-1 and class-2 patients before and after SMOTE.

Table 6.1: Feature distribution of alive and dead patients before and after SMOTE.

		Alive	Dead (Before)	Dead (After)
Size		156	34	156
Continuous features		Average (SD)	Average (SD)	Average (SD)
	Age	67.9 (17.1)	81.3 (9.6)	81.4 (8.2)
	Pre-NIHSS	12.2 (7.2)	18.2 (5.9)	18.2 (4.8)
	Systolic blood pressure	151.0 (28.2)	147.3 (31.0)	145.5 (27.5)
	Diastolic blood pressure	82.4 (17.7)	79.1 (19.1)	77.9 (16.3)
	Blood glucose	131.7 (48.1)	164.8 (74.7)	160.8 (63.0)
	Blood platelet count	2245.0 (75.6)	181.5 (54.9)	179.4 (50.5)
	Hematocrit	39.0 (5.3)	38.7 (5.4)	38.4 (4.8)
	Time difference between first image and symptom (mins)	178.8 (152.8)	166.2 (109.3)	171.4 (96.3)
	Time difference between first image and admission (mins)	44.7 (27.7)	49.4 (34.7)	48.9 (30.7)
Binary features		% of presence	% of presence	% of presence
	Gender (Male)	44.9	52.9	37.8
	Hypertension	66.7	82.4	84.6
	Diabetes	16.7	20.6	6.4
	Hyperlipidemia	27.6	35.3	18.0
	Atrial fib	30.1	55.9	41.7
	Myocardial infarction	12.12	38.2	25.0
	Coronary artery bypass surgery	7.7	11.8	2.6
	Congestive heart failure	3.2	26.5	8.3
	Peripheral vascular disease	0.0	2.9	0.6
	Carotid endarterectomy angioplasty	1.9	5.9	1.3
	Brain aneurysm	0.0	2.9	0.6
	Active internal bleeding	0.0	0.0	0.0
	Low platelet count	0.0	0.0	0.0
	Abnormal glucose	0.0	0.0	0.0
	Diabetes medication	14.7	32.4	10.9
	Hypertension medication	53.2	82.4	86.5
	Hyperlipidemia medication	16.7	23.5	7.7
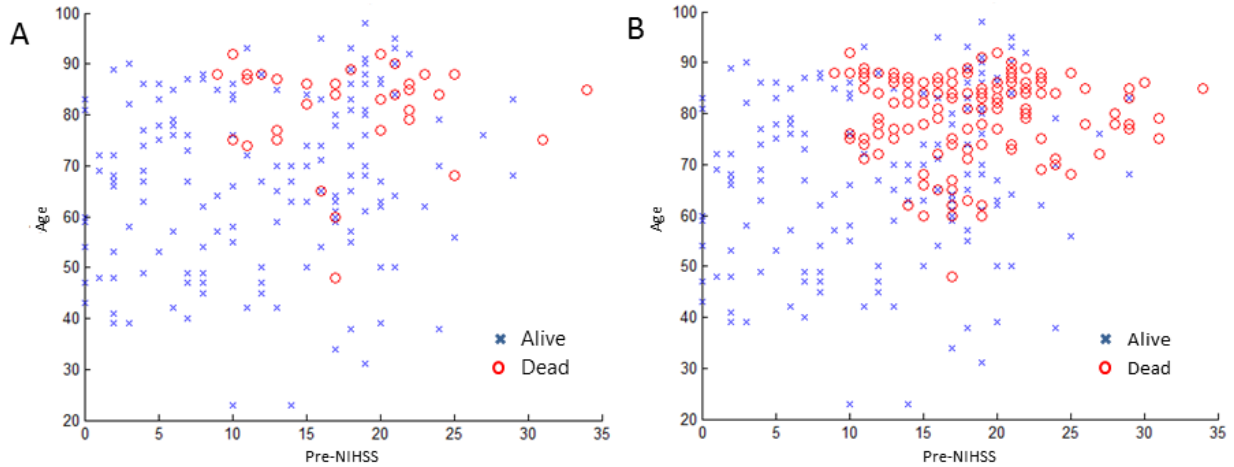
Figure 6.2: Data distribution before and after SMOTE sampling. (A) Before SMOTE, (B) After SMOTE.

Figure 6.2A shows the two-dimensional data distribution before SMOTE sampling and Figure 6.2B shows the data distribution after SMOTE sampling. Pre-treatment National Institutes of Health Stroke Scale (pre-NIHSS) and age were used in the x-axis and y-axis respectively because a previous work found that they were predictive of stroke outcome [102]. The original class-2 distribution was scattered, and it was hard to determine where the decision boundary should be. After oversampling by SMOTE, it was clear to observe that class-2 patients clustered in the region of high age and high Pre-NIHSS.

6.4.2   Feature selection

Chi-squared statistics (weights) were calculated to obtain a ranking of feature importance. Table 6.2 shows the top-10 normalized chi-squared statistics. Six out of 10 statistics belonged to continuous features. Among the top-10 features, pre-NIHSS and age were the first two features that were correlated the most to the discharge mRS.

Table 6.2: Top ten normalized weights by chi-squared statistic.

Feature	Weight
Pre-NIHSS	1.000
Age	0.631
Patient history of congestive heart failure	0.565
Platelet count	0.450
Patient history of myocardial Infarction	0.346
Serum glucose	0.325
Patient on hypertension medication	0.249
Time difference between the first MRI image and admission	0.243
Systolic blood pressure	0.223
Patient history of atrial fibrillation	0.209

6.4.3 Discharge mRS prediction models

Ten-fold cross-validation was performed to compare different models. Model performance and model bias were evaluated using AUROC and F1-score respectively [229]. For the original dataset, nine folds of data were used to train classifiers and one data fold was classified. For the SMOTE-dataset, nine folds of data were balanced using SMOTE first and then classifiers were trained using the balanced data. The held-out imbalanced one-fold data was then classified. Features were added sequentially to train a classifier in the order of largest weight to smallest weight.

The AUROCs of each classifier with respect to features is shown in Figure 6.3. The performance of classifiers with optimal feature set is summarized in Table 6.3. Among all of the classifiers trained with the SMOTE-dataset, SVM performed the best with an AUROC of 0.865
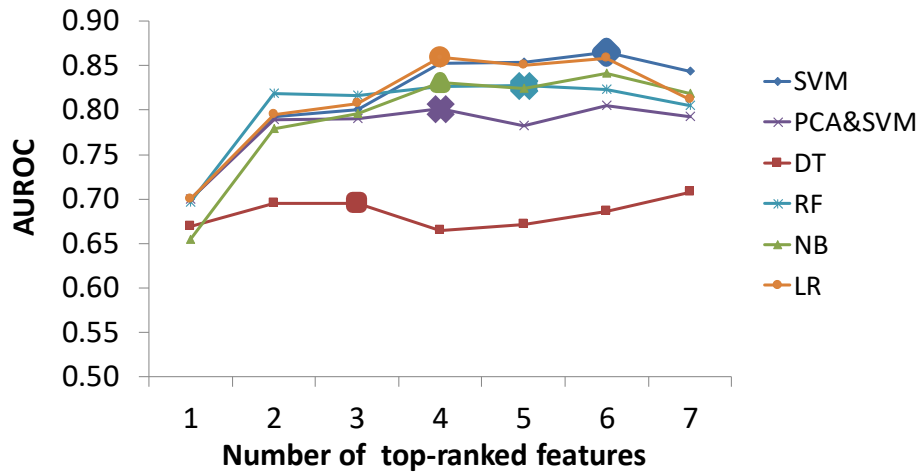
Figure 6.3: The AUROCs of classifiers with respect to different number of top-ranked features. The size of the optimal feature set for each classifier was indicated by the enlarged marker. Each classifier has different size of optimal feature set. Among all classifiers, SVM has the highest AUROC with six features being used.

Table 6.3: The size of the optimal feature set for each classifier and the cross-validation result.

Classifier	Optimal feature set size	SMOTE-dataset		Original dataset	
		AUROC	F1-score	AUROC	F1-score
SVM	6	**0.865**	**0.594**	**0.868**	0.336
PCA&SVM	4	0.802	0.488	0.856	**0.368**
DT	2	0.695	0.441	0.709	0.061
RF	5	0.828	0.401	0.789	0.249
NB	4	0.831	0.560	0.839	0.354
LR	4	0.859	0.582	0.867	0.226

and a F1-score of 0.594. PCA+SVM did not perform well, which could be due to a scenario where two classes overlapped more after projection [136]. RF and NB both generally performed more poorly than SVM in all feature sizes. DT had the worst performance with the first optimal AUROC

Table 6.4: Performance comparison of six-variable SVMs.

	SVM 1	SVM 2
Number of testing data	32 alive, 7 dead	
Number of features	6	
Applied SMOTE?	No	Yes
c-statistic	0.750	**0.781**
F1-score	0.400	**0.500**

at 0.695. This is mostly because more features were required to build a good classifier based on the nature of DT. LR had the closest performance to SVM. However, previous research has shown that SVM performs better than LR in the case of multivariate and mixture of distributions with a better (or equivalent) misclassification rate [230]. Therefore, SVM is determined to be the most suitable classifier for acute ischemic stroke mortality prediction at discharge.

After identifying the top six features for the SVM model, the database was revisited, and an additional 39 patients were retrieved and used as an independent test dataset to test the classifier. Six-variable SVM classifier trained with the SMOTE-dataset achieved higher AUROC and F1-score (Table 6.4). Both SVM1 (trained without SMOTE) and SVM2 (trained with SMOTE) were statistically better than random ($p=0.020$ and $p=0.011$, respectively). The performance of SVM2 was significantly better than the performance of SVM1 ($p=0.039$). These results suggest that balancing data (i.e., SMOTE) yields an improved classifier for predicting mortality at discharge in acute stroke patients.

