Confocal Laser Endomicroscopy Image Analysis

with Deep Convolutional Neural Networks

by

Mohammadhassan Izady Yazdanabadi

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2019 by the
Graduate Supervisory Committee:

Mark C. Preul, Co-Chair
Yezhou Yang, Co-Chair
Peter Nakaji
Brent Vernon

ARIZONA STATE UNIVERSITY

May 2019

ABSTRACT

Rapid intraoperative diagnosis of brain tumors is of great importance for planning treatment and guiding the surgeon about the extent of resection. Currently, the standard for the preliminary intraoperative tissue analysis is frozen section biopsy that has major limitations such as tissue freezing and cutting artifacts, sampling errors, lack of immediate interaction between the pathologist and the surgeon, and time consuming.

Handheld, portable confocal laser endomicroscopy (CLE) is being explored in neurosurgery for its ability to image histopathological features of tissue at cellular resolution in real time during brain tumor surgery. Over the course of examination of the surgical tumor resection, hundreds to thousands of images may be collected. The high number of images requires significant time and storage load for subsequent reviewing, which motivated several research groups to employ deep convolutional neural networks (DCNNs) to improve its utility during surgery. DCNNs have proven to be useful in natural and medical image analysis tasks such as classification, object detection, and image segmentation.

This thesis proposes using DCNNs for analyzing CLE images of brain tumors. Particularly, it explores the practicality of DCNNs in three main tasks. First, off-the shelf DCNNs were used to classify images into diagnostic and non-diagnostic. Further experiments showed that both ensemble modeling and transfer learning improved the classifier's accuracy in evaluating the diagnostic quality of new images at test stage. Second, a weakly-supervised learning pipeline was developed for localizing key features of diagnostic CLE images from gliomas. Third, image style transfer was used to improve the diagnostic quality of CLE images from glioma tumors by transforming the histology patterns in CLE images of fluorescein sodium-stained tissue into the ones in conventional hematoxylin and eosin-stained tissue slides.

These studies suggest that DCNNs are opted for analysis of CLE images. They may assist surgeons in sorting out the non-diagnostic images, highlighting the key regions and enhancing their appearance through pattern transformation in real time. With recent advances in deep learning such as generative adversarial networks and semi-supervised learning, new research directions need to be followed to discover more promises of DCNNs in CLE image analysis.

DEDICATION

To my loving mom and dad who taught me to empathize with those suffering, be curios to find the cause, and work hard toward resolving it.

To my wife who always encouraged me to believe in myself and think big.

To my mentors Dr. Mark Preul and Dr. Yezhou Yang for their support and supervision.

ACKNOWLEDGMENTS

I would like to thank my mentors at Barrow Neurological Institute, Drs. Mark Preul and Peter Nakaji, whose advices were a true source of inspiration for my research and were always supportive. Thank you Dr. Preul for allowing me to be a member of the neurosurgery research lab and thank you Dr. Nakaji for allowing me to observe brain tumor surgeries and understand the treatment process better.

I would like to thank my computer scientist advisor Dr. Yezhou Yang for his fresh and intuitive ideas in my research. Thank you Dr. Yang for introducing me to the active perception lab which was a positive environment, full of smart enthusiastic members.

I would like to thank Dr. Jennifer Eschbacher for accepting that I do research on confocal microscopy in the pathology lab. Thank you Dr. Eschbacher for taking the time to explain to me what I didn’t know despite your busy schedule.

I would like to thank the previous co-director of neuroscience program at Barrow, Dr. Eric Vu, who introduced me to the astounding research labs on brain tumors.

I would like to thank Drs. Evgenii Belykh, Nikolay Martirosyan, and Joseph Georges for providing the CLE data and their collaboration. Further, I am thankful to the following colleagues for their contribution in my research: Drs. Michael Mooney, Xiaochun Zhao, Sirin Gandhi, Leandro Moreira, and Claduio Cavallo.

I would like to thank Drs. Jianming Liang and Nima Tajbakhsh for their valuable directions about recent technologies in medical image analysis. I am thankful to all my lab mates and coworkers at ASU and BNI, to my dearest friends and family members who supported me in the last five years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

INTRODUCTION

This dissertation is focused on deep convolutional neural networks (DCNN) – an advanced technology with immense success in computer vision – and its applications to analyze confocal laser endomicroscopy (CLE) images from brain tumors. Intricacies involved in examining CLE images due to their unique appearance, high noise, blood and motion artifacts, and the enormous number of images acquired during each surgery, mandates a computer-aided system to support their analysis. This aid could be in the form of selecting important frames (e.g. diagnostic classification), highlighting interesting areas (e.g. segmentation, feature localization) or quality enhancement (e.g. noise removal).

The rest of this thesis is organized as follows. Chapter 1 reviews the current studies using DCNN and other quantitative methods for analyzing CLE data from tumors in the brain and other organs. Chapter 2 describes an ensemble model developed for diagnostic classification. This model can give a diagnostic score for a given CLE image based on the previous examples provided during training process. Chapter 3 presents a novel architecture to localize diagnostic features in CLE images from glioma tumors. The novelty of this study is that despite supervised learning approaches that utilize pixel-level annotated datasets, it learns the features from an image-level annotated dataset (which is easier to acquire). Chapter 4 studies how DCNN based style transfer may transform CLE images to reduce the noise and make the histopathological features of gliomas more identifiable. Quality assessment of modified images (performed by neurosurgeons) confirmed improvement through both removing artifacts and amplifying critical structures. Chapter 5 summarizes all the findings and discusses limitations and future directions for this research.

CHAPTER 1

The following chapter has been published in the Journal of Frontiers in Oncology,

Theranostic Imaging in Cancer Precision Medicine.

(DOI: 10.3389/fonc.2018.00240)

CHAPTER 1

PROSPECTS FOR THERANOSTICS IN NEUROSURGICAL IMAGING: EMPOWERING CONFOCAL LASER ENDOMICROSCOPY DIAGNOSTICS VIA DEEP LEARNING

Izadyyazdanabadi M, Belykh E, Mooney M, Eschbacher J, Nakaji P, Yang Y, Preul MC

ABSTRACT

Confocal laser endomicroscopy (CLE) is an advanced optical fluorescence imaging technology that has potential to increase intraoperative precision, extend resection, and tailor surgery for malignant invasive brain tumors because of its subcellular dimension resolution. Despite its promising diagnostic potential, interpreting the gray tone fluorescence images can be difficult for untrained users. CLE images can be distorted by motion artifacts, fluorescence signals out of detector dynamic range, or may be obscured by red blood cells, and thus interpreted as nondiagnostic. However, just a single CLE image with a detectable pathognomonic histological tissue signature can suffice for intraoperative diagnosis. Dealing with the abundance of images from CLE is not unlike sifting through a myriad of genes, proteins, or other structural or metabolic markers to find something of commonality or uniqueness in cancer that might indicate a potential treatment scheme or target. In this review we provide a detailed description of bioinformatical analysis methodology of CLE images that begins to assist the neurosurgeon and pathologist to rapidly connect on-the-fly intraoperative imaging, pathology, and surgical observation into a conclusionary system within the concept of theranostics. We present an overview and discuss deep learning models for automatic detection of the diagnostic CLE images and discuss various training regimes and ensemble modeling effect on power of deep learning predictive models. Two major approaches reviewed in this paper include the models that can automatically classify CLE images into diagnostic/nondiagnostic, glioma/nonglioma, tumor/injury/normal categories and models that can localize histological features on the CLE images using weakly supervised methods. We also briefly review advances in the deep learning approaches used for CLE image analysis in other organs. Significant advances in speed and precision of automated diagnostic frame selection would augment the diagnostic potential of CLE, improve operative workflow and integration into brain tumor surgery. Such technology and bioinformatics analytics lend themselves to improved precision, personalization, and theranostics in brain tumor treatment.

Introduction

According to the American Cancer Society (American Cancer Society, 2018), in 2018 nearly 24,000 patients will be diagnosed with brain or other nervous system cancer and about 17,000 patients will die of the disease. Gliomas represent about 25% of all primary brain tumors and about 80% of all malignant tumors of the central nervous system (Ostrom et al., 2015). Over half of gliomas are glioblastoma multiforme (GBM), which is the most malignant primary brain tumor. GBMs are infiltrative and normally lack a clear margin making complete resection nearly impossible. Maximal resection of gliomas has been associated with improved prognosis (Almeida, Chaichana, Rincon-Torroella, & Quinones-Hinojosa, 2015; Sanai, Polley, McDermott, Parsa, & Berger, 2011), although invasion and the bounds of functional cortex often limit extensive removal. Currently, technology for extending the limits of the tumor resection relies on intraoperative image-guided surgical navigation platforms, intraoperative magnetic resonance imaging (MRI), and intraoperative ultrasound (Sanai & Berger, 2018). Wide-field fluorescence illumination through the operative microscope has been utilized more recently in an attempt to identify the margins of infiltrating tumors (Maugeri et al., 2018).

Regardless of the means for identifying the tumor margin, examining tissue samples during surgery is paramount, especially for neurosurgery. Rapid intraoperative assessment of tumor tissue remains key for planning the treatment and for guiding the surgeon to areas of suspected tumor tissue during the operation, or planning adjunct intraoperative or post-operative therapy. The standard for the preliminary intraoperative histopathological interpretation is frozen section biopsy. However, the frozen section biopsy method has inherent complications such as sampling error, tissue freezing and cutting artifacts, lack of immediate pathologist interactivity with the surgeon, time spent for tissue delivery, processing, and analysis reporting back to the operating room (Martirosyan et al., 2014; Tofte, Berger, Torp, & Solheim, 2014).

Handheld (i.e., size of a pen), portable confocal laser endomicroscopy (CLE) is undergoing exploration in brain tumor surgery because of its ability to produce precise histopathological information of tissue with subcellular resolution *in vivo* in real time during tumor resection (Belykh et al., 2016; Charalampaki et al., 2015; Foersch et al., 2012; Martirosyan et al., 2014, 2016; Sanai, Eschbacher, et al., 2011). CLE is a fluorescence imaging technology that is used with a combination of fluorescent drugs or probes. While a wide range of fluorophores have been used for CLE in gastroenterology and other medical specialties, fluorophore options are limited for in vivo human brain use due to potential toxicity (Belykh et al., 2016; Foersch et al., 2012; Martirosyan et al., 2014; Zehri et al., 2014). Fluorescent dyes currently approved for use *in vivo* in the human brain include fluorescein sodium (FNa), indocyanine green (ICG), and 5-aminolevulinic acid (5-ALA) (Belykh et al., 2016; Liu, Meza, & Sanai, 2014; Mooney, Zehri, Georges, & Nakaji, 2014). Other fluorescent dyes such as acridine orange (AO), acriflavine (AF), cresyl violet, etc., can be used on human brain tissue *ex vivo* (Martirosyan et al., 2016, 2018). In neurosurgical oncology, CLE has been used to rapidly obtain optical cellular and cytoarchitectural information about tumor tissue as the resection progresses and to interrogate the resection cavity (Martirosyan et al., 2016; Sanai, Eschbacher, et al., 2011). The details of system operation have been previously described in detail (J. Eschbacher et al., 2012; Martirosyan et al., 2014, 2016; Sanai, Eschbacher, et al., 2011). Briefly, the neurosurgeon may hold the CLE probe by the hand, fixate it with a flexible instrument holder in place, or may glide the probe across the tissue surface to obtain an "optical biopsy" with an image acquisition speed ranging between 0.8 and 20 frames per second dependent on operation of the particular CLE system. The surgeon may place the probe in a resting position at any time and proceed with the tumor resection, then take up the probe conveniently as desired. CLE imaging is believed to be potentially advantageous for appraisal of tumor margin regions or to examine suspected invasion into functional cortex near the final phases of tumor resection. The images display on a touchscreen monitor attached to the system. The neurosurgeon uses a foot pedal module to control depth of scanning and image acquisition. An assistant can also control the acquisition of images using a touchscreen. CLE images can be processed and presented as still images, digital video loops showing motion, or 3-dimensional digital imaging volumes. CLE is a promising technology with the strategy to optimize or maximally increase the resection of malignant infiltrating brain tumors and/or to increase the positive yield of tissue biopsy. CLE may be of especial value during surgery when interrogating tissue at the tumor border regions or within the surgical resection bed that may harbor remnant malignant or spreading tumor.

Cancer is the subject of intense investigation into how theranostics may improve care and survival. As oncology is continually refined in its quest to understand and treat malignant brain tumors, such as GBMs, with which it has had very little success, utilization of precision and personalized surgical techniques would seem to be a logical step forward, especially as tumor resection is usually the first definitive treatment step. Dealing with the abundance of images from CLE is not unlike sifting through a myriad of genes, proteins, or other structural or metabolic markers to find something of commonality or uniqueness in cancer that might indicate a potential treatment scheme or target. CLE data acquisition is vast, burdened with a near-overwhelming number of images, many of which appear not useful at first inspection, although they may have unrecognized informative image subregions or characteristics. Mathematical algorithms and computer-based technology may rapidly assist making decisions upon an incredible number of images, such as CLE produces, that has never been encountered in neurosurgery.

Critical success in theranostics relies on the analytical method. Finding meaning can be elusive, and what may seem at first meaningful may only be superficial or even a spurious result, thus the analytical methodology is critical. In this review we provide a detailed description of bioinformatical analysis methodology of CLE images that begins to assist the neurosurgeon and pathologist to rapidly connect on-the-fly intraoperative imaging, pathology, and surgical observation into a conclusionary system within the concept of theranostics. We describe methodology of deep convolutional neural networks (DCNNs) applied to CLE imaging focusing on neurosurgical application and review current modeling outcomes, elaborating and discussing studies aiming to suggest a more precise and tailored surgical approach and workflow for brain tumor surgery.

Demanding imaging information load of confocal laser endomicroscopy

Although the number of non-diagnostic CLE images has been shown to be high, the first diagnostic frames were acquired at an average after the 14th frame (about 17 sec) *in vivo* (Martirosyan et al., 2016). This is certainly faster than an intraoperative frozen section biopsy preparation and diagnostic interpretation. Nevertheless, the high number of non-diagnostic images imposed a significant time requirement and image storage load for subsequent image reviews, leading us and other groups to employ deep learning algorithms and neural networks that could potentially sort out non-diagnostic frames, while retaining only the diagnostic ones (M. Izadyyazdanabadi et al., 2017; Mohammadhassan Izadyyazdanabadi, Belykh, Mooney, et al., 2018). Attempts to use advanced feature coding schemes to classify cellular CLE images of brain tumor samples stained *ex vivo* with AF have been reported (Kamen et al., 2016). Advantageously, acquired CLE images may be exchanged and translated for off-site digital histopathology review. However, large amounts of data may create an information overload that requires novel solutions for data CLE management and storage.

While CLE has obvious benefits of rapid on-the-fly digital imaging of tissue that can obviate long wait times for tissue interpretation and be quickly communicated between surgeons and pathologists, there are challenges to manage the amount of information provided. Current CLE systems can generate hundreds to thousands of images over the course of examination of the tumor or resection cavity which may take only a few minutes. It has been estimated that since CLE technology was put into use in 2011 for gastrointestinal diagnosis, over 100 million images have been created, with 30 million images created in the past year (Loiseau, 2017). The number of images may become rapidly overwhelming for the neurosurgeon and neuropathologist when trying to review and select a diagnostic or meaningful image or group of images as the surgical inspection progresses. CLE is designed to be used in real time while the surgeon operates on the brain, but overcoming the barriers of image selection for diagnosis is a key component for making CLE a practical and advantageous technology for the neurosurgical operating room.

Other barriers for revealing underlying meaningful histology are motion and blood artifacts (especially) that are present in some of the CLE images, especially for CLE systems functioning in the blue laser range versus near-infrared (Belykh et al., 2016; Martirosyan et al., 2011, 2014, 2016; Sankar et al., 2010). In addition, the neuropathologist must begin to work in a world of fluorescence images showing shades of gray, black and white or artificial colorization, where before natural colored stains existed. The display of suboptimal nondiagnostic frames interferes with the selection of and focus upon diagnostic images by the neurosurgeon and pathologist throughout the surgery to make a correct intraoperative interpretation. A previous study of CLE in human brain tumor surgeries found that about half of the acquired images were interpreted as nonuseful (i.e., nondiagnostic) due to an inherent nature of the handheld microscopic probe with a narrow field of view that is subject to motion and blood artifacts or lack of discernible or characteristic features of the tissue itself (Martirosyan et al., 2016). These artifacts or inherent aspects of operation of the probe include unsteady hand movements, moving the probe while in imaging mode across the tissue surface, and irregularities of the tissue surface such as a tumor resection bed in the cortex that includes tissue crevices, surface irregularities, bleeding, movement of the cortex with arterial pressure and respiration, etc.

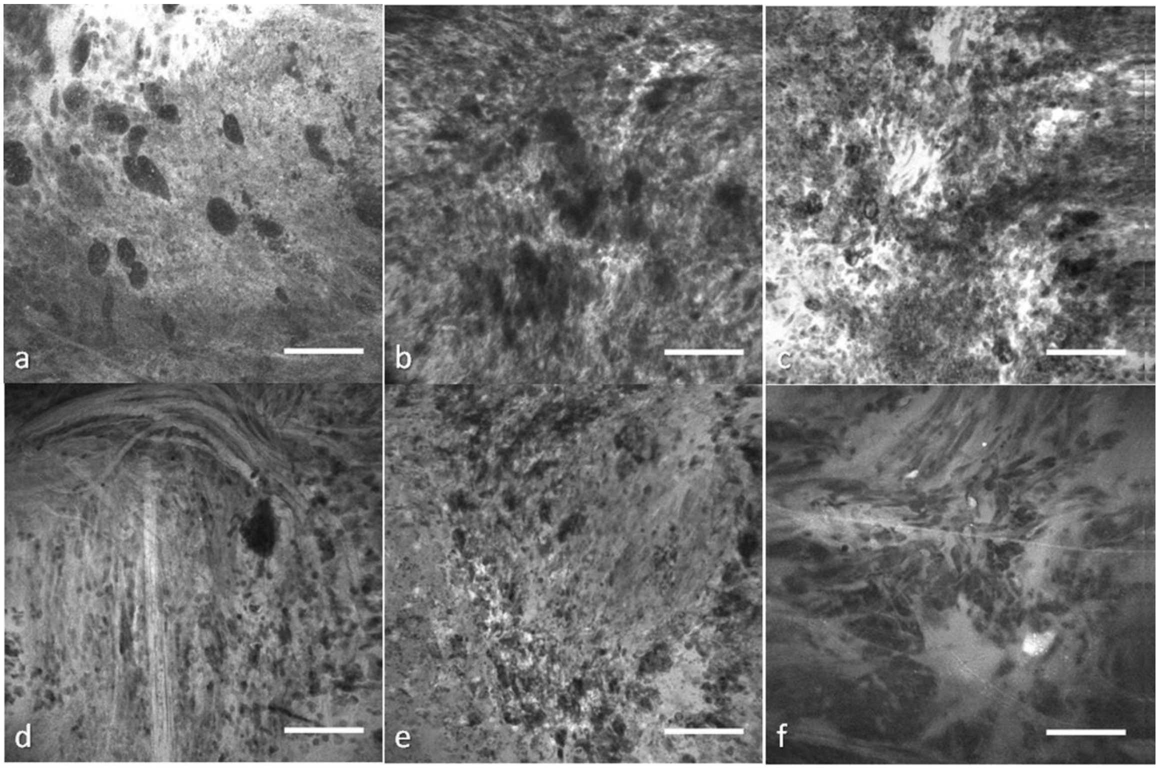


Figure 1.1

Representative Confocal Laser Endomicroscopy (CLE) images from glioma and meningioma acquired with Optiscan 5.1, Optiscan Pty., Ltd. (a) anaplastic oligodendroglioma, (b) recurrent astrocytoma, (c) glioblastoma multiforme (GBM), (d) fibrous meningioma (grade I), (e) chordoid meningioma (grade II), (f) atypical meningioma (grade II). Field of view = 475 × 475 μm, resolution = 1024 × 1024 pixels, bar = 100 µm. (Glioblastomas are a brain malignancy of astrocytic cell origin, show wild pleomorphism, proliferation of abnormal tumor-associated vasculature, necrosis, and vast brain invasion. Meningiomas arise from meningothelial cells and are usually attached to the dura. Although of a common origin, meningiomas have histological pattern subtypes and more aggressive types show atypical or anaplastic features. They do not display malignant brain infiltration.(Louis et al., 2007))

Thus, although imaging is acquired on-the-fly, an image discrimination system to optimally sift out and identify useful images would substantially improve the performance of the CLE. Manually filtering out the nondiagnostic images before making an intraoperative decision is challenging due to the large number of images acquired, the novel and frequently unfamiliar appearance of fluorescent stained tissue features compared to conventional histology. Interpretation of fluorescence CLE images for routine clinical pathology has only recently been trained and studied. Great variability among images from the same tumor type, and potential similarity between images from other tumor types for the untrained interpreter (Figure 1.1) make simple image filters and thresholding unreliable, thus requiring advanced computational methods.

Theranostics and confocal laser endomicroscopy

Investigations into cancer genetics have produced treatment pathways by the application of bioinformatics methods leading to the concept of theranostics. As advances in molecular science have enabled “fingerprinting” of individual tumor with genomic and proteomic profiling, personalized theranostic agents can be developed to target specific tumor microenvironment compartments (Stasinopoulos et al., 2011). Although theranostic imaging provides new opportunities for personalized cancer treatment through the interface of chemistry, molecular biology, and imaging, quantitative image analysis remains as one of its challenges (M. F. Penet, Chen, Kakkad, Pomper, & Bhujwalla, 2012). More sophisticated image analysis methods are required to visualize and target every aspect of the tumor microenvironment in combination with molecular agents (M.-F. Penet, Krishnamachary, Chen, Jin, & Bhujwalla, 2014). CLE technology potentially allows a more personalized, precise, or tailored approach to the surgical procedure to remove an invasive brain tumor because of its capability to image at cell resolution intraoperatively on-the-fly. Fluorescent stains or markers allow the imaging and potential targeting of cells – nearly we are at the surgery of the "cell" as specific stains or fluorescent markers are developed. Whether, this technology has practicality or yields survival benefit for malignant invasive brain tumors awaits the results of the first substantial *in vivo* explorations.

Deep learning application in CLE brain tumor diagnostics

Deep convolutional neural networks (DCNNs) are a subset of "deep learning" technology, a machine learning subfield that has achieved immense recognition in the field of medical image analysis. Advances in computer-aided detection (CADe) and diagnosis (CADx) systems in ultrasound, magnetic resonance imaging, and computed tomography (CT) have been reviewed previously (Greenspan, van Ginneken, & Summers, 2016). There have been only a few studies that investigated deep learning application to enhance the diagnostic utility of CLE imaging in brain tumors. The utilization of deep learning approaches is mainly focused around three goals: diagnostic image detection, tumor classification, and feature localization originating from image segmentation. Here we provide an overview of the basics of a deep learning methodology applicable to CLE images, summarize current results of brain tumor CLE image analysis using DCNNs, and juxtapose these with related works in other cancers.

A deep convolutional neural network (DCNN) consists of several layers, each having multiple units called feature maps (Yann LeCun, Bengio, & Hinton, 2015). The first layer includes the input images that will be analyzed. To produce the second layer's feature maps, each pixel of layer 2 is connected to a local patch of pixels in layer 1 through a filter bank followed by an activation function, which is usually a rectified linear unit (ReLU) because of its fast-to-compute property compared to other functions (Glorot, Bordes, & Bengio, 2011). Model parameters are learned by minimizing the loss (the error between model prediction and the ground truth) using an optimization algorithm in two steps: forward and backward propagations (Le Cun Jackel, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, Cun, Denker, & Henderson, 1990; Rumelhart, Hinton, & Williams, 1986). To adjust the weights of filter banks, after each iteration of forward propagating the network, the derivatives of the loss function with respect to different weights are calculated to update the weights in a backward propagation (Le Cun Jackel, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard et al., 1990). A pooling layer accumulates the features in a smaller region by replacing windows of a feature map with their maximum or the average value. By stacking several convolutional, pooling, and fully connected layers, a DCNN can learn a hierarchy of visual representations to recognize class-specific features in images (Yann LeCun et al., 2015). Figure 1.2 shows an example network architecture and how the feature maps are calculated to perform diagnostic brain tumor image classification.

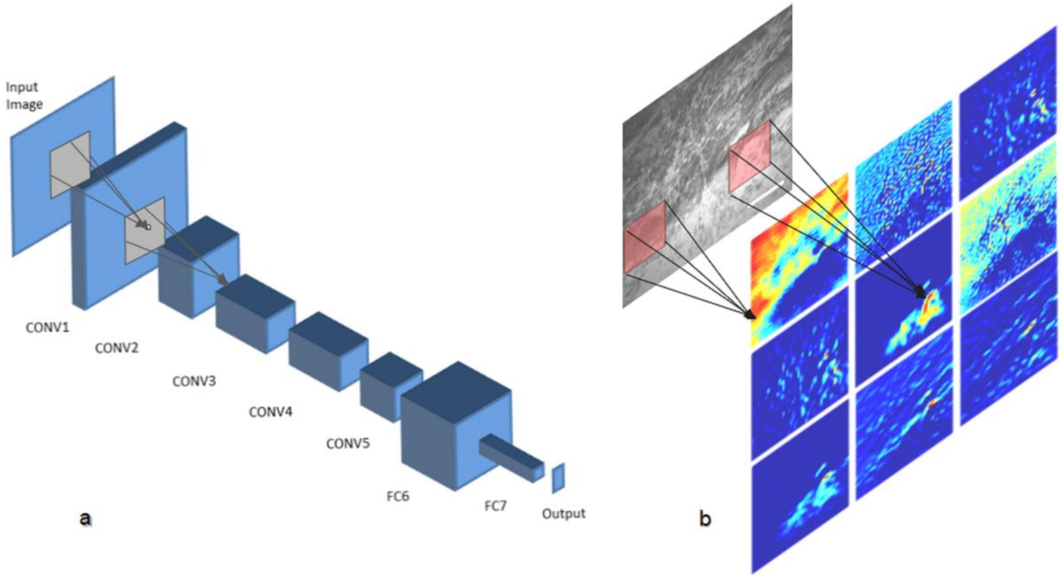


Figure 1.2

Deep convolutional neural network (DCNN) architecture. A schematic diagram of AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), a famous DCNN architecture that was trained on CLE images for diagnostic classification by Izadyyazdanabadi et al. (Mohammadhassan Izadyyazdanabadi, Belykh, Mooney, et al., 2018), is shown in (a). Different feature maps of the first convolutional layer (color images) were calculated by convolving different filters (red squares) with the corresponding regions of the input image (illustrated in (b)).

The validity of recommendations resulted from the DCNN analysis greatly depends upon the ground truth established by the expert professional. Unlike other conventional surgical tissue examination modalities like hematoxylin and eosin (H&E) stained histopathological slides, the CLE images are novel to neuropathologists and neurosurgeons. Since the beginning of CLE investigation in brain tumor surgery, the ground truth was established by surgical biopsy and subsequent standard histopathology analysis acquired from the same location as the CLE "optical biopsy" and correlating the features on CLE images to the histopathological sections. Neuropathologists and neurosurgeons at a few select centers are correlating CLE features to histopathology in order to establish an expertise in reading CLE images, and such investigations are ongoing (Charalampaki et al., 2015; J. Eschbacher et al., 2012; Martirosyan et al., 2016). The experience in CLE image interpretation is imperative for meaningful DCNN analysis. However, as described later, delving deeper into the DCNN analysis of the CLE while using ground truth established by the standard histopathology, results in identification of novel CLE features and allows many more images that may be termed suboptimal to in fact become useful. The improvement in workflow and diagnostics, and thus theranostics in CLE, will be dependent upon robust computer learning architecture.

