ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

PhD. THESIS

Abdela Ahmed MOSSA

FULLY AUTOMATED DEEP LEARNING AND MACHINE LEARNING–BASED PROGNOSIS MODELS FOR SURVIVAL PREDICTION OF BRAIN TUMOR PATIENTS USING MULTI-MODAL MRI IMAGES

DEPARTMENT OF COMPUTER ENGINEERING

ADANA-2021

ABSTRACT

PhD. THESIS

FULLY AUTOMATED DEEP LEARNING AND MACHINE LEARNING –BASED PROGNOSIS MODELS FOR SURVIVAL PREDICTION OF BRAIN TUMOR PATIENTS USING MULTI-MODAL MRI IMAGES

Abdela Ahmed MOSSA

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF COMPUTER ENGINEERING

Supervisor : Prof. Dr. Ulus ÇEVİK
Year: 2021, Pages: 183
Jury : Prof. Dr.Ulus ÇEVİK
: Prof. Dr. Mustafa GÖK
: Prof. Dr. Turgay İBRİKÇİ
: Assoc. Prof. Dr. Mustafa ORAL
: Asst. Prof. Dr. Gökay DİŞKEN

Brain tumor is one of the most deadly types of cancer diseases. Accurate assessment of pre-surgical prognosis for patients with this disease can lead to better patient management. While Biopsy is the most commonly used diagnostic technique in routine clinical applications of prognosis estimation, it has several disadvantages such as it is invasive, and prone to tissue trauma. Consequently, automated pre-operative prognosis estimation techniques based on MRI images are recently getting attention, so noninvasive. However, most of the recently developed automated techniques are based on the handcrafted image features extracted from the manually segmented tumor regions in MRI, which is tedious & time-consuming. This study aimed to develop fully automated pre-operative prognostic models for the survival time, and glioma grade predictions in multi-modal MRI images of patients with brain tumors by using two-stage learning-based methods. In the first stage, we developed novel CNN architectures using pre-trained deep learning models as backend. In the second stage, the outputs of CNN models were fused using various classical machine learning methods to get the final prediction results. The experimental results demonstrate that the proposed prognostic models achieve AUC values of 99.7%, and 93% in glioma grading, and survival time predictions, respectively, outperforming current state-of-the-art results.

Keywords: Brain Tumor Diagnosis, *Deep Learning, Machine Learning, MRI*

ÖZ

DOKTORA TEZİ

MULTİMODAL MRI GÖRÜNTÜLERİ KULLANARAK BEYİN TÜMÖRÜ HASTALARININ HAYATTA KALMA TAHMİNİ İÇİN TAM OTOMATİK DERİN ÖĞRENME VE MAKİNE ÖĞRENME TABANLI PROGNOZ MODELLERİ

Abdela Ahmed MOSSA

ÇUKUROVA ÜNIVERSITESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Danışman : Prof. Dr. Ulus ÇEVİK
Year: 2021, Pages: 183
Jüri : Prof. Dr. Ulus ÇEVİK
: Prof. Dr. Mustafa GÖK
: Prof. Dr. Turgay İBRİKÇİ
: Doç. Dr. Mustafa ORAL
: Dr. Öğr.Üyesi Gökay DİŞKEN

Beyin tümörü, en ölümcül kanser hastalıkları türlerinden biridir. Bu hastalığa sahip hastalar için ameliyat öncesi prognozun doğru değerlendirilmesi, daha iyi hasta yönetimine yol açabilir. Biyopsi, prognoz tahminin rutin klinik uygulamalarında en sık kullanılan tanı tekniği olmasıyla birlikte, invazif ve doku travmasına yatkınlık gibi birçok dezavantaja sahiptir. Sonuç olarak, non invaziv olan MRI görüntülerine dayanan otomatik preoperatif prognoz tahmin teknikleri son zamanlarda dikkat çekmektedir. Bununla birlikte, son zamanlarda geliştirilen otomatik tekniklerin çoğu, sıkıcı ve zaman alıcı olan MRI'da manuel olarak segmentlere ayrılmış tümör bölgelerinden çıkarılan özel yapımı görüntü özelliklerine dayanmaktadır. Bu çalışma, beyin tümörlü hastaların multi-modal MR görüntülerinde sağkalım süresi ve glioma derecesi tahminleri için iki aşamalı öğrenme tabanlı yöntemler kullanarak tam otomatik preoperatif prognostik modeller geliştirmeyi amaçlamıştır. İlk aşamada, arka uç olarak önceden eğitilmiş derin öğrenme modellerini kullanarak yeni CNN mimarileri geliştirildi. İkinci aşamada, CNN modellerinin çıktıları, nihai tahmin sonuçlarını elde etmek için çeşitli klasik makine öğrenimi yöntemleri kullanılarak birleştirildi. Deneysel sonuçlar, önerilen prognostik modellerin glioma derecelendirme ve hayatta kalma süresi tahminlerinde sırasıyla % 99,7 ve% 93 AUC değerlerine ulaşıldığını gösterdi ve mevcut son-teknoloji sonuçlarından daha iyi bir performans gösterdi.