6.5 Discussion

The performance of SVM, PCA-SVM, DT, RF, NB, and LR models were compared for predicting stroke patient mortality at discharge and SVM performed the best. With only one feature (Pre-NIHSS), all classifiers had moderate AUROCs (0.650-0.700) in the training dataset. The

performance gradually increased as more features were added. Each classifier reached to the first maximum AUROCs at a different number of features, with performance for each leveling off with additional features. This validates the hypothesis that more features may not necessarily improve the performance.

The size of the optimal feature set for SVM was six: Pre-NIHSS, age, platelet count, serum glucose, congestive heart failure, and myocardial infarction. These features are routinely collected in acute stroke patients. Pre-NIHSS is a standard measure of impairment caused by a stroke, with most patients receiving Pre-NIHSS assessment immediately after hospital admission. Pre-NIHSS, age, platelet count, and serum glucose level have also been shown to be relevant to stroke outcomes [102], [231], [232]. Therefore, it is reasonable to expect other institutions may maintain these variables and could validate the model.

There are a few limitations in this study. First, the distributions of binary features became more imbalanced after SMOTE (Table 6.1). A possible solution to this problem would be to assign values to binary features by sampling the probability distribution of nearest neighbors, rather than a majority vote. Second, mortality is a standard measure that most existing stroke models aim to predict. However, patients have different degrees of disability even though they are still alive. One clear difference is whether a patient can live independently or dependently [233]. A possible improvement is to extend the prediction ability to more classes (independent, dependent, and death) when a larger dataset is available. More details about future works will be presented in Chapter 7.

CHAPTER 7

Conclusion

7.1 Overview

This chapter summarizes the results and contributions of this dissertation. It also suggests future work and directions as a result of this research. This chapter concludes by giving a concluding remark in stroke treatment guidance.

7.2 Summary of Contributions and Results

This dissertation presents methods that improve acute ischemic stroke imaging analysis to help guide stroke treatment. The approaches and analyses focus on several aspects of stroke, including tissue outcome classification, time since stroke classification, perfusion parameter estimation, and stroke mortality prediction. With these methods, researchers and clinicians can gain more insight of a stroke patient status and therefore may make better judgments on the stroke treatment plan. The specific contributions of this dissertation are as follows:

- *A deep convolutional neural network to predict acute ischemic stroke tissue outcome.* We presented a novel deep CNN architecture that used contralateral information to better predict stroke tissue outcome using source perfusion images.

- *A robust machine learning approach to classify stroke onset time using MR images*. We presented an end-to-end machine learning framework, which included deep learning feature generation, to classify acute ischemic stroke onset time (<4.5 hrs).

- *A bi-input convolutional neural network for perfusion parameter estimation*. We presented a convolution neural network that adapted a denoising architecture to estimate perfusion parameters and demonstrated that the perfusion parameter estimation could be achieved without deconvolution.

- *A machine learning model for acute stroke outcome prediction with an imbalanced data set.* We described the imbalance data problem in stroke data set and developed a machine learning model for stroke patient mortality prediction.

Predicting ischemic tissue outcome is a challenging and important task for better stroke evaluation and treatment planning. Knowing the potential tissue outcome before the use of an intervention has the potential to provide important information to clinicians about the relative value of interventions. In this dissertation, we demonstrated the use of deep CNNs in ischemic tissue outcome prediction (infarcted vs non-infarcted) and proposed a deep CNN that outperformed state-of-the-art models in Chapter 3. There are two merits of the proposed model. First, the proposed algorithm can automatically learn hierarchical imaging features from only source pre-treatment perfusion images, which eliminates the use of AIF and deconvolution. The learned features represent more complicated characteristics than just the summary perfusion parameters (e.g., time-to-peak), and the results in Section 3.4.4.2 shows that these features are predictive of tissue outcome. Second, the deep learning algorithm can better capture non-linear relationships than other models. Such non-linear relationships cannot be captured by linear models, such as GLMs. In Section 3.4.4.2, compared to models such as SR-KDA and SVM, the deep CNNs automatically learned spatio-temporal features from the source perfusion images that are more complex and predictive than perfusion parameters (e.g., Tmax). In Sections 3.4.4.2 and 3.4.4.3, the results show

104

that the use of contralateral patches (as matched controls) leads to the greater learning capability of the proposed CNNs. Overall, the proposed model may benefit the weight initialization of the deep learning models for body parts other than the brain when MR perfusion images are available.

Determining stroke onset time independent of patient history is a challenging and important task for better stroke evaluation and stroke treatment decision-making. We proposed and performed a rigorous evaluation of the machine learning approach on TSS classification. This approach consisted of novel baseline and deep features generation using autoencoders, and the results in Section 4.5.1 show that the proposed deep feature generation approach achieved the best performance. In Sections 4.5.2 and 4.5.4, the ROI sensitivity analysis and the subgroup classification analysis demonstrate that the deep feature generation technique is robust to the change in ROI mask and the image acquisition parameters. To tackle the criticism of deep learning technique as a "black-box" approach, we suggested a feature correlation analysis in Section 4.4.2 and the results in Section 4.5.3 show that the deep hidden representations correlated with certain tissue perfusion status. Overall, the proposed machine learning framework is highly applicable to any task when high dimensional data is used (e.g., four-dimensional PWIs) while the imaging data set is limited.

In Chapter 5, the bi-input convolutional neural network approach builds the basis for estimating perfusion parameters without using deconvolution. This works provides a methodology to estimate perfusion parameters using pattern recognition. The results in Section 5.4.3 demonstrate the deep learning approach works well and shows that it is possible to extract patterns from input patches to estimate perfusion parameters. This work serves as the foundation to move towards better perfusion parameters estimation, which are often used for stroke evaluation, such

as penumbra calculation [234]. In Chapter 6, we provided an analysis on machine learning models trained with imbalance data and suggested a systemic approach to tackle the imbalance problem for stroke outcome prediction. The results in Section 6.4 show that the six-variable SVM is the best model to predict stroke patient discharge mortality. This work illustrates the model bias prediction caused by using imbalanced data set and points out the significance of using balanced data set. The insights from this analysis are important to any task that involves an imbalance clinical dataset. Because of this work, the proposed methods for tissue outcome prediction, TSS classification, and perfusion parameter estimation all include a data balancing step, which is critical for model performance.

7.3 Future Work

While a substantial amount of work has been done to develop approaches presented in Chapters 3-6, there are ways to improve the existing methodologies. A sample of such improvements are briefly discussed in the following sections.

7.3.1 Developing a multimodal framework

Because of the strict imaging criterion to define final infarct volume (3-7 days post-FLAIR), the collected data set was not big enough for patient subgroup analysis, and this limits the development of the deep CNNs model on tissue outcome prediction using clinical and treatment variables. Multi-modal deep learning frameworks have been developed that use different modalities of data (e.g., clinical and imaging data) to predict an outcome [51], [235], and this kind of framework can be adapted in stroke image analysis. In future work, multi-modal frameworks

may be combined with multi-task learning [236] to perform tissue outcome classification, stroke mortality prediction, and TSS classification simultaneously.

7.3.2   Improving the deep models by advanced architectures

Recent developments in deep model building blocks have raised significant interest. Some of these building blocks are slowly becoming standard layers in CNNs, such as the residual block [149], fire block [237], dense block [189], and squeeze-and-excitation block [150]. One possible extension of this dissertation is to apply these blocks or adapt the concepts of these blocks on the current proposed networks in tissue outcome prediction, TSS classification, and perfusion parameter estimation. The advantage of using these building blocks is not only to learn better features, but also to provide extra information to help interpret the models. For example, the squeeze-and-excitation block may provide the intra-correlation of feature maps within a convolutional layer.

7.3.3   Interpreting deep models

Deep learning approaches have been criticized as "black-boxes," in which the learning and classification mechanisms are too complicated and difficult to understand, engendering doubt in medical applicability because clinical decision making is ideally evidence-based [238]. In this dissertation, we provided a methodology to interpret the complex deep AE features for TSS classification (Section 4.4.2). This is an important first step because it shows clinicians what these deep AE features may represent, helping them to understand more about how the classifiers make the classification and why they can achieve better performance. The next step may be using visualization techniques [148], [239]–[241] to bring further insight into stroke image analysis, like

highlighting important brain regions or signals that drive a specific classification. Through a comprehensive visualization tool, clinicians may then be able to associate clinical reasoning (e.g., the location and the strength of the highlighted signals) with classification, making the deep learning approach more intuitive and therefore integral to the medical decision-making process.

7.3.4 Generalizing the TSS classification

The proposed machine learning approach was only evaluated using MR images. Computed tomography (CT) perfusion imaging is cheaper, faster, and more readily available than MRI and could become the imaging modality of choice for acute ischemic stroke patients if TSS analysis on CT images was accurate and independent of clinical history. Future work could collect the CT perfusion images and validate the robustness of the model on this new imaging modality. Additionally, the TSS classification may not have to be limited to <4.5hrs only. Future work could change the classification model to classify different cutoff time points (e.g., <6hrs [71] or <24hrs [72]), or even make the classification model into a regression model to make it more generalizable.

7.4 Concluding Remarks

The goal of stroke treatments is to limit mortality and reduce the long-term effects of stroke. Making stroke treatment decisions is a complicated process that involves several factors, including the patient's risk of hemorrhage and the findings in the perfusion images (e.g., the size of a penumbra). Yet the relationships between these factors and the patient outcomes (e.g., tissue fate, discharge mortality) are not fully understood. These challenges highlight the need for assisting clinicians with better understanding of stroke images to provide better stroke treatments. The contributions of this dissertation lay the foundation towards better understanding and utilizing

stroke images. First, the contributions include developing a pattern recognition approach to estimate perfusion parameters, which are often used to evaluate the size of a penumbra. This creates a base for the future research to improve perfusion parameter estimation using machine learning models. Second, this dissertation presents models using patient imaging and clinical features to predict tissue fate and patient discharge mortality. These models may generate extra information to assist clinicians for stroke treatment guidance. Lastly, the contributions also include a machine learning approach to classify TSS, which may serve as alternative to the DWI-FLAIR mismatch method. This new way provides a foundation to use imaging data in TSS classification, which could ultimately provide decision-making guidance for clinicians in acute stroke intervention treatment.