Tumor classification

One of the first deep learning approaches for making a diagnosis of a brain tumor type based on the CLE images was a cascaded deep decision network (DDN), a type of DCNN (N. Murthy et al., 2017). A network was trained for classification of glioma and meningioma images using their previously proposed multi-stage DDN architecture (Murthy, Singh, Chen, Manmatha, & Comaniciu, 2016) for developing the model. The training process was as follows: LeNet, a relatively shallow CNN architecture initially proposed by LeCun, et al. (Y. LeCun et al., 1995) for handwritten digit recognition, was trained on the training dataset until it produced descent classification results on validation images. Then, the images were divided into two categories: *easy images* (classified correctly by the model with high confidence) and *challenging images* (classified either wrongly or even correctly yet with a low confidence). The challenging images were passed to the next stage for retraining. In the second stage, a convolutional stage and two fully connected layers and a softmax layer were stacked to the previous network and trained, while freezing the previous layers' parameters. After training the second stage on the challenging images from stage 1, the same process was repeated (finding confidence threshold, filtering the easy images, passing the challenging images to next stage, stacking the new layers to the previous network) until the model fails to improve on the validation dataset. After removing uninformative images using image entropy, a dataset was created of about 14,000 GBM and 12,000 meningioma images. The final proposed DDN could classify the GBM images with 86% accuracy while outperforming other methods such as SVM classifier applied on manual feature extraction, pre-trained networks, and shallow CNNs. (Murthy et al., 2016)

We have previously developed an architecture to classify CLE images from experimental brain gliomas into three classes (Belykh et al., 2018): tumor tissue, injured brain cortex tissue (no tumor) and normal brain cortical tissue. This study was undertaken to examine the ability of CLE image analysis to discriminate between tumor tissue, tissue subjected to the minor tissue trauma that surgical resection produces, and normal brain tissue. FNa may extravasate in the first two situations potentially causing surgeon confusion. This classification model was inspired by Inception, which is a DCNN for classifying generic images. (Szegedy et al., 2015) Due to the small size of our training dataset (663 diagnostic images selected from 1,130 images acquired), we used fine-tuning to train the model with a learning rate of 0.001. We used a nested left-out validation method to estimate the model performance on images from new biopsies. Images were divided into 3 data sets based on biopsy level: training (n=446), validation (n=217), and test set (n=40). Model performance increased to 88% when images were classified using 2 classes only (tumor tissue or non-tumor tissue) which was only slightly lower than the neuropathologists' mean accuracy (90%). The sensitivity and specificity of the model in discriminating a tumor region from non-tumor tissue were 78% and 100%, respectively. The Area Under the ROC Curve (AUC) value for tumor/non-tumor tissue classification was 93%. Subgroup analysis showed that the model could discriminate CLE images from tumor and injury with 85% accuracy (mean of accuracy for neuropathologists was 88%), 78% sensitivity and 100% specificity. We expect that performance of the model will be improved in terms of accuracy and speed by going forth from a small experimental data set to operation on large clinical data sets.

Diagnostic image classification

Entropy-based filtering is one of the simplest ways to filter out non-diagnostic CLE images. In a study by Kamen, et al. (Kamen et al., 2016), CLE images obtained from brain tumors were classified automatically. An entropy-based approach was used to remove the noninformative images from their dataset and two common brain tumors (meningioma and glioma) were differentiated using bag of words and other sparse coding methods. However, entropy might not be an ideal method since many nondiagnostic images have nearly as high entropy as diagnostic ones, as shown in Figure 1.3. Due to the large number of CLE images produced during surgery, importance of data pruning, as has been shown in a previous study (Belykh et al., 2018), and the incompetency of entropy method, we developed a deep learning model for reliable classification of images into diagnostic and non-diagnostic categories (Mohammadhassan Izadyyazdanabadi, Belykh, Mooney, et al., 2018).

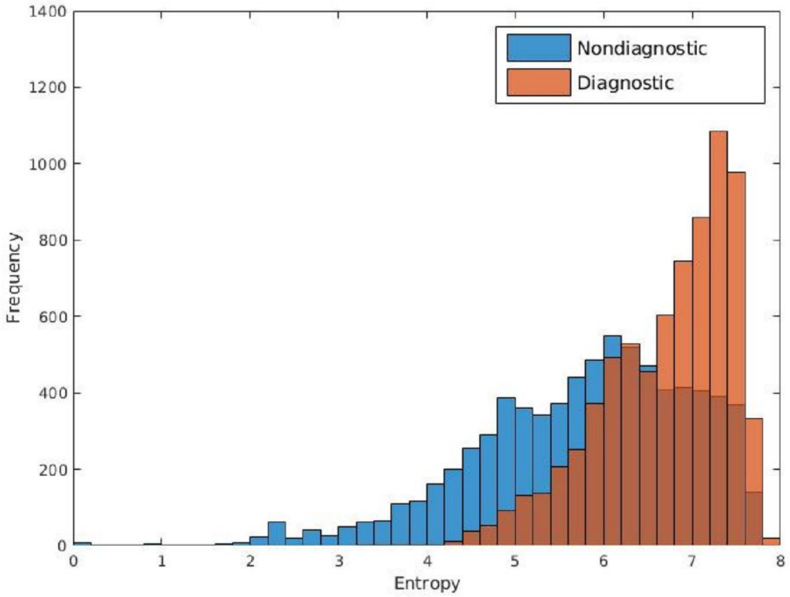


Figure 1.3

Entropy of diagnostic (orange) and nondiagnostic (blue) images. The overlap between the entropy of diagnostic and nondiagnostic CLE frames limits its feasibility for precise discrimination between the two classes(M. Izadyyazdanabadi et al., 2017).

A blinded neuropathologist and 2 neurosurgeons proficient with CLE image interpretation individually annotated all the images in our dataset. We developed single and ensemble models using two network architectures (Mohammadhassan Izadyyazdanabadi, Belykh, Mooney, et al., 2018). The ensemble of DCNN models for detecting diagnostic CLE images achieved 85% agreement with the ground truth. These results indicated that when only CLE images were provided, the model could detect the diagnostic CLE images with better agreement to the H&E-aided annotation. The example CLE images assessed with our diagnostic analysis model are presented on Figure 1.4. In order to compare the power of deep learning models with filtering approaches used in other related studies, we used entropy as a baseline (Kamen et al., 2016; N. Murthy et al., 2017). Subsequent evaluation of our test dataset of CLE images suggested that DCNN-based diagnostic evaluation has a higher agreement with the ground truth compared to the entropy-based quality assessment (Table 1.1).

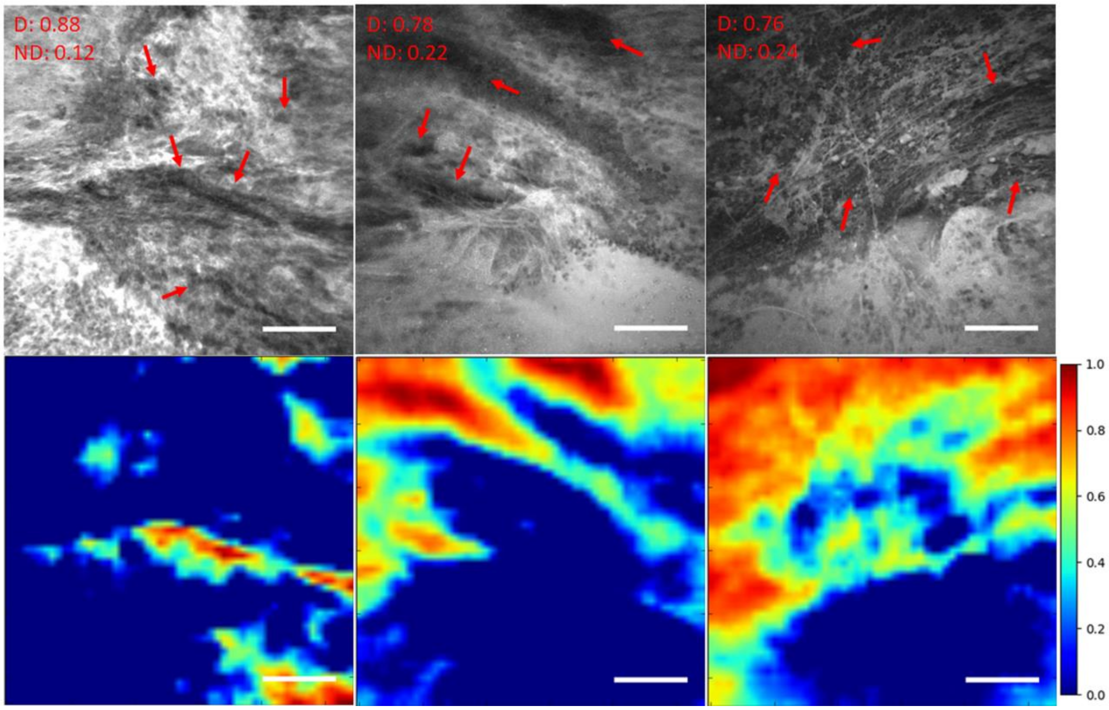


Figure 1.4

Unsupervised semantic localization of the CLE histopathological features(Mohammadhassan Izadyyazdanabadi, Belykh, Mooney, et al., 2018). First row displays the original CLE images, along with the probability of each image being diagnostic (D) and nondiagnostic (ND), estimated by the model. Red arrows mark the cellular regions recognized by a neurosurgeon. Second row shows the corresponding activation of neurons from the first layer (conv1, neuron 24) (shallow features learned by the model); it highlights some of the cellular areas (in warm colors) present in the images which were identified as diagnostic regions by the neurosurgeon reviewer. The color bars show the relative diagnostic value for each color: red marks the most diagnostic regions

(1.0) and blue marks the nondiagnostic regions (0.0). Field of view = 475×475 µm, resolution = 1024×1024 pixels, bar = 100 µm.

Model	Accuracy (%)	AUC
DCNN 1	78.8	0.87
DCNN 2	81.8	0.89
Entropy-based	57.2	0.71

Table 1.1

DCNN and entropy-based performance in diagnostic image classification(Mohammadhassan Izadyyazdanabadi, Belykh, Mooney, et al., 2018). DCNN methods showed higher agreement with the neurosurgeons’ evaluation. DCNN2(Szegedy et al., 2015) has a deeper architecture with fewer parameters than DCNN1(Krizhevsky et al., 2012).

Feature localization and image segmentation

Most of the current object localization studies in medical imaging use supervised learning that requires an annotated dataset for the training process. Physicians need to review the images and mark the location of interesting areas for each image, thus making it a costly and time-consuming process. Weakly supervised localization (WSL) methods have been proposed in computer vision to localize features using a weaker annotation, i.e., image-level labels instead of pixel-level labels.

We have previously investigated feature localization on brain tumor CLE images.(Mohammadhassan Izadyyazdanabadi, Belykh, Mooney, et al., 2018) Following the training and testing of the DCNN model for diagnostic image classification, 8 out of 384 reviewed colored neuron activation maps from the first layer of the model were selected for 4 diagnostic CLE images representative for glioma. Selected activation maps highlighted diagnostic tissue architecture patterns in warm colors. Particularly, selected maps emphasized regions of optimal image contrast, where hypercellular and abnormal nuclear features could be identified, and could serve as diagnostic features for image classification (Figure 1.4, bottom row). Additionally, a sliding window method was successfully applied to highlight diagnostic aggregates of abnormally large malignant glioma cells and atypically hypercellular areas (Mohammadhassan Izadyyazdanabadi, Belykh, Mooney, et al., 2018) (Figure 1.4). Such feature localization from the hidden layers makes the interpretation of the model results more illustrative and objective, especially from a clinical point of view where diagnosis cannot be made without sufficient evidence. Additionally, model-based feature localization can be performed considerably faster than human inspection and interpretation.

In another study, we applied a state of the art WSL approach (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2016) to localize glioma tumor features in CLE images (Mohammadhassan Izadyyazdanabadi, Belykh, Cavallo, et al., 2018). In this method, a global average pooling (GAP) layer was stacked to the convolutional layers of the network to create diagnostic feature maps. Representative localization and segmentation results are shown in Figures 1.5 and 1.6. A neurosurgeon with expertise in CLE imaging identified and highlighted the cellular areas in each CLE image (first column of both figures). By inserting the images and their labels (i.e., overall diagnostic quality:

diagnostic and nondiagnostic) to the network, the model automatically learns the primary diagnostic features of gliomas (e.g., cellular areas).

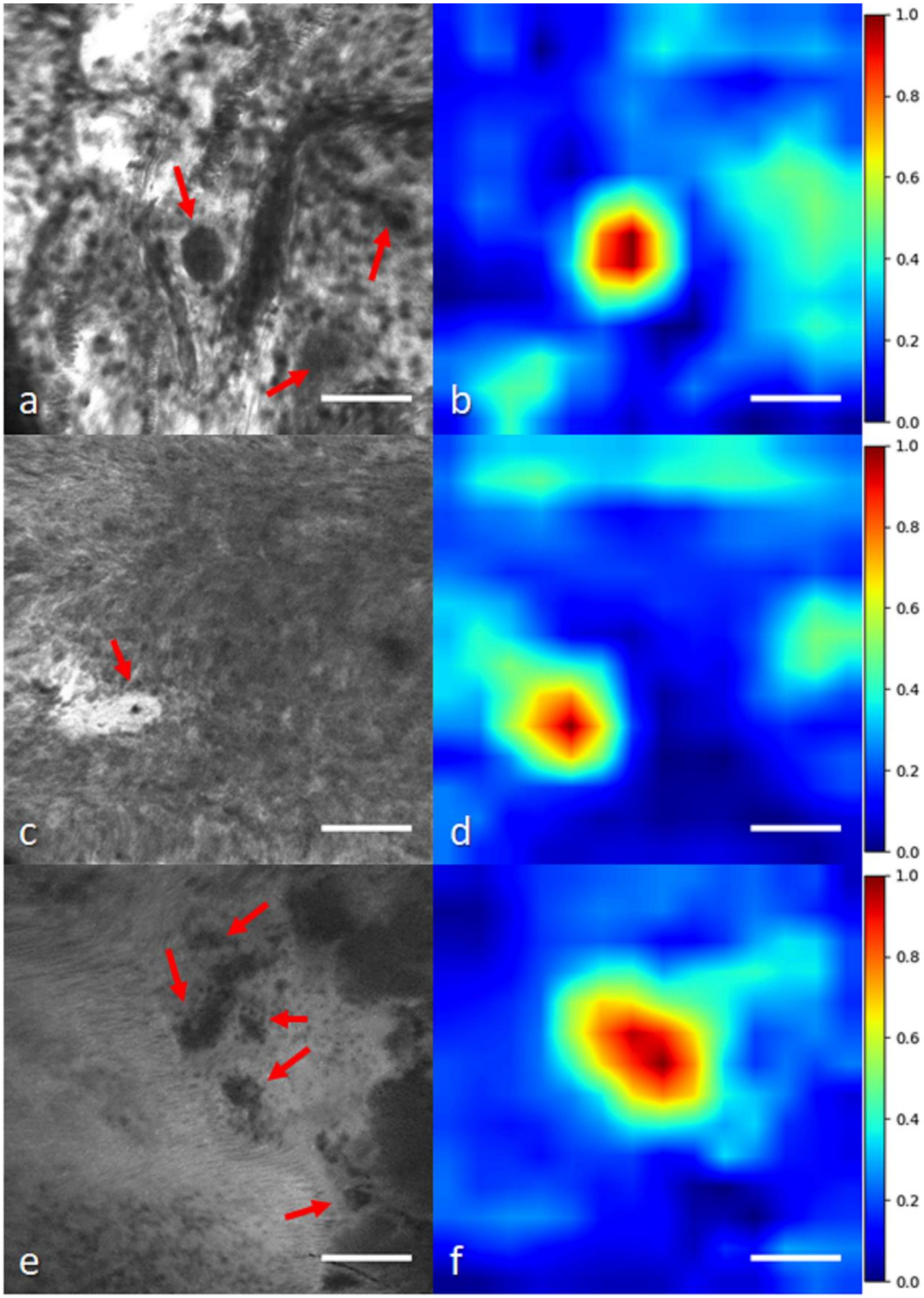

1.0
0.8
0.6
0.4
0.2
0.0
a
b
c
d
e
f

Figure 1.5

Histological glioma feature localization with a weakly supervised approach: global average pooling (GAP)(Mohammadhassan Izadyyazdanabadi, Belykh, Cavallo, et al., 2018). Left column shows CLE images from glioma cases ((a, c) recurrent infiltrating astrocytoma (e) oligodendroglioma). Red arrows mark the cellular regions recognized by a neurosurgeon. Second column shows the important regions detected with the model (highlighted in warm colors). The color bars show the relative diagnostic value for each color: red marks the most diagnostic regions (1.0) and blue marks the nondiagnostic regions (0.0). Field of view = 475×475 µm, resolution = 1024×1024 pixels, bar = 100 µm.

In Figure 1.5, the first column shows three CLE images along with the annotated diagnostic areas (red arrows) by a neurosurgeon, while the second column presents the diagnostic areas that the model highlighted with warm colors. The color bar near to each intensity map shows the relative diagnostic value for each color -- red marks the most diagnostic regions and blue marks the nondiagnostic regions. In Figure 1.6, after producing the diagnostic feature maps, each image was then segmented into diagnostic and nondiagnostic regions by thresholding (highlighted in green and purple); the recognized diagnostic regions correlated well with the neurosurgeon’s annotation. This method has two potential benefits: 1) improvement of the efficiency of glioma CLE imaging by recognizing the present diagnostic features and guiding the surgeon in tumor resection; and, 2) further investigation of the detected diagnostic regions may extend the physician’s perceptions about the glioma appearance and its phenotypes in CLE images.

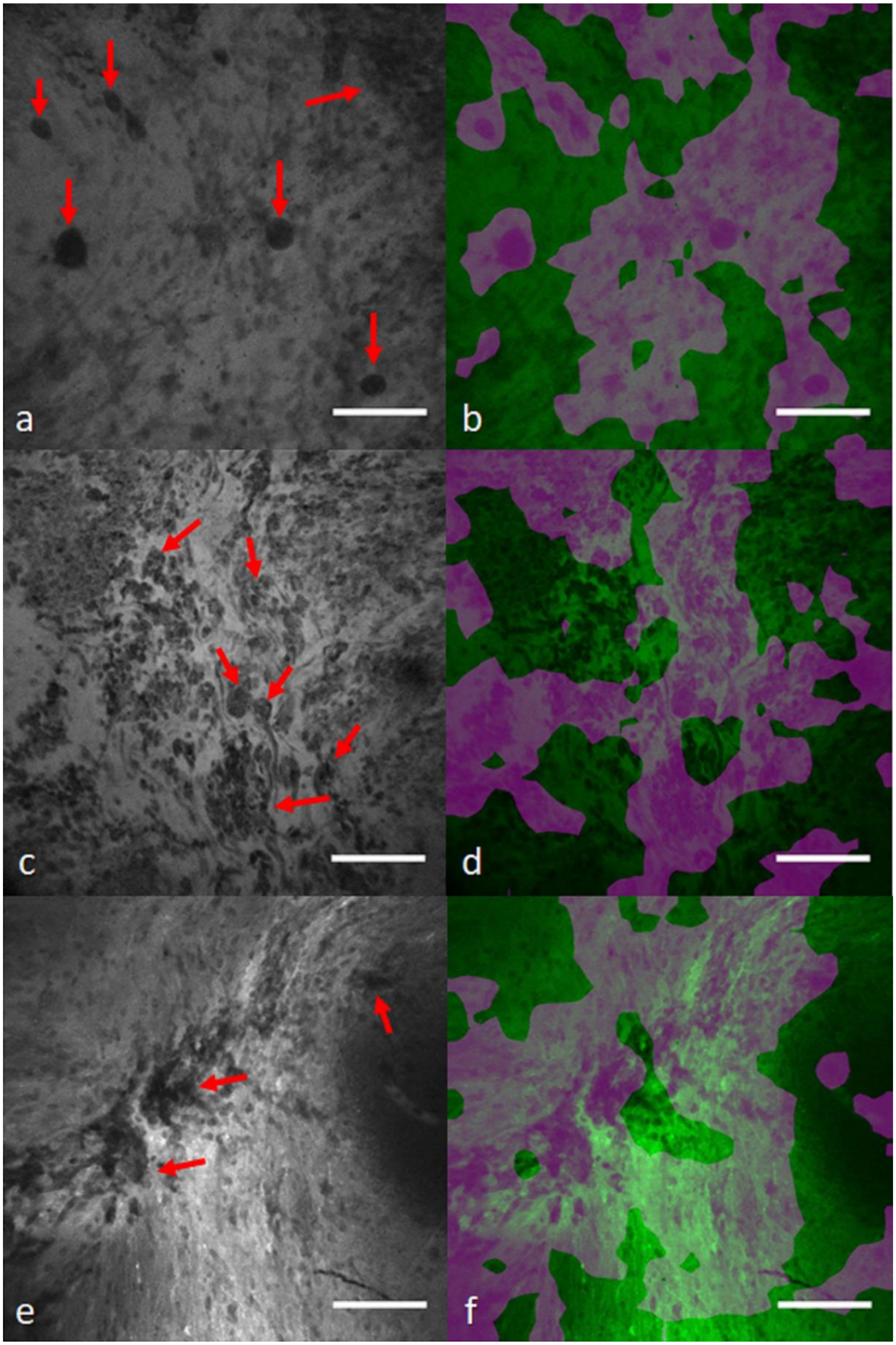

a
b
c
d
e
f

Figure 1.6

Diagnostic image segmentation in CLE images of gliomas with GAP approach(Mohammadhassan Izadyyazdanabadi, Belykh, Cavallo, et al., 2018). First column (a, d, e) shows the original diagnostic images from glioma cases ((a) recurrent GBM, (d) recurrent infiltrating astrocytoma, (e) anaplastic oligodendroglioma). Red arrows mark the cellular regions recognized by a neurosurgeon. Second column shows the segmented key features highlighted in purple. Field of view = 475×475 µm, resolution = 1024×1024 pixels, bar = 100 µm.

Deep learning-empowered CLE diagnostics in other cancers

Oral squamous cell carcinoma

Oral squamous cell carcinoma (OSCC) is a common cancer affecting 1.3 million cases worldwide annually (Ferlay et al., 2015). Because of the insufficient precision in current screening methods, most OSCC cases are unfortunately diagnosed at advanced stages leading to poor clinical outcome. CLE has allowed in vivo examination of OSCC which may lead to earlier and more effective therapeutic outcomes during examination (Thong et al., 2007). In a study by Aubreville, et al. (Aubreville et al., 2017), a CNN was trained to classify normal and carcinogenic CLE image patches. A dataset of 11,000 CLE images was evenly distributed between the two classes. The images were acquired from 12 patients and images with artifact (motion, noise, mucus or blood) were excluded, leading to 7,894 good quality images. Consequently, each image was divided into 21 overlapping patches, all of which were labeled the same as the whole image. The artifact patches were removed from images and the remaining ones were normalized to have zero mean and unit standard deviation. Image rotation was used to augment the image dataset size.

LeNet was used to train the model for patch classification. This network has only two convolutional and one fully connected layers with drop out. The model combined the probability scores from each constituent patch as being carcinogenic to arrive at the final prediction for the whole image. The network was trained from scratch with initial learning rate of 0.001 and Adam optimizer to minimize the cross-entropy.

To compare the proposed method with conventional textural feature-based classification approaches, two feature extraction methods (Gray-Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP)) and a classification approach (Random Forest (RF)) methods were combined to discriminate images at two scales (1.0 x and 0.5 x). Furthermore, CNN transfer learning was explored by shallow fine-tuning the last fully connected layer of the pretrained Inception network (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), using the original dataset. For cross validation, a leave-one-patient-out cross validation was followed, meaning images were used from one patient for testing the model and the remaining cases for training the model.

Both the patch-based and whole image CNN approaches outperformed the textural feature extraction and classification methods. The proposed CNN method could differentiate the normal and carcinogenic CLE images with 88% accuracy, 87% sensitivity and 90% specificity when applied at 0.5x scale (the 1x scale produced suboptimal results). The shallow fine-tuned Inception model could also achieve 87% accuracy, 91% sensitivity and 84% specificity. The AUC values for the two methods (the proposed CNN and Inception) were roughly similar (95%). The AUC values for feature extraction methods and RF classifier was significantly lower than CNN methods (RF-GLCM = 81%, RF-LBP = 89%). Interestingly, the trained model on OSCC CLE images was successfully applied for classification of CLE images from a different organ site, vocal cord squamous cell carcinoma.

In transfer learning with pretrained Inception the authors only modified the weights of the last layer, while keeping the previous layers parameters stationary. However, studies (Mohammadhassan Izadyyazdanabadi, Belykh, Mooney, et al., 2018; Tajbakhsh et al., 2016) have shown that deeply fine-tuning the pretrained networks may help the network adapt better to the new dataset by upgrading the feature extraction layers as well. However, shallow fine tuning only allows updating the classification layer, which might not be sufficient for optimal performance.

Vocal cord cancer

To differentiate between healthy and cancerous tissue of vocal cords, Vo, et al. (Vo, Jaremenko, Bohr, Neumann, & Maier, 2017) developed a Bag of Words (BoW) based on textural and CNN features using a dataset of 1,767 healthy and 2,657 carcinogenic images from five patients. Small patches with 105×105 pixel size were extracted and augmented (with rotation), leading to 374,972 patches. For the textural feature-based classification, each image was represented by the concatenation of all its constituent patch-driven feature descriptors. For the CNN features, a LeNet shallow CNN was trained on the patch dataset for a binary classification (with SGD optimizer; momentum = 0.9 and learning rate = 0.0005). To create the visual vocabulary, two feature encoding (Fisher vector(Sánchez, Perronnin, Mensink, & Verbeek, 2013) and Vector of Locally Aggregated Descriptors (VLAD) (Jégou, Douze, Schmid, & Pérez, 2010)) and two classification methods (SVM and RF) were tested for comparing their classification performance.

A Leave-One-Sequence-Out (LOSO) cross-validation was used to evaluate these methods. The CNN features combined with VLAD encoder and RF classifier achieved an accuracy of 82% and sensitivity of 82% on the test images that surpassed other approaches. However, despite its promising accuracy, the proposed multi-stage approach (patch creation, feature extraction, feature encoding, clustering and classification) is much more complicated than the current end-to-end DCNN architectures, which have all these procedures embedded in their stacked layers. However even with this approach, the CNN features could outperform textural features extracted manually (Vo et al., 2017).

Lung cancer

Gil, et al. (Gil, 2017) investigated visual patterns in bronchoscopic CLE images for discriminating benign and malignant lesions and aiding lung cancer diagnosis. A pretrained network developed by the Visual Geometry Group (VGG) (Chatfield, Simonyan, Vedaldi, & Zisserman, 2014) on a large generic image dataset was used for feature extraction and reduced the resulting feature vector dimension from 4,096 to 100, while preserving roughly 90% of the original feature vector energy for computational efficiency. Three different methods (k-Means, k-Nearest Neighbor (kNN), and their proposed topology-based approach) were applied on the feature codes to group images with similar features together and intrinsically discriminate images from benign and malignant tissue. Model predictive power was compared with the final diagnosis on 162 images from 12 cases (6 with malignant and 6 with benign lesions) and achieved 85% accuracy, 88% sensitivity, and 81% specificity.