Anahtar Kelimeler: Beyin tümörü teşhisi, Derin Öğrenme, Makine Öğrenmesi, MRI

EXTENDED ABSTRACT

A brain tumor is one of the most deadly types of cancer diseases characterized by the uncontrolled growth and spread of abnormal cells in the central nervous system (brain and spinal cord) (Stewart & Wild, 2014). As defined by the WHO, there are more than 120 types of brain tumors that differ in the tumor cell origin, location, size, cancerous, and characteristics of the tumor tissues. However, they can be coarsely classified as primary or metastatic. A brain tumor is considered as primary if there is an abnormal growth of cells that starts in the central nervous system, and does not spread to other parts of the body. However, the metastatic brain tumors begin as cancer elsewhere in the body (lung, breast, prostate, etc.), and spread to the brain through the bloodstream or lymphatic vessels.

Gliomas are a broad category of primary brain tumors that grow from the glial cells (Laws & Thapar, 1993). Gliomas are responsible for substantial morbidity and mortality across the globe that can strike all human races at various ages, and the overall incidence is increasing every year (Goodenberger & Jenkins, 2012). According to the report by CBTRUS, in 2016 alone, there were a total of 330,000 incident cases and 227, 000 deaths of glioma patients across worldwide. Thus, diagnosis and treatment of brain tumor patients require huge resource allocation and sophisticated diagnostic and therapeutic technologies (Patel et al., 2019).

Prognosis is a branch of medicine that encompasses many aspects but classically specializes in predicting the likely or expected development of a disease. Knowledge of patients' prognosis is a key element because it does not only assist physicians in deciding on appropriate treatment options but also in discussing the opinions of patients or relatives concerning management in palliative care settings if the life expectancy is minimal. Therefore, prognostic models for pre-operative grading and survival time estimation are critical steps in the management

of patients with glioma. There are four grades of glioma that range from grade I, which are slow-growing, to grade IV, which are the most malignant and almost incurable. Historically, the slower-growing lesions, corresponding to grade I and grade II, have been commonly referred to as low-grade glioma (LGG), while the more rapidly progressive tumors corresponding to grade III and grade IV are commonly referred to as high-grade glioma (HGG) (Bauer, Wiest, Nolte, & Reyes, 2013; Louis et al., 2007). Overall survival time of patients with HGG is also further categorized into two (long vs. short), or three (short, medium and long) groups of survival for valuable insights in advancing surgical and treatment plans (Adeberg et al., 2014; Gandía-González et al., 2019; Macyszyn et al., 2016).

An accurate assessment of pre-surgical prognosis for patients with glioma can lead to better management of the patients. While Biopsy is the most commonly used diagnostic technique in routine clinical applications of prognosis estimation, it has several disadvantages such as it is invasive and prone to tissue trauma. Most importantly, recent studies indicated that it is also possible to do it on non-invasive medical imaging techniques, such as multi-modal magnetic resonance images (MRIs). Consequently, an MRI-based automated medical image analysis technique for pre-operative brain tumor prognosis estimation is gaining attention due to an increased need for efficient and objective evaluation of the large amounts of images generated by the MRI scanners. However, most of the recently developed automated techniques are based on the handcrafted image features extracted from the manually segmented tumor regions in MRI, which is tedious & time-consuming.

The main objective of this study was to develop fully automated prognostic models for the pre-operative survival time, and glioma grade predictions based on multi-modal MRI images of patients with brain tumors by using a two-stage learning-based method. In the first stage, inspired by Su et al. (2015), and N. Wu et al. (2019), three novel convolutional neural network architectures by using state-of-

the-art pre-trained deep learning models as backends were developed. The proposed architectures were named Mv-CNN, Triplanar-CNN, and Mc-CNN.