REFERENCES

[1]   E. J. Benjamin, S. S. Virani, C. W. Callaway, A. M. Chamberlain, A. R. Chang, S. Cheng, S. E. Chiuve, M. Cushman, F. N. Delling, R. Deo, and others, "Heart disease and stroke statistics—2018 update: a report from the American Heart Association," *Circulation*, vol. 137, no. 12, pp. e67--e492, 2018.

[2]   W. Hacke, M. Kaste, E. Bluhmki, M. Brozman, A. Dávalos, D. Guidetti, V. Larrue, K. R. Lees, Z. Medeghri, T. Machnig, and others, "Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke," *N. Engl. J. Med.*, vol. 359, no. 13, pp. 1317–1329, 2008.

[3]   R. G. Nogueira, H. L. Lutsep, R. Gupta, T. G. Jovin, G. W. Albers, G. A. Walker, D. S. Liebeskind, and W. S. Smith, "Trevo versus Merci retrievers for thrombectomy revascularisation of large vessel occlusions in acute ischaemic stroke (TREVO 2): A randomised trial," *Lancet*, vol. 380, no. 9849, pp. 1231–1240, 2012.

[4]   J. L. Saver, R. Jahan, E. I. Levy, T. G. Jovin, B. Baxter, R. G. Nogueira, W. Clark, R. Budzik, and O. O. Zaidat, "Solitaire flow restoration device versus the Merci Retriever in patients with acute ischaemic stroke (SWIFT): A randomised, parallel-group, non-inferiority trial," *Lancet*, vol. 380, no. 9849, pp. 1241–1249, 2012.

[5]   A. M. Spiotta, M. I. Chaudry, F. K. Hui, R. D. Turner, R. T. Kellogg, and A. S. Turk, "Evolution of thrombectomy approaches and devices for acute stroke: a technical review.," *J. Neurointerv. Surg.*, pp. 1–6, 2014.

[6]   A. J. Furlan, "Endovascular Therapy for Stroke --- It's about Time," *N. Engl. J. Med.*, pp. 1–3, 2015.

[7]   C. S. Kidwell, R. Jahan, J. Gornbein, J. R. Alger, V. Nenov, Z. Ajani, L. Feng, B. C. Meyer, S. Olson, L. H. Schwamm, A. J. Yoo, R. S. Marshall, P. M. Meyers, D. R. Yavagal, M. Wintermark, J. Guzy, S. Starkman, and J. L. Saver, "A trial of imaging selection and endovascular treatment for ischemic stroke.," *N. Engl. J. Med.*, vol. 368, no. 10, pp. 914–23, 2013.

[8]   C. Counsell, M. Dennis, M. McDowall, and C. Warlow, "Predicting outcome after acute and subacute stroke: Development and validation of new prognostic models," *Stroke*, vol. 33, no. 4, pp. 1041–1047, 2002.

[9]   C. Counsell, M. Dennis, and M. Mcdowall, "Predicting functional outcome in acute stroke: comparison of a simple six variable model with other predictive systems and informal clinical prediction," *J. Neurol. Neurosurg. Psychiatry*, vol. 75, no. 3, pp. 401–405, 2004.

[10]   M. Dennis, "Predictions Models in Acute Stroke. Potential Uses and Limitations," *Stroke*, vol. 39, no. 6, pp. 1665–1666, 2008.

[11]   Q. Gu, Z. Cai, L. Zhu, and B. Huang, "Data Mining on Imbalanced Data Sets," pp. 1020–1024, 2008.

[12]   H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[13]   Q. Wei and R. L. Dunbrack, "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics," *PLoS One*, vol. 8, no. 7, 2013.

[14]   G. W. Albers, V. N. Thijs, L. Wechsler, S. Kemp, G. Schlaug, E. Skalabrin, R. Bammer, W. Kakuda, M. G. Lansberg, A. Shuaib, W. Coplin, S. Hamilton, M. Moseley, and M. P. Marks, "Magnetic resonance imaging profiles predict clinical response to early reperfusion: The diffusion and perfusion imaging evaluation for understanding stroke evolution (DEFUSE) study," *Ann. Neurol.*, vol. 60, no. 5, pp. 508–517, 2006.

[15]   W. Kakuda, M. G. Lansberg, V. N. Thijs, S. M. Kemp, R. Bammer, L. R. Wechsler, M. E. Moseley, M. P. Marks, M. P. Parks, and G. W. Albers, "Optimal definition for PWI/DWI mismatch in acute ischemic stroke patients.," *J. Cereb. Blood Flow Metab.*, vol. 28, no. 5, pp. 887–91, May 2008.

[16]   K. S. Butcher, M. Parsons, L. MacGregor, P. A. Barber, J. Chalk, C. Bladin, C. Levi, T. Kimber, D. Schultz, J. Fink, and others, "Refining the perfusion--diffusion mismatch hypothesis," *Stroke*, vol. 36, no. 6, pp. 1153–1159, 2005.

[17]   H. Ay, E. M. Arsava, S. C. Johnston, M. Vangel, L. H. Schwamm, K. L. Furie, W. J. Koroshetz, and A. G. Sorensen, "Clinical- and imaging-based prediction of stroke risk after transient ischemic attack: The CIP model," *Stroke*, vol. 40, no. 1, pp. 181–186, 2009.

[18]   A. J. Yoo, E. R. Barak, W. A. Copen, S. Kamalian, L. R. Gharai, M. A. Pervez, L. H. Schwamm, R. G. González, and P. W. Schaefer, "Combining acute diffusion-weighted imaging and mean transmit time lesion volumes with national institutes of health stroke scale score improves the prediction of acute stroke outcome," *Stroke*, vol. 41, no. 8, pp. 1728–1735, 2010.

[19]   J. Doshi, G. Erus, Y. Ou, B. Gaonkar, and C. Davatzikos, "Multi-Atlas Skull-Stripping," *Acad. Radiol.*, vol. 20, no. 12, pp. 1566–1576, 2013.

[20]   J. Ashburner, G. Barnes, C. Chen, J. Daunizeau, G. Flandin, K. Friston, S. Kiebel, J. Kilner, V. Litvak, R. Moran, and others, "SPM12 Manual The FIL Methods Group (and honorary members)," 2014.

[21]   M. Straka, G. W. Albers, and R. Bammer, "Real-time Diffusion-Perfusion Mismatch Analysis in Acute Stroke," *J. Magn. Reson. Imaging*, vol. 32, no. 5, pp. 1024–1037, 2010.

[22]   F. Scalzo, Q. Hao, J. R. Alger, X. Hu, and D. S. Liebeskind, "Regional prediction of tissue fate in acute ischemic stroke," *Ann. Biomed. Eng.*, vol. 40, no. 10, pp. 2177–2187, 2012.

[23]   I. Rekik, S. Allassonniere, T. Carpenter, and J. M. Wardlaw, "Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: Segmentation, prediction and insights into dynamic evolution simulation models. A critical appraisal," *NeuroImage Clin.*, vol. 1, no. 1, pp. 164–178, 2012.

[24]   B. K. Menon, B. C. V Campbell, C. Levi, and M. Goyal, "Role of Imaging in Current Acute Ischemic Stroke Workflow for Endovascular Therapy," *Stroke*, vol. 46, no. 6, pp. 1453–1461, 2015.

[25] M. Goyal, B. K. Menon, and C. P. Derdeyn, "Perfusion imaging in acute ischemic stroke: let us improve the science before changing clinical practice.," *Radiology*, vol. 266, no. 1, pp. 16–21, 2013.

[26] V. Lai, "Application of Diffusion--And Perfusion--Weighted Imaging in Acute Ischemic Stroke," in *Advanced Brain Neuroimaging Topics in Health and Disease-Methods and Applications*, InTech, 2014.

[27] F. Calamante, "Arterial input function in perfusion MRI: A comprehensive review," *Prog. Nucl. Magn. Reson. Spectrosc.*, vol. 74, pp. 1–32, 2013.

[28] F. Calamante, S. Christensen, P. M. Desmond, L. Østergaard, S. M. Davis, and A. Connelly, "The physiological significance of the time-to-maximum (Tmax) parameter in perfusion MRI," *Stroke*, vol. 41, no. 6, pp. 1169–1174, 2010.

[29] S. Christensen, K. Mouridsen, O. Wu, N. Hjort, H. Karstoft, G. Thomalla, J. R??ther, J. Fiehler, T. Kucinski, and L. ??stergaard, "Comparison of 10 Perfusion MRI Parameters in 97 Sub-6-Hour Stroke Patients Using Voxel-Based Receiver Operating Characteristics Analysis," *Stroke*, vol. 40, no. 6, pp. 2055–2061, 2009.

[30] G. Thomalla, P. Rossbach, M. Rosenkranz, S. Siemonsen, A. Krützelmann, J. Fiehler, and C. Gerloff, "Negative fluid-attenuated inversion recovery imaging identifies acute ischemic stroke at 3 hours or less," *Ann. Neurol.*, vol. 65, no. 6, pp. 724–732, 2009.

[31] G. Thomalla, B. Cheng, M. Ebinger, Q. Hao, T. Tourdias, O. Wu, J. S. Kim, L. Breuer, O. C. Singer, and S. Warach, "DWI-FLAIR mismatch for the identifi cation of patients with acute ischaemic stroke within 4 · 5 h of symptom onset (PRE-FLAIR): a multicentre observational study," vol. 10, no. November, 2011.