Inter-observer studies were performed with three observers to compare the subjective visual assessment of images with the model performance (the observers were blinded to the final diagnosis). Interestingly, the three observers could make a correct diagnosis only for 60% of the selected CLE images (sensitivity: 73% for malignant and 36% for benign images) on average. In the second experiment, two observers made a final diagnosis after examining all the images from each case. The model was also supplied with all the images from each of the 12 cases and rendered a final decision for each case. While the model could differentiate malignant and benign cases with 100% accuracy (12/12), the two observers could confirmatively make the correct diagnosis only in 67% (8/12) of cases.

Although the observers' knowledge in the domain might have affected their performance, the objective results suggest that the bronchoscopic CLE images contain enough visual information for determining the malignancy of the tumor and the VGG network is an excellent candidate for extracting these discriminative features. The proposed topology-based clustering method could outperform common clustering and classification methods (K-means and kNN) in differentiating the two classes of images.

Despite its advantages, the proposed method had two limitations. First, even though it can differentiate images from the two classes, it cannot predict the label for each cluster. The method can separate the images into two groups, but it is not able to give information about their labels. Second, it is unclear if there was independent development (for determining model parameters) and test datasets to avoid bias in model development.

Gastrointestinal tract cancer

Hong, et al. (Hong, Park, & Park, 2017) proposed a CNN architecture for classifying CLE images from three subcategories of Barret's esophagus: intestinal metaplasia (IM), gastric metaplasia (GM), and neoplasia (NPL). The network was composed of four convolutional layers and two max-pooling and fully connected layers. The size of convolutional kernels was 3×3 and zero padding was also used. Stride of max-pooling was 2×2 which was applied in layers two and four. Fully connected layers followed the fourth convolutional layer, and each had 1,024 neurons. The output label was determined by a softmax layer which produced 3 probabilities for each subcategory.

The network was trained on the augmented CLE images of Barret's esophagus (155 IM, 26 GM and 55 NPL) for 15,000 iterations with the batch size of 20 images. Cross-entropy was used as a cost function in their experiment. The trained model was then tested on 26 independent images (17 IM, 4 GM and 5 NPL) for validation. The imbalance in size of different subcategories caused the model to observe more frequent instances of IM and NPL compared to GM during training. This created a bias in the model prediction which can be seen in the high accuracy for predicting IM and NPL instances (100% and 80%) and very low accuracy for predicting GM instances (0%). However, CLE is being used with increasing frequency for detecting pre-cancerous and cancerous lesions in the gastrointestinal (GI) tract. The highest numbers of CLE images have been acquired from the GI tract where such imaging technology has been approved for use clinically for a few years.

CONCLUSIONS

Precision, personalization, and improved therapeutics in medicine can only progress with improved technology, analysis, and logic. The science and philosophy of theranostics is the nexus of these. Several studies have emphasized the importance of theranostic imaging in personalized treatment of cancer (M.-F. Penet et al., 2014; M. F. Penet et al., 2012; Stasinopoulos et al., 2011). Medical data acquired on patients has become more voluminous, and it will continue in such manner. The amount of data available and necessary for analysis has already eclipsed human capabilities. For example, the new technology of handheld surgical tools that can rapidly image at the cellular resolution on-the-fly produces more images than a pathologist can possibly examine. As CLE technology develops, there will not be one fluorophore, but multiple fluorophores applied directly to the tumor or administered to the patient varying from nonspecific to specifically identifying cell structures or processes used simultaneously and presented in a myriad of image combinations for greatly varying histopathology. Analytic methods for selection and interpretation of the CLE images is already being explored to be incorporated into CLE operating systems so that the unit display can differentiate tissue and label the image as well with near-on-the-fly capabilities. Computational hardware power and effective analytic model infrastructure are the only two limits. CLE systems and other related systems are being produced by several imaging technology companies and groups and are close to approval with European and American medical device regulatory agencies. However, it seems prudent given the enormous numbers of images already produced and those projected with adoption of such technology, that there is immediate exploration into such image analysis methods to allow the pathologist and neurosurgeon to make optimal decisions based on the CLE imaging, and to take advantage of the on-the-fly technology proposition.

Success or meaningful diagnostic and therapeutic indication in the burgeoning field of theranostics is only as good as the data incorporated and the methodology employed for analysis and to extract meaning, including its validation. In many cases relatively simple statistics have been used for analysis, while pattern recognition or neural network techniques may be used in more complicated scenarios. For images such as from CLE, the whole image may be important, or perhaps only certain subregions, or crucial data may lie in regions on cursory inspection deemed to be nonuseful, such as in areas of motion artifact. Complicating this situation are the overwhelming numbers of images yielded from the CLE application. Clinical decision environments currently require assistance to not only access and categorize the collection of images, but to also draw conclusions and inferences that have critical diagnostic and treatment consequences. A pathologist and neurosurgeon will not have time to inspect a thousand images per case, especially in the midst of CLE use intraoperatively. Therefore, a theranostics approach, i.e., the nexus of biological data, rapid informatics scrutiny and evaluation, and tailored human decision, must be employed as we venture into realms of ever increasing information in neurosurgery in search of personalization and precision, especially as we have encountered it first in the surgery and treatment of malignant invasive brain tumors. Additionally, pathologists and neurosurgeons will need to become versed in the methodology of the CLE decision making processes to have confidence in diagnostic labels and to base treatment decisions upon them, thus the reason for presenting details of analytical architectures in this review.

Two DCNN based approaches are reviewed in this paper: models that can automatically classify CLE images (classifications of images that are diagnostic/nondiagnostic (Mohammadhassan Izadyyazdanabadi, Belykh, Mooney, et al., 2018) (M. Izadyyazdanabadi et al., 2017), tumor/injury/normal (Belykh et al., 2018)) and models that can localize histological features from diagnostic images using weakly supervised methods (Mohammadhassan Izadyyazdanabadi, Belykh, Cavallo, et al., 2018). Manually annotated in-house datasets were used to train and test these approaches in most of the studies. For the tumor classification purpose, data pruning could enhance the results for both DCNN models and outperformed manual feature extraction and classification (Kamen et al., 2016). Fine-tuning and ensemble modeling could enhance the model performance in the diagnostic image classification. The ensemble effect was stronger in DT and DFT than SFT developed models.

Despite extensive research on CLE clinical application in neurosurgery (Belykh et al., 2016; Charalampaki et al., 2015; J. Eschbacher et al., 2012; Foersch et al., 2012; Martirosyan et al., 2016; Sanai, Eschbacher, et al., 2011; Zehri et al., 2014), there have been few attempts in the automatic analysis of these images to enhance CLE clinical utility. Deep learning could be beneficial in filtering the nondiagnostic images with higher speed and reasonable accuracy compared to subjective assessment. (M. Izadyyazdanabadi et al., 2017) Our inter-rater agreement evaluation (Mohammadhassan Izadyyazdanabadi, Belykh, Mooney, et al., 2018) showed that the proposed model could achieve promising agreement with the gold-standard defined by a majority assessment by neurosurgeon reviewers. Overall, results suggest that DCNN-based diagnostic evaluation has a higher agreement with the ground truth than the entropy-based quality assessment used in other studies (Kamen et al., 2016; N. Murthy et al., 2017). Furthermore, such methods suggest that semantic histological features may be highlighted in CLE images as confirmed by a neurosurgeon reviewer. This shows that the DCNN structure could learn semantic concepts like tumor type or diagnostic value of CLE images through different levels of feature representation. Early results show that WSL-based glioma feature localization was able to precisely mark the cells in the images. DCNNs are also much faster than handcrafted methods at deployment phase. Our deeply trained models could classify about 40 new images in a second, while the handcrafted method takes 5.4 seconds to process single image (Kamen et al., 2016).

Other confocal imaging techniques may be aided by such deep learning models. Confocal reflectance microscopy (CRM) has been studied (J. M. Eschbacher et al., 2017; Mooney et al., 2018) for rapid, fluorophore-free evaluation of brain biopsy specimen *ex vivo*. CRM allows preserving the biopsy tissue for future permanent analysis, immunohistochemical studies, and molecular studies. Proposed DCNN classification and localization approaches are well-suited for analysis and interpretation of CRM images as well as CLE. Further studies on application of DCNN on CRM images are needed to further validate their utility for intraoperative diagnosis.

Continued use of unsupervised image segmentation methods to detect meaningful histological features from confocal brain tumor images will likely allow for more rapid and detailed diagnosis. With the large rate of images produced, a technology-free and unassisted approach to analyze the CLE images would impede the exploitation of maximal pathological information during surgery. Accessible databases of CLE images would allow various image analysis methods to be tried on large numbers of images. Such image collection strategies are part of the platform of the relatively new International Society for Endomicroscopy. DCNNs can enhance extraction and recognition of CLE diagnostic features that may be integrated into the standard brain tumor classification protocols similarly to the current research flow in the whole-slide digital image analysis for personalized cancer care (Djuric, Zadeh, Aldape, & Diamandis, 2017; Madabhushi & Lee, 2016). This may refine current diagnostic criteria and potentially aid the discovery of novel related features. With such technology, neurosurgery truly enters the realm of theranostics in the operating room itself—we are on the verge of highly tailored and precise surgery at the cellular level. Such an approach is critical for neurosurgery because surgery and treatment for an invasive brain tumor frequently deals with spread into eloquent cortex – the areas that make us “human.” In fact, even before entering the operating room the neurosurgeon can begin to discuss strategy with the patient if tumor is located or not located in eloquent cortex based on a CLE “optical biopsy”. Thus, theranostics also involves treatment strategies and decisions of when to “stop”, especially true when the CLE system intraoperatively reveals cells invading for example primary motor or language cortex. With analytical pathologists uniting different clinical and morphological information for an integrated diagnosis, such a computer-aided CLE analysis workflow would improve imaging (diagnostics) and achieve maximal, more precise removal of tumor mass (therapy) as the initial treatment goals toward greater precision, personalization and success in the surgery and treatment of malignant invasive brain tumors (theranostics).

References

Almeida, J. P., Chaichana, K. L., Rincon-Torroella, J., & Quinones-Hinojosa, A. (2015). The Value of Extent of Resection of Glioblastomas: Clinical Evidence and Current Approach. Current Neurology and Neuroscience Reports. https://doi.org/10.1007/s11910-014-0517-x

American Cancer Society. (2018). Cancer Facts and Statistics. Retrieved May 29, 2018, from https://cancerstatisticscenter.cancer.org/

Aubreville, M., Knipfer, C., Oetter, N., Jaremenko, C., Rodner, E., Denzler, J., … Maier, A. (2017). Automatic Classification of Cancerous Tissue in Laserendomicroscopy Images of the Oral Cavity using Deep Learning. Scientific Reports, 7. https://doi.org/10.1038/s41598-017-12320-8

Belykh, E., Martirosyan, N. L., Yagmurlu, K., Miller, E. J., Eschbacher, J. M., Izadyyazdanabadi, M., … Preul, M. C. (2016). Intraoperative fluorescence imaging for personalized brain tumor resection: Current state and future directions. Frontiers in Surgery, 3.

Belykh, E., Miller, E. J., Patel, A. A., IzadyYazdanabadi, M., Martirosyan, N. L., Yagmurlu, K., … Preul, M. C. (2018). Diagnostic accuracy of the confocal laser endomicroscope for in vivo differentiation between normal and tumor tissue during fluorescein-guided glioma resection: Laboratory investigation. World Neurosurgery, In press.

Charalampaki, P., Javed, M., Daali, S., Heiroth, H.-J., Igressa, A., & Weber, F. (2015). Confocal Laser Endomicroscopy for Real-time Histomorphological Diagnosis: Our Clinical Experience With 150 Brain and Spinal Tumor Cases. Neurosurgery, 62, 171–176.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the Devil in the Details: Delving Deep into Convolutional Nets. BMVC, 1–11. https://doi.org/10.5244/C.28.6

Djuric, U., Zadeh, G., Aldape, K., & Diamandis, P. (2017). Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. Npj Precision Oncology, 1(1), 22. https://doi.org/10.1038/s41698-017-0022-1

Eschbacher, J. M., Georges, J. F., Belykh, E., Yazdanabadi, M. I., Martirosyan, N. L., Szeto, E., … others. (2017). Immediate Label-Free Ex Vivo Evaluation of Human Brain Tumor Biopsies With Confocal Reflectance Microscopy. Journal of Neuropathology & Experimental Neurology, 76(12), 1008–1022.

Eschbacher, J., Martirosyan, N. L., Nakaji, P., Sanai, N., Preul, M. C., Smith, K. A., … Spetzler, R. F. (2012). In vivo intraoperative confocal microscopy for real-time histopathological imaging of brain tumors: Clinical article. Journal of Neurosurgery, 116(4), 854–860.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., … Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. International Journal of Cancer, 136(5), E359–E386. https://doi.org/10.1002/ijc.29210

Foersch, S., Heimann, A., Ayyad, A., Spoden, G. A., Florin, L., Mpoukouvalas, K., … Charalampaki, P. (2012). Confocal laser endomicroscopy for diagnosis and histomorphologic imaging of brain tumors in vivo. PLoS One, 7(7), e41760.

Gil, D. et al. (2017). Classification of Confocal Endomicroscopy Patterns for Diagnosis of Lung Cancer. In Cardoso M. et al. (eds) Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures. CARE 2017, CLIP 2017. Lecture Notes in Computer Science (Vol. 10550, pp. 151–159).

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. AISTATS '11: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 15, 315–323. https://doi.org/10.1.1.208.6449

Greenspan, H., van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging, 35(5), 1153–1159.

Hong, J., Park, B., & Park, H. (2017). Convolutional neural network classifier for distinguishing Barrett's esophagus and neoplasia endomicroscopy images. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 2892–2895). IEEE. https://doi.org/10.1109/EMBC.2017.8037461

Izadyyazdanabadi, M., Belykh, E., Cavallo, C., Zhao, X., Gandhi, S., Moreira, L. B., … Yang, Y. (2018). Weakly-Supervised Learning-Based Feature Localization in Confocal Laser Endomicroscopy Glioma Images. arXiv Preprint arXiv:1804.09428.

Izadyyazdanabadi, M., Belykh, E., Martirosyan, N., Eschbacher, J., Nakaji, P., Yang, Y., & Preul, M. C. (2017). Improving utility of brain tumor confocal laser endomicroscopy: Objective value assessment and diagnostic frame detection with convolutional neural networks. In Progress in Biomedical Optics and Imaging - Proceedings of SPIE (Vol. 10134). https://doi.org/10.1117/12.2254902

Izadyyazdanabadi, M., Belykh, E., Mooney, M., Martirosyan, N., Eschbacher, J., Nakaji, P., … Yang, Y. (2018). Convolutional Neural Networks: Ensemble Modeling, Fine-Tuning and Unsupervised Semantic Localization for Neurosurgical CLE Images. Journal of Visual Communication and Image Representation, 54, 10–20.

Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 3304–3311). https://doi.org/10.1109/CVPR.2010.5540039

Kamen, A., Sun, S., Wan, S., Kluckner, S., Chen, T., Gigler, A. M., … others. (2016). Automatic Tissue Differentiation Based on Confocal Endomicroscopic Images for Intraoperative Guidance in Neurosurgery. BioMed Research International, 2016.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097–1105).

Le Cun Jackel, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D., Cun, B. Le, Denker, J., & Henderson, D. (1990). Handwritten Digit Recognition with a Back-Propagation Network. Advances in Neural Information Processing Systems, 396–404. https://doi.org/10.1111/dsu.12130

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., … Vapnik, V. (1995). Learning algorithms for classification: A comparison on handwritten digit recognition. In Neural networks: the statistical mechanics perspective (pp. 261–276).

Liu, J. T. C., Meza, D., & Sanai, N. (2014). Trends in fluorescence image-guided surgery for gliomas. Neurosurgery. https://doi.org/10.1227/NEU.0000000000000344

Loiseau, S. (2017). Presentation at International Society for Endomicroscopy, Paris, France. Paris.

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., … Kleihues, P. (2007). The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathologica. https://doi.org/10.1007/s00401-007-0243-4

Madabhushi, A., & Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical Image Analysis, 33, 170–175. https://doi.org/10.1016/j.media.2016.06.037

Martirosyan, N. L., Cavalcanti, D. D., Eschbacher, J. M., Delaney, P. M., Scheck, A. C., Abdelwahab, M. G., … Preul, M. C. (2011). Use of in vivo near-infrared laser confocal endomicroscopy with indocyanine green to detect the boundary of infiltrative tumor. Journal of Neurosurgery, 115(6), 1131–1138. https://doi.org/10.3171/2011.8.JNS11559

Martirosyan, N. L., Eschbacher, J. M., Kalani, M. Y. S., Turner, J. D., Belykh, E., Spetzler, R. F., … Preul, M. C. (2016). Prospective evaluation of the utility of intraoperative confocal laser endomicroscopy in patients with brain neoplasms using fluorescein sodium: experience with 74 cases. Neurosurgical Focus, 40(3), E11.

Martirosyan, N. L., Georges, J., Eschbacher, J. M., Belykh, E., Carotenuto, A., Spetzler, R. F., … Preul, M. C. (2018). Confocal scanning microscopy provides rapid, detailed intraoperative histological assessment of brain neoplasms: Experience with 106 cases. Clinical Neurology and Neurosurgery, 169, 21–28.

Martirosyan, N. L., Georges, J., Eschbacher, J. M., Cavalcanti, D. D., Elhadi, A. M., Abdelwahab, M. G., … Preul, M. C. (2014). Potential application of a handheld confocal endomicroscope imaging system using a variety of fluorophores in experimental gliomas and normal brain. Neurosurgical Focus, 36(2), E16.

Maugeri, R., Villa, A., Pino, M., Imperato, A., Giammalva, G. R., Costantino, G., … others. (2018). With a Little Help from My Friends: The Role of Intraoperative Fluorescent Dyes in the Surgical Management of High-Grade Gliomas. Brain Sciences, 8(2), 31.

Mooney, M. A., Georges, J., Yazdanabadi, M. I., Goehring, K. Y., White, W. L., Little, A. S., … Eschbacher, J. M. (2018). Immediate ex-vivo diagnosis of pituitary adenomas using confocal reflectance microscopy: a proof-of-principle study. Journal of Neurosurgery, 128, 1072–1075.

Mooney, M. A., Zehri, A. H., Georges, J. F., & Nakaji, P. (2014). Laser scanning confocal endomicroscopy in the neurosurgical operating room: a review and discussion of future applications. Neurosurgical Focus, 36(2), E9. https://doi.org/10.3171/2013.11.FOCUS13484

Murthy, V. N., Singh, V., Chen, T., Manmatha, R., & Comaniciu, D. (2016). Deep Decision Network for Multi-class Image Classification. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2240–2248). https://doi.org/10.1109/CVPR.2016.246

N. Murthy, V., Singh, V., Sun, S., Bhattacharya, S., Chen, T., & Comaniciu, D. (2017). Cascaded deep decision networks for classification of endoscopic images. In M. A. Styner & E. D. Angelini (Eds.), Medical Imaging 2017: Image Processing (Vol. 10133, p. 101332B). https://doi.org/10.1117/12.2254333

Ostrom, Q. T., Gittleman, H., Fulop, J., Liu, M., Blanda, R., Kromer, C., … Barnholtz-Sloan, J. S. (2015). CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012. Neuro-Oncology, 17, iv1-iv62. https://doi.org/10.1093/neuonc/nov189

Penet, M.-F., Krishnamachary, B., Chen, Z., Jin, J., & Bhujwalla, Z. M. (2014). Molecular imaging of the tumor microenvironment for precision medicine and theranostics. Advances in Cancer Research, 124, 235–256. https://doi.org/10.1016/B978-0-12-411638-2.00007-0

Penet, M. F., Chen, Z., Kakkad, S., Pomper, M. G., & Bhujwalla, Z. M. (2012). Theranostic imaging of cancer. European Journal of Radiology, 81(SUPPL1). https://doi.org/10.1016/S0720-048X(12)70051-7

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536. https://doi.org/10.1038/323533a0

Sanai, N., & Berger, M. S. (2018). Surgical oncology for gliomas: the state of the art. Nature Reviews Clinical Oncology, 15(2), 112.

Sanai, N., Eschbacher, J., Hattendorf, G., Coons, S. W., Preul, M. C., Smith, K. A., … Spetzler, R. F. (2011). Intraoperative confocal microscopy for brain tumors: a feasibility analysis in humans. Neurosurgery, 68, ons282--ons290.

Sanai, N., Polley, M.-Y., McDermott, M. W., Parsa, A. T., & Berger, M. S. (2011). An extent of resection threshold for newly diagnosed glioblastomas: clinical article. Journal of Neurosurgery, 115(1), 3–8.

Sánchez, J., Perronnin, F., Mensink, T., & Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. International Journal of Computer Vision, 105(3), 222–245. https://doi.org/10.1007/s11263-013-0636-x

Sankar, T., Delaney, P. M., Ryan, R. W., Eschbacher, J., Abdelwahab, M., Nakaji, P., … Preul, M. C. (2010). Miniaturized handheld confocal microscopy for neurosurgery: Results in an experimental glioblastoma model. Neurosurgery, 66(2), 410–417. https://doi.org/10.1227/01.NEU.0000365772.66324.6F

Stasinopoulos, I., Penet, M. F., Chen, Z., Kakkad, S., Glunde, K., & Bhujwalla, Z. M. (2011). Exploiting the tumor microenvironment for theranostic imaging. NMR in Biomedicine. https://doi.org/10.1002/nbm.1664

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2818–2826). IEEE. https://doi.org/10.1109/CVPR.2016.308

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Transactions on Medical Imaging, 35(5), 1299–1312.

Thong, P. S.-P., Olivo, M., Kho, K.-W., Zheng, W., Mancer, K., Harris, M., & Soo, K.-C. (2007). Laser confocal endomicroscopy as a novel technique for fluorescence diagnostic imaging of the oral cavity. Journal of Biomedical Optics, 12(1), 14007. https://doi.org/10.1117/1.2710193

Tofte, K., Berger, C., Torp, S. H., & Solheim, O. (2014). The diagnostic properties of frozen sections in suspected intracranial tumors: A study of 578 consecutive cases. Surgical Neurology International, 5.

Vo, K., Jaremenko, C., Bohr, C., Neumann, H., & Maier, A. (2017). Automatic Classification and Pathological Staging of Confocal Laser Endomicroscopic Images of the Vocal Cords. In Bildverarbeitung für die Medizin 2017 (pp. 312–317). Springer.

Zehri, A., Ramey, W., Georges, J., Mooney, M., Martirosyan, N., Preul, M., & Nakaji, P. (2014). Neurosurgical confocal endomicroscopy: A review of contrast agents, confocal systems, and future imaging modalities. Surgical Neurology International, 5, 60.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2921–2929). https://doi.org/10.1109/CVPR.2016.319

CHAPTER 2

The following chapter has been published in the Journal of Visual Communication and Image Representation.

(DOI: 10.1016/j.jvcir.2018.04.004)

CHAPTER 2

CONVOLUTIONAL NEURAL NETWORKS: ENSEMBLE MODELING, FINE-TUNING AND UNSUPERVISED SEMANTIC LOCALIZATION FOR NEUROSURGICAL CLE IMAGES

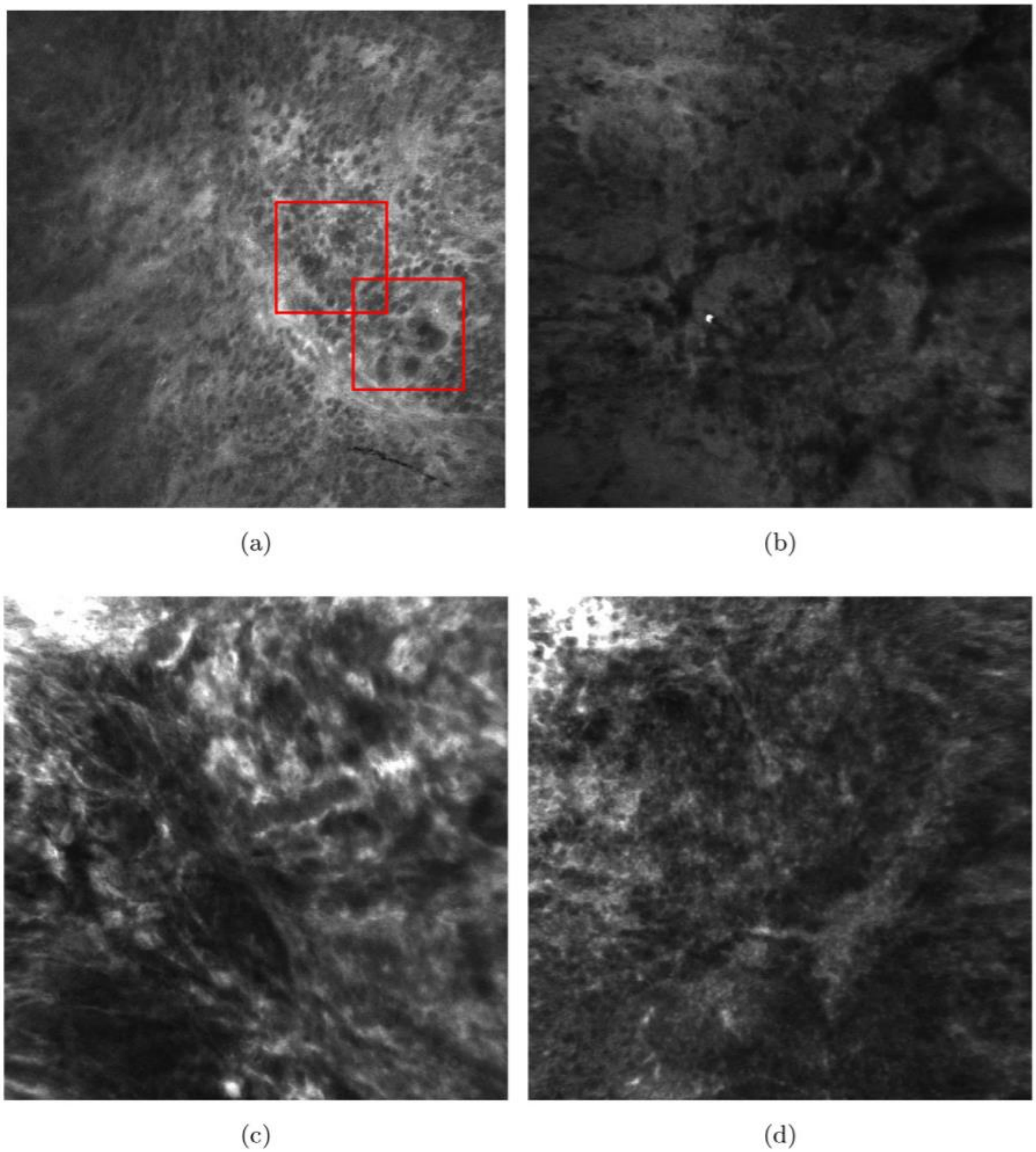
Izadyyazdanabadi M, Belykh E, Mooney M, Martirosyan N, Eschbacher J, Nakaji P, Preul MC, Yang Y

ABSTRACT

Confocal laser endomicroscopy (CLE) is an advanced optical fluorescence technology undergoing assessment for applications in brain tumor surgery. Many of the CLE images can be distorted and interpreted as nondiagnostic. However, just one neat CLE image might suffice for intraoperative diagnosis of the tumor. While manual examination of thousands of nondiagnostic images during surgery would be impractical, this creates an opportunity for a model to select diagnostic images for the pathologists or surgeons review. In this study, we sought to develop a deep learning model to automatically detect the diagnostic images. We explored the effect of training regimes and ensemble modeling and localized histological features from diagnostic CLE images. The developed model could achieve promising agreement with the ground truth. With the speed and precision of the proposed method, it has potential to be integrated into the operative workflow in the brain tumor surgery.