The proposed architectures use a stack of 2D MRI slices as input. Specifically, the volumetric MRI images were reconstructed into 2D stacked slices in the axial, coronal, and sagittal planes, which are subsequently used to train the proposed CNN models. Transfer learning techniques consisting of a fine-tuned deep learning models based on natural images and classical data augmentation techniques that increase the available training data sets were used to improve the performance of the proposed deep learning models. In the meanwhile, we examined both whether large pre-trained CNN architectures perform better than smaller pre-trained architectures and whether higher performance on natural images guarantees higher performance on medical image classification problems. Moreover, another novel pre-processing technique, which is a concatenation of two images, was used to leverage pre-trained deep learning models trained on natural images without enlarging them to benefit from the fine-grained details in each slice of the coronal and sagittal projections.

In the second stage, the predicted probability outputs by the CNN models were fused using different classical machine learning methods to improve the performances of the proposed prognostic models and get the final prediction results. Moreover, majority voting, weighted majority voting, averaging, and majority averaging ensemble strategies were also implemented and compared with the classical machine learning methods.

We applied our methods to the most widely used public dataset of brain multi-modal MRI images (FLAIR MRI, T1Ce MRI, T1 MRI, and T2 MRI), BraTS'17 dataset (Bakas et al., 2017; Menze et al., 2015). The dataset consists of multi-modal volumetric MRI exams of 75 LGG, and 210 HGG glioma patients who underwent the examination before surgery, and obtained at different hospitals with a pathologically confirmed diagnosis. The experimental results demonstrate that our best proposed prognostic models achieve AUC, Sensitivity, and Specificity

values of 99.7%, 98.6%, and 94.7%, and 93%, 92.9% and 92.3% in glioma grading prediction, and survival time prediction, respectively, outperforming current state-of-the-art results.

In this study, the individual classification performance of the four MRI modalities both for glioma grading, and survival time predictions were also compared. The experimental results show that T1Ce coronal slices, and FLAIR axial slices have more discriminative deep features for glioma grading, and survival time prediction, respectively. Moreover, with the aim of improving the proposed prognostic model for the survival time prediction, the effecting of using segmented glioma region (ROIs), determined by a deep learning-based segmentation model (Wang et al., 2018), or extended glioma region (EROIs) that contains the surrounding tumor tissues, or whole slide image (without segmentation) as input while training deep learning model were analyzed. The conclusion is that incorporating the surrounding tumor tissues while training the deep learning models for survival time prediction improves the model performance.

To conclude, prognostic models in glioma are useful for health professionals in guiding when choosing the right treatment options for each patient, end of life planning for seriously ill patients, and avoid unnecessary treatments. Thus, the developed fully automated prognostic models can have huge contributions to an early diagnosis of patients with a brain tumor through providing valuable insights for physicians in advancing treatment planning using only MRI at the time of diagnosis.

GENİŞLETİLMİŞ ÖZET

Beyin tümörü, merkezi sinir sistemindeki (beyin ve omurilik) anormal hücrelerin kontrolsüz büyümesi ve yayılmasıyla karakterize edilen en ölümcül kanser hastalığı türlerinden biridir (Stewart & Wild, 2014). Dünya Sağlık Örgütü'nün tanımladığı gibi, tümör hücresi orijini, yeri, boyutu, kanserli ve tümör dokularının özellikleri bakımından farklılık gösteren 120'den fazla beyin tümörü türü vardır. Bununla birlikte, kabaca birincil veya metastatik olarak sınıflandırılabilirler. Bir beyin tümörü, merkezi sinir sisteminde başlayan ve vücudun diğer bölgelerine yayılmayan anormal bir hücre büyümesi varsa, birincil olarak kabul edilir. Bununla birlikte, metastatik beyin tümörleri vücudun başka bir yerinde (akciğer, meme, prostat vb.) kanser olarak başlar ve kan dolaşımı veya lenfatik damarlar yoluyla beyne yayılır.

Gliomlar, glial hücrelerden büyüyen geniş bir birincil beyin tümörleri kategorisidir (Laws & Thapar, 1993). Gliomalar, dünya genelinde çeşitli yaşlarda tüm insan ırklarını etkileyebilecek önemli morbidite ve mortaliteden sorumludur ve genel insidans her yıl artmaktadır (Goodenberger ve Jenkins, 2012). CBTRUS'un raporuna göre, yalnızca 2016 yılında dünya çapında toplam 330.000 vaka ve 227.000 beyin tümörü hastası ölümü olmuştur. Bu nedenle, glioma hastalarının teşhis ve tedavisi, büyük kaynak tahsisi, sofistike teşhis ve tedavi teknolojileri gerektirir (Patel et al., 2019).