[32] M. Ebinger, I. Galinovic, M. Rozanski, P. Brunecker, M. Endres, and J. B. Fiebach, "Fluid-attenuated inversion recovery evolution within 12 hours from stroke onset: A reliable tissue clock?," *Stroke*, vol. 41, no. 2, pp. 250–255, 2010.

[33] S. Emeriau, I. Serre, O. Toubas, F. Pombourcq, C. Oppenheim, and L. Pierot, "Can Diffusion-Weighted Imaging--Fluid-Attenuated Inversion Recovery Mismatch (Positive Diffusion-Weighted Imaging/Negative Fluid-Attenuated Inversion Recovery) at 3 Tesla Identify Patients With Stroke at< 4.5 Hours?," *Stroke*, vol. 44, no. 6, pp. 1647–1651, 2013.

[34] A. Ziegler, M. Ebinger, J. B. Fiebach, H. J. Audebert, and S. Leistner, "Judgment of FLAIR signal change in DWI-FLAIR mismatch determination is a challenge to clinicians," *J. Neurol.*, vol. 259, no. 5, pp. 971–973, 2012.

[35] I. Galinovic, J. Puig, L. Neeb, J. Guibernau, A. Kemmling, S. Siemonsen, S. Pedraza, B. Cheng, G. Thomalla, J. Fiehler, and J. B. Fiebach, "Visual and region of interest-based inter-rater agreement in the assessment of the diffusion-weighted imaging-fluid-attenuated inversion recovery mismatch," *Stroke*, vol. 45, no. 4, pp. 1170–1172, 2014.

[36] O. Wu, W. J. Koroshetz, L. Ostergaard, F. S. Buonanno, W. a Copen, R. G. Gonzalez, G. Rordorf, B. R. Rosen, L. H. Schwamm, R. M. Weisskoff, and a G. Sorensen, "Predicting tissue outcome in acute human cerebral ischemia using combined diffusion- and perfusion-weighted MR imaging.," *Stroke.*, vol. 32, pp. 933–942, 2001.

[37]   O. Wu, T. Sumii, M. Asahi, M. Sasamata, L. Ostergaard, B. R. Rosen, E. H. Lo, and R. M. Dijkhuizen, "Infarct prediction and treatment assessment with MRI-based algorithms in experimental stroke models.," *J. Cereb. Blood Flow Metab.*, vol. 27, no. 1, pp. 196–204, 2007.

[38]   V. Nguyen, H. Pien, and N. Menenzes, "Stroke tissue outcome prediction using a spatially-correlated model," *Pan Pacific Imaging Conf.*, no. 3, pp. 238–241, 2008.

[39]   H. Bagher-Ebadian, K. Jafari-Khouzani, P. D. Mitsias, M. Lu, H. Soltanian-Zadeh, M. Chopp, and J. R. Ewing, "Predicting final extent of ischemic infarction using artificial Neural network analysis of Multi-Parametric mri in patients with stroke," *PLoS One*, vol. 6, no. 8, 2011.

[40]   S. Huang, Q. Shen, and T. Q. Duong, "Artificial neural network prediction of ischemic tissue fate in acute stroke imaging.," *J. Cereb. Blood Flow Metab.*, vol. 30, no. 9, pp. 1661–70, 2010.

[41]   K. M. A. Welch, J. Windham, R. A. Knight, V. Nagesh, J. W. Hugg, M. Jacobs, D. Peck, P. Booker, M. O. Dereski, and S. R. Levine, "A model to predict the histopathology of human stroke using diffusion and T2-weighted magnetic resonance imaging," *Stroke*, vol. 26, no. 11, pp. 1983–1989, 1995.

[42]   M. A. Jacobs, P. Mitsias, H. Soltanian-Zadeh, S. Santhakumar, A. Ghanei, R. Hammond, D. J. Peck, M. Chopp, and S. Patel, "Multiparametric MRI tissue characterization in clinical stroke with correlation to clinical outcome," *Stroke*, vol. 32, no. 4, pp. 950–957, 2001.

[43]   P. D. Mitsias, J. R. Ewing, M. Lu, M. M. Khalighi, M. Pasnoor, H. B. Ebadian, Q. Zhao, S. Santhakumar, M. A. Jacobs, N. Papamitsakis, and others, "Multiparametric iterative self-organizing MR imaging data analysis technique for assessment of tissue viability in acute cerebral ischemia," *Am. J. Neuroradiol.*, vol. 25, no. 9, pp. 1499–1508, 2004.

[44]   H. Soltanian-Zadeh, H. Bagher-Ebadian, J. R. Ewing, P. D. Mitsias, A. Kapke, M. Lu, Q. Jiang, S. C. Patel, and M. Chopp, "Multiparametric iterative self-organizing data analysis of ischemic lesions using pre-or post-Gd T1 MRI," *Cerebrovasc. Dis.*, vol. 23, no. 2–3, pp. 91–102, 2007.

[45]   A. Zavaljevski, A. P. Dhawan, M. Gaskil, W. Ball, and J. D. Johnson, "Multi-level adaptive segmentation of multi-parameter MR brain images," *Comput. Med. Imaging Graph.*, vol. 24, no. 2, pp. 87–98, 2000.

[46]   J. Bernarding, J. Braun, J. Hohmann, U. Mansmann, M. Hoehn-Berlage, C. Stapf, K.-J. Wolf, and T. Tolxdorff, "Histogram-based characterization of healthy and ischemic brain tissues using multiparametric MR imaging including apparent diffusion coefficient maps and relaxometry," *Magn. Reson. Med.*, vol. 43, no. 1, pp. 52–61, 2000.

[47]   V. S. Parekh, J. R. Jacobs, and M. A. Jacobs, "Unsupervised nonlinear dimensionality reduction machine learning methods applied to multiparametric MRI in cerebral ischemia: preliminary results," in *SPIE Medical Imaging*, 2014, p. 90342O--90342O.

[48]   Y. Moradiya and N. Janjua, "Presentation and outcomes of 'wake-up strokes' in a large randomized stroke trial: analysis of data from the International Stroke Trial," *J. Stroke Cerebrovasc. Dis.*, vol. 22, no. 8, pp. e286--e292, 2013.

[49]   A. Odland, P. Særvoll, R. Advani, M. W. Kurz, and K. D. Kurz, "Are the current MRI criteria using the DWI-FLAIR mismatch concept for selection of patients with wake-up stroke to thrombolysis excluding too many patients?," *Scand. J. Trauma. Resusc. Emerg. Med.*, vol. 23, p. 22, 2015.

[50]   O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[51]   J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," *Proc. 28th Int. Conf. Mach. Learn.*, pp. 689–696, 2011.

[52]   L. Y., B. Y., and H. G., "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[53]   S. Jung, M. Gilgen, J. Slotboom, M. El-Koussy, C. Zubler, C. Kiefer, R. Luedi, M.-L. Mono, M. R. Heldner, A. Weck, and others, "Factors that determine penumbral tissue loss in acute ischaemic stroke," *Brain*, vol. 136, no. 12, pp. 3554–3560, 2013.

[54]   J. L. Saver, "Time is brain—quantified," *Stroke*, vol. 37, no. 1, pp. 263–266, 2006.

[55]   J. P. Broderick, Y. Y. Palesch, A. M. Demchuk, S. D. Yeatts, P. Khatri, M. D. Hill, E. C. Jauch, T. G. Jovin, B. Yan, F. L. Silver, and others, "Endovascular therapy after intravenous t-PA versus t-PA alone for stroke," *N. Engl. J. Med.*, vol. 368, no. 10, pp. 893–903, 2013.

[56]   A. Ciccone, L. Valvassori, M. Nichelatti, A. Sgoifo, M. Ponzio, R. Sterzi, and E. Boccardi, "Endovascular treatment for acute ischemic stroke," *N. Engl. J. Med.*, vol. 368, no. 10, pp. 904–913, 2013.

[57]   O. A. Berkhemer, P. S. S. Fransen, D. Beumer, L. A. Van Den Berg, H. F. Lingsma, A. J. Yoo, W. J. Schonewille, J. A. Vos, P. J. Nederkoorn, M. J. H. Wermer, and others, "A randomized trial of intraarterial treatment for acute ischemic stroke," *N. Engl. J. Med.*, vol. 372, no. 1, pp. 11–20, 2015.

[58]   B. C. V Campbell, P. J. Mitchell, T. J. Kleinig, H. M. Dewey, L. Churilov, N. Yassi, B. Yan, R. J. Dowling, M. W. Parsons, T. J. Oxley, and others, "Endovascular therapy for ischemic stroke with perfusion-imaging selection," *N. Engl. J. Med.*, vol. 372, no. 11, pp. 1009–1018, 2015.

[59]   M. Goyal, A. M. Demchuk, B. K. Menon, M. Eesa, J. L. Rempel, J. Thornton, D. Roy, T. G. Jovin, R. A. Willinsky, B. L. Sapkota, and others, "Randomized assessment of rapid endovascular treatment of ischemic stroke," *N. Engl. J. Med.*, vol. 372, no. 11, pp. 1019–1030, 2015.

[60]   J. L. Saver, M. Goyal, A. Bonafe, H.-C. Diener, E. I. Levy, V. M. Pereira, G. W. Albers, C. Cognard, D. J. Cohen, W. Hacke, and others, "Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke," *N. Engl. J. Med.*, vol. 372, no. 24, pp. 2285–2295, 2015.

[61]   T. G. Jovin, A. Chamorro, E. Cobo, M. A. de Miquel, C. A. Molina, A. Rovira, L. San Román, J. Serena, S. Abilleira, M. Ribó, and others, "Thrombectomy within 8 hours after symptom onset in ischemic stroke," *N. Engl. J. Med.*, vol. 372, no. 24, pp. 2296–2306, 2015.