INTRODUCTION

Handheld, portable Confocal Laser Endomicroscopy (CLE) is being explored in neurosurgery because of its ability to image histopathological features of tissue in real time (Belykh et al., 2016; Charalampaki et al., 2015; Foersch et al., 2012; Sanai et al., 2011). CLE provides cellular resolution imaging during brain tumor surgery and thus may provide the surgeon with precise histopathological information during tumor resection in order to interrogate regions that may harbor malignant or spreading tumor, especially at the tumor border.



(a) (b) (c) (d)

Figure 2.1

Typical diagnostic and nondiagnostic CLE images (field of view = 475×475µm). (a) A diagnostic CLE image from a glioma case (anaplastic oligodendroglioma). Red boxes highlight the histopathological features of gliomas such as pleomorphism and hypercellularity detected by our model. For more results please see Figure 2.5. (b) A nondiagnostic CLE image from a glioma case. (c) A diagnostic image from a meningioma case (transitional meningioma). (d) Another diagnostic CLE image from a different case with anaplastic oligodendroglioma.

Current CLE systems are able to image more than one image per second, and thus over the course of examination of the surgical tumor resection or inspection area, hundreds to thousands of images may be collected. The number of images may become rapidly overwhelming for the neurosurgeon and neuropathologist when trying to quickly select a diagnostic or meaningful image or group of images as the surgical inspection progresses. CLE is designed to be used on the fly in real time while the surgeon operates the brain. Thus, overcoming the barriers involved in image selection is a key component for making CLE a practical and advantageous technology for the neurosurgical operating room.

A wide range of fluorophores are able to be used for CLE in gastroenterology, but fluorophore options are limited for in vivo human brain use due to potential toxicity (Belykh et al., 2016; Foersch et al., 2012; Zehri et al., 2014). In addition, motion and blood artifacts that are present in many of the images acquired with CLE using fluorescein sodium (FNa) are a barrier for revealing underlying meaningful histology. The display of suboptimal images or nondiagnostic frames interferes with the selection of and focus upon diagnostic images by the neurosurgeon and pathologist throughout the operation in order to make a correct intraoperative diagnosis. Previous assessment (Martirosyan et al., 2016) of CLE in human brain tumor surgery found that about half of the acquired images were interpreted as nondiagnostic due to abundance of motion and blood artifacts or lack of discernible or characteristic histopathological features. Filtering out the nondiagnostic images before making an intraoperative diagnosis is challenging due to the high number of images acquired, the novel and frequently unfamiliar appearance of tissue features compared to conventional histology, great variability among images from the same tumor type (Figure 2.1 (a, b, d)), and potential similarity between images from other tumor types for the untrained interpreter (Figure 2.1 (c, d)).

Applications of machine learning in medical imaging have greatly increased in the last ten years, resulting in numerous computer-aided detection (CADe) and diagnosis (CADx) systems in ultrasound, magnetic resonance imaging (MRI), and computed tomography (CT) (Greenspan, van Ginneken, & Summers, 2016). Applications of machine learning for CLE imaging in neurosurgery are yet to be performed. In this study, we developed an ensemble of deep convolutional neural networks that can automatically evaluate the diagnostic value of CLE frames within milliseconds. Compared to our previous work (Izadyyazdanabadi et al., 2017), the following contributions and advancements were provided:

- **Dataset.** Our dataset contains CLE images which is a novel technology in contrast to commonly used MRI or CT scans. The dataset used includes 20,734 CLE images from intracranial neoplasms.

- **Patient-based nested cross-validation for testing.** The model validation was performed in a patient-based nested cross-validation form. We divided our dataset into two sections: development and test. The images from patients which were used in the development stage were isolated from the test set. The development dataset is used for finding the model hyperparameters, training, initial evaluation of the trained models and model selection. After the final model is selected, the test dataset is used for the final estimation of model performance on images from new patients.
- **Deep training, shallow fine-tuning or deep fine-tuning?** The CNN architectures were trained in three regimes: I. deeply trained (train the network from scratch with model weights randomly initialized) II. shallow fine-tuned (fine-tune only the fully connected layer(s) of the model which are responsible for the classification) III. deeply fine-tuned (fine-tune the whole network using our dataset). In this study we report model accuracy on the test dataset for the best 5 models from each network architecture and training regime. Our work is different from (Tajbakhsh et al., 2016) since it considers the fine tuning effect on two different network architectures and its effect on the ensemble models.
- **Ensemble modeling.** Prior to the test phase, we created an ensemble of the best 5 models from each network and training regime. We explored the effect of ensemble modeling in all circumstances by comparing the ensemble performance with the average of single models. Our work is different from (Kumar, Kim, Lyndon, Fulham, & Feng, 2017) since our ensemble generates diversity in single models by using different training and validation data achieved from nested-cross

validation. Further, we studied the effect of ensemble modeling on different training scenarios rather than one.

- **Unsupervised localizing of histological features.** We visualized the shallow neurons' activation to depict the broad histological patterns; visualization of deep neurons’ activation could localize specific histopathological lesions for diagnostic images. The neural response of convolutional layers to the diagnostic images are visualized and analyzed by a neurosurgeon. We also extracted the CNN's deepest neural activation in response to patches of the diagnostic images using the sliding-window. Qualitative assessment of the localized regions was performed by a neurosurgeon with further analysis of the histopathological features.
- **Interobserver study.** We compared the interobserver agreement between physician-physician and ensemble of models-physician to compare our ensemble model performance with human performance. We also reported the kappa statistic for this observer study.

Due to the limited number of images in our dataset, we sought to transfer learning benefits by using pretrained models, fine-tuning them in shallow and deep manner and compare results with the models trained from scratch. Our results demonstrated that a shallow fine-tuned model, although performs better than trained from scratch, is not enough for the optimal performance and that a deep fine-tuned model detects the diagnostic CLE frames better. We also investigated the effect of ensemble modeling by creating an ensemble of models which were crafted at the development stage and produced the minimum loss on validation dataset. Finally, we compared the performance among the ensemble of models and each single model.

RELATED WORKS

Convolutional neural networks

Convolutional neural networks (CNNs), a subcategory of deep learning methods, have proven useful in visual imagery analysis from numerous fields, including medical images. This is mainly due to the deep multilayer architecture of CNNs which enables extracting abstract discriminant features, both local and global, present in the images (LeCun, Bengio, & Hinton, 2015).

In the recent years, deep learning has been vastly applied in medical image or exam classification. According to a recent survey (Litjens et al., 2017), exam and object classification together make up the number one task of interest in medical image analysis followed by object detection and organ segmentation (exam classification alone is the third task of interest). Most of these studies in medical imaging field use one of the three following imaging modalities: MRI, microscopy or CT.

Histopathological microscopic images and brain MRI scans were the first two areas where deep learning has been explored in medical imaging (Litjens et al., 2017). In histopathology, deep learning has been used for mitosis detection (Ciresan, Giusti, Gambardella, & Schmidhuber, 2012), classification of leukocytes (J. Zhao, Zhang, Zhou, Chu, & Cao, 2016) and nuclei detection and classification (Sirinukunwattana et al., 2016). In brain MRI, several studies have concluded that CNN benefits the diagnosis of Alzheimer’s disease (Shi, Zheng, Li, Zhang, & Ying, 2017; Suk & Shen, 2016) as well as brain extraction (Salehi, Erdogmus, & Gholipour, 2017) and lesion detection, classification, and tumor grading (Ghafoorian et al., 2017; L. Zhao & Jia, 2016).

No-reference image quality assessment has been formulated as a classification problem as employed in retinal (Mahapatra, Roy, Sedai, & Garnavi, 2016) and echocardiographic (Abdi et al., 2017) images. CNNs may also be exercised in the detection of key frames from a temporal sequence of frames in a video. Two studies demonstrated the use of classification scheme on ultrasound (US) stream video to label the frames (Gao, Maraci, & Noble, 2016; Kumar et al., 2016).

Transfer learning vs. deep training

One of the major limitations in medical imaging is the small size of datasets. The number of images employed for deep learning applications in medical imaging is usually much smaller than those in computer vision. Therefore, two forms of transfer learning have gained great interest: 1. Application of a pretrained network on large-size natural images (i.e. ImageNet) as a feature extractor. 2. Initializing model parameters (weights and biases) using the data from a pretrained model (Yosinski, Clune, Bengio, & Lipson, 2014) instead of random initialization. A previous study (Tajbakhsh et al., 2016) showed that a sufficiently fine-tuned AlexNet model could produce equal or better results than a deeply trained one for colonoscopy image quality assessment and few other medical applications. Here, we'll study the fine-tuning effect by extending it to Inception network architectures in single and ensemble mode.

Ensemble modeling

Ensemble modeling is an established method for increasing the model performance and reducing its variance in machine learning (Ciregan, Meier, & Schmidhuber, 2012; Dietterich & others, 2000; Zhou, Wu, & Tang, 2002).

Kumar et al. A recent study (Kumar et al., 2017) created an ensemble of 5 different models to classify the image modality from ImageCLEF 2016 medical image dataset. Specifically, 2 classifiers were created by fine-tuning AlexNet and GoogLeNet with softmax and 3 other classifiers by training an SVM on top of the features extracted by AlexNet, GoogLeNet, and their combination. Their results showed the ensemble could improve the top-1 accuracy of the classifier compared to single models, however it is not clear that how much of the improvement was because of the AlexNet and GoogLeNet combination or the 5 classifiers ensemble.

Another study (Christodoulidis, Anthimopoulos, Ebner, Christe, & Mougiakakou, 2017) created an ensemble of multi-source transfer learning using an automatic model selection approach. After creating a pool of pretrained CNNs on several public texture datasets and fine-tuning them on the lung CT dataset, the top models which iterative grouping would produce the highest F-scores on the validation dataset were aggregated, creating an ensemble model. 5 ensemble models were developed and their output was then averaged to make the final output. Despite its computational complexity, it enhanced the lung disease pattern classification accuracy only by 2%.

To generate diversity in our models while using the whole training dataset, we trained different neural networks on different data using cross-validation. Although previous studies have tried to create variant deep learning models using different network architectures, none of them have employed training data diversification through cross-validation as described in (Krogh & Vedelsby, 1995). Our proposed ensemble employed model diversification both in the network architectures and in the training and validation datasets following (Krogh & Vedelsby, 1995).

Confocal laser endomicroscopy in neurosurgery

Handheld, portable CLE has demonstrated its value for brain tumor surgery due to its ability to provide rapid intraoperative information regarding histopathological features of the tumor tissue (Martirosyan et al., 2016). Convenience, portability, and speed of CLE are significant advantages in surgery. A decision support system aiding and accelerating analysis of CLE images by the neuropathologist or neurosurgeon would improve the workflow in the neurosurgical operating room (Izadyyazdanabadi et al., 2017).

Potentially used at any time during the surgery, CLE interrogation of the tissue generates images with a speed of 0.8 - 1.2 frames per second. The frames are considered nondiagnostic when the histological features are obscured by the red blood cells or motion artifacts, are out of focus, or lack any useful histopathological information. Acquired images are then exported from the instrument as JPEG or TIFF files for review. Currently, the pathologist reviews all images, including nondiagnostic ones, trying to explore the diagnostic frames for the diagnosis. Manual selection and review of thousands of images acquired at some point in surgery by the CLE operator is tedious and impractical for widespread use. Previously, we have presented (Izadyyazdanabadi et al., 2017) the first deeply trained CNN model for automatic detection of diagnostic CLE frames.

METHODS

Image acquisition

In the following sections we briefly explain the confocal imaging instrument specifications and the intraoperative data collection process.

Instrument specifications

The CLE image acquisition system (Optiscan 5.1, Optiscan Pty, Ltd.) included a rigid pen-sized optical laser scanner with a 6.3 mm outer diameter and a working length of 150 mm. A 488 nm diode laser provided excitation light and the fluorescent emission signal was detected with a ~505-585 nm band-pass filter. A single optical fiber acted as both the excitation pinhole and the detection pinhole for confocal isolation of the image plane. The detector signal was digitized synchronously with the scanning to construct images parallel to the tissue surface (en face optical sections).

Laser power was typically set to 550-900 µW and maximum power was limited to 1000 µW when applied to the brain tissue. A field of view of 475×475 µm was scanned at 1024×1024 (1.2/second frame rate), with a lateral resolution of 0.7 µm and an axial resolution (i.e., effective optical slice thickness) of approximately 4.5 µm.

Intraoperative CLE imaging

Seventy-four adult patients (31 male and 43 female) were enrolled in the study (mean age 47.5 years). Intraoperative CLE images were acquired both in vivo and ex vivo by 4 neurosurgeons. For in vivo imaging, multiple locations of the tissue around the lesion were imaged and excised from the patient. For ex vivo imaging, tissue samples suspicious for tumor were excised, placed on gauze and imaged on a separate work station in the operating room. Multiple images were obtained from each biopsy location.

Co-registration of the CLE probe with the image guided surgical system allowed precise intraoperative mapping of CLE images with regard to the site of the biopsy. The only fluorophore administered was FNa (5 mL, 10%) that was injected intravenously during the surgery.

Precise location of the areas imaged with the CLE was marked with tissue ink. Imaged tissue was sent to the pathology laboratory for formalin fixation, paraffin embedding and histological sections preparation. Final histopathological assessment was performed by standard light microscopic evaluation of 10-μm-thick hematoxylin and eosin (H & E)-stained sections.

Image annotation

The image annotation process was done in two distinct stages: initial review and validation review.

Initial review

Initially all images were reviewed. A neuropathologist and 2 neurosurgeons who were not involved in the surgeries reviewed the CLE images. For each patient, the histopathological features of corresponding CLE images and H & E-stained frozen and permanent sections were reviewed and the diagnostic value of each image was examined. When CLE image revealed clear identifiable histopathological feature, it was labeled as diagnostic; otherwise it was labeled as nondiagnostic.

Validation review

The database of images was divided into development and test datasets (explained in dataset preparation). Test dataset composed of 4171 CLE images randomly chosen from various patients. The *validation review* (val-review) dataset consists of 540 images randomly chosen from the test dataset. Following this separation, two neurosurgeons reviewed val-review dataset without having access to the corresponding H & E-stained slides and labeled them as diagnostic or nondiagnostic.

Convolutional neural networks

Convolutional Neural Networks (CNNs) are multilayer learning frameworks and may consist of an input layer, a few convolutional layers, pooling layers, fully connected layers and the output. The goal of a CNN is to learn the hierarchy of underlying feature representations. We explain the fundamental elements of a CNN below.

Convolutional layer

Convolutional layers, first introduced in (LeCun & Bengio, 1998) are the substitute of previous hand-crafted feature extractors. At each convolutional layer three dimensional matrices (kernels) are slid over the input and set the dot product of kernel weights with the receptive field of the input as the corresponding local output. This helps to retain the relative position of features to each other. The multi-kernel characteristic of convolutional layers enables them to prospectively extract several distinct feature maps from the same input image.

Activation layer

The convolutional layer output then goes through an activation function to adjust the negative values. We employed the rectified linear unit (ReLU) which is usually the preferred choice because of its simplicity, higher speed, reduced likelihood of vanishing gradients (especially in deep networks) and tendency to add sparsity over other nonlinear functions such as sigmoid function. The output of j^{th} ReLU layer a_j^{out}, given its input a_j^{in}, was calculated in-place (to consume less memory) by following:

$$a_j^{out} = \max(a_j^{in}, 0)$$

Normalization layer

Following the ReLU layer, a local response normalization (LRN) map is applied after the initial convolutional layers. This layer inhibits the local ReLU neurons' activations since there's no bound to limit them. By using the Caffe (Jia et al., 2014) implemented LRN, the local regions are expanded across neighbor feature maps at each spatial location. The output of j^{th} LRN layer a_j^{out}, given its input a_j^{in}, is calculated as:

$$\mathrm{a}_\mathrm{j}^\mathrm{out} = \frac{\mathrm{a}_\mathrm{j}^\mathrm{in}}{\left(1 + \frac{\alpha}{\mathrm{L}} \sum_{\mathrm{n}=1}^{\mathrm{L}} \mathrm{a}_\mathrm{j}^\mathrm{in}(\mathrm{n})^2\right)^\beta}$$

where $a_j^{in}(n)$ is the n^{th} element of the a_j^{in} and L is the length of a_j^{in} vector (number of neighbor maps employed in the normalization). α, β, and L are the layer's hyperparameters and are set to their default values obtained from (Krizhevsky, Sutskever, & Hinton, 2012) (α=1, β=0.75 and L=5).

Pooling layer

After rectification and normalization of convolutional layer output, it's further down-sampled by pooling operations. Pooling operations accumulate values in a smaller region by subsampling operations such as max, min, and average sampling. Here, max pooling was applied with a kernel size of 3, stride 2 for network 1 and stride 1 for network 2.

Fully connected layer

Following several convolutional and pooling layers, the network lateral layers are fully connected. Each neuron of the layer's output is greedily connected to all the layer's input neurons. It can be thought of as a convolutional layer with kernel size of the layer input. The layer output is also passed through a ReLU layer.

The fully connected layers are generally thought of as the classifier of a CNN model because they intake the most abstract features extracted in convolutional layers and make the output, which is the model prediction.

Dropout layer

Fully connected layers are usually followed by a dropout layer, except the last fully connected layer that produces the class-specific probabilities. In dropout layers, a subset of input neurons as well as all their connections are temporarily removed from the network. Srivastava et al. have demonstrated this method efficiency at improving the CNN performance in numerous computer vision tasks through reducing the overfitting (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014).

Learning

The learning of a CNN is by loss minimization using an optimization algorithm in two steps: Forward and Back Propagation. In this study, Stochastic Gradient Descent (SGD) was used as the optimization algorithm. In forward propagation, the model makes predictions using the images in the training batch and the current model parameters. Once the prediction for all training images is made, the loss is calculated using the truth label provided by the experts in the initial review. In this work we adopt the softmax loss function given by:

$$L(t, y) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{C} t_k^n log\left(\frac{e^{y_k^n}}{\sum_{m=1}^{C} e^{y_m^n}}\right)$$

where t_k^n is the n^{th} training image's k^{th} ground truth output, and y_k^n is the value of the k^{th} output layer unit in response to the n^{th} input training image. N is the number of training images in the minibatch, and since we consider 2 diagnostic value categories, C=2.

Through the back propagation, the loss gradient with respect to all model weights aids upgrading the weights as follows:

$$\mathrm{W(j, i+1) = W(j, i) + \mu\Delta W(j, i) - \alpha(j, i)\frac{\partial L}{\partial W(j)}}$$

Where $W(j, i)$, $W(j, i+1)$ and $\Delta W(j, i)$ are the weights of j^{th} convolutional layer at iteration i and i+1 and the weight update of iteration i, μ is the momentum and $\alpha(j, i)$ is the learning rate and is dynamically lowered as the training progresses.

Evaluation metrics

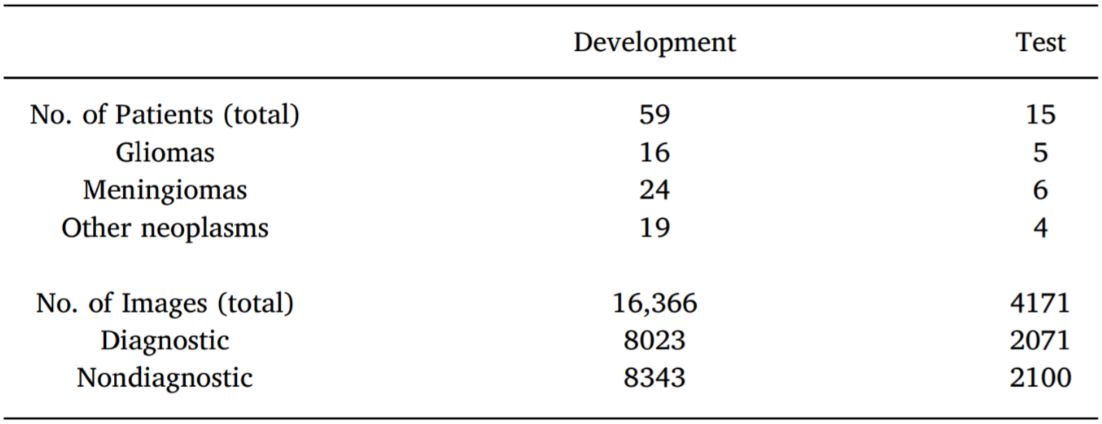
In model performance estimation, we calculated the *loss, accuracy, sensitivity, specificity and area under receiver operating characteristics (ROC) curve (AUC)*. Assuming the state of being a diagnostic image as positive and being nondiagnostic as negative, *sensitivity* determines the model ability to detect diagnostic images and *specificity* determines its ability to detect nondiagnostic images. *Accuracy* determines general capability of a model to detect diagnostic and nondiagnostic images correctly (Metz, 1978).

EXPERIMENTAL SETUP

Dataset preparation

Our dataset included 20,734 CLE images from 74 brain tumor cases. For each CLE image, the diagnostic quality was determined by the experts in the initial review. To match the size of CLE images and the network input shape, images were resized to 256×256. The dataset was then divided into two main subsets on patient level: *development (dev)* and *test*. All results reported in the following sections were based on the model prediction on images from new patients.

The total number of patients and images used at each stage are presented in Table 2.1. Each subset contains images from various tumor types (mainly from gliomas and meningiomas). The dev set will be available online. The test set was isolated all through the model development and was accessed only in the test phase.



	Development	Test
No. of Patients (total)	59	15
Gliomas	16	5
Meningiomas	24	6
Other neoplasms	19	4
No. of Images (total)	16,366	4171
Diagnostic	8023	2071
Nondiagnostic	8343	2100

Table 2.1

The composition of images. Patient-based allocation of diagnostic and nondiagnostic images from various neoplasms to model development and testing. Number of patients for each tumor type is also provided.

Model development

After the initial data split, we employed a patient-based 5-fold cross validation for model development. Fifty-nine cases that were allocated for model development were divided into 5 groups. Since CNNs require a large set of hyperparameters to be defined optimally (i.e. initial value of the learning rate and its lowering policy, momentum, batch size, etc.), we used different values with grid searching throughout the model development process. For every set of feasible parameters, we trained the model on 4 folds and validated on the fifth left-out group of patients. The set of hyperparameters which produced the minimum average loss was employed for each set of experiments.

The small dataset size was a main limitation of our study for using CNNs, especially with the patient-level data preparation. Therefore, we counterbalanced this limitation by fine-tuning the pretrained publicly available CNN architectures trained on large computer vision datasets (i.e. ImageNet).

Though the question about *how deep should we fine-tune the pretrained models for optimal results* still remains unanswered, one study tried to answer this question using endoscopy and ultrasound images (Tajbakhsh et al., 2016). Due to the substantial intrinsic dissimilarities between the images in the 2 studies, we performed a similar investigation. Our confocal images have a much higher spatial resolution and are fluorescent images from the brain.

In total, we developed 42 models (30 single models and 12 ensemble models) using two network architectures and three training regimes (deep training, shallow fine-tuning and deep fine-tuning). The experiments are designed in order to practically find the optimal model development pathway that produces the highest performance in the considered clinical application.

Network architectures

Two CNN architectures that have achieved promising results in general computer vision challenges (Krizhevsky et al., 2012; Szegedy et al., 2015) were applied in this study. Other studies have used these networks for analyzing medical images for other applications including endoscopy, ultrasonography (Tajbakhsh et al., 2016), and CT (Shin et al., 2016), confirming the generality of their pretrained features. However, transferability of these features on brain CLE images had yet to be studied. Deeper states of the art networks (e.g. ResNet (He, Zhang, Ren, & Sun, 2015)) were excluded, because the size of our dataset is relatively small and training very deep models would result in overfitting (Yosinski et al., 2014). The input layer of both networks had a size of 256×256.

Network 1 had 5 convolutional layers. The first two convolutional layers had 96 and 256 filters of size 11 and 5 with maximum pooling over 3 and stride of 2. The third, fourth and fifth convolutional layers were connected back to back without any pooling in between. The third convolutional layer had 384 filters of size 3×3×256, the fourth layer had 384 filters of size 3×3×192 and the fifth layer had 256 filters of size 3×3×192 with maximum pooling. For more details please refer to (Krizhevsky et al., 2012).

Network 2 had 22 layers with parameters and 9 inception modules. Each inception module was a combination of filters of size 1, 3, 5 and max pooling with a kernel size of 3 and stride of 1, put together in parallel and the output filter banks concatenated into an input single vector for the next stage. For more details please refer to (Szegedy et al., 2015).

The pretrained model for network 1, exploited in fine-tuning experiments, was the iteration 360,000 snapshot of training the model on ImageNet classification with 1000 classes. The pretrained model for network 2 was iteration 2,400,000 of training on ImageNet classification dataset. Both models are publicly available in Caffe libraries (Jia et al., 2014).

Training regimes

We exercised various training regimes to see how deep fine-tuning should be done in CLE image classification for optimal results. Depending on which layers of the network are being learned through training, we had three regimes.

In regime 1, ***deep training (DT)***, the whole model weights were initialized randomly (training from scratch) and got modified all through the training with nonzero learning rates.

In regime 2, ***shallow fine-tuning (SFT)***, the whole model weights, except the last fully connected layer, were initialized with the corresponding values from the pretrained model and their values were fixed for the period of training. The last fully connected layer was initialized randomly and got tuned during training.

In regime 3, ***deep fine-tuning (DFT)***, all model weights, except for the last fully connected layer, were initialized with the corresponding values from the pretrained model and last fully connected layer was initialized randomly. Throughout the training, all the CNN layers, including the last fully connected layer, were tuned with nonzero learning rates. Our hyperparameter optimization showed that the SFT and DFT experiments required 10 times smaller initial learning rates (0.001) compared to the DT regime (0.01).

We used Gaussian distribution with zero mean and standard deviation of 0.01 to initialize the weights randomly. A previous investigation of different weight initialization techniques (Tajbakhsh et al., 2016) showed no significant performance gain for colonoscopic image classification and polyp detection.

The training process was stopped after 3 epochs of consistent loss increment on the validation dataset to avoid overfitting. We also used a dropout layer (ratio=0.5) and L2 regularization (lambda=0.005).

Ensemble modeling

Let's assume $y_k^n(j)$ is the the value of the k^{th} output layer unit of the j^{th} CNN model in response to the n^{th} input test image. The linear and log-linear ensemble classifier output for the same input would be:

$$Ens_{linear}^n = \arg\ \max 1_k \sum_{j=1}^{l} y_k^n(j)$$

$$Ens_{log-linear}^n = \arg\ \max 1_k \prod_{j=1}^{l} y_k^n(j)$$

where l is the number of CNN models combined to generate the ensemble models.

Model selection was done in two forms: single models and ensemble of models. We selected the top model (with minimum loss on the validation dataset) from each fold of the 5-fold cross validation (**Model 1-5** in Table 2.2). In each network architecture and training regime, we combined the top-5 developed single models to produce two ensembles of models using the arithmetic and geometric mean of their outputs. We created 12 ensemble models ($2^{\text{network architectures}} \times 3^{\text{ training regimes}} \times 2^{\text{ensemble types}}$) in total and compared their performance with single models.

Interobserver study

Each solo and ensemble model developed was tested on the test dataset. The ensemble of network 2 models trained with DFT was also tested on the val-review images to compare human-human and model-human interobserver agreements. The resulting agreement rate (val-rater 1 and 2) was further compared with the initial image review results. The agreement of the model prediction with the initial review was also calculated. The general agreements are compared and discussed later. Kappa analysis was also done for further validation.

Gold standard ground truth for the val-review images was defined by majority voting (see Figure 2.3). The agreements of the third rater with the gold standard and the proposed ensemble model with the gold standard is calculated and compared as well in Table 2.3.