Prognoz, birçok yönü kapsayan ancak klasik olarak bir hastalığın olası veya beklenen gelişimini tahmin etmede uzmanlaşmış bir tıp dalıdır. Hastaların prognozuna ilişkin bilgiler, hekimlere sadece uygun tedavi seçeneklerine karar vermede yardımcı olmakla kalmaz, aynı zamanda yaşam beklentisi minimum ise palyatif bakım ortamlarında tedavi ile ilgili hasta veya yakınlarının görüşlerini tartışmada da yardımcı olur. Bu nedenle, ameliyat öncesi derecelendirme ve hayatta kalma süresi tahmini için prognostik modeller, gliomalı hastaların yönetiminde kritik adımlardır. Yavaş büyüyen I. evreden en kötü huylu ve

neredeyse tedavi edilemez olan evre IV'e kadar değişen dört derece glioma vardır. Tarihsel olarak, derece I ve derece II'ye karşılık gelen daha yavaş büyüyen lezyonlar, yaygın olarak düşük dereceli glioma (LGG) olarak adlandırılırken, derece III ve derece IV'e karşılık gelen daha hızlı ilerleyen tümörler genellikle yüksek dereceli gliomalar (HGG) olarak adlandırılır (Bauer, Wiest, Nolte ve Reyes, 2013; Louis ve diğerleri, 2007). HGG'li hastaların ameliyat öncesi sağkalım süresi ayrıca, cerrahi ve tedavi planlarında ilerlemeyi kaydedilebilmesi için iki (uzun vs. kısa) veya üç (kısa, orta ve uzun) sağkalım grubuna ayrılmıştır (Adeberg ve diğerleri, 2014; Gandía-González ve diğerleri, 2019).

Gliomalı hastalar için ameliyat öncesi prognozun doğru bir şekilde değerlendirilmesi, hastaların daha iyi yönetilmesine yol açabilir. Biyopsi, prognoz tahminin rutin klinik uygulamalarında en sık kullanılan tanı tekniği olmakla birlikte, invazif ve doku travmasına yatkınlık gibi birçok dezavantaja sahiptir. En önemlisi, son çalışmalar, bunu multi-modal manyetik rezonans görüntüler (MRI) gibi invazif olmayan tıbbi görüntüleme tekniklerinde de yapmanın mümkün olduğunu gösterdi. Sonuç olarak, ameliyat öncesi beyin tümörü prognoz tahmini için MRI tabanlı otomatik tıbbi görüntü analizi tekniği, MRI tarayıcıları tarafından oluşturulan büyük miktarlarda görüntünün verimli ve objektif değerlendirilmesine olan artan ihtiyaç nedeniyle dikkat çekmektedir. Bununla birlikte, son zamanlarda geliştirilen otomatik tekniklerin çoğu, sıkıcı ve zaman alıcı olan MRI'da manuel olarak segmentlere ayrılmış tümör bölgelerinden çıkarılan özel yapımı görüntü özelliklerine dayanmaktadır.

Bu çalışmanın temel amacı, iki aşamalı öğrenme tabanlı bir yöntem kullanarak, beyin tümörü olan hastaların çok modlu MRI görüntülerine dayanan ameliyat öncesi sağkalım süresi ve glioma derecesi tahminleri için tam otomatik prognostik modeller geliştirmektir. İlk aşamada Su ve ark. (2015) ve N. Wu ve ark. (2019), arka uç geliştirilirken son teknoloji ürünü önceden eğitilmiş derin öğrenme modellerini kullanarak üç yeni evrişimli sinir ağı mimarisi geliştirmişlerdir. Önerilen mimariler Mv-CNN, Triplanar-CNN ve Mc-CNN olarak adlandırılmıştır.

Önerilen mimariler, girdi olarak bir 2D MRI dilimi yığını kullanmakttadır. Spesifik olarak, hacimsel MRI görüntüleri, daha sonra önerilen CNN modellerini eğitmek için kullanılan eksenel, koronal ve sagital düzlemlerde 2D yığılmış dilimler halinde yeniden yapılandırımıştır. Önerilen derin öğrenme modellerinin performansını iyileştirmek için doğal görüntülere dayalı ince ayarlı derin öğrenme modellerinden ve mevcut eğitim veri setlerini artıran klasik veri artırma tekniklerinden oluşan aktarım öğrenme teknikleri (Transfer Learning) kullanılmıştır. Ayrıca, hem önceden eğitilmiş büyük CNN mimarilerinin daha küçük önceden eğitilmiş mimarilerden daha iyi performans gösterip göstermediğini hem de doğal görüntülerde daha yüksek performansın tıbbi görüntü sınıflandırma problemlerinde daha yüksek performansı garanti edip etmediğini inceledik. Ayrıca, iki görüntünün bir araya getirilmesi olan bir başka yeni ön işleme tekniği, koronal ve sagittalin her bir dilimindeki ince taneli ayrıntılardan yararlanmak için onları büyütmeden, doğal görüntüler üzerinde önceden eğitilmiş derin öğrenme modellerinden yararlanmak için kullanıldı.