[62]   B. C. V Campbell, N. Yassi, H. Ma, G. Sharma, S. Salinas, L. Churilov, A. Meretoja, M. W. Parsons, P. M. Desmond, M. G. Lansberg, and others, "Imaging selection in ischemic stroke: feasibility of automated CT-perfusion analysis," *Int. J. Stroke*, vol. 10, no. 1, pp. 51–54, 2015.

[63]   A. Mnyusiwalla, R. I. Aviv, and S. P. Symons, "Radiation dose from multidetector row CT imaging for acute stroke," *Neuroradiology*, vol. 51, no. 10, pp. 635–640, 2009.

[64]   L. Lin, A. Bivard, C. R. Levi, and M. W. Parsons, "Comparison of computed tomographic and magnetic resonance perfusion measurements in acute ischemic stroke: back-to-back quantitative analysis," *Stroke*, vol. 45, no. 6, pp. 1727–1732, 2014.

[65]   O. Y. Bang, J. L. Saver, B. H. Buck, J. R. Alger, S. Starkman, B. Ovbiagele, D. Kim, R. Jahan, G. R. Duckwiler, S. R. Yoon, and others, "Impact of collateral flow on tissue fate in acute ischaemic stroke," *J. Neurol. Neurosurg. Psychiatry*, vol. 79, no. 6, pp. 625–629, 2008.

[66]   M. P. Marks, M. G. Lansberg, M. Mlynash, J.-M. Olivot, M. Straka, S. Kemp, R. McTaggart, M. Inoue, G. Zaharchuk, R. Bammer, and others, "Effect of collateral blood flow on patients undergoing endovascular therapy for acute ischemic stroke," *Stroke*, vol. 45, no. 4, pp. 1035–1039, 2014.

[67]   D. S. Liebeskind, R. Jahan, R. G. Nogueira, O. O. Zaidat, J. L. Saver, and S. Investigators, "Impact of collaterals on successful revascularization in SWIFT." Am Heart Assoc, 2013.

[68]   E. Cheng-Ching, J. A. Frontera, S. Man, J. Aoki, Y. Tateishi, F. K. Hui, D. Wisco, P. Ruggieri, M. S. Hussain, and K. Uchino, "Degree of collaterals and not time is the determining factor of core infarct volume within 6 hours of stroke onset," *Am. J. Neuroradiol.*, 2015.

[69]   R. Hakimelahi, B. A. Vachha, W. A. Copen, G. D. E. Papini, J. He, M. M. Higazi, M. H. Lev, P. W. Schaefer, A. J. Yoo, L. H. Schwamm, and others, "Time and diffusion lesion size in major anterior circulation ischemic strokes," *Stroke*, vol. 45, no. 10, pp. 2936–2941, 2014.

[70]   N. T. Cheng and A. S. Kim, "Intravenous thrombolysis for acute ischemic stroke within 3 hours versus between 3 and 4.5 hours of symptom onset," *The Neurohospitalist*, vol. 5, no. 3, pp. 101–109, 2015.

[71]   A. Hameed, H. Zafar, D. Mylotte, and F. Sharif, "Recent trends in clot retrieval devices: a review," *Cardiol. Ther.*, vol. 6, no. 2, pp. 193–202, 2017.

[72]   F. N. Release, "FDA expands treatment window for use of clot retrieval devices in certain stroke patients," *fda.gov*, 2018. .

[73]   J. M. Wardlaw, V. Murray, E. Berge, and G. J. Del Zoppo, "Thrombolysis for acute ischaemic stroke," *Cochrane database Syst. Rev.*, no. 4, 2009.

[74]   A. Furlan, R. Higashida, L. Wechsler, M. Gent, H. Rowley, C. Kase, M. Pessin, A. Ahuja, F. Callahan, W. M. Clark, and others, "Intra-arterial prourokinase for acute ischemic stroke: the PROACT II study: a randomized controlled trial," *Jama*, vol. 282, no. 21, pp. 2003–2011, 1999.

[75] A. Ogawa, E. Mori, K. Minematsu, W. Taki, A. Takahashi, S. Nemoto, S. Miyamoto, M. Sasaki, and T. Inoue, "Randomized trial of intraarterial infusion of urokinase within 6 hours of middle cerebral artery stroke: the middle cerebral artery embolism local fibrinolytic intervention trial (MELT) Japan," *Stroke*, vol. 38, no. 10, pp. 2633–2639, 2007.

[76] E. C. Jauch, J. L. Saver, H. P. Adams Jr, A. Bruno, J. J. Connors, B. M. Demaerschalk, P. Khatri, P. W. McMullan Jr, A. I. Qureshi, K. Rosenfield, and others, "Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association," *Stroke*, vol. 44, no. 3, pp. 870–947, 2013.

[77] M. Brooks, "New Clot Retrievers Outperform Old," *Medscape.com*, 2012. .

[78] A. Bose, H. Henkes, K. Alfke, W. Reith, T. E. Mayer, A. Berlis, V. Branca, and S. Po Sit, "The penumbra system: A mechanical device for the treatment of acute stroke due to thromboembolism," *Am. J. Neuroradiol.*, vol. 29, no. 7, pp. 1409–1413, 2008.

[79] Stryker.com, "Stryker's DAWN Trial results contribute to updated acute ischemic stroke guidelines from the American Heart Association and American Stroke Association," *stryker.com*, 2018. .

[80] M. Wintermark, G. W. Albers, J. P. Broderick, A. M. Demchuk, J. B. Fiebach, J. Fiehler, J. C. Grotta, G. Houser, T. G. Jovin, K. R. Lees, and others, "Acute stroke imaging research roadmap II," *Stroke*, vol. 44, no. 9, pp. 2628–2639, 2013.

[81] J.-H. Rha and J. L. Saver, "The impact of recanalization on ischemic stroke outcome: a meta-analysis," *stroke*, vol. 38, no. 3, pp. 967–973, 2007.

[82] J. P. Tsai and G. W. Albers, "Reperfusion versus recanalization: the winner is….." American Heart Association, Inc., 2015.

[83] A. D. Horsch, J. W. Dankbaar, J. M. Niesten, T. van Seeters, I. C. van der Schaaf, Y. van der Graaf, W. P. T. M. Mali, and B. K. Velthuis, "Predictors of reperfusion in patients with acute ischemic stroke," *Am. J. Neuroradiol.*, 2015.

[84] Y.-H. Hwang, D.-H. Kang, Y.-W. Kim, Y.-S. Kim, S.-P. Park, and D. S. Liebeskind, "Impact of time-to-reperfusion on outcome in patients with poor collaterals," *Am. J. Neuroradiol.*, vol. 36, no. 3, pp. 495–500, 2015.

[85] J. G. Merino and S. Warach, "Imaging of acute stroke," *Nat. Rev. Neurol.*, vol. 6, no. 10, pp. 560–571, 2010.

[86] M. Poustchi-Amin, S. A. Mirowitz, J. J. Brown, R. C. McKinstry, and T. Li, "Principles and applications of echo-planar imaging: a review for the general radiologist," *Radiographics*, vol. 21, no. 3, pp. 767–779, 2001.

[87] M. Hammer, "MRI Physics: Diffusion-Weighted Imaging," 2013. .

[88] M.-A. Labeyrie, G. Turc, A. Hess, P. Hervo, J.-L. Mas, J.-F. Meder, J.-C. Baron, E. Touzé, and C. Oppenheim, "Diffusion lesion reversal after thrombolysis: a MR correlate of early neurological improvement," *Stroke*, vol. 43, no. 11, pp. 2986–2991, 2012.

[89]   M. Luby, S. J. Warach, Z. Nadareishvili, and J. G. Merino, "Immediate changes in stroke lesion volumes post thrombolysis predict clinical outcome," *Stroke*, vol. 45, no. 11, pp. 3275–3279, 2014.

[90]   N. Asdaghi, B. C. V Campbell, K. S. Butcher, J. I. Coulter, J. Modi, A. Qazi, M. Goyal, A. M. Demchuk, and S. B. Coutts, "DWI reversal is associated with small infarct volume in patients with TIA and minor stroke," *Am. J. Neuroradiol.*, vol. 35, no. 4, pp. 660–666, 2014.

[91]   B. M. Dale, M. A. Brown, and R. C. Semelka, "MRI: Basic principles and applications." Publics Corporate Publishing, 2003.

[92]   J. Aoki, K. Kimura, Y. Iguchi, K. Shibazaki, K. Sakai, and T. Iwanaga, "FLAIR can estimate the onset time in acute ischemic stroke patients," *J. Neurol. Sci.*, vol. 293, no. 1–2, pp. 39–44, 2010.

[93]   T. Tourdias, P. Renou, I. Sibon, J. Asselineau, L. Bracoud, M. Dumoulin, F. Rouanet, J. M. Orgogozo, and V. Dousset, "Final cerebral infarct volume is predictable by MR imaging at 1 week.," *AJNR. Am. J. Neuroradiol.*, vol. 32, no. 2, pp. 352–8, 2011.

[94]   A. Fieselmann, M. Kowarschik, A. Ganguly, J. Hornegger, and R. Fahrig, "Deconvolution-Based CT and MR Brain Perfusion Measurement : Theoretical Model Revisited and Practical Implementation Details," vol. 2011, 2011.

[95]   O. Wu, L. Østergaard, R. M. Weisskoff, T. Benner, B. R. Rosen, and A. G. Sorensen, "Tracer arrival timing-insensitive technique for estimating flow in MR perfusion-weighted imaging using singular value decomposition with a block-circulant deconvolution matrix," *Magn. Reson. Med.*, vol. 50, no. 1, pp. 164–174, 2003.

[96]   J. M. Olivot, M. Mlynash, V. N. Thijs, S. Kemp, M. G. Lansberg, L. Wechsler, R. Bammer, M. P. Marks, and G. W. Albers, "Optimal tmax threshold for predicting penumbral tissue in acute stroke," *Stroke*, vol. 40, no. 2, pp. 469–475, 2009.