Unsupervised histological feature localization

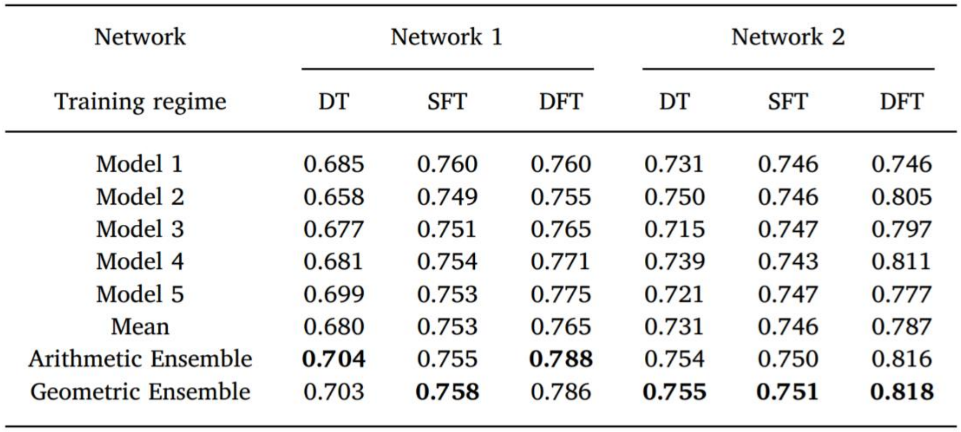
For localization of the histological features, we examined the neural activation at two sites. First, the activation of neurons in the first convolutional layer of the network 1 were visualized and the 96 feature planes were saved for review by a neurosurgeon.

To this end, the CLE image was given to the trained model as an input. A feedforward of the model was run consequently in Caffe to activate neurons at different layers while predicting the diagnostic value of the input image. Using Caffe Python interface, each neuron (feature plane) of the first layer was accessed with indexing and saved as a separate image for review. Neurons in the first layer that presented high activation to the location of cellular structures in the input image were selected and seemed to be consistent with diverse diagnostic images.

Secondly, we applied a sliding window of size 227×227 pixels (size of network 1 input after input cropping) with stride of 79 pixels over the diagnostic CLE images 1024×1024 pixels). The result was a 10×10 matrix that provided the diagnostic value of different locations of the input image (*diagnostic map*). The locations of input images corresponding to the highest activations of the diagnostic map were detected and marked with a bounding box. The detected features using each of these two ways were further reviewed by a neurosurgeon.

RESULTS AND DISCUSSION

We developed 42 models and tested them on 4,171 test images; accuracy rates (agreement with the initial review) are presented in Table 2.2. We found that network 2 resulted in more precise predictions about the diagnostic quality of images than network 1 when DT and DFT training regimes were used, while SFT training regime resulted in slightly better accuracy of network 1, compared to network 2. Therefore, network 2 architecture is a better feature extraction tool for CLE images, since it concatenates multi-scale features inside its inception modules.



Network	Network 1			Network 2		
Training regime	DT	SFT	DFT	DT	SFT	DFT
Model 1	0.685	0.760	0.760	0.731	0.746	0.746
Model 2	0.658	0.749	0.755	0.750	0.746	0.805
Model 3	0.677	0.751	0.765	0.715	0.747	0.797
Model 4	0.681	0.754	0.771	0.739	0.743	0.811
Model 5	0.699	0.753	0.775	0.721	0.747	0.777
Mean	0.680	0.753	0.765	0.731	0.746	0.787
Arithmetic Ensemble	**0.704**	0.755	**0.788**	0.754	0.750	0.816
Geometric Ensemble	0.703	**0.758**	0.786	**0.755**	**0.751**	**0.818**

Table 2.2

Accuracy analysis on the test dataset. For each network, the ensemble of DFT models makes the most accurate predictions.

Ensemble or solo model?

We did an ROC analysis for each of the two networks and three training regimes to see how the ensemble of models performed compared to the single models. Figure 2.2 presents the ROC curves and corresponding AUC values for each ensemble model and the mean of single models. The AUC value increased by 2% for both networks with DT and DFT when the ensemble is applied instead of the single model. This effect gets smaller with network 1 SFT and becomes negligible with network 2 SFT. The two arithmetic and geometric ensemble models produced roughly similar results (paired t-test: *P value $<$ 0.05*).

SFT models displayed less sensitivity to the ensemble effect compared to DT and DFT. This is not surprising since they represented identical models except in the softmax layer which has been adjusted during training. This minimized the diversity of the models and consequently the ensemble effect compared to DT and DFT. Further, the larger degrees of freedom for changing the parameters in DT and DFT regimes (parameters of all the layers in each network) resulted in a more diverse set of models compared to SFT which allowed parameter tuning only for the last layer. We speculate that the larger diversity of DT and DFT was responsible for their higher sensitivity to ensemble modeling. Though, further experiments on more datasets are required to assess this speculation.

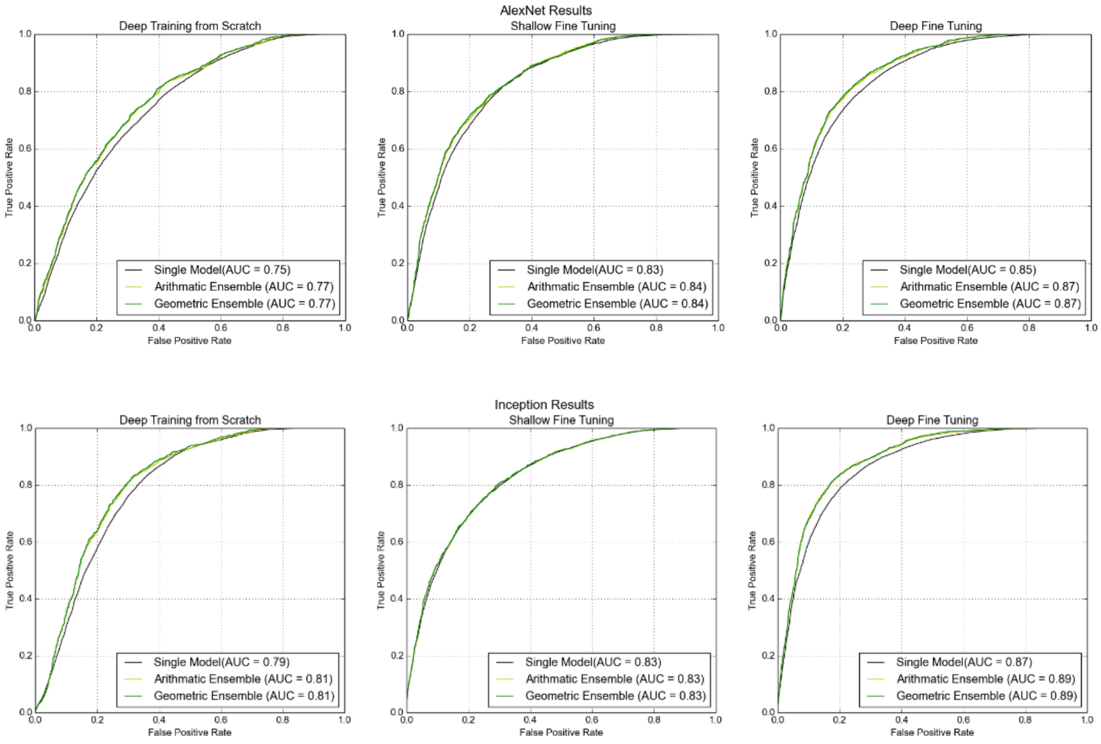


Figure 2.2

ROC analysis of ensemble effect on different training regimes. For both networks, the improvement was more noticeable with DT and DFT regimes. The arithmetic and geometric ensemble performed similarly. Neither of the two ensembles could improve network 2 trained with SFT.

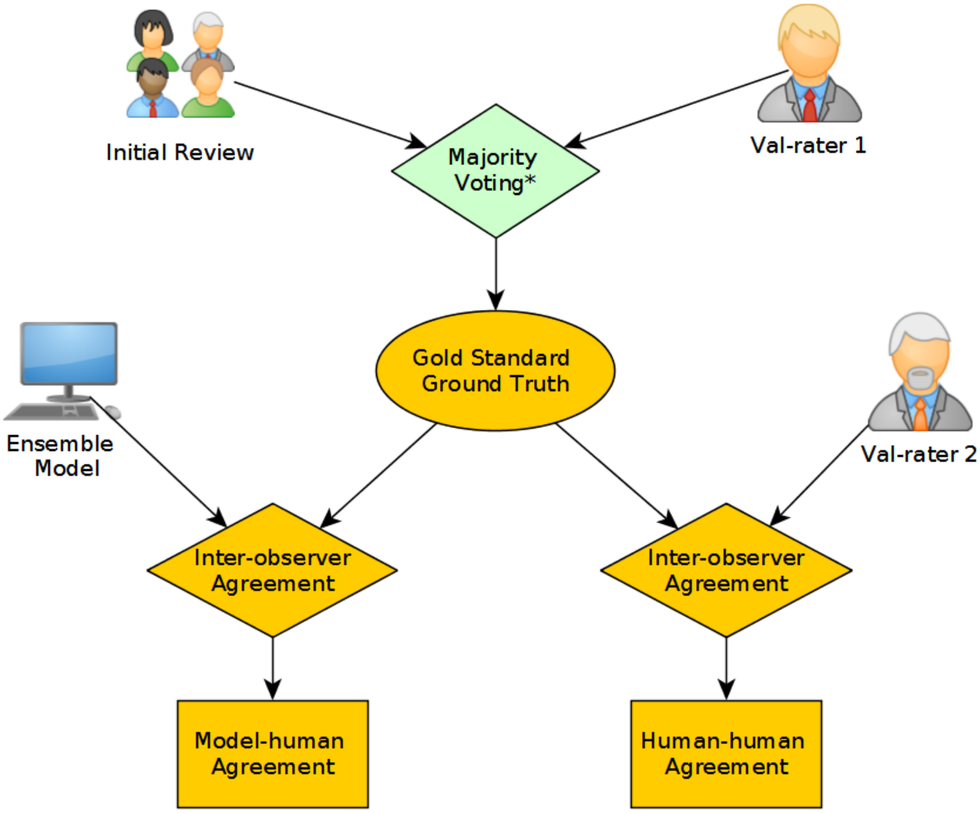


Figure 2.3

A schematic diagram of interobserver study. Gold standard was defined using the initial review and one of the val-raters (here val-rater 1). Then, the agreement of the ensemble model and the other val-rater (here val-rater 2) with the gold standard is calculated to compare the human-human with model-human agreement. *If the initial review and the val-rater 1 agreed on an image, it is added to the gold standard, otherwise it is disregarded.

Which training regime: DT, SFT or DFT?

Figure 2.4 displays the results of ROC analysis when comparing the three training regimes in each network architecture and single/ensemble states. In all paired comparisons, DFT outperformed SFT and SFT outperformed the DT regime (paired t-test: *P value $<$ 0.05*).

We traced the AUC elevation from DT to DFT regime to see how much of it corresponded to the transformation of DT to SFT and SFT to DFT. For network 1, 70% - 80% of the improvement occurred in the DT to SFT transformation, depending on whether the model is single or ensemble. For network 2 ensemble model (right bottom of Figure 2.4), however, the AUC improvement caused by transforming the training regime from DT to SFT (2%) is only 25% of the total improvement from DT to DFT. For network 2 single model the AUC improvement was evenly divided between the two transformations.

Our results from this experiment indicated that for our dataset, fine-tuning the classification layer contributed the most to the improvement in network 1, and fine-tuning feature extractor layers had a smaller contribution. However, for network 2, fine-tuning the feature extractors contributed equally or more than modifying the classification layer. A previous study on thoraco-abdominal lymph node detection (Shin et al., 2016) observed similar results: deep fine tuning GoogLeNet produced better improvement in detecting lymph nodes than deep fine tuning AlexNet. The relationship between different training regimes and the network architectures has yet to be clarified in future studies, however, the difference between how these two networks responded to SFT and DFT might be partially linked to their architectural differences. Yosinski et al. studied the transferability of CNN features and concluded that deep fine tuning a model with a large number of parameters on a small dataset would make the model overfit (Yosinski et al., 2014). Since the number of parameters in network 1 is 12 times larger than network 2 (Szegedy et al., 2015), it would overfit more easily on our relatively small dataset. Therefore, the lower performance of network 1 compared to network 2 in DT and DFT might be due to its overfitting.

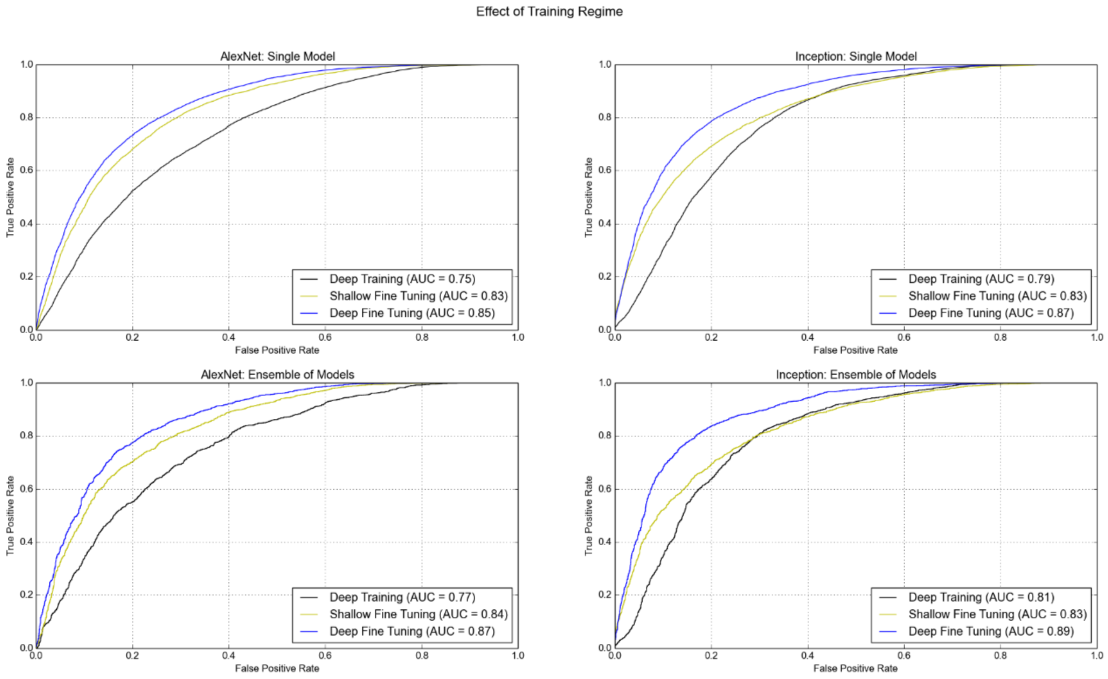


Figure 2.4

ROC analysis of training regime effect on single and ensemble models. The AUC value for DFT was greater than the SFT and SFT is greater than DT, although the effect size varied.

Two factors may explain why network 2 produced lower performance compared with network 1 in SFT: fragile co-adaptation of network 2 neurons and higher generality of network 1 features. If neighboring neurons of a CNN become co-adapted during pretraining (i.e. fragile co-adaptation), the network performance drops when few layers are frozen and the others are fine-tuned (Yosinski et al., 2014). The generality of CNN features is determined by how well they transfer from one dataset/task to another dataset/task. This confirms previous studies on generality of network 1 features in previous studies (Razavian, Azizpour, Sullivan, & Carlsson, 2014; Shin et al., 2016; Tajbakhsh et al., 2016). However, further experiments on larger datasets and deeper networks are essential to verify this.

Histological features localization

After being reviewed by a neurosurgeon, 8 out of the total 384 reviewed colored neuron activation maps from the first layer were selected for 4 diagnostic CLE images representative for glioma. The selected activation maps were from the same neurons (i.e. neuron 22 and 24) and highlighted diagnostic tissue architecture patterns in warm colors. Particularly, several maps emphasized regions of optimal image contrast, where hypercellular and abnormal nuclear features could be identified, and would serve as diagnostic features for image classification (Figure 2.5, columns 2 and 4). Additionally, the sliding window method was able to identify diagnostic aggregates of abnormally large malignant glioma cells and atypically hypercellular areas (Figure 2.5, third column).

Activation of the neurons 22 and 24 in the first convolutional layer (conv1) were found to highlight areas with increased fluorescein signal, a sign specific to brain tumor regions. Increased fluorescent signal on CLE images represent areas with blood brain barrier disruption which correspond to the tumor areas visible on a contrast enhanced MR imaging. In Figure 2.5, (d) highlights the blood-brain-barrier disruption and (b) marks the cluster of cells which are both indicative of cancer. The red boxes in (c) identify some other diagnostic regions. Although these areas might have some overlaps, not everything detected in (b) and (d) was successfully localized in (c).

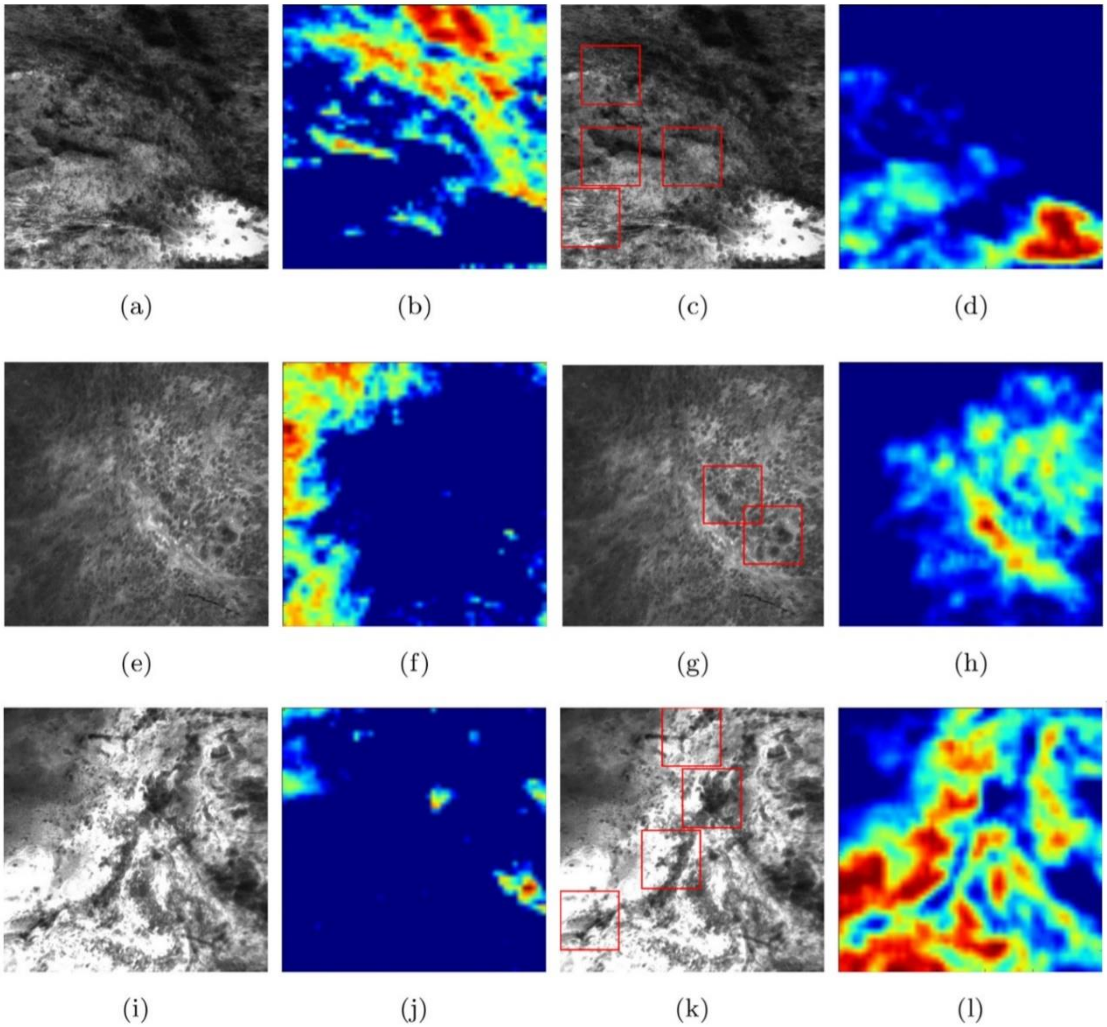


Figure 2.5

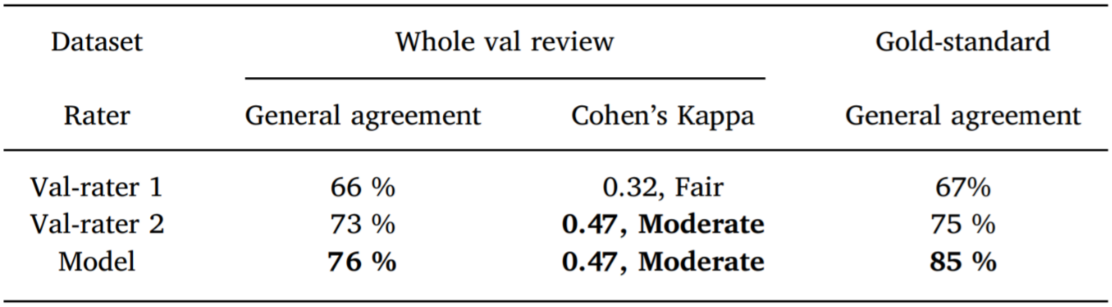
Unsupervised semantic localization of the CLE histopathological features. First column (a, e, i) shows the input CLE images from human glioblastoma obtained intraoperatively. Second column (b, f, j) displays activation of neurons from the first layer (conv1, neuron 24) (shallow features); it highlights some of the cellular areas present in the image. Third column (c, g, k) illustrates diagnostic regions of interest identified with the sliding window approach. The boxed regions represent high activation of the deepest network neuron. Fourth column (d, h, l) contains images extracted from conv1 activation (neuron 22), representative of the high fluorescence signal, a diagnostic sign of blood-brain barrier disruption and leakage of fluorescent agent from the vessels into the extracellular space.

Interestingly, the sliding window method and selected colored activation maps were not distracted or deceived by the red blood cells contamination, as they mostly highlighted tumor and brain cells rather than hypercellular areas due to bleeding. The proposed feature localization approach may be useful in the future to aid in the identification of not only the diagnostic frames, but also directing the surgeon's attention to the image parts containing major histopathological features.

Inter-rater agreement

Table 2.3 demonstrates the agreement between each of the val-raters and the initial review on the whole val review dataset and the gold standard subset (explained in Figure 2.3). The model agreement with the initial review is larger than each val-rater's agreement with the initial review. This suggests that the model has successfully learned the histological features of the CLE images that are more probable to be noticed by the neurosurgeons when the corresponding H & E-stained histological slides were also provided for reference.

To consider images from the val review set that the majority of raters agreed on, that is one of the val-raters agreed on with the initial review, we used the gold standard subset. The gap between the model-human and human-human agreements became even more evident (19% for val-rater 1 and 9% for val-rater 2) with the gold standard subset (Table 2.3, column 4).



Dataset	Whole val review		Gold-standard
Rater	General agreement	Cohen’s Kappa	General agreement
Val-rater 1	66 %	0.32, Fair	67%
Val-rater 2	73 %	**0.47, Moderate**	75 %
Model	**76 %**	**0.47, Moderate**	**85 %**

Table 2.3

Interobserver study results. The model reached promising agreement with the ground truth both on the val review dataset and the gold standard subset.

CONCLUSION AND FUTURE WORK

This paper presents a deep CNN based approach that can automatically detect the diagnostic CLE images from brain tumor surgery. We used a manually annotated in-house dataset to train and test this approach. Our results showed that both deep fine-tuning and creating an ensemble of models could enhance the performance; but only their combination could reach the maximum accuracy. The ensemble effect was stronger in DT and DFT than SFT developed models. The proposed method was also able to localize some histological features of diagnostic images. Ultimately, Table 2.3 indicates that the proposed ensemble of deeply fine-tuned models could detect the diagnostic images with a promising agreement to the gold-standard. Other confocal imaging techniques may be aided by such deep learning models. Confocal reflectance microscopy (CRM) has been studied (Mooney et al., 2017) for rapid, fluorophore-free evaluation of pituitary adenoma biopsy specimens ex vivo. CRM allows preserving the biopsy tissue for future permanent analysis, immunohistochemical studies, and molecular studies.

Parallel computational frameworks (Yan, Zhang, Xu, Dai, Li, et al., 2014; Yan, Zhang, Xu, Dai, Zhang, et al., 2014) for high definition video coding, could speed up our ensemble model, especially when the ensemble contains larger number of parallel models. Recent studies have proposed to jointly learn different levels of abstract features and combine them for visual recognition (X. Zhang et al., 2016) and categorization (Yao, Zhang, Zhang, Li, & Tian, 2016). Supervised and unsupervised hash coding methods have also received increasing attention in image content analysis (Yan, Xie, Yang Dongbao, et al., 2017; L. Zhang, Zhang, Gu, Tang, & Tian, 2014). (Yan, Xie, Liu, et al., 2017) proposed an effective method for text detection in images with complex background. In future work, we will investigate the usefulness of these methods in understanding CLE images.

Continued use of unsupervised image segmentation methods to detect meaningful histological features from confocal brain tumor images will likely allow for more rapid and detailed diagnosis.

References

Abdi, A., Luong, C., Tsang, T., Jue, J., Hawley, D., Fleming, S., … Abolmaesumi, P. (2017). Automatic Quality Assessment of Echocardiograms Using Convolutional Neural Networks: Feasibility on the Apical Four-chamber View. *IEEE Transactions on Medical Imaging*.

Belykh, E., Martirosyan, N. L., Yagmurlu, K., Miller, E. J., Eschbacher, J. M., Izadyyazdanabadi, M., … Preul, M. C. (2016). Intraoperative fluorescence imaging for personalized brain tumor resection: Current state and future directions. *Frontiers in Surgery*, *3*.

Charalampaki, P., Javed, M., Daali, S., Heiroth, H.-J., Igressa, A., & Weber, F. (2015).

Confocal Laser Endomicroscopy for Real-time Histomorphological Diagnosis: Our Clinical Experience With 150 Brain and Spinal Tumor Cases. *Neurosurgery*, *62*, 171–176.

Christodoulidis, S., Anthimopoulos, M., Ebner, L., Christe, A., & Mougiakakou, S. (2017). Multisource Transfer Learning With Convolutional Neural Networks for Lung Pattern Analysis. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 76–84.

Ciregan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 3642–3649).

Ciresan, D., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems* (pp. 2843–2851).

Dietterich, T. G., & others. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems*, *1857*, 1–15.

Foersch, S., Heimann, A., Ayyad, A., Spoden, G. A., Florin, L., Mpoukouvalas, K., … Charalampaki, P. (2012). Confocal laser endomicroscopy for diagnosis and histomorphologic imaging of brain tumors in vivo. *PLoS One*, *7*(7), e41760.

Gao, Y., Maraci, M. A., & Noble, J. A. (2016). Describing ultrasound video content using deep convolutional neural networks. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on* (pp. 787–790).

Ghafoorian, M., Karssemeijer, N., Heskes, T., Bergkamp, M., Wissink, J., Obels, J., … others. (2017). Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. *NeuroImage: Clinical*, *14*, 391–399.

Greenspan, H., van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, *35*(5), 1153–1159.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv Preprint arXiv:1512.03385*.