İkinci aşamada, CNN modellerinin tahmin ettiği olasılık çıktıları, önerilen prognostik modellerin performanslarını iyileştirmek ve nihai tahmin sonuçlarını elde etmek için farklı klasik makine öğrenimi yöntemleri kullanılarak birleştirilmiştir. Ayrıca, çoğunluk oylaması, ağırlıklı çoğunluk oyu, ortalama ve çoğunluk ortalamalı topluluk stratejileri de uygulandı ve klasik makine öğrenimi yöntemleriyle karşılaştırımıştır.

Bu çalışmada yöntemlerimizi, beyin çok modlu MRI görüntülerinin (FLAIR MRI, T1Ce MRI, T1 MRI ve T2 MRI), BraTS'17 veri setinin (Bakas ve diğerleri, 2017; Menze ve diğerleri, 2015) en yaygın kullanılan kamu veri setine uyguladık). Veri seti, ameliyat öncesi muayeneye giren ve farklı hastanelerde patolojik olarak doğrulanmış tanı ile alınan 75 LGG ve 210 HGG glioma hastasının multi-modal volümetrik MR incelemelerinden oluşmaktadır. Deneysel sonuçlar, en iyi önerilen prognostik modellerimizin, glioma derecelendirme tahmininde ve hayatta kalma süresi tahmininde sırasıyla% 99,7,% 98,6 ve% 94,7 ve% 93,% 92,9

ve% 92,3 AUC, Duyarlılık ve Özgüllük değerlerine ulaşıldığını göstermiş ve mevcut son-teknoloji sonuçlarından daha iyi bir performans göstermiştir.

Bu çalışmada, hem glioma derecelendirme hem de sağkalım süresi tahminleri için dört MRG modalitesinin bireysel sınıflandırma performansı da karşılaştırılmiştir. Deneysel sonuçlar, T1Ce koronal dilimlerin ve FLAIR eksenel dilimlerinin sırasıyla glioma derecelendirme ve hayatta kalma süresi tahmini için daha ayırt edici derin özelliklere sahip olduğunu göstermektedir. Ayrıca, hayatta kalma süresi tahmini için önerilen prognostik modeli iyileştirmek amacıyla, derin öğrenmeye dayalı bir segmentasyon modeli (Wang ve diğerleri, 2018) veya genişletilmiş glioma bölgesi ile belirlenen segmentli glioma bölgesi (ROI'ler) kullanımının etkisi Derin öğrenme modeli eğitilirken girdi olarak çevreleyen tümör dokularını içeren (EROI'ler) veya tüm slayt görüntüsü (segmentasyonsuz) analiz edilmiştir. Hayatta kalma süresi tahmini için derin öğrenme modellerini eğitirken çevreleyen tümör dokularını dahil etmenin model performansını iyileştirdiği görülmüştür.

Sonuç olarak, gliomadaki prognoz modelleri, her hasta için doğru tedavi seçeneklerini seçerken, ağır hastalar için yaşam sonu planlaması yaparken ve gereksiz tedavilerden kaçınırken rehberlik etmede sağlık uzmanları için yararlıdır. Bu nedenle, geliştirilmiş tam otomatik prognostik modeller, tanı anında sadece MRI kullanarak tedavi planlamasını ilerletmede doktorlara değerli bilgiler sağlayarak beyin tümörü olan hastaların erken teşhisine katkıda bulunabilir.

ACKNOWLEDGEMENTS

First and foremost, I am extremely grateful to Allah the Almighty for giving me the opportunity and capability to conduct this research and for getting me through exciting and challenging situations. This achievement wouldn't have been possible without His will.

I would like to express my sincere gratitude to my supervisor Prof. Dr. Ulus ÇEVİK for having given me the opportunity to work under his guidance and learn from his experience. He always supported and guided me throughout the research work from title selection to finding the results and compiling the thesis. He had been an outstanding source of motivation and inspiration for me. Working with him has indeed upgraded my professionalism which will definitely help me in my future endeavor.