[97]   C. B. Grandin, T. P. Duprez, A. M. Smith, C. Oppenheim, A. Peeters, A. R. Robert, and G. Cosnard, "Which MR-derived perfusion parameters are the best predictors of infarct growth in hyperacute stroke? Comparative study between relative and quantitative measurements," *Radiology*, vol. 223, no. 2, pp. 361–370, 2002.

[98]   I. Kane, T. Carpenter, F. Chappell, C. Rivers, P. Armitage, P. Sandercock, and J. Wardlaw, "Comparison of 10 different magnetic resonance perfusion imaging processing methods in acute ischemic stroke: Effect on lesion size, proportion of patients with diffusion/perfusion mismatch, clinical scores, and radiologic outcomes," *Stroke*, vol. 38, no. 12, pp. 3158–3164, 2007.

[99]   O. Zaro-Weber, W. Moeller-Hartmann, W.-D. Heiss, and J. Sobesky, "Maps of time to maximum and time to peak for mismatch definition in clinical stroke studies validated with positron emission tomography," *Stroke*, vol. 41, no. 12, pp. 2817–2821, 2010.

[100]   M. Stineman, G. Maislin, R. Fiedler, and C. Granger, "A prediction model for functional recovery in stroke," *Stroke*, vol. 28, no. 3, pp. 550–6, 1997.

[101]   H. Nam, K. Lee, S. Han, S. Kim, J. Y. Lee, S. H. Ahn, D. J. Kim, C. M. Nam, and J. H. Heo, "Prediction of long-term outcome by percent improvement after the first day of

thrombolytic treatment in stroke patients," *J. Neurol. Sci.*, vol. 281, no. 1–2, pp. 69–73, 2009.

[102]   C. Weimar, I. R. König, K. Kraywinkel, a Ziegler, and H. C. Diener, "Age and National Institutes of Health Stroke Scale Score within 6 hours after onset are accurate predictors of outcome after cerebral ischemia: development and external validation of prognostic models.," *Stroke.*, vol. 35, no. 1, pp. 158–62, Jan. 2004.

[103]   I. R. König, A. Ziegler, E. Bluhmki, W. Hacke, P. M. W. Bath, R. L. Sacco, H. C. Diener, and C. Weimar, "Predicting long-term outcome after acute ischemic stroke: a simple index works in patients from controlled clinical trials.," *Stroke.*, vol. 39, no. 6, pp. 1821–6, Jun. 2008.

[104]   E. Smith, N. Shobha, D. Dai, D. Olson, M. J. Reeves, J. L. Saver, A. F. Hernandez, E. D. Peterson, G. C. Fonarow, and L. H. Schwamm, "Risk Score for In-Hospital Ischemic Stroke Mortality Derived and Validated Within the Get With The Guidelines–Stroke ProgramClinical Perspective," *Circulation*, vol. 122, no. 15, pp. 1496–1504, 2010.

[105]   G. Fonarow, J. Saver, … E. S.-J. of the, and U. 2012, "Relationship of national institutes of health stroke scale to 30-day mortality in medicare beneficiaries with acute ischemic stroke," *J. Am. Heart Assoc.*, vol. 1, no. 1, pp. 42–50, 2012.

[106]   G. Saposnik, M. K. Kapral, Y. Liu, R. Hall, M. O'donnell, S. Raptis, J. V Tu, M. Mamdani, and P. C. Austin, "IScore: a risk score to predict death early after hospitalization for an acute ischemic stroke," *Circulation*, vol. 123, pp. 739–749, 2011.

[107]   D. Stribian, A. Meretoja, F. Ahlhelm, J. Pitkaniemi, P. Lyrer, M. Kaste, S. Engelter, and T. Tatlisumak, "Predicting outcome of IV thrombolysis-treated ischemic stroke patients: the DRAGON score," *Stroke*, vol. 78, no. 6, pp. 427–432, 2012.

[108]   A. Sarraj, K. Albright, A. D. Barreto, A. K. Boehme, C. W. Sitton, J. Choi, S. L. Lutzker, C. H. J. Sun, W. Bibars, C. B. Nguyen, O. Mir, F. Vahidy, T. C. Wu, G. A. Lopez, N. R. Gonzales, R. Edgell, S. Martin-Schild, H. Hallevi, P. R. Chen, M. Dannenbaum, J. L. Saver, D. S. Liebeskind, R. G. Nogueira, R. Gupta, J. C. Grotta, and S. I. Savitz, "Optimizing prediction scores for poor outcome after intra-arterial therapy in anterior circulation acute ischemic stroke," *Stroke*, vol. 44, no. 12, pp. 3324–3330, 2013.

[109]   D. Cai, X. He, and J. Han, "Speed up kernel discriminant analysis," *VLDB J.*, vol. 20, no. 1, pp. 21–33, 2011.

[110]   R. A. D. Carano, K. Takano, K. G. Helmer, T. Tatlisumak, K. Irie, J. D. Petruccelli, M. Fisher, and C. H. Sotak, "Determination of focal ischemic lesion volume in the rat brain using multispectral analysis," *J. Magn. Reson. Imaging*, vol. 8, no. 6, pp. 1266–1278, 1998.

[111]   S. E. Rose, J. B. Chalk, M. P. Griffin, A. L. Janke, F. Chen, G. J. McLachan, D. Peel, F. O. Zelaya, H. S. Markus, D. K. Jones, and others, "MRI based diffusion and perfusion predictive model to estimate stroke evolution," *Magn. Reson. Imaging*, vol. 19, no. 8, pp. 1043–1053, 2001.

[112]   S. E. Rose, A. L. Janke, M. Griffin, M. W. Strudwick, S. Finnigan, J. Semple, and J. B. Chalk, "Improving the prediction of final infarct size in acute stroke with bolus delay-- corrected perfusion MRI measures," *J. Magn. Reson. Imaging An Off. J. Int. Soc. Magn.*

Reson. Med., vol. 20, no. 6, pp. 941–947, 2004.

[113] Q. Shen, H. Ren, M. Fisher, J. Bouley, and T. Q. Duong, "Dynamic tracking of acute ischemic tissue fates using improved unsupervised ISODATA analysis of high-resolution quantitative perfusion and diffusion data," *J. Cereb. Blood Flow Metab.*, vol. 24, no. 8, pp. 887–897, 2004.

[114] N. H. Montiel, C. Rosso, N. Chupin, S. Deltour, E. Bardinet, D. Dormont, Y. Samson, and S. Baillet, "Automatic prediction of infarct growth in acute ischemic stroke from MR apparent diffusion coefficient maps," *Acad. Radiol.*, vol. 15, no. 1, pp. 77–83, 2008.

[115] C. Gottrup, K. Thomsen, P. Locht, O. Wu, A. G. Sorensen, W. J. Koroshetz, and L. Østergaard, "Applying instance-based techniques to prediction of final outcome in acute stroke," *Artif. Intell. Med.*, vol. 33, no. 3, pp. 223–236, 2005.

[116] M. Ibaraki, E. Shimosegawa, H. Toyoshima, K. Takahashi, S. Miura, and I. Kanno, "Tracer delay correction of cerebral blood flow with dynamic susceptibility contrast-enhanced MRI.," *J. Cereb. Blood Flow Metab.*, vol. 25, no. 3, pp. 378–90, 2005.

[117] T. Boutelier, K. Kudo, F. Pautot, and M. Sasaki, "Bayesian hemodynamic parameter estimation by bolus tracking perfusion weighted imaging," *IEEE Trans. Med. Imaging*, vol. 31, no. 7, pp. 1381–1395, 2012.

[118] I. K. Andersen, A. Szymkowiak, C. E. Rasmussen, L. G. Hanson, J. R. Marstrand, H. B. W. Larsson, and L. K. Hansen, "Perfusion quantification using Gaussian process deconvolution," *Magn. Reson. Med.*, vol. 48, no. 2, pp. 351–361, 2002.

[119] F. Calamante, D. G. Gadian, and A. Connelly, "Quantification of Bolus-Tracking MRI: Improved Characterization of the Tissue Residue Function Using Tikhonov Regularization," *Magn. Reson. Med.*, vol. 50, no. 6, pp. 1237–1247, 2003.

[120] M. Petkova, S. Rodrigo, C. Lamy, G. Oppenheim, E. Touzé, J.-L. Mas, J.-F. Méder, and C. Oppenheim, "MR Imaging Helps Predict Time from Symptom Onset in Patients with Acute Stroke: Implications for Patients with Unknown Onset Time," *Radiology*, vol. 257, no. 3, pp. 782–792, 2010.

[121] G. Thomalla, J. B. Fiebach, L. Østergaard, S. Pedraza, V. Thijs, N. Nighoghossian, P. Roy, K. W. Muir, M. Ebinger, B. Cheng, I. Galinovic, T. H. Cho, J. Puig, F. Boutitie, C. Z. Simonsen, M. Endres, J. Fiehler, and C. Gerloff, "A multicenter, randomized, double-blind, placebo-controlled trial to test efficacy and safety of magnetic resonance imaging-based thrombolysis in wake-up stroke (WAKE-UP)," *Int. J. Stroke*, vol. 9, no. 6, pp. 829–836, 2014.

[122] M. Koga, K. Toyoda, K. Kimura, H. Yamamoto, M. Sasaki, T. Hamasaki, T. Kitazono, J. Aoki, K. Seki, K. Homma, S. Sato, and K. Minematsu, "THrombolysis for Acute Wake-up and unclear-onset Strokes with alteplase at 0.6 mg/kg (THAWS) Trial," *Int. J. Stroke*, vol. 9, no. 8, pp. 1117–1124, 2014.