Izadyyazdanabadi, M., Belykh, E., Martirosyan, N., Eschbacher, J., Nakaji, P., Yang, Y., & Preul, M. C. (2017). Improving utility of brain tumor confocal laser endomicroscopy: Objective value assessment and diagnostic frame detection with convolutional neural networks. In *Progress in Biomedical Optics and Imaging - Proceedings of SPIE* (Vol. 10134). https://doi.org/10.1117/12.2254902

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., … Darrell, T.

(2014). Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv Preprint arXiv:1408.5093*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In *Advances in neural information processing systems* (pp. 231–238).

Kumar, A., Kim, J., Lyndon, D., Fulham, M., & Feng, D. (2017). An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 31–40.

Kumar, A., Sridar, P., Quinton, A., Kumar, R. K., Feng, D., Nanan, R., & Kim, J. (2016). Plane identification in fetal ultrasound images using saliency maps and convolutional neural networks. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on* (pp. 791–794).

LeCun, Y., & Bengio, Y. (1998). The handbook of brain theory and neural networks. In M. A. Arbib (Ed.) (pp. 255–258). Cambridge, MA, USA: MIT Press. Retrieved from http://dl.acm.org/citation.cfm?id=303568.303704

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *arXiv Preprint arXiv:1702.05747*.

Mahapatra, D., Roy, P. K., Sedai, S., & Garnavi, R. (2016). Retinal Image Quality Classification Using Saliency Maps and CNNs. In *International Workshop on Machine Learning in Medical Imaging* (pp. 172–179).

Martirosyan, N. L., Eschbacher, J. M., Kalani, M. Y. S., Turner, J. D., Belykh, E., Spetzler, R. F., … Preul, M. C. (2016). Prospective evaluation of the utility of intraoperative confocal laser endomicroscopy in patients with brain neoplasms using fluorescein sodium: experience with 74 cases. *Neurosurgical Focus*, *40*(3), E11.

Metz, C. E. (1978). Basic principles of ROC analysis. In *Seminars in nuclear medicine* (Vol. 8, pp. 283–298).

Mooney, M. A., Georges, J., Yazdanabadi, M. I., Goehring, K. Y., White, W. L., Little, A. S., … Eschbacher, J. M. (2017). Immediate ex-vivo diagnosis of pituitary adenomas using confocal reflectance microscopy: a proof-of-principle study. *Journal of Neurosurgery*.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on* (pp. 512–519).

Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging. *IEEE Transactions on Medical Imaging*.

Sanai, N., Eschbacher, J., Hattendorf, G., Coons, S. W., Preul, M. C., Smith, K. A., … Spetzler, R. F. (2011). Intraoperative confocal microscopy for brain tumors: a feasibility analysis in humans. *Neurosurgery*, *68*, ons282--ons290.

Shi, J., Zheng, X., Li, Y., Zhang, Q., & Ying, S. (2017). Multimodal Neuroimaging Feature Learning with Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer's Disease. *IEEE Journal of Biomedical and Health Informatics*.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., … Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285–1298.

Sirinukunwattana, K., Raza, S. E. A., Tsang, Y.-W., Snead, D. R. J., Cree, I. A., & Rajpoot, N. M. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, *35*(5), 1196–1206.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Suk, H.-I., & Shen, D. (2016). Deep Ensemble Sparse Regression Network for Alzheimer's Disease Diagnosis. In *International Workshop on Machine Learning in Medical Imaging* (pp. 113–121).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Transactions on Medical Imaging*, *35*(5), 1299–1312.

Yan, C., Xie, H., Liu, S., Yin, J., Zhang, Y., & Dai, Q. (2017). Effective Uyghur language text detection in complex background images for traffic prompt identification. *IEEE Transactions on Intelligent Transportation Systems*.

Yan, C., Xie, H., Yang Dongbao, Yin, J., Zhang, Y., & Dai, Q. (2017). Supervised Hash Coding With Deep Neural Network for Environment Perception of Intelligent Vehicles. *IEEE Transactions on Intelligent Transportation Systems*.

Yan, C., Zhang, Y., Xu, J., Dai, F., Li, L., Dai, Q., & Wu, F. (2014). A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors. *IEEE Signal Processing Letters*, *21*(5), 573–576.

Yan, C., Zhang, Y., Xu, J., Dai, F., Zhang, J., Dai, Q., & Wu, F. (2014). Efficient parallel framework for HEVC motion estimation on many-core processors. *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(12), 2077–2089.

Yao, H., Zhang, S., Zhang, Y., Li, J., & Tian, Q. (2016). Coarse-to-fine description for fine-grained visual categorization. *IEEE Transactions on Image Processing*, *25*(10), 4858–4872.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).

Zehri, A., Ramey, W., Georges, J., Mooney, M., Martirosyan, N., Preul, M., & Nakaji, P. (2014). Neurosurgical confocal endomicroscopy: A review of contrast agents, confocal systems, and future imaging modalities. *Surgical Neurology International*, *5*, 60.

Zhang, L., Zhang, Y., Gu, X., Tang, J., & Tian, Q. (2014). Scalable similarity search with topology preserving hashing. *IEEE Transactions on Image Processing*, *23*(7), 3025–3039.

Zhang, X., Zhang, H., Zhang, Y., Yang, Y., Wang, M., Luan, H., … Chua, T.-S. (2016). Deep fusion of multiple semantic cues for complex event recognition. *IEEE Transactions on Image Processing*, *25*(3), 1033–1046.

Zhao, J., Zhang, M., Zhou, Z., Chu, J., & Cao, F. (2016). Automatic detection and classification of leukocytes using convolutional neural networks. *Medical & Biological Engineering & Computing*, 1–15.

Zhao, L., & Jia, K. (2016). Multiscale cnns for brain tumor segmentation and diagnosis. *Computational and Mathematical Methods in Medicine*, *2016*.

Zhou, Z.-H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artificial Intelligence*, *137*(1–2), 239–263.

CHAPTER 3

The following chapter has been published as a chapter in the book Medical Image Computing and Computer Assisted Intervention – MICCAI 2018.

(DOI: 10.1007/978-3-030-00934-2_34)

CHAPTER 3

WEAKLY-SUPERVISED LEARNING-BASED FEATURE LOCALIZATION IN CONFOCAL LASER ENDOMICROSCOPY GLIOMA IMAGES

Izadyyazdanabadi M, Belykh E, Cavallo C, Zhao X, Gandhi S, Moreira L, Eschbacher J,
Nakaji P, Preul MC, Yang Y

ABSTRACT

Confocal Laser Endomicroscopy (CLE) is novel handheld fluorescence imaging technology that has shown promise for rapid intraoperative diagnosis of brain tumor tissue. Currently CLE is capable of image display only and lacks an automatic system to aid the surgeon in diagnostically analyzing the images. The goal of this project was to develop a computer-aided diagnostic approach for CLE imaging of human glioma with feature localization function. Despite the tremendous progress in object detection and image segmentation methods in recent years, most of such methods require large annotated datasets for training. However, manual annotation of thousands of histopathology images by physicians is costly and time consuming. To overcome this problem, we constructed a Weakly-Supervised Learning (WSL)-based model for feature localization that trains on image-level annotations, and then localizes incidences of a class-of-interest in the test image. We developed a novel convolutional neural network for diagnostic features localization from CLE images by employing a novel multiscale activation map that is laterally inhibited and collaterally integrated. To validate our method, we compared the model output to the manual annotation performed by four neurosurgeons on test images. The model achieved 88% mean accuracy and 86% mean

intersection over union on intermediate features and 87% mean accuracy and 88% mean intersection over union on restrictive fine features, while outperforming other state of the art methods tested. This system can improve accuracy and efficiency in characterization of CLE images of glioma tissue during surgery, and may augment intraoperative decision making regarding the tumor margin and improve brain tumor resection.

INTRODUCTION

Rapid intraoperative interpretation of suspected brain tumor tissue is of paramount importance for planning the treatment and guiding the neurosurgeon towards the optimal extent of tumor resection. Handheld, portable Confocal Laser Endomicroscopy (CLE) is being explored as a fluorescence imaging technique for its ability to image histopathological features of tissue at cellular resolution in real time during brain tumor surgery (Belykh et al., 2016; Eschbacher et al., 2012; Foersch et al., 2012; Martirosyan et al., 2016). CLE systems can acquire up to 20 images per second, with areas in the tumor resection bed interrogated as an “optical biopsy”. Hundreds of images may be acquired showing thousands of cells, but the images may be affected with artifacts such as red blood cells (for CLE systems operating in the blue laser range) and motion distortion, making them complicated to analyze. Although images may be interpreted as largely artefactual, detailed inspection often reveals image areas that may be diagnostic. CLE images present a new fluorescent image environment for the pathologist. Augmenting CLE technology with a computer aided system that can rapidly highlight image regions that may reveal malignant or spreading tumor would have great impact on intraoperative diagnosis. This is relevant for tumors such as gliomas where discrimination of margin regions is key to achieve maximal safe resection, which has been correlated with increased patient survival duration (Almeida, Chaichana, Rincon-Torroella, & Quinones-Hinojosa, 2015; Sanai, Polley, McDermott, Parsa, & Berger, 2011).

Recent studies have shown that off-the-shelf Convolutional Neural Networks (CNNs) can be used effectively for classifying CLE images based on their diagnostic value (M. Izadyyazdanabadi et al., 2017; Mohammadhassan Izadyyazdanabadi et al., 2018) and tumor type (N. Murthy et al., 2017). However, feature localization models have not been previously applied to CLE images. Feature localization models based on fully supervised learning require large number of images for object-level annotation of the features, which is expensive and time consuming. To overcome this limitation, we used a weakly-supervised localization (WSL) approach. A WSL approach allowed the model to learn and localize the class-specific features from image-level labels.

A few groups have recently applied WSL approaches to medical images, including placenta scans (Qi, Collins, & Noble, 2017), whole-slide images of colorectal cancer (Korbar et al., 2017), diabetic retinopathy (Gondal, Köhler, Grzeszick, Fink, & Hirsch, 2017), microscopic cellular images (Sailem, Arias--Garcia, Bakal, Zisserman, & Rittscher, 2017), and lung computed tomography scans (Feng, Yang, Laine, & Angelini, 2017). Here, we present a novel model for detection of histological features of glioma on CLE images trained on a dataset of CLE images acquired during brain surgery for this invasive tumor. The architecture included end-to-end Multi-Layer Class Activation Map (MLCAM) with Lateral Inhibition (LI) and Collateral Integration (CI) of the glioma feature localizer neurons. The model was able to segment the CLE images semantically by disentangling class-specific discriminative features that can complement interpretation by the physicians. Performance of the model was assessed by comparing its output to CLE image segmentations performed by neurosurgeons and other deep learning models. Additionally, we validated the significance of the MLCAM, LI and CI architecture components on the overall performance of the model. The model localized known diagnostic CLE features and revealed new CLE features that correlated with the final classification and importantly which were not previously recognized by the expert reviewers.

Unlike previous models that require patch labeling (Korbar et al., 2017) or an extra step for creating the activation maps during testing (Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2016), our model is solely trained based on the whole image-level labels. Furthermore, we did not limit the network to localize features that are already known phenotypes to the physicians (Feng et al., 2017; Sailem et al., 2017). CLE images are relatively novel to the pathology tissue diagnosis workflow. Although the tissue architecture suggestive for a certain tumor type can be identified on CLE images (Belykh et al., 2016; Eschbacher et al., 2012; Foersch et al., 2012; Martirosyan et al., 2016), detailed characteristic brain tumor patterns for CLE images are not yet well described. Therefore, we used a more general concept (glioma diagnostic vs. nondiagnostic) that includes a range of known histological diagnostic elements (i.e., large nucleus, mitotic figures, hypercellularity, etc.) and allows for discovery of previously unrecognized features that may correlate with final image classification. Further investigation of detected features may deepen the understanding of glioma histopathological phenotypes in CLE images, consequently improving their theranostic implications. We intend to make our dataset along with the codes publicly available online to facilitate further research.

METHODS

We constructed a WSL-based model to generate glioma Diagnostic Feature Maps (DFM) from CLE images, which includes three main components (Figure 3.1): 1) Customized CNN architecture with new design of CAM at different CNN layers. 2) Lateral inhibition (LI) mechanism that suppresses the activation of DFM at locations where its competitor, nondiagnostic feature map (NFM), also exhibit high activation. 3) Collateral integration (CI) mechanism that amplifies activation of DFM at locations where its allies at other layers also have high activations.

For an input image I_m supplied to the CNN, the class scores (S_{D} for diagnostic and S_{N} for nondiagnostic) are defined from three layers via global pooling of discriminative regions estimated in each activation map (DFM, NFM). The class scores achieved from each layer, are then passed to independent softmax layers. The three predictions (probability of I_m being diagnostic (D) and nondiagnostic (ND)) achieved from the softmax layers are streamed into three multinomial logistic loss layers and inject the weight update into the CNN during backpropagation. The total loss is calculated by summing the three loss values.

New design of class activation map (CAM)

To produce the CAM from each layer, a new convolutional layer is stacked to sum its weighted feature planes. Formally, the DFM and NFM at location (x, y) achieved from layer z^j, are defined as:

$$DFM\left(x, y, z^j\right) = \sum_l w_{k^1}^{z^j} f_l\left(x, y, z^j\right), \quad (1)$$

$$NFM\left(x,y,z^j\right) = \sum_l w_{k^0}^{z^j} f_l\left(x,y,z^j\right), \quad (2)$$

where $f_l\left(x,y,z^j\right)$ is the activation of l^{th} feature plane of layer z^j at location (x, y) and $w_{k^1}^{z^j}$ and $w_{k^0}^{z^j}$ are the weights to produce the DFM and NFM, respectively. By applying GAP and then softmax function on DFM and NFM, the classification scores for different classes are calculated at each layer. Therefore, the softmax input for diagnostic (S_{D}) and nondiagnostic (S_{N}) class at layer z^j can be formulated as:

$$S_{\mathrm{D}} = \frac{1}{W^{z^j} \times H^{z^j}} \sum_{x,y} \mathrm{DFM}\ (\mathrm{x}, \mathrm{y}, z^j) = \frac{1}{W^{z^j} \times H^{z^j}} \sum_{x,y} \sum_l w_{k^1}^{z^j} f_l\left(x,y,z^j\right), \qquad (3)$$

$$S_{\mathrm{N}} = \frac{1}{W^{z^j} \times H^{z^j}} \sum_{x,y} \mathrm{NFM}\ (\mathrm{x}, \mathrm{y}, z^j) = \frac{1}{W^{z^j} \times H^{z^j}} \sum_{x,y} \sum_l w_{k^0}^{z^j} f_l\left(x,y,z^j\right), \qquad (4)$$

where W^{z^j} and H^{z^j} are the width and height of DFM and NFM at layer z^j. With the novel design of MLCAM, DFM, and NFM are produced in every forward pass and are updated through backpropagation. Furthermore, producing DFM from deeper layers empowers the overall predictive power of the model (i.e. labeling the detected region as diagnostic or nondiagnostic), while DFM from shallower layers allows larger spatial resolution and more precise detection of fine regions.

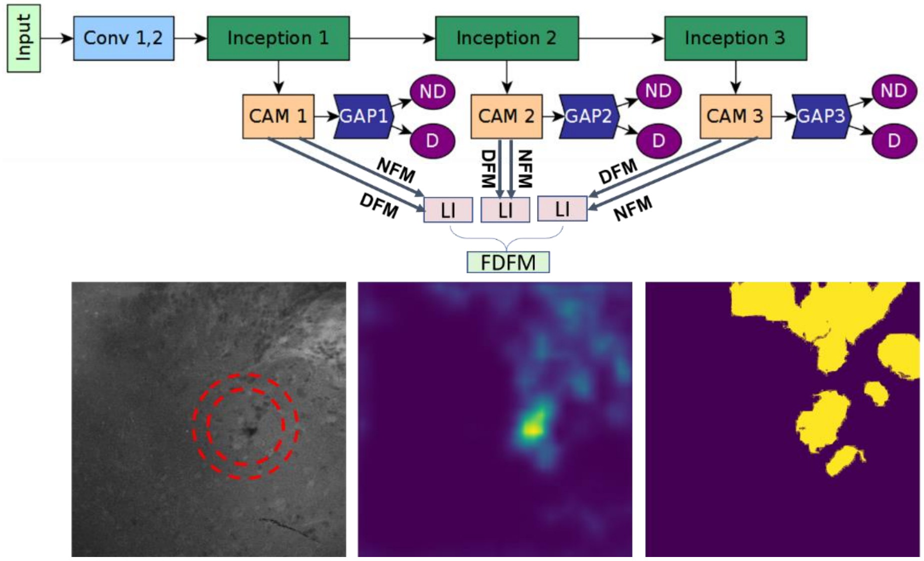


Figure 3.1

Network architecture with Lateral Inhibition (LI) and Collateral Integration (CI) components for weakly supervised localization of glioma diagnostic features. Bottom image shows a CLE image along with the final diagnostic feature map generated by the model.

Lateral inhibition and collateral integration of localizer neurons

During the computation of DFM and NFM, some locations might be activated in both feature maps, which indicates the model's confusion about the diagnostic value of those regions. The activation of DFM is downregulated in these regions, using NFM activations. This mechanism is known as neuronal lateral inhibition in neurobiology (Baars & Gage, 2010)). Furthermore, we upregulate the activation of regions which had higher recurrence of activation by integrating DFMs achieved from different layers. To combine these two neural interactions, we compose the following equation to produce the Final DFM (FDFM):

$$FDFM(x,y) = \sum_{z^i,\, z^j (i \neq j)} \left[DFM'\left(x,y,z^i\right) - DFM'\left(x,y,z^i\right). NFM'\left(x,y,z^i\right)\right].\ \left[DFM'\left(x,y,z^j\right) - DFM'(x,y,z^j). NFM'(x,y,z^j)\right], \tag{5}$$

where $DFM'(x,y,z^i)$ and $NFM'(x,y,z^i)$ are the value of normalized diagnostic and nodiagnostic feature maps achieved from layer z^i, after up-sampling to the original input image size. As shown in Eq. (5), the downregulation for layer z^i is implemented by subtracting the $DFM(x,y,z^i). NFM(x,y,z^i)$ term, which represents the confusing regions at this layer, from $DFM(x,y,z^i)$**.** Lastly, $FDFM(x,y)$ is also normalized. Figure

3.1 presents the developed network's architecture. The three inception modules have the same architecture, each combines filters of size 1 × 1, 3 × 3, 5 × 5 in parallel, and concatenates the outputs from each filter into a single tensor (Szegedy et al., 2015).

EXPERIMENTAL SETUP AND RESULTS

To train our model on image-level annotations, first, a "classification dataset" was created. The CLE images were acquired with an Optiscan 5.1 CLE as described previously (Martirosyan et al., 2016). The classification dataset included 6,287 CLE images (3,126 diagnostic and 3,161 nondiagnostic) from 20 patients with glioma brain tumors. If the CLE image depicted any distinguishable diagnostic features, it was labeled as diagnostic and otherwise as nondiagnostic. Table 3.1 shows the composition of the classification dataset and the number of images used in each stage.

	D	ND	All
Train	1714	1729	3443
Val	487	511	998
Test	925	921	1846
Total	3126	3161	6287

Table 3.1

Number of Diagnostic (D) and Nondiagnostic (ND) images used for training, validation (Val), and test stage is presented.

The classification dataset was divided on a patient level for model development and test (12 cases for training, 4 cases for validation and 4 cases isolated for testing). Stochastic Gradient Descent (SGD) with an initial learning rate of 0.001 and momentum of 0.9 was used to optimize the model's parameters. Learning rate decay policy was set to step function with a gamma of 0.9 and step size of 500 iterations. Image cropping and rotation were not used for augmentation because these might harm the validity of images. Since the diagnostic features could be very small, not every crop of a diagnostic image would be diagnostic. Also, there is no guarantee that the acquired CLE images are rotation invariant (e.g. the surgeons' preference for holding the CLE probe). Training batch size was set to 15 images and it took 22,000 iterations to achieve the model with the minimum loss on classification of validation images. All the experiments were performed in Caffe (Jia et al., 2014) deep learning framework, using a GeForce GTX 980 Ti GPU (6 GB memory).

The classification accuracy of the model was 84% on the test set (sensitivity = 83.8%, specificity = 84.1%). To validate the efficacy of the WSL model, we tested the following three hypotheses. *First*, the model can correctly segment the image regions which have features that are indicative of glioma, confirmed by physicians at different scales (i.e., medium-sized intermediate and small-sized restrictive scales) and without much reliance on previous exposure (i.e., images from training, validation and test stages). *Second*, the new components utilized (MLCAM, LI, and CI) increase the performance of the model in detecting the features (especially restrictive features) compared to the other state of the art WSL methods that lack them and removing any of these would affect the model performance negatively. *Third*, the developed method can detect novel features in CLE images that were not previously recognized by the physicians. The three hypotheses were tested empirically, using image semantic segmentation task with the following evaluation metrics: mean accuracy (mean_acc), mean intersection over union (mean_IU), and frequency-weighted intersection over union (fw_IU).

A segmentation dataset including 310 CLE images was acquired from images annotated by four neurosurgeons. Each observer highlighted the diagnostic glioma features of each CLE images, independently. We used majority voting to process the annotation variations from the neurosurgeons. For rigorous assessment of the first hypothesis, the segmentation dataset included diagnostic regions at different scales. (145 images were annotated for both Intermediate (Set2-I) and Restrictive (Set2-R) features). Also, to study the effect of previous exposure of CLE image to the model, we used images from all three stages: 30 images from training (Set1), 145 images form validation (Set2), and 135 images from test set (Set3 and Set4)). To appraise the second hypothesis, we sequentially altered components of the designed architecture and assessed the resulting performance of the model ("ablation study"). All models were trained and tested on the same data with the same parameters to avoid any bias. Finally, to test the third hypothesis, our dataset included 55 CLE images that were known to be from glioma tumors but were initially classified as nondiagnostic (Set4). The model generated the segmentation mask by creating the FDFM of the input image with one forward pass and then thresholding (threshold value of 0.03 for intermediate and 0.2 for restrictive features).

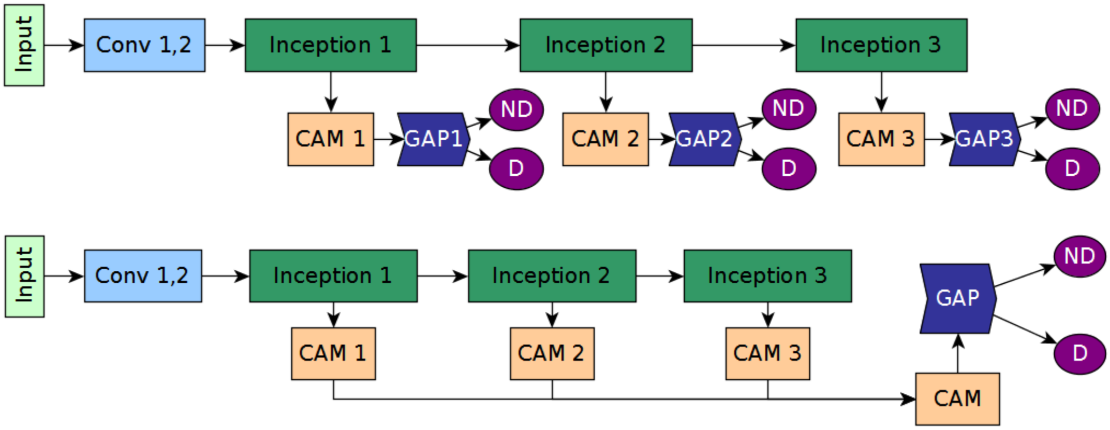


Figure 3.2

Network architectures used for the ablation study. Top network shows the developed architecture without the LI and CI components. Bottom network shows the MLGAP architecture (Feng et al., 2017) which combines the three CAMs and then uses a GAP layer for classification.

Table 3.2 shows experimental results of segmentation performance by ten different models with respect to the annotators. Each model constructs a DFM to create a segmentation map: M1, similar to (Feng et al., 2017); M2 – DFM and NFM of CAM 1,2, and 3 are first laterally inhibited and then collaterally integrated; M3 – CAM 1,2, and 3 are collaterally integrated; M4, M5, M6 – by laterally inhibiting the DFM and NFM of CAM 1, 2, and 3, respectively; M7, M8, M9 – by using the DFMs from CAM 1, 2, and 3 without any further processing; M10, similar to (Zhou et al., 2016). The first hypothesis proved to be true, since our developed model, M2, produced high mean_acc, mean_IU, and fw_IU for all the intermediate features from diagnostic images (Set1, Set2-I, and Set3). Moreover, it could segment the images from Set3 without significant change in mean_acc, while producing better fw_IU and mean_IU values on images that were previously revealed to it (Set1). Results from Set2-I and Set2-R images showed that all models generated much lower mean_IU and fw_IU on restrictive features compared to intermediate features, except for M1 and M2 models, both of which utilize shallower layers for enhancing the DFM's spatial resolution. In all experiments, M2 made the best performance for three measures (except in mean_acc for Set2-R), supporting the second hypothesis about the significance of the utilized components (MLCAM, LI, and CI). Specifically, M4-M6 models outperformed other ablated models (M7-M9), highlighting the significant value of LI. The higher mean_IU value of M6 and M9 compared to M4,5 and M7,8, respectively, indicates that more abstract features were learned by inception 3 than by inception 1,2. In the first round of review, clinicians labeled Set4 images as nondiagnostic, however, after features were highlighted by the developed model, the clinicians re-classified Set4 images as diagnostic. The highest performance in Set4 belonged to M2 (mean_acc = 88% and mean_IU = 89%). High mean_IU value achieved by the model and clinical feedback emphasize significance and novelty of the features.

	Set	**M1**	**$M2^{*}$**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**
mean_acc	Set1	0.71	**0.88**	0.71	0.75	0.75	0.77	0.71	0.71	0.71	0.7
	Set2-I	0.76	**0.85**	0.74	0.76	0.76	0.77	0.74	0.74	0.74	0.74
	Set3	0.72	**0.86**	0.72	0.75	0.75	0.76	0.72	0.72	0.72	0.72
	Set2-R	0.78	0.87	0.79	**0.88**	**0.88**	0.85	0.78	0.79	0.81	0.78
	Set4	0.74	**0.88**	0.74	0.76	0.76	0.78	0.74	0.74	0.74	0.72
mean_IU	Set1	0.65	**0.9**	0.61	0.69	0.69	0.72	0.61	0.61	0.61	0.63
	Set2-I	0.69	**0.86**	0.67	0.69	0.71	0.73	0.65	0.67	0.67	0.69
	Set3	0.57	**0.82**	0.56	0.59	0.61	0.63	0.56	0.56	0.56	0.59
	Set2-R	0.77	**0.88**	0.29	0.57	0.63	0.59	0.27	0.29	0.31	0.63
	Set4	0.48	**0.89**	0.48	0.52	0.55	0.57	0.48	0.48	0.48	0.5
fw_IU	Set1	0.8	**0.99**	0.8	0.83	0.85	0.87	0.8	0.8	0.8	0.8
	Set2-I	0.88	**0.98**	0.86	0.88	0.9	0.92	0.86	0.86	0.86	0.86
	Set3	0.65	**0.88**	0.65	0.69	0.71	0.73	0.65	0.65	0.65	0.67
	Set2-R	0.9	**0.97**	0.18	0.5	0.61	0.58	0.14	0.16	0.2	0.67
	Set4	0.38	**0.79**	0.35	0.42	0.44	0.46	0.35	0.35	0.35	0.4

Table 3.2

Segmentation performance by different models. M2* is the developed model.