Apart from my supervisor, I would like to thank members of the thesis committee Prof..Dr. Turgay IBRİKÇİ and Assoc.Prof.Dr. Mustafa ORAL for their regular support and collaboration in the process of undertaking the research work and I am really glad for the final results we have accomplished.

I also extend my gratitude to all the professors, research assistants, and administrative staff that I met at both the Department of Computer Engineering and Department of Electronics and Electrical Engineering, Çukurova University for letting me fulfill my dream of completing a Ph.D. degree through enabling me to take advantage of the departments' facilities. Especially, a special thank goes to the research assistants Mr.Halit ERİŞ and Mr.Emre KILINÇ who are exceptional persons both professionally and humanly speaking.

I am deeply thankful to the Çukurova University, Faculty of Engineering, Computer laboratory administrative staff, Mr.Önder KARAKAYA for having been helpful to me from installing the GPU to maintaining my computer and to Mr.Metin KURT for allowing me to use his lab room, including weekends, for running the experiments smoothly without which doing this research would have been inconceivable. I'm also grateful to the Scientific Research Projects Center of Çukurova University for their financial support.

Big good lucks to my roommates Mr.Nevzat Dik and Mr.Hüseyin Yağan, and to all international students in Adana. We shared together many moments and I am very glad to have you on my side to laugh and to support each other. I thank you from the bottom of my heart for helping me a lot when I was sick and not be able to walk for more than 3 months. Especially, a special thanks to the Computer Engineering research assistant Mr.Emre KILINÇ, to the ADANA KYK Men's Dormitory Manager Mr. Ismail YİĞİT, to all Ethiopian student's in Adana, who have been taking care of me at the hospital, always there day and night when no one else was. You gave me all the strength, courage and love when I needed the most to keep going and get better. I really appreciate it.

Finally, I must express my deep-rooted gratitude to all members of my family, and the University of Gondar colleagues Aster Gebeyaw and Ali Muhammed. From distance, you have been always there to provide me priceless moral support and continuous encouragement throughout my years of study.

Thank you!! Teşekkürler!! አ ማሰ ግና ለ ሁ!!

TABLE OF CONTENTS **PAGE**

LIST OF TABLES **PAGE**

LIST OF FIGURES **PAGE**

LIST OF SYMBOLS AND ABBREVIATIONS

%	: Percentage
3D	: Three-Dimensional
ACS	: Americal Cancer Society
Adam	: Adaptive moment estimation
AI	: Artificial Intelligence
ANN	: Artificial Neural Network
ASCO	: American Society of Clinical Oncology
CAD	: Computer Aided Diagnosis
CapsNet	: Capsule Network
CBTRUS	: Central Brain Tumor Registry of the United States
CNN	: Convolutional Neural Network
CMV	: CytoMegaloVirus
ConvNet	: Convolutional Neural Network
CT	: Computed Tomography
CSF	: Cerebrospinal Fluid
DDBN	: Discriminative Deep Belief Network
DBM	: Deep Boltzmann Machine
DBN	: Deep Belief Network
DL	: Deep Learning
DNN	: Deep Neural Network
EL	: Ensemble Learning
EMR	: Electronic Medical Records
FLAIR	: Fluid-Attenuated Inversion Recovery MRI
GAN	: Generative Adversarial Network
GPU	: Graphical Processing Unit
GRU	: Gated Recurrent Unit
HGG	: High-Grade Glioma
HPC	: High-Performance Computing

KNN : K-Nearest Neighbors
ILSVRC : ImageNet Large-Scale Visual Recognition Challenge
LGG : Low-Grade Glioma
LR : Logistic Regression
LSTM : Long Short Term Memory
Mc-CNN : Multi-Column CNN
ML : Machine Learning
MLP : Multi Layer Perceptron
MRI : Magnetic Resonance Imaging
MRS : Magnetic Resonance Spectroscopy
Mv-CNN : Multi-View CNN
NB : Naïve Bayes
NCI : National Cancer Institute
NLP : Natural Language Processing
OS : Overall Survival
PFS : Progression-Free Survival
PET : Positron Emission Tomography
RBM : Restricted Boltzmann Machine
ReLU : Rectified Linear Unit
RF : Random Forest
RT : Random Tree
RNN : Recurrent Neural Network
SGD : Stochastic Gradient Descent
SVM : Suport Vector Machine
TE : Time to Echo
TPU : Tensor Processing Unit
TR : Repetition Time
T1 : T1-Weighted MRI
T1Ce : Contrast-Enhanced T1-Weighted MRI
T2 : T2-Weighted MRI
WHO : World Health Organization