[123] L. Schwamm, "MR WITNESS: A Study of Intravenous Thrombolysis With Alteplase in MRI-Selected Patients (MR WITNESS)," *ClinicalTrials.gov*, 2011. .

[124] J. Minnerup, G. Broocks, J. Kalkoffen, S. Langner, M. Knauth, M. N. Psychogios, H.

Wersching, A. Teuber, W. Heindel, B. Eckert, and others, "Computed tomography--based quantification of lesion water uptake identifies patients within 4.5 hours of stroke onset: A multicenter observational study," *Ann. Neurol.*, vol. 80, no. 6, pp. 924–934, 2016.

[125] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Trans. Signal Inf. Process.*, vol. 3, 2014.

[126] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[127] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *J. Mach. Learn. Res.*, vol. 11, no. 3, pp. 3371–3408, 2010.

[128] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *Acoust. Speech Signal Process. (ICASSP), 2013 IEEE Int. Conf. on. IEEE*, no. 3, pp. 6645–6649, 2013.

[129] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[130] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets."

[131] H.-I. Suk, S.-W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *Neuroimage*, vol. 101, pp. 569–582, 2014.

[132] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *AISTATS '11 Proc. 14th Int. Conf. Artif. Intell. Stat.*, vol. 15, pp. 315–323, 2011.

[133] I. Arel, D. C. Rose, T. P. Karnowski, and others, "Deep machine learning-a new frontier in artificial intelligence research," *IEEE Comput. Intell. Mag.*, vol. 5, no. 4, pp. 13–18, 2010.

[134] Q. Le, J. Ngiam, Z. Chen, D. J. hao Chia, P. W. Koh, and A. Ng, "Tiled convolutional neural networks," *Adv. Neural Inf. Process. Syst. 23*, pp. 1279–1287, 2010.

[135] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2012.

[136] C. M. Bishop, "Pattern recognition," *Mach. Learn.*, vol. 128, pp. 1–58, 2006.

[137] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[138] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv:1502.03167*, vol. 37, pp. 1–11, 2015.

[139] Y. Bengio, *Learning Deep Architectures for AI*, vol. 2, no. 1. Now Publishers, Inc., 2009.

[140] Y. LeCun, "Generalization and network design strategies," *Connect. Perspect.*, pp. 143–155, 1989.

[141] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient Based Learning Applied to Document Recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[142] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, Springer, 2010, pp. 177–186.

[143] J. Duchi, E. Hazan, and Y. Singer, "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011.

[144] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *arXiv Prepr. arXiv1212.5701*, 2012.

[145] D. P. Kingma and J. L. Ba, "Adam: a Method for Stochastic Optimization," *Int. Conf. Learn. Represent. 2015*, pp. 1–15, 2015.

[146] S. Shi, Q. Wang, P. Xu, and X. Chu, "Benchmarking state-of-the-art deep learning software tools," in *Cloud Computing and Big Data (CCBD), 2016 7th International Conference on*, 2016, pp. 99–104.

[147] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *ICLR*, 2015.

[148] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *Comput. Vision–ECCV 2014*, vol. 8689, pp. 818–833, 2014.

[149] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[150] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," *Comput. Vis. pattern Recognit.*, 2018.

[151] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3361–3368, 2011.

[152] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, U. C. B. Eecs, A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D:Generic Features for video analysis," *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 675–678, 2014.

[153] A. Karpathy and T. Leung, "Large-scale Video Classification with Convolutional Neural Networks," *Proc. 2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1725–1732, 2014.

[154] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *Adv. Neural Inf. Process. Syst.*, pp. 568–576, 2014.

[155] A. Davy, M. Havaei, D. Warde-Farley, A. Biard, L. Tran, P.-M. Jodoin, A. Courville, H. Larochelle, C. Pal, and Y. Bengio, "Brain Tumor Segmentation With Deep Neural Networks," *Med. Image Anal.*, vol. 35, p. 1, 2017.

[156] J. Li, W. Speier, K. C. Ho, K. V Sarma, A. Gertych, B. S. Knudsen, and C. W. Arnold, "An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies," *Comput. Med. Imaging Graph.*, 2018.

[157] W. Li, J. Li, K. V. Sarma, K. C. Ho, S. Shen, Beatrice, S. Knudsen, A. Gertych, and C. W. Arnold, "Path R-CNN for Prostate Cancer Diagnosis and Gleason Grading of Histological Images," *IEEE Trans. Med. Imaging*, 2018.

[158] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[159] H. Il Suk, S. W. Lee, and D. Shen, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *Neuroimage*, vol. 101, pp. 569–582, 2014.

[160] H. C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1930–1943, 2013.

[161] H. R. Roth, J. Yao, L. Lu, J. Stieger, J. E. Burns, and R. M. Summers, "Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications," *Recent Adv. Comput. methods Clin. Appl. spine imaging*, pp. 3–12, 2015.

[162] M. G. Ertosun and D. L. Rubin, "Automated Grading of Gliomas using Deep Learning in Digital Pathology Images : A modular approach with ensemble of convolutional neural networks," in *AMIA Annual Symposium proceedings*, 2015.

[163] A. Love, C. W. Arnold, S. El-Saden, D. S. Liebeskind, L. Andrada, J. Saver, and A. A. T. Bui, "Unifying acute stroke treatment guidelines for a Bayesian belief network.," *Stud. Health Technol. Inform.*, vol. 192, p. 1012, Jan. 2013.

[164] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 1, pp. 436–444, May 2015.

[165] K. C. Ho, S. El-Saden, F. Scalzo, A. A. Bui, and C. W. Arnold, "Abstract WP41: Predicting Acute Ischemic Stroke Tissue Fate Using Deep Learning on Source Perfusion MRI," *Stroke*, vol. 47, no. Suppl 1, pp. AWP41--AWP41, 2016.

[166] P. a. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde, "Research Electronic Data Capture (REDCap) - A metadata driven methodology and workflow process for providing translational research informatict support.," *J. Biomed. Inform.*, vol. 42, no. 2, pp. 377–81, 2009.

[167] M. J. McAuliffe, F. M. Lalonde, D. McGarry, W. Gandler, K. Csaky, and B. L. Trus, "Medical image processing, analysis and visualization in clinical research," in *Computer-Based Medical Systems, 2001. CBMS 2001. Proceedings. 14th IEEE Symposium on*, 2001, pp. 381–386.

[168] R. Fang, T. Chen, and P. C. Sanelli, "Towards robust deconvolution of low-dose perfusion CT : Sparse perfusion deconvolution using online dictionary learning," *Med. Image Anal.*,

vol. 17, no. 4, pp. 417–428, 2013.

[169] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. J. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, R. K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J. M. Brady, and P. M. Matthews, "Advances in functional and structural MR image analysis and implementation as FSL," *Neuroimage*, vol. 23, no. SUPPL. 1, pp. 208–219, 2004.

[170] J. Kim, E. C. Leira, R. C. Callison, B. Ludwig, T. Moritani, V. A. Magnotta, and M. T. Madsen, "Toward fully automated processing of dynamic susceptibility contrast perfusion MRI for acute ischemic cerebral stroke," *Comput. Methods Programs Biomed.*, vol. 98, no. 2, pp. 204–213, 2010.

[171] K. Mouridsen, S. Christensen, L. Gyldensted, and L. Østergaard, "Automatic selection of arterial input function using cluster analysis," *Magn. Reson. Med.*, vol. 55, no. 3, pp. 524–531, 2006.

[172] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*, Springer, 2012, pp. 9–48.

[173] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," *Proc. 27th Int. Conf. Mach. Learn.*, no. 3, pp. 807–814, 2010.

[174] B. C. V Campbell, S. Christensen, C. R. Levi, P. M. Desmond, G. A. Donnan, S. M. Davis, and M. W. Parsons, "Cerebral blood flow is the optimal CT perfusion parameter for assessing infarct core," *Stroke*, vol. 42, no. 12, pp. 3435–3440, 2011.

[175] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout : A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[176] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A Matlab-like Environment for Machine Learning," *BigLearn, NIPS Work.*, pp. 1–6, 2011.

[177] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[178] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the Youden Index and its associated cutoff point," *Biometrical J.*, vol. 47, no. 4, pp. 458–472, 2005.

[179] K. C. Ho, W. Speier, S. El-Saden, D. S. Liebeskind, J. L. Saver, A. A. T. Bui, and C. W. Arnold, "Predicting Discharge Mortality after Acute Ischemic Stroke Using Balanced Data," in *AMIA Annual Symposium Proceedings*, 2014, vol. 2014, p. 1787.

[180] Developers.google.com, "Classification: ROC and AUC," 2018. .

[181] C. D. Manning, H. Schütze, and others, *Foundations of statistical natural language processing*, vol. 999. MIT Press, 1999.

[182] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases.," *Radiology*, vol. 148, no. 3, pp. 839–43, 1983.

[183]   J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*. Springer, 2011.

[184]   C. Stinear, "Prediction of recovery of motor function after stroke," *Lancet Neurol.*, vol. 9, no. 12, pp. 1228–1232, 2010.

[185]   L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[186]   L. Willats, A. Connelly, S. Christensen, G. A. Donnan, S. M. Davis, and F. Calamante, "The role of bolus delay and dispersion in predictor models for stroke," *Stroke*, vol. 43, no. 4, pp. 1025–1031, 2012.

[187]   B. Friedrich, O. Kertels, D. Bach, S. Wunderlich, C. Zimmer, S. Prothmann, and A. Förschler, "Fate of the penumbra after mechanical thrombectomy," *Am. J. Neuroradiol.*, vol. 35, no. 5, pp. 972–977, 2014.

[188]   A. Krizhevsky, Ii. Sulskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.*, pp. 1–9, 2012.

[189]   G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *Comput. Vis. pattern Recognit.*, pp. 2261–2269, 2017.