CONCLUSIONS

In this study, a WSL model was developed to localize the diagnostic features of gliomas in CLE images. It utilizes three fundamental components for creating the final glioma DFM: multi-scale DFM, LI for removing confusing regions, and CI to spatially infuse diagnostic areas from DFMs with different spatial resolutions. The model could detect the diagnostic regions with high agreement compared with annotation by neurosurgeon, from both diagnostic and nondiagnostic images (i.e., images that were initially designated as lacking diagnostic features) in intermediate and restrictive features, while outperforming other methods. Such an approach should be tested on larger datasets. Initial testing demonstrated that WSL has the potential to identify not only relevant, but novel or unrecognized diagnostic features in CLE images that were not previously discriminated by human inspection, requiring further investigation. This approach can be augmented with active learning and patch clustering to create an atlas of glioma phenotypes in CLE images. Further detailed studies correlating regular histology and CLE images are necessary for better understanding of glioma histopathological features on CLE images.

References

Almeida, J. P., Chaichana, K. L., Rincon-Torroella, J., & Quinones-Hinojosa, A. (2015). The Value of Extent of Resection of Glioblastomas: Clinical Evidence and Current Approach. *Current Neurology and Neuroscience Reports*. https://doi.org/10.1007/s11910-014-0517-x

Baars, B. J., & Gage, N. M. (2010). *Cognition, Brain and Consciousness*. *Cognition, Brain and Consciousness*. https://doi.org/10.1016/C2009-0-01556-6

Belykh, E., Martirosyan, N. L., Yagmurlu, K., Miller, E. J., Eschbacher, J. M., Izadyyazdanabadi, M., … Preul, M. C. (2016). Intraoperative fluorescence imaging for personalized brain tumor resection: Current state and future directions. *Frontiers in Surgery, 3*.

Eschbacher, J., Martirosyan, N. L., Nakaji, P., Sanai, N., Preul, M. C., Smith, K. A., … Spetzler, R. F. (2012). In vivo intraoperative confocal microscopy for real-time histopathological imaging of brain tumors: Clinical article. *Journal of Neurosurgery, 116*(4), 854–860.

Feng, X., Yang, J., Laine, A. F., & Angelini, E. D. (2017). Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp.

568–576).

Foersch, S., Heimann, A., Ayyad, A., Spoden, G. A., Florin, L., Mpoukouvalas, K., … Charalampaki, P. (2012). Confocal laser endomicroscopy for diagnosis and histomorphologic imaging of brain tumors in vivo. *PLoS One*, *7*(7), e41760.

Gondal, W. M., Köhler, J. M., Grzeszick, R., Fink, G. A., & Hirsch, M. (2017). Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. *arXiv Preprint arXiv:1706.09634*.

Izadyyazdanabadi, M., Belykh, E., Martirosyan, N., Eschbacher, J., Nakaji, P., Yang, Y., & Preul, M. C. (2017). Improving utility of brain tumor confocal laser endomicroscopy: Objective value assessment and diagnostic frame detection with convolutional neural networks. In *Progress in Biomedical Optics and Imaging - Proceedings of SPIE* (Vol. 10134). https://doi.org/10.1117/12.2254902

Izadyyazdanabadi, M., Belykh, E., Mooney, M., Martirosyan, N., Eschbacher, J., Nakaji, P., … Yang, Y. (2018). Convolutional Neural Networks: Ensemble Modeling, Fine-Tuning and Unsupervised Semantic Localization for Neurosurgical CLE Images. *Journal of Visual Communication and Image Representation*, *54*, 10–20.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., … Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv Preprint arXiv:1408.5093*.

Korbar, B., Olofson, A. M., Miraflor, A. P., Nicka, C. M., Suriawinata, M. A., Torresani, L., … Hassanpour, S. (2017). Looking Under the Hood: Deep Neural Network Visualization to Interpret Whole-Slide Image Analysis Outcomes for Colorectal Polyps. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on* (pp. 821–827).

Martirosyan, N. L., Eschbacher, J. M., Kalani, M. Y. S., Turner, J. D., Belykh, E., Spetzler, R. F., … Preul, M. C. (2016). Prospective evaluation of the utility of intraoperative confocal laser endomicroscopy in patients with brain neoplasms using fluorescein sodium: experience with 74 cases. *Neurosurgical Focus, 40*(3), E11.

N. Murthy, V., Singh, V., Sun, S., Bhattacharya, S., Chen, T., & Comaniciu, D. (2017). Cascaded deep decision networks for classification of endoscopic images. In M. A. Styner & E. D. Angelini (Eds.), *Medical Imaging 2017: Image Processing* (Vol. 10133, p. 101332B). https://doi.org/10.1117/12.2254333

Qi, H., Collins, S., & Noble, A. (2017). Weakly Supervised Learning of Placental Ultrasound Images with Residual Networks. In *Annual Conference on Medical Image Understanding and Analysis* (pp. 98–108).

Sailem, H., Arias--Garcia, M., Bakal, C., Zisserman, A., & Rittscher, J. (2017). Discovery of Rare Phenotypes in Cellular Images Using Weakly Supervised Deep

Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 49–55).

Sanai, N., Polley, M.-Y., McDermott, M. W., Parsa, A. T., & Berger, M. S. (2011). An extent of resection threshold for newly diagnosed glioblastomas: clinical article. *Journal of Neurosurgery*, *115*(1), 3–8.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2921–2929). https://doi.org/10.1109/CVPR.2016.319

CHAPTER 4

FLUORESCENCE DIGITAL IMAGE HISTOLOGY PATTERN TRANSFORMATION USING IMAGE STYLE TRANSFER

ABSTRACT

Fluorescein sodium (FNa) has been recently used with confocal laser endomicroscopy (CLE) to aid neurosurgeons differentiate between brain normal tissue and neoplasm during tumor dissection. FNa-driven CLE imaging has numerous advantages over conventional intraoperative diagnostic tools such as frozen section hematoxylin and eosin (H&E)-stained histology slide (e.g., more interactive, rapid, and portable). However, it has several limitations: CLE images are usually contaminated with artifacts (motion, red blood cells, noise), FNa stains the tissue architecture indistinctly, and neuropathologists are mainly trained on colorful stained histology slides like H&E while the CLE images are gray. To improve the diagnostic quality of CLE images, we used a micrograph of an H&E slide from a glioma tumor biopsy and image style transfer, a neural network method for integrating the content and style of two images. This was done through minimizing the deviation of the target image from both the content (CLE) and style (H&E) images. Five neurosurgery fellows examined how much artifacts were removed and how much unclear structures were enhanced to validate the quality enhancement in 100 pairs of original and transformed images. Average score of the reviewers on test images showed that 84 out of 100 transformed images had lower artifacts and more noticeable critical structures compared to their original form. By providing images that are more interpretable than the original CLE images and faster to

acquire than H&E slides, this method allows a real-time, cellular-level tissue examination toward a shorter, more accurate, and interactive surgery.

INTRODUCTION

Confocal laser endomicroscopy (CLE) is undergoing rigorous research for its potential to assist neurosurgeons to examine tissue in situ during brain surgery (Belykh et al., 2018; Foersch et al., 2012; Mohammadhassan Izadyyazdanabadi et al., 2018; Martirosyan et al., 2014, 2016). The ability to scan tissue or surgical resection bed on-the-fly essentially producing optical biopsies, compatibility with different fluorophores, miniature size of the probe and the portability of the system are essential features of this promising technology. Currently, the most frequent technique used for neurosurgical intraoperative diagnosis is examination of frozen section hematoxylin and eosin (H&E)-stained histology slides.

Figure 4.1 (a) shows an example image from a glioma acquired by CLE (left) and a micrograph of an H&E slide (right), acquired by conventional light microscopy. Although generating CLE images is much faster than H&E slides (1 second per image compared to about 20 minutes per slide), many CLE images may be non-optimal and can be obscured with artifacts including background noise, blur, and red blood cells (M. Izadyyazdanabadi et al., 2017). The histopathological features of gliomas are often more identifiable in the H&E slide images compared to the CLE images generated using nonspecific fluorescent dyes such as fluorescein sodium (FNa). Neuropathologists as well are used to evaluating detailed histoarchitecture colorfully stained with H&E for diagnoses, especially for frozen section biopsies. Fluorescent images from intraoperative neurosurgical application present a new digital gray scale (monochrome) imaging environment to the neuropathologist for diagnosis that may include hundreds of images from one case. Recently, the U.S. FDA has approved a blue laser range CLE system primarily utilizing FNa for use in neurosurgery.

Countervailing these diagnostic and visual deficiencies in CLE images requires a rapid, automated transformation that can: 1) remove the occluding artifacts, and 2) add (amplify) the histological patterns that are difficult to recognize in the CLE images. Finally, this transformation should avoid removing the critical details (e.g., cell structures, shape) or adding unreal patterns to the image, to maintain the integrity of the image content. Such a method may present "transformed" CLE images to the neuropathologist and neurosurgeon that may resemble images based on familiar and standard, even colored, appearances from histology stains, such as H&E.

One method for implementing this transformation could be supervised learning, however, supervised learning requires paired images (from the same object and location) to learn the mapping between the two domains (CLE and H&E). Creating a dataset of colocalized H&E and CLE images is infeasible because of problems in exact co-localization and intrinsic tissue movements, although small, and artifacts introduced during H&E slide preparation. "Image style transfer", first introduced by Gatys et al. (Gatys, Ecker, & Bethge, 2016), is an image transformation technique that blends the content and style of two separate images to produce the target image. This process minimizes the distance between feature maps of the source and target images using a pretrained convolutional neural network (CNN).

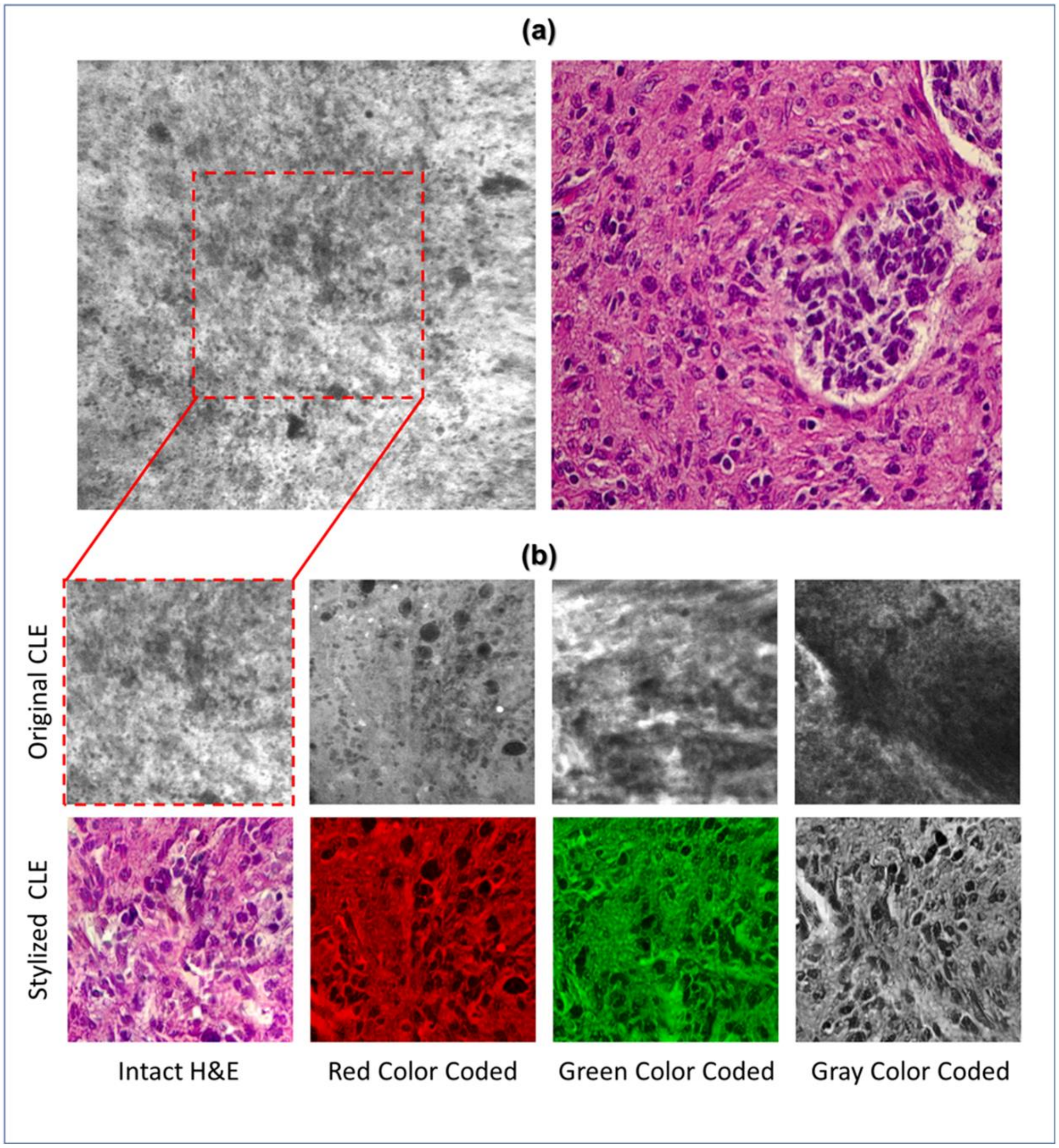


Figure 4.1

(a) Representative CLE (Optiscan 5.1, Optiscan Pty., Ltd.) and H&E images from glioma tumors. (b) Original and stylized CLE images from glioma tumors, in 4 color coding: gray, green, red, intact H&E.

In this study, we aimed to remove the inherent occlusions and enhance the structures that were problematical to recognize in CLE images. Essentially, we attempted to make CLE images generated from non-specific FNa application during glioma surgery appear like standard H&E-stained histology and evaluate the accuracy and usefulness. We used the image style transfer method since it extracts abstract features from the CLE and H&E image that are independent of their location in the image and thus can operate on the images that are not from the same location. More details about the image style transfer algorithm and the quality assessment protocol follow in section 2. Our results from a test dataset showed that on average, the diagnostic quality of stylized images was higher than the original CLE images, although there were some cases where the transformed image showed new artifacts.

METHODS

Image Style Transfer

Image style transfer takes a content and style image as input and produces a target image that shows the structures of the content image and the general appearance of the style image. This is achieved through four main components: 1) a pretrained CNN that extracts feature maps from source and target images, 2) quantitative calculation of the content and style representations for source and target images, 3) a loss function to capture the difference between the content and style representation of source and target images, and 4) an optimization algorithm to minimize the loss function. In contrast to CNN supervised learning, where the model parameters are updated to minimize the prediction error, image style transfer modifies the pixel values of the target image to minimize the loss function while the model parameters are fixed.

A 19-layer visual geometry group network (VGG-19), that is pretrained on ImageNet dataset, extracts feature maps from CLE, H&E, and target images. Feature maps in layer "Conv4_2" of VGG-19 are used to calculate the image content representation, and a list of gram matrices from feature maps of five layers (“ReLU1_1”, “ReLU2_1”, …, “ReLU5_1”) are used to calculate the image style representation. To examine the difference between the target and source images, the following loss function was used:

$$Loss_{Total} = \frac{1}{2}\underbrace{\sum\left(C_{CLE} - C_{Target}\right)^2}_{\boxed{\text{Content Loss: difference between content representation of the CLE and target image}}} + \alpha \times \underbrace{\sum_{i=1}^{5} w^i \times \sum(S^i_{H\&E} - S^i_{Target})^2}_{\boxed{\text{Style Loss: difference between style representation of the H\&E and target image}}}$$

C_{CLE} and C_{Target} are the content representations of the CLE and target image, $S^i_{H\&E}$ and S^i_{Target} are the style representations of the H&E and target image based on the feature maps of the i^{th} layer, and w^i (weight of i^{th} layer in the style representation) equals 0.2. The parameter α determines relative weight of style loss in the total loss and is set to 100. A limited memory optimization algorithm (L-BFGS (Zhu, Byrd, Lu, & Nocedal, 1997)) minimizes this loss.

For the experiment, 100 CLE images (from a recent study by Martirosyan et al.(Martirosyan et al., 2016)) were randomly selected from 15 subjects with glioma tumors as content images. A single micrograph of an H&E slide from a glioma tumor biopsy of a different patient (not one of the 15 subjects) was used as the style image (Figure 4.1 (a), right). For each CLE image, the optimization process was run for 1600 iterations and the target image was saved for evaluation and referred to as the “stylized image” in the following sections.

Evaluation

Although the stylized images presented the same histological patterns as H&E images and seemed to contain similar structures to those present in the corresponding original CLE images, a quantitative image quality assessment was performed to rigorously evaluate the stylized images. Five neurosurgeons independently assessed the diagnostic quality of the 100 pairs of original and stylized CLE images. For each pair, the reviewers sought to examine two properties in each stylized image and provided a score for each property: 1) “what’s removed”: whether the stylization process removed any critical structures (negative impact) or artifacts (positive impact) that were present in the original CLE image, and 2) “what’s added”: whether the stylization process added new structures that were not present (negative impact) or were difficult to notice (positive impact) in the original CLE image. The scores are between 0 and 6 with the following annotations: 0, extreme negative impact; 1, moderate negative impact; 2, slight negative impact; 3, no significant impact; 4, slight positive impact; 5, moderate positive impact; and 6, extreme positive impact. Further information and instructions about the quality assessment survey is available in Table 4.1 and 4.2.

Score	Description
0	Negative* impact; Severe structures are removed
1	Negative impact; Moderate structures are removed
2	Negative impact; Slight structures are removed
3	No significant structures are removed

4	Positive** impact; Slight artifacts are removed
5	Positive impact; Moderate artifacts are removed
6	Positive impact; Severe artifacts are removed

Table 4.1

Scoring protocol for evaluating the diagnostic quality of stylized images based on the “what’s removed” criterion. *** Negative impact** denotes that the transformed image has a lower diagnostic quality than the original image (e.g. removing cells or other preexisting diagnostic features). **** Positive impact** denotes that the transformed image has a higher diagnostic quality than the original image. (e.g. less artifacts)

Score	Description
0	Negative* impact; Severe artifacts are added
1	Negative impact; Moderate artifacts are added
2	Negative impact; Slight artifacts are added
3	No significant structures are added
4	Positive** impact; Slight structures are added
5	Positive impact; Moderate structures are added
6	Positive impact; Severe structures are added

Table 4.2

Scoring protocol for evaluating the diagnostic quality of stylized images based on the “what’s added” criterion. *** Negative impact** denotes that the transformed image has a lower diagnostic quality than the original image (e.g. hallucinating cells, misleading structures, and artifacts). **** Positive impact** denotes that the transformed image has a higher diagnostic quality than the original image (e.g. highlighting cells or other structures that were hard to notice).

Since the physicians were more familiar with the H&E than the CLE images, it was possible that the reviewers would overestimate the quality of stylized images merely due to their color resemblance to H&E images (during style transfer the color of CLE image is also changed to pink and purple). To explore how the reviewers' scores would change if the stylized images were presented in a different color other than the pink and purple (the common color for H&E images), the stylized images were processed in four different ways: I) 25 stylized images were converted into gray-scale images (averaging the three red, green, and blue channels), II) 25 stylized images were color-coded in green (first converted the image to gray-scale and then set red and blue channels to zero), III) 25 stylized images were color-coded as red (similar approach), and IV) 25 stylized images were used without any further changes (intact H&E). Since there are too many structures in each CLE image, and to examine the images more precisely, we used the center-crop of each original CLE and its stylized version for evaluation. Figure 4.1 (b) shows some example stylized images used for evaluation.

Results and Analysis

Figure 4.2 (a) shows a histogram of all reviewers' scores for the removed artifacts (blue bars) and added structures (orange bars) in the stylized images with different colors. Overall, the number of stylized CLE images that have higher diagnostic quality than the original images (score greater than 3) was significantly larger than those with equal or lower diagnostic quality for both removed artifacts and added structures scores (one-way chi square test p-value<0.001). Results from stylized images that were color-coded (gray, green, red) showed the same trend for the added structures scores, indicating that the improvement was not just because of color resemblance.

There was significant difference between how much the model added structures and removed artifacts. For all the color-coded and intact stylized images, the average of added structures scores was larger than the removed artifacts scores (t-test p-value <0.001). This suggests that the model was better at enhancing the structures that were challenging to recognize than removing the artifacts.

Figure 4.2 (b) shows the frequency of different combinations of scores for removed artifacts and added structures in an intensity map. Each block represents how many times a rater scored an image with the corresponding values on the x (improvement by added structures) and y (improvement by removed artifacts) axes for that block. The most frequent incident across all the stylized images is the coordinates (5,4), which means moderately adding structures and slightly removing artifacts, followed by (5,5) meaning moderately adding structures and removing artifacts. Although the intensity maps derived from different color-coded images were not exactly similar, the most frequent combination in each group still indicated positive impact in both properties. The most frequent combination of scores, for each of the color-coded images, was as follows: gray = (5,4), green = (5,5), red = (5,4), and intact = (5,4).

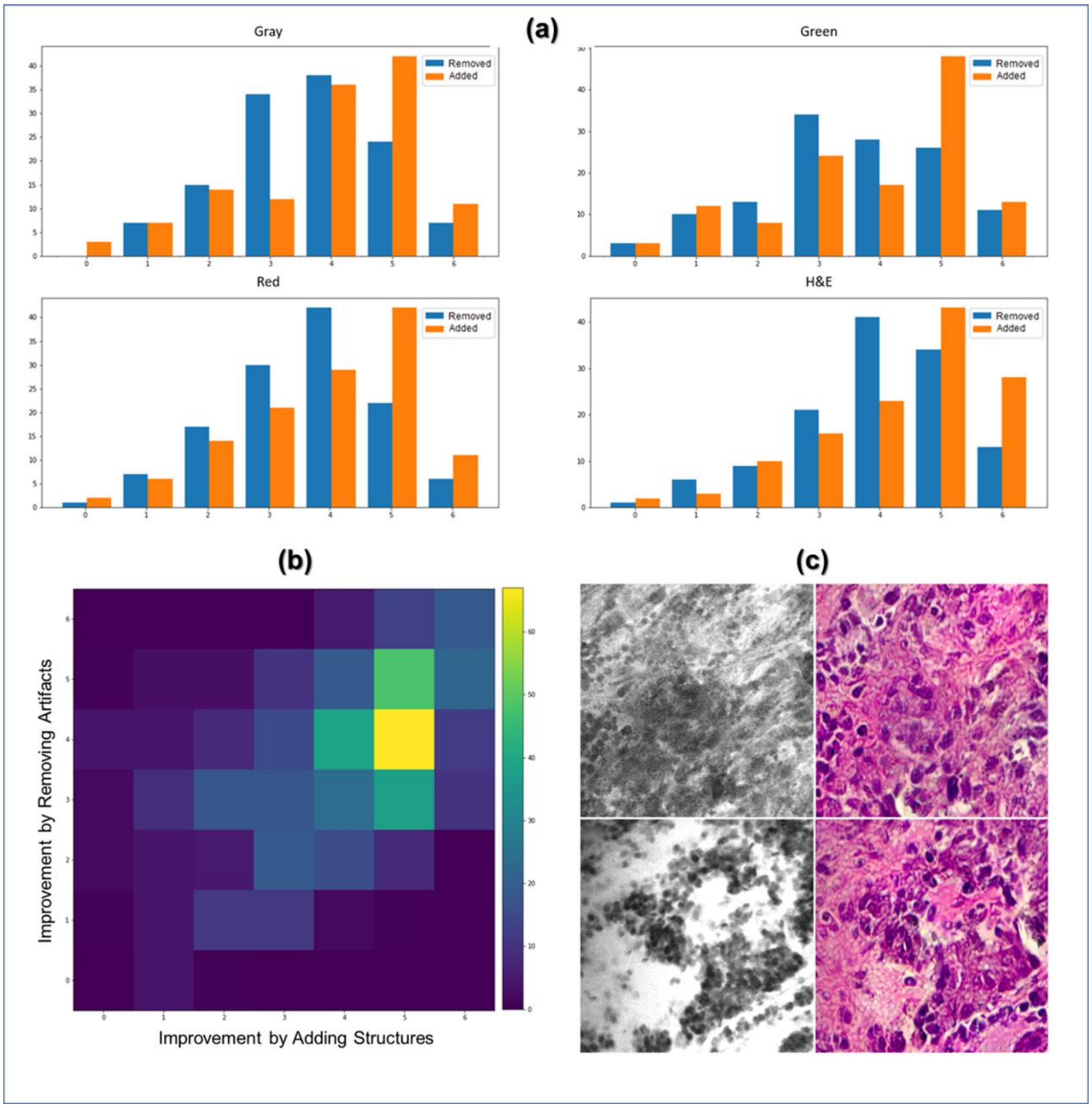


Figure 4.2.

(a) Histogram of scores for added structures and removed artifacts from different color- coded images. **(b)** An intensity map showing the frequency of different combinations of scores for adding structures (x axis) and removing artifacts (y axis). **(c)** Two example images that the stylization process removed some critical details (top) or added unreal structures (bottom).

As a further analysis, we counted the number of images that had an average score of below 3 to see how often the algorithm removed critical structures or added artifacts that were misleading to the neurosurgeons. From the 100 tested images, 3 images had only critical structures removed, 4 images had only artifacts or unreal structures added, and 2 images had both artifacts added and critical structures removed. On the contrary, 84 images showed improved diagnostic quality through both removed artifacts and added structures that were hard to recognize, 6 images had only artifacts removed, and 5 images had only critical structures added. Figure 4.1 (b) shows some example stylized images with improved quality compared to the original CLE, and Figure 4.2 (c) shows two stylized images with decreased diagnostic quality through removed critical structures (top) and added artifacts (bottom).

Conclusions

In this study, image style transfer was applied to CLE images from gliomas to enhance their diagnostic quality. Style transfer with an H&E-stained slide image had an overall positive impact on the diagnostic quality of CLE images. The improvement was not solely because of the colorization of CLE images; even the stylized images that were converted to gray, red, and green, reported improved diagnostic quality compared to the original CLE images. Employment of more specific clinical tasks to explore the advantage of stylization in diagnosing gliomas and other tumor types is underway based on this preliminary success. In the future, application of more advanced methods for transferring patterns in the histology slides to the CLE images will be used to improve their interpretability. Because of the high number of CLE images acquired during a single case, style transfer could add value to such fluorescence images and allows for computer-aided techniques to play a meaningful, convenient, and efficient role to aid the neurosurgeon and neuropathologist in analysis of CLE images and to more rapidly determine diagnosis.

References

Belykh, E., Miller, E. J., Patel, A. A., IzadyYazdanabadi, M., Martirosyan, N. L., Yagmurlu, K., … Preul, M. C. (2018). Diagnostic accuracy of the confocal laser endomicroscope for in vivo differentiation between normal and tumor tissue during fluorescein-guided glioma resection: Laboratory investigation. *World Neurosurgery*, *In press*.

Foersch, S., Heimann, A., Ayyad, A., Spoden, G. A., Florin, L., Mpoukouvalas, K., … Charalampaki, P. (2012). Confocal laser endomicroscopy for diagnosis and histomorphologic imaging of brain tumors in vivo. *PLoS One*, *7*(7), e41760.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image Style Transfer Using Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2016.265

Izadyyazdanabadi, M., Belykh, E., Martirosyan, N., Eschbacher, J., Nakaji, P., Yang, Y., & Preul, M. C. (2017). Improving utility of brain tumor confocal laser endomicroscopy: Objective value assessment and diagnostic frame detection with convolutional neural networks. In *Progress in Biomedical Optics and Imaging - Proceedings of SPIE* (Vol. 10134). https://doi.org/10.1117/12.2254902

Izadyyazdanabadi, M., Belykh, E., Mooney, M., Martirosyan, N., Eschbacher, J., Nakaji, P., … Yang, Y. (2018). Convolutional Neural Networks: Ensemble Modeling, Fine-Tuning and Unsupervised Semantic Localization for Neurosurgical CLE Images. *Journal of Visual Communication and Image Representation*, *54*, 10–20.