1. INTRODUCTION

1.1. Artificial Intelligence

One of the main goals in the field of computer engineering is to develop algorithms that can solve problems computationally. Human ingenuity has been playing a key role in developing this kind of algorithms. However, there are problems whose solutions seem extremely challenging to solve using human ingenuity alone. Also, there are some problems that are easy for people to solve but hard to articulate in a formal way. Collectively, these kinds of problems are often named the “AI-set” in which systems claiming to be intelligent ought to be able to have a larger impact in complementing and augmenting human ingenuity in solving them. Problems categorized under the “AI-set” include finding and recognizing objects in an image, make a diagnosis in medicine, natural language understanding, and providing customer services.

The concept of using computers to simulate intelligent behavior and critical thinking was first coined in the "Computing Machinery and Intelligence" seminal paper written in 1950 by Alan Turing (Turing, 1950). However, the term artificial intelligence (AI) as a field was officially introduced for the first time by John McCarthy in 1955 at the conference that took place in Hanover, New Hampshire, Dartmouth College. During the conference, AI was used to describe a new field of computer science that studies about machines in principle performing as good as humans in all aspects of cognition, including problem-solving, decision-making, reasoning, planning, and learning. In addition, the participants of the conference had expected that AI would fulfill the principle in no more than a few decades if a carefully selected group of scientists work on it together (McCarthy, Minsky, Rochester, & Shannon, 2006). However, failure to meet the intense optimism set during the conference and the astonishing programs developed in the years after the conference characterized the life cycle of the field which led to

several “AI-winters and AI-summers’ in which the intense optimism spurred by brief progress (Russell & Norvig, 2002).

AI is now a broad scientific discipline with its roots in computer science, philosophy, neuroscience and mathematics that has a definite goal of understanding intelligence and develop intelligent systems for the “AI-set” problems. However, the methods and formalisms used on the way to this goal are not firmly set (yet), which has resulted in AI consisting of a bunch of tools or subdisciplines. The most tangible form of AI is machine learning (ML), which are AI techniques that give computers the ability to learn without being explicitly programmed to do so. Another subdiscipline of AI which is most important these days is deep learning (DL). It is a subset of machine learning which makes the computation of multi-layer neural networks feasible. Supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, graphical models, knowledge graph, and so on are also the most commonly used buzzwords in the field of AI. The easiest way to understand the relationship between AI and its subdisciplines is to visualize them as a Venn diagram as shown in Figure 1.1. This thesis contributes to computer aided diagnosis (CAD) models for brain tumor diagnosis using ML and DL algorithms. Hence, the remainder of this chapter describes the background and summarizes the contributions and goals set to this thesis.

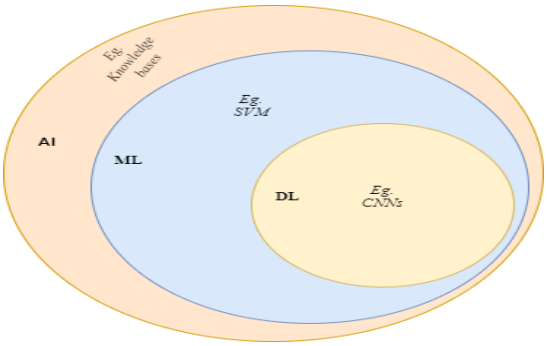

Figure 1.1. The Venn diagram showing how deep learning (DL) is a kind of Machine Learning (ML), which is in turn a kind of Artificial Intelligence (AI).

1.1.1. Machine Learning

The main reason that early researchers found the “AI-set” problem to be much harder is that it wasn’t amenable both to the hard-coded or rule-based algorithms defined by human experts. The solution for these kinds of problems turned out to be machine learning techniques, which are used to accomplish tasks through a provided predefined model of the task with training data examples and let it figure things out through experience similar to how humans do. These examples can come from nature, generated by another algorithm, or handcrafted by humans. For instance, Naïve Bayes, which is a simple machine learning technique can now distinguish between legitimate e-mail and spam e-mail, which was not possible earlier.

Arthur Samuel (1901-1990), who was one of the attendees at the Dartmouth conference and one of the most famous pioneers in the AI field, defined machine learning as a “field of computer science that gives computers the ability to learn without being explicitly programmed” (Samuel, 2000). Machine learning algorithms have some inputs and outputs. Moreover, they are equipped with parameters that can be adjusted using optimization algorithms by anticipating the output of every training example in the data set in such a manner that its expected future performance improves. The main goal of every ML model is to be able to generalize their learned expertise and deliver correct predictions for new or unseen data during training, often named testing dataset.