[190]   J. Mackey, D. Kleindorfer, H. Sucharew, C. J. Moomaw, B. M. Kissela, K. Alwell, M. L. Flaherty, D. Woo, P. Khatri, O. Adeoye, S. Ferioli, J. C. Khoury, R. Hornung, and J. P. Broderick, "Population-based study of wake-up strokes," *Neurology*, vol. 76, no. 19, pp. 1662–1667, 2011.

[191]   J. M. Reid, D. Dai, B. Cheripelli, C. Christian, Y. Reidy, G. J. Gubitz, and S. J. Phillips, "Differences in wake-up and unknown onset stroke examined in a stroke registry.," *Int. J. Stroke*, vol. 10, no. 3, pp. 331–5, Apr. 2015.

[192]   Y. Moradiya and N. Janjua, "Presentation and Outcomes of '" Wake-Up Strokes "' in a Large Randomized Stroke Trial : Analysis of Data from the International Stroke Trial," vol. 22, no. 8, pp. 286–292, 2013.

[193]   M. B. Maas and A. B. Singhal, "Unwitnessed stroke: impact of different onset times on eligibility into stroke trials.," *J. Stroke Cerebrovasc. Dis.*, vol. 22, no. 3, pp. 241–3, Apr. 2013.

[194]   K. C. Ho, W. Speier, S. EL-Saden, and W. C. Arnold, "Classifying Acute Ischemic Stroke Onset Time using Deep Imaging Features," in *AMIA Annual Symposium Proceedings*, 2017.

[195]   T. Messay, R. C. Hardie, and S. K. Rogers, "A new computationally efficient CAD system for pulmonary nodule detection in CT imagery," *Med. Image Anal.*, vol. 14, no. 3, pp. 390–406, 2010.

[196]   C. Frindel, A. Rouanet, M. Giacalone, and T. Cho, "Validity of Shape as a Predictive Biomarker of Final Infarct Volume in Acute Ischemic Stroke," *Stroke*, vol. 46, pp. 976–981, 2015.

[197]   A. Krizhevsky, I. Sutskever, G. E. G. E. Hinton, Ii. Sulskever, and G. E. G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[198] V. Bewick, L. Cheek, and J. Ball, "Statistics review 14: Logistic regression," *Crit. care*, vol. 9, no. 1, p. 112, 2005.

[199] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.*, pp. 1189–1232, 2001.

[200] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[201] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 2014.

[202] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Int. Conf. Learn. Represent.*, pp. 1–14, 2015.

[203] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS Workshop*, 2017.

[204] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and others, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[205] T. Chen and C. Guestrin, "XGBoost : Reliable Large-scale Tree Boosting System," *arXiv*, pp. 1–6, 2016.

[206] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models."

[207] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding Neural Networks Through Deep Visualization," *Int. Conf. Mach. Learn. - Deep Learn. Work. 2015*, p. 12, 2015.

[208] R. Grech, P. L. Galvin, S. Power, A. O'Hare, S. Looby, P. Brennan, and J. Thornton, "Outcome prediction in acute stroke patients considered for endovascular treatment: a novel tool," *Interv Neuroradiol.*, vol. 20, no. 1591–0199 (Print), pp. 312–324, 2014.

[209] M. E. Mayerhoefer, P. Szomolanyi, D. Jirak, A. Materka, and S. Trattnig, "Effects of MRI acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: an application-oriented study," *Med. Phys.*, vol. 36, no. 4, pp. 1236–1243, 2009.

[210] J. A. Hanley and K. O. Hajian-Tilaki, "Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update," *Acad. Radiol.*, vol. 4, no. 1, pp. 49–58, 1997.

[211] L. Lu, H. Shin, H. R. Roth, M. Gao, L. Lu, S. Member, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer Learning Deep Convolutional Neural Networks for Computer-Aided Detection : CNN Architectures , Dataset Characteristics and Transfer," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[212] B. D. Murphy, X. Chen, and T.-Y. Lee, "Serial changes in CT cerebral blood volume and

flow after 4 hours of middle cerebral occlusion in an animal model of embolic cerebral ischemia," *Am. J. Neuroradiol.*, vol. 28, no. 4, pp. 743–749, 2007.

[213] D. D. Mcleod, M. W. Parsons, C. R. Levi, S. Beautement, D. Buxton, B. Roworth, and N. J. Spratt, "Establishing a rodent stroke perfusion computed tomography model," *Int. J. Stroke*, vol. 6, no. 4, pp. 284–289, 2011.

[214] K. Kudo, "Accuracy and Reliability Assessment of CT and MR Perfusion Analysis Software," vol. 267, no. 1, pp. 201–211, 2013.

[215] F. Zhu, T. Carpenter, D. Rodriguez Gonzalez, M. Atkinson, and J. Wardlaw, "Computed tomography perfusion imaging denoising using gaussian process regression," *Phys Med Biol*, vol. 57, no. 12, pp. N183-98, 2012.

[216] K. C. Ho, F. Scalzo, K. Sarma, S. El-Saden, A. Bui, and C. Arnold, "A Novel Bi-Input Convolutional Neural Network for Deconvolution-Free Estimation of Stroke MR Perfusion Parameters," in *2016 Annual Meeting of the Radiological Society of North America*, 2016.

[217] K. C. Ho, F. Scalzo, V. K. Sarma, S. EL-Saden, and W. C. Arnold, "A Temporal Deep Learning Approach for MR Perfusion Parameter Estimation in Stroke," in *International Conference of Pattern Recognition*, 2016.

[218] K. Kudo, S. Christensen, M. Sasaki, L. Østergaard, H. Shirato, K. Ogasawara, M. Wintermark, and S. Warach, "Accuracy and reliability assessment of CT and MR perfusion analysis software using a digital phantom," *Radiology*, vol. 267, no. 1, pp. 201–211, 2013.

[219] L. Xu, J. S. J. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Advances in Neural Information Processing Systems*, 2014, pp. 1790–1798.

[220] A. M. Smith, C. B. Grandin, T. Duprez, F. Mataigne, and G. Cosnard, "Whole brain quantitative CBF, CBV, and MTT measurements using MRI bolus tracking: implementation and application to data acquired from hyperacute stroke patients," *J. Magn. Reson. Imaging*, vol. 12, no. 3, pp. 400–410, 2000.

[221] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. C. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index 1: Scientific reports," *Acad. Radiol.*, vol. 11, no. 2, pp. 178–189, 2004.

[222] C. R. Gomez, "Time Is Brain: The Stroke Theory of Relativity," *J. Stroke Cerebrovasc. Dis.*, 2018.

[223] A. Yoo, E. Barak, W. Copen, S. Kamalian, L. G.- Stroke, and undefined 2010, "Combining acute diffusion-weighted imaging and mean transmit time lesion volumes with National Institutes of Health Stroke Scale Score improves the," *Am Hear. Assoc*.

[224] G. C. Fonarow, W. Pan, J. L. Saver, E. E. Smith, M. J. Reeves, J. P. Broderick, D. O. Kleindorfer, R. L. Sacco, D. M. Olson, A. F. Hernandez, E. D. Peterson, and L. H. Schwamm, "Comparison of 30-day mortality models for profiling hospital performance in acute ischemic stroke with vs without adjustment for stroke severity," *Jama*, vol. 308, pp. 257–264, 2012.

[225] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[226] M. B. Sesen, T. Kadir, R.-B. Alcantara, J. Fox, and M. Brady, "Survival prediction and treatment recommendation with Bayesian techniques in lung cancer," in *AMIA Annual Symposium Proceedings*, 2012, vol. 2012, p. 838.

[227] E. Gumus, N. Kilic, A. Sertbas, and O. N. Ucan, "Evaluation of face recognition techniques using PCA, wavelets and SVM," *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6404–6408, 2010.

[228] H. Kosorus, J. Honigl, and J. Kung, "Using r, weka and rapidminer in time series analysis of sensor data for structural health monitoring," in *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on*, 2011, pp. 306–310.

[229] D. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[230] D. A. Salazar, J. I. Vélez, and J. C. Salazar, "Comparison between SVM and logistic regression: Which one is better to discriminate?," *Rev. Colomb. Estad{\'\i}stica*, vol. 35, no. SPE2, pp. 223–237, 2012.

[231] T. O'Malley, P. Langhorne, Elton R.A., and S. C, "Platelet Size in Stroke Patients," *Stroke*, vol. 26, pp. 995–999, 1995.

[232] J. Alvarez-Sabín, C. A. Molina, M. Ribó, J. F. Arenillas, J. Montaner, R. Huertas, E. Santamarina, and M. Rubiera, "Impact of admission hyperglycemia on stroke outcome after thrombolysis: Risk stratification in relation to time to reperfusion," *Stroke*, vol. 35, no. 11, pp. 2493–2498, 2004.

[233] A. Bruno, N. Shah, C. Lin, B. Close, D. C. Hess, K. Davis, V. Baute, J. A. Switzer, J. L. Waller, and F. T. Nichols, "Improving modified rankin scale assessment with a simplified questionnaire," *Stroke*, vol. 41, no. 5, pp. 1048–1050, 2010.

[234] J. Astrup, B. K. Siesjö, and L. Symon, "Thresholds in cerebral ischemia-the ischemic penumbra.," *Stroke*, vol. 12, no. 6, pp. 723–725, 1981.

[235] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Rob. Res.*, vol. 34, no. 4–5, pp. 705–724, 2015.

[236] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European Conference on Computer Vision*, 2014, pp. 94–108.

[237] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and <1MB Model Size," *arXiv*, pp. 1–5, 2016.

[238] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[239] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," 2017.

[240] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS One*, vol. 10, no. 7, pp. 1–46, 2015.

[241] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *Iclr*, p. 1-, 2014.

[242] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Comput. Vis. Image Underst.*, vol. 106, no. 1, pp. 59–70, 2007.