Martirosyan, N. L., Eschbacher, J. M., Kalani, M. Y. S., Turner, J. D., Belykh, E., Spetzler, R. F., … Preul, M. C. (2016). Prospective evaluation of the utility of intraoperative confocal laser endomicroscopy in patients with brain neoplasms using fluorescein sodium: experience with 74 cases. *Neurosurgical Focus*, *40*(3), E11.

Martirosyan, N. L., Georges, J., Eschbacher, J. M., Cavalcanti, D. D., Elhadi, A. M., Abdelwahab, M. G., … Preul, M. C. (2014). Potential application of a handheld confocal endomicroscope imaging system using a variety of fluorophores in experimental gliomas and normal brain. *Neurosurgical Focus*, *36*(2), E16.

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, *23*(4), 550–560.

CHAPTER 5

DISCUSSION

This dissertation studies three computer aided systems to enhance neurosurgical CLE imaging with deep convolutional neural networks. In the following sections, a discussion of methods and the significance of each task is provided. Then, the limitations and future works are presented.

CONTRIBUTIONS

First, an ensemble of deep convolutional neural networks was developed to evaluate the diagnostic quality of CLE images rapidly. This was critical since almost half of the acquired images during surgery are blurred with motion and occluded with blood artifacts and lack useful histopathological features for an accurate diagnosis. To countervail this issue, thousands of images can be fed to the model and be quantified based on their diagnostic quality, allowing the neurosurgeon and pathologist review images with highest quality much faster. Experimental results in chapter 2 showed that with the current size of our CLE dataset (~20,000), combining transfer learning (initializing the weights with a pre-trained model) and ensemble modeling (aggregating the outputs of diverse models) produced better classification accuracy and AUC than training from scratch (random weight initialization) or a single model. Ensemble modeling positive impact was more evident when the individual models were more diverse (e.g. deep fine tuning and training from scratch) and less when there was less diversity (shallow fine tuning). Different DCNN architectures responded differently to shallow fine tuning and deep fine tuning, but overall, both showed improvement from training from scratch to deep fine tuning. An interesting observation was that the ensemble of models had higher agreement with the defined gold standard ground truth than the two neurosurgery fellows. This could be because the model was trained on images that were labeled with physicians while they had access to both CLE and H&E slides, and the model might have learnt to consider more details when evaluating CLE images than the secondary reviewers.

Second, a weakly-supervised learning approach was adopted to train a customized DCNN architecture for segmenting the CLE images into regions containing features indicative of glioma. Image-level labels were used to train the network on classification task. Two feature maps representing the diagnostic and nondiagnostic areas of the input were produced at three different resolutions and globally averaged to produce the probability scores for classification. During the test stage, diagnostic feature maps were extracted from the images and compared with the overall annotation provided by four neurosurgery fellows. To lower the false positives in model output, confusing regions of the diagnostic feature maps (i.e. areas that were activated in both diagnostic and nondiagnostic feature maps) were laterally inhibited. To increase the model sensitivity regarding small critical structures, a consequent layer integrated the feature maps from the shallow layers to produce the final segmentation map. Ablation study showed that removing any of these three components (multi-layer feature maps, lateral inhibition, and collateral integration) lowered the performance and the developed model outperformed two states of the arts in the field of weakly supervised learning; the evaluation was based on two common metrics in image segmentation: accuracy and intersection over union. The fact that the model could detect structures in images that were tentatively labeled as nondiagnostic by reviewers, suggest that such systems may aid the reviewing process of CLE images through marking suspicious areas that might otherwise be missed by the reviewer.

Third, image style transfer was applied to CLE images to improve their quality. A micrograph of an H&E-stained slide obtained from a glioma tumor biopsy was used as the “style image” to amplify the diagnostic features and reduce the artifacts in the CLE image (“content image”), resulting in the “target image”. To blend the style of the H&E image with the structures in the CLE image, a pre-trained DCNN was used to find a quantitative representation of content and style properties. An optimization algorithm minimized the weighted sum of the distance between the content of target and CLE and the distance between style of target and H&E images. Five neurosurgery fellows evaluated the diagnostic quality of the target and CLE images and confirmed quality enhancement in both intact H&E and pseudo-colored target images. The rationale behind pseudo-coloring style transferred images was to inspect if the perceived higher quality is solely due to the colorization which was denied since even gray color-coded style transferred images showed lower artifacts and more identifiable structures than the original images.

LIMITATIONS AND FUTURE WORKS

In the first study on diagnostic classification of CLE images, the ensemble model averaged the outputs from five models that were trained on different portions of the dataset. This allowed reducing the overfitting effect and better generalization on test images from new cases. However, this approach intensified the computational cost and required longer time to process. Future studies can explore other approaches for developing ensemble models that address this issue. Obtaining more CLE images and creating a larger dataset allows training deeper networks and learning more comprehensive and accurate mappings between CLE images and their diagnostic quality. With the introduction of new generations of CLE imaging instruments, transfer learning from the current models (that are based on previous generations of the instrument) can be used to address the low number of images available at the beginning.

Second study showed the promise of a weakly-supervised learning approach for image segmentation and detecting novel features in CLE images from gliomas. A semi-supervised approach may augment it in the future to combine the knowledge in image-level annotated dataset with the knowledge in a smaller pixel-level annotated dataset. A systematic organization of the features detected by the developed model may help understanding of glioma phenotypes in CLE images and allow better characterization of CLE images with an atlas of glioma attributes. However, this requires a larger dataset with annotations and further experiments with supervised object detection along rigorous validation by neurosurgeons. Moreover, clustering algorithms (e.g. K-means) can be used for categorizing the glioma features in an unsupervised way.

One of the limitations in the current study is that each frame is analyzed independent of the other frames. The reason for pursuing this approach in this study is the small size of the dataset. In future when more images are available, the whole sequence of frames can be fed to the network; this may aid the model to differentiate red blood cells from tumor cells considering that red blood cells tend to move more than tumor cells in a sequence of frames. An alternative could be preprocessing the images with a more conventional method such as optical flow and feed the motion information along the frames to the network.

Third study applied an optimization-based method for enhancing the quality of CLE images through reducing their artifacts and amplifying the critical structures. First limitation is that for every new image it needs to run the optimization to find the target stylized image which make it slow. The reason for choosing this approach was that due to the intrinsic differences between CLE and H&E image acquisition process, it is infeasible to acquire paired images for training an end-to-end algorithm to reconstruct H&E from CLE. One alternative is to first create a dataset of synthesized H&E images with style transfer and then train a CNN on the paired CLE-H&E images; the drawback would be that the new CNN will inherit any deficiencies in the current approach. Also, further diagnostic tasks are needed to rigorously validate the usefulness of this approach in CLE guided neurosurgery.

Overall, both CLE and DCNN are powerful new technologies with promising achievements in the last couple of years. Future studies may reveal further benefits of applying DCNN to CLE for other purposes, such as: differential diagnosis of tumors with customized classification networks (e.g., benign versus malignant), noise removal with autoencoders, increasing the resolution with generative models, cell detection with supervised object recognition, image captioning to annotate features in the image, multi-channel image analysis (i.e., staining the tissue with multiple fluorophores instead of single FNa), and categorizing diagnostic features of brain neoplasms in CLE images in the form of an atlas (with clustering methods). After all, since DCNN is a data-driven approach and acquiring large annotated datasets is time consuming, more images need to be acquired from different tumors and efficient weakly-supervised and semi-supervised methods should be developed to resolve the annotation challenge.

References

Abdi, A., Luong, C., Tsang, T., Jue, J., Hawley, D., Fleming, S., … Abolmaesumi, P. (2017). Automatic Quality Assessment of Echocardiograms Using Convolutional Neural Networks: Feasibility on the Apical Four-chamber View. *IEEE Transactions on Medical Imaging*.

Almeida, J. P., Chaichana, K. L., Rincon-Torroella, J., & Quinones-Hinojosa, A. (2015). The Value of Extent of Resection of Glioblastomas: Clinical Evidence and Current Approach. *Current Neurology and Neuroscience Reports*. https://doi.org/10.1007/s11910-014-0517-x

American Cancer Society. (2018). Cancer Facts and Statistics. Retrieved May 29, 2018, from https://cancerstatisticscenter.cancer.org/

Aubreville, M., Knipfer, C., Oetter, N., Jaremenko, C., Rodner, E., Denzler, J., … Maier, A. (2017). Automatic Classification of Cancerous Tissue in Laserendomicroscopy Images of the Oral Cavity using Deep Learning. *Scientific Reports*, *7*. https://doi.org/10.1038/s41598-017-12320-8

Baars, B. J., & Gage, N. M. (2010). *Cognition, Brain and Consciousness*. *Cognition, Brain and Consciousness*. https://doi.org/10.1016/C2009-0-01556-6

Belykh, E., Martirosyan, N. L., Yagmurlu, K., Miller, E. J., Eschbacher, J. M., Izadyyazdanabadi, M., … Preul, M. C. (2016). Intraoperative fluorescence imaging for personalized brain tumor resection: Current state and future directions. *Frontiers in Surgery*, *3*.

Belykh, E., Miller, E. J., Patel, A. A., IzadyYazdanabadi, M., Martirosyan, N. L., Yagmurlu, K., … Preul, M. C. (2018). Diagnostic accuracy of the confocal laser endomicroscope for in vivo differentiation between normal and tumor tissue during fluorescein-guided glioma resection: Laboratory investigation. *World Neurosurgery*, *In press*.

Charalampaki, P., Javed, M., Daali, S., Heiroth, H.-J., Igressa, A., & Weber, F. (2015). Confocal Laser Endomicroscopy for Real-time Histomorphological Diagnosis: Our Clinical Experience With 150 Brain and Spinal Tumor Cases. *Neurosurgery*, *62*, 171–176.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the Devil in the Details: Delving Deep into Convolutional Nets. *BMVC*, 1–11. https://doi.org/10.5244/C.28.6

Christodoulidis, S., Anthimopoulos, M., Ebner, L., Christe, A., & Mougiakakou, S. (2017). Multisource Transfer Learning With Convolutional Neural Networks for

Lung Pattern Analysis. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 76–84.

Ciregan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 3642–3649).

Ciresan, D., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems* (pp. 2843–2851).

Dietterich, T. G., & others. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems*, *1857*, 1–15.

Djuric, U., Zadeh, G., Aldape, K., & Diamandis, P. (2017). Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *Npj Precision Oncology*, *1*(1), 22. https://doi.org/10.1038/s41698-017-0022-1

Eschbacher, J. M., Georges, J. F., Belykh, E., Yazdanabadi, M. I., Martirosyan, N. L., Szeto, E., … others. (2017). Immediate Label-Free Ex Vivo Evaluation of Human Brain Tumor Biopsies With Confocal Reflectance Microscopy. *Journal of Neuropathology & Experimental Neurology*, *76*(12), 1008–1022.

Eschbacher, J., Martirosyan, N. L., Nakaji, P., Sanai, N., Preul, M. C., Smith, K. A., … Spetzler, R. F. (2012). In vivo intraoperative confocal microscopy for real-time histopathological imaging of brain tumors: Clinical article. *Journal of Neurosurgery*, *116*(4), 854–860.

Feng, X., Yang, J., Laine, A. F., & Angelini, E. D. (2017). Discriminative localization in CNNs for weakly-supervised segmentation of pulmonary nodules. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 568–576).

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., … Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, *136*(5), E359–E386. https://doi.org/10.1002/ijc.29210

Foersch, S., Heimann, A., Ayyad, A., Spoden, G. A., Florin, L., Mpoukouvalas, K., … Charalampaki, P. (2012). Confocal laser endomicroscopy for diagnosis and histomorphologic imaging of brain tumors in vivo. *PLoS One*, *7*(7), e41760.

Gao, Y., Maraci, M. A., & Noble, J. A. (2016). Describing ultrasound video content using deep convolutional neural networks. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on* (pp. 787–790).

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image Style Transfer Using

Convolutional Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2016.265

Ghafoorian, M., Karssemeijer, N., Heskes, T., Bergkamp, M., Wissink, J., Obels, J., … others. (2017). Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin. *NeuroImage: Clinical*, *14*, 391–399.

Gil, D. et al. (2017). Classification of Confocal Endomicroscopy Patterns for Diagnosis of Lung Cancer. In *Cardoso M. et al. (eds) Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures. CARE 2017, CLIP 2017. Lecture Notes in Computer Science* (Vol. 10550, pp. 151–159).

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. *AISTATS '11: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, *15*, 315–323. https://doi.org/10.1.1.208.6449

Gondal, W. M., Köhler, J. M., Grzeszick, R., Fink, G. A., & Hirsch, M. (2017). Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. *arXiv Preprint arXiv:1706.09634*.

Greenspan, H., van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, *35*(5), 1153–1159.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv Preprint arXiv:1512.03385*.

Hong, J., Park, B., & Park, H. (2017). Convolutional neural network classifier for distinguishing Barrett's esophagus and neoplasia endomicroscopy images. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 2892–2895). IEEE. https://doi.org/10.1109/EMBC.2017.8037461

Izadyyazdanabadi, M., Belykh, E., Cavallo, C., Zhao, X., Gandhi, S., Moreira, L. B., … Yang, Y. (2018). Weakly-Supervised Learning-Based Feature Localization in Confocal Laser Endomicroscopy Glioma Images. *arXiv Preprint arXiv:1804.09428*.

Izadyyazdanabadi, M., Belykh, E., Martirosyan, N., Eschbacher, J., Nakaji, P., Yang, Y., & Preul, M. C. (2017). Improving utility of brain tumor confocal laser endomicroscopy: Objective value assessment and diagnostic frame detection with convolutional neural networks. In *Progress in Biomedical Optics and Imaging - Proceedings of SPIE* (Vol. 10134). https://doi.org/10.1117/12.2254902

Izadyyazdanabadi, M., Belykh, E., Mooney, M., Martirosyan, N., Eschbacher, J., Nakaji, P., … Yang, Y. (2018). Convolutional Neural Networks: Ensemble Modeling, Fine-Tuning and Unsupervised Semantic Localization for Neurosurgical CLE Images.

Journal of Visual Communication and Image Representation, *54*, 10–20.

Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3304–3311). https://doi.org/10.1109/CVPR.2010.5540039

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., … Darrell, T. (2014). Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv Preprint arXiv:1408.5093*.

Kamen, A., Sun, S., Wan, S., Kluckner, S., Chen, T., Gigler, A. M., … others. (2016). Automatic Tissue Differentiation Based on Confocal Endomicroscopic Images for Intraoperative Guidance in Neurosurgery. *BioMed Research International*, *2016*.

Korbar, B., Olofson, A. M., Miraflor, A. P., Nicka, C. M., Suriawinata, M. A., Torresani, L., … Hassanpour, S. (2017). Looking Under the Hood: Deep Neural Network Visualization to Interpret Whole-Slide Image Analysis Outcomes for Colorectal Polyps. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on* (pp. 821–827).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In *Advances in neural information processing systems* (pp. 231–238).

Kumar, A., Kim, J., Lyndon, D., Fulham, M., & Feng, D. (2017). An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 31–40.

Kumar, A., Sridar, P., Quinton, A., Kumar, R. K., Feng, D., Nanan, R., & Kim, J. (2016). Plane identification in fetal ultrasound images using saliency maps and convolutional neural networks. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on* (pp. 791–794).

Le Cun Jackel, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D., Cun, B. Le, Denker, J., & Henderson, D. (1990). Handwritten Digit Recognition with a Back-Propagation Network. *Advances in Neural Information Processing Systems*, 396–404. https://doi.org/10.1111/dsu.12130

LeCun, Y., & Bengio, Y. (1998). The handbook of brain theory and neural networks. In M. A. Arbib (Ed.) (pp. 255–258). Cambridge, MA, USA: MIT Press. Retrieved from http://dl.acm.org/citation.cfm?id=303568.303704

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–

444. https://doi.org/10.1038/nature14539

LeCun, Y., Jackel, L. D., Bottou, L., Cortes, C., Denker, J. S., Drucker, H., … Vapnik, V. (1995). Learning algorithms for classification: A comparison on handwritten digit recognition. In *Neural networks: the statistical mechanics perspective* (pp. 261–276).

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *arXiv Preprint arXiv:1702.05747*.

Liu, J. T. C., Meza, D., & Sanai, N. (2014). Trends in fluorescence image-guided surgery for gliomas. *Neurosurgery*. https://doi.org/10.1227/NEU.0000000000000344

Loiseau, S. (2017). *Presentation at International Society for Endomicroscopy, Paris, France*. Paris.

Louis, D. N., Ohgaki, H., Wiestler, O. D., Cavenee, W. K., Burger, P. C., Jouvet, A., … Kleihues, P. (2007). The 2007 WHO classification of tumours of the central nervous system. *Acta Neuropathologica*. https://doi.org/10.1007/s00401-007-0243-4

Madabhushi, A., & Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, *33*, 170–175. https://doi.org/10.1016/j.media.2016.06.037

Mahapatra, D., Roy, P. K., Sedai, S., & Garnavi, R. (2016). Retinal Image Quality Classification Using Saliency Maps and CNNs. In *International Workshop on Machine Learning in Medical Imaging* (pp. 172–179).

Martirosyan, N. L., Cavalcanti, D. D., Eschbacher, J. M., Delaney, P. M., Scheck, A. C., Abdelwahab, M. G., … Preul, M. C. (2011). Use of in vivo near-infrared laser confocal endomicroscopy with indocyanine green to detect the boundary of infiltrative tumor. *Journal of Neurosurgery*, *115*(6), 1131–1138. https://doi.org/10.3171/2011.8.JNS11559

Martirosyan, N. L., Eschbacher, J. M., Kalani, M. Y. S., Turner, J. D., Belykh, E., Spetzler, R. F., … Preul, M. C. (2016). Prospective evaluation of the utility of intraoperative confocal laser endomicroscopy in patients with brain neoplasms using fluorescein sodium: experience with 74 cases. *Neurosurgical Focus*, *40*(3), E11.

Martirosyan, N. L., Georges, J., Eschbacher, J. M., Belykh, E., Carotenuto, A., Spetzler, R. F., … Preul, M. C. (2018). Confocal scanning microscopy provides rapid, detailed intraoperative histological assessment of brain neoplasms: Experience with 106 cases. *Clinical Neurology and Neurosurgery*, *169*, 21–28.

Martirosyan, N. L., Georges, J., Eschbacher, J. M., Cavalcanti, D. D., Elhadi, A. M., Abdelwahab, M. G., … Preul, M. C. (2014). Potential application of a handheld confocal endomicroscope imaging system using a variety of fluorophores in experimental gliomas and normal brain. *Neurosurgical Focus*, *36*(2), E16.

Maugeri, R., Villa, A., Pino, M., Imperato, A., Giammalva, G. R., Costantino, G., … others. (2018). With a Little Help from My Friends: The Role of Intraoperative Fluorescent Dyes in the Surgical Management of High-Grade Gliomas. *Brain Sciences*, *8*(2), 31.

Metz, C. E. (1978). Basic principles of ROC analysis. In *Seminars in nuclear medicine* (Vol. 8, pp. 283–298).

Mooney, M. A., Georges, J., Yazdanabadi, M. I., Goehring, K. Y., White, W. L., Little, A. S., … Eschbacher, J. M. (2017). Immediate ex-vivo diagnosis of pituitary adenomas using confocal reflectance microscopy: a proof-of-principle study. *Journal of Neurosurgery*.

Mooney, M. A., Georges, J., Yazdanabadi, M. I., Goehring, K. Y., White, W. L., Little, A. S., … Eschbacher, J. M. (2018). Immediate ex-vivo diagnosis of pituitary adenomas using confocal reflectance microscopy: a proof-of-principle study. *Journal of Neurosurgery*, *128*, 1072–1075.

Mooney, M. A., Zehri, A. H., Georges, J. F., & Nakaji, P. (2014). Laser scanning confocal endomicroscopy in the neurosurgical operating room: a review and discussion of future applications. *Neurosurgical Focus*, *36*(2), E9. https://doi.org/10.3171/2013.11.FOCUS13484

Murthy, V. N., Singh, V., Chen, T., Manmatha, R., & Comaniciu, D. (2016). Deep Decision Network for Multi-class Image Classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2240–2248). https://doi.org/10.1109/CVPR.2016.246

N. Murthy, V., Singh, V., Sun, S., Bhattacharya, S., Chen, T., & Comaniciu, D. (2017). Cascaded deep decision networks for classification of endoscopic images. In M. A. Styner & E. D. Angelini (Eds.), *Medical Imaging 2017: Image Processing* (Vol. 10133, p. 101332B). https://doi.org/10.1117/12.2254333

Ostrom, Q. T., Gittleman, H., Fulop, J., Liu, M., Blanda, R., Kromer, C., … Barnholtz-Sloan, J. S. (2015). CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012. *Neuro-Oncology*, *17*, iv1-iv62. https://doi.org/10.1093/neuonc/nov189

Penet, M.-F., Krishnamachary, B., Chen, Z., Jin, J., & Bhujwalla, Z. M. (2014). Molecular imaging of the tumor microenvironment for precision medicine and theranostics. *Advances in Cancer Research*, *124*, 235–256. https://doi.org/10.1016/B978-0-12-411638-2.00007-0

Penet, M. F., Chen, Z., Kakkad, S., Pomper, M. G., & Bhujwalla, Z. M. (2012).

Theranostic imaging of cancer. *European Journal of Radiology, 81*(SUPPL1). https://doi.org/10.1016/S0720-048X(12)70051-7

Qi, H., Collins, S., & Noble, A. (2017). Weakly Supervised Learning of Placental Ultrasound Images with Residual Networks. In *Annual Conference on Medical Image Understanding and Analysis* (pp. 98–108).

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on* (pp. 512–519).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Sailem, H., Arias--Garcia, M., Bakal, C., Zisserman, A., & Rittscher, J. (2017). Discovery of Rare Phenotypes in Cellular Images Using Weakly Supervised Deep Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 49–55).

Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging. *IEEE Transactions on Medical Imaging*.

Sanai, N., & Berger, M. S. (2018). Surgical oncology for gliomas: the state of the art. *Nature Reviews Clinical Oncology*, *15*(2), 112.

Sanai, N., Eschbacher, J., Hattendorf, G., Coons, S. W., Preul, M. C., Smith, K. A., … Spetzler, R. F. (2011). Intraoperative confocal microscopy for brain tumors: a feasibility analysis in humans. *Neurosurgery*, *68*, ons282--ons290.

Sanai, N., Polley, M.-Y., McDermott, M. W., Parsa, A. T., & Berger, M. S. (2011). An extent of resection threshold for newly diagnosed glioblastomas: clinical article. *Journal of Neurosurgery*, *115*(1), 3–8.

Sánchez, J., Perronnin, F., Mensink, T., & Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, *105*(3), 222–245. https://doi.org/10.1007/s11263-013-0636-x

Sankar, T., Delaney, P. M., Ryan, R. W., Eschbacher, J., Abdelwahab, M., Nakaji, P., … Preul, M. C. (2010). Miniaturized handheld confocal microscopy for neurosurgery: Results in an experimental glioblastoma model. *Neurosurgery*, *66*(2), 410–417. https://doi.org/10.1227/01.NEU.0000365772.66324.6F

Shi, J., Zheng, X., Li, Y., Zhang, Q., & Ying, S. (2017). Multimodal Neuroimaging Feature Learning with Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer's Disease. *IEEE Journal of Biomedical and Health*

Informatics.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., … Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285–1298.

Sirinukunwattana, K., Raza, S. E. A., Tsang, Y.-W., Snead, D. R. J., Cree, I. A., & Rajpoot, N. M. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, *35*(5), 1196–1206.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Stasinopoulos, I., Penet, M. F., Chen, Z., Kakkad, S., Glunde, K., & Bhujwalla, Z. M. (2011). Exploiting the tumor microenvironment for theranostic imaging. *NMR in Biomedicine*. https://doi.org/10.1002/nbm.1664

Suk, H.-I., & Shen, D. (2016). Deep Ensemble Sparse Regression Network for Alzheimer’s Disease Diagnosis. In *International Workshop on Machine Learning in Medical Imaging* (pp. 113–121).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2818–2826). IEEE. https://doi.org/10.1109/CVPR.2016.308

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Transactions on Medical Imaging*, *35*(5), 1299–1312.

Thong, P. S.-P., Olivo, M., Kho, K.-W., Zheng, W., Mancer, K., Harris, M., & Soo, K.-C. (2007). Laser confocal endomicroscopy as a novel technique for fluorescence diagnostic imaging of the oral cavity. *Journal of Biomedical Optics*, *12*(1), 14007. https://doi.org/10.1117/1.2710193

Tofte, K., Berger, C., Torp, S. H., & Solheim, O. (2014). The diagnostic properties of frozen sections in suspected intracranial tumors: A study of 578 consecutive cases. *Surgical Neurology International*, *5*.

Vo, K., Jaremenko, C., Bohr, C., Neumann, H., & Maier, A. (2017). Automatic Classification and Pathological Staging of Confocal Laser Endomicroscopic Images of the Vocal Cords. In *Bildverarbeitung für die Medizin 2017* (pp. 312–317). Springer.

Yan, C., Xie, H., Liu, S., Yin, J., Zhang, Y., & Dai, Q. (2017). Effective Uyghur language text detection in complex background images for traffic prompt identification. *IEEE Transactions on Intelligent Transportation Systems*.

Yan, C., Xie, H., Yang Dongbao, Yin, J., Zhang, Y., & Dai, Q. (2017). Supervised Hash Coding With Deep Neural Network for Environment Perception of Intelligent Vehicles. *IEEE Transactions on Intelligent Transportation Systems*.

Yan, C., Zhang, Y., Xu, J., Dai, F., Li, L., Dai, Q., & Wu, F. (2014). A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors. *IEEE Signal Processing Letters*, *21*(5), 573–576.

Yan, C., Zhang, Y., Xu, J., Dai, F., Zhang, J., Dai, Q., & Wu, F. (2014). Efficient parallel framework for HEVC motion estimation on many-core processors. *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(12), 2077–2089.

Yao, H., Zhang, S., Zhang, Y., Li, J., & Tian, Q. (2016). Coarse-to-fine description for fine-grained visual categorization. *IEEE Transactions on Image Processing*, *25*(10), 4858–4872.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (pp. 3320–3328).

Zehri, A., Ramey, W., Georges, J., Mooney, M., Martirosyan, N., Preul, M., & Nakaji, P. (2014). Neurosurgical confocal endomicroscopy: A review of contrast agents, confocal systems, and future imaging modalities. *Surgical Neurology International*, *5*, 60.

Zhang, L., Zhang, Y., Gu, X., Tang, J., & Tian, Q. (2014). Scalable similarity search with topology preserving hashing. *IEEE Transactions on Image Processing*, *23*(7), 3025–3039.

Zhang, X., Zhang, H., Zhang, Y., Yang, Y., Wang, M., Luan, H., … Chua, T.-S. (2016). Deep fusion of multiple semantic cues for complex event recognition. *IEEE Transactions on Image Processing*, *25*(3), 1033–1046.

Zhao, J., Zhang, M., Zhou, Z., Chu, J., & Cao, F. (2016). Automatic detection and classification of leukocytes using convolutional neural networks. *Medical & Biological Engineering & Computing*, 1–15.

Zhao, L., & Jia, K. (2016). Multiscale cnns for brain tumor segmentation and diagnosis.

Computational and Mathematical Methods in Medicine, *2016*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2921–2929). https://doi.org/10.1109/CVPR.2016.319

Zhou, Z.-H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artificial Intelligence*, *137*(1–2), 239–263.

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, *23*(4), 550–560.