There are four major settings in which we wish to learn the machine learning models. The first one in which most of the recently developed applications are based is called the supervised learning approach. In this approach, the values of the output samples in the training data set are known and the model learns from these input-output pairs to yield correct outputs on the previously unseen data. Localization and segmentation of brain tumors from MRI images using human-labeled data (Bien et al., 2018) or cyberbullying detection using Twitter users’ psychological features and manually annotated by human experts (Balakrishnan,

Khan, & Arabnia, 2020), are some of the two recent tasks tackled using the supervised ML algorithms.

In contrast to the supervised learning algorithms, the unsupervised learning algorithms consume the unlabeled training data as input for the machine learning models. These unsupervised learning approaches use techniques for finding traits in the training data that have common natural partitions of patterns. Clustering, anomaly detection, and dimensionality reduction problems are major examples that can be tackled using the unsupervised learning approach (Alloghani, Al-Jumeily, Mustafina, Hussain, & Aljaaf, 2020; Hinton, Sejnowski, & Wang, 2001).

Machine learning algorithms that are intermediate between the supervised and unsupervised approaches in which a few labeled and a large collection of unlabeled training examples are given to train the models are often called the semisupervised learning approach (Alloghani et al., 2020). Moreover, any ML technique that can be applied in unknown environments in which the models learn from a series of reinforcement rewards or penalties for the action it performs to maximize the reward is named reinforcement learning. AlphaGo and AlphaZero application developed by DeepMind is the most famous application of reinforcement learning (Mnih et al., 2015; Silver et al., 2017).

Machine learning techniques can also be coarsely classified as traditional or representational taking into account the way features are extracted from data sets. Machine learning techniques that learn from the datasets based on expert-based extracted features or based on features obtained by applying an additional several feature extraction algorithms are categorized as conventional or traditional ML algorithms. Most importantly, traditional machine learning models expect a small amount of data for training, and require all the input data in the format of structured data like numbers. Logistic Regression (LR), Linear Regression, Support Vector Machine (SVM), Random Forest (RF), and Naive Bayes (NB) fall under this category.

ML techniques where discriminative features are learned directly from underlying data without experts explicitly identifying them are called representational machine learning models. This automatic feature extraction capability is the strong point of these representational machine learning techniques against the traditional machine learning approaches. Besides, when large datasets are presented, the performance of the conventional machine learning algorithms becomes steady, whereas the representational machine learning algorithms increase with respect to the increment of the amount of data. Deep learning techniques are the quintessential example of these representational ML algorithms.

In addition, when the outputs of any ML models are an instance of a predefined set of classes (such as high-grade cancer vs low-grade cancer), the task is referred to as a classification problem and is called binary classification if there are only two values. On the other hand, if the outputs are continuous (such as the survival time of cancer patients in days), it is known as a regression problem.

1.1.2. Deep Learning

The introduction of ML algorithms enabled computers to solve many complex real-world problems and make correct decisions that appear subjective. However, it is since 2010 where machine learning has witnessed a tremendous amount of attention when deep artificial neural networks, often termed deep learning, began outperforming other established automated models on several benchmarks. In fact, it was when a deep learning architecture named convolutional neural network (CNN) halved the second-best error rate on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (Krizhevsky, Sutskever, & Hinton, 2012). Now, deep learning is so widespread that it has effectively become synonymous with the term artificial intelligence itself.

In a physiologic analogy, a neuron is a unit that takes electrical and chemical signals as input and subsequently produces and sends an output signal to others. Artificial neural networks (ANN) are engineered systems inspired by the

computation mechanism of the biological brains that learn complicated patterns in data in a hierarchical manner by changing the strengths of synaptic connections between neurons. The main idea of an artificial neural network is that if a bunch of artificial neurons can be linked together in the right way, analogous to how neurons are linked in the brain, then we can build models that can be able to learn a variety of tasks. Artificial Neural Networks are discussed in more detail in Section 3.2.1

Deep learning is part of a broader family of ML methods based on artificial neural networks with millions or sometimes billions of artificial neurons organized into hierarchical layers between the input and the output. The hidden layers amplify aspects of the input that are important for discrimination and suppress irrelevant variations. Deep learning needs a machine to be fed with large quantities of raw data and it subsequently discovers discriminative features necessary for detection or classification using a general-purpose learning procedure. The general-purpose learning procedure represents the raw data as a nested hierarchy of features or concepts in many stages between the input and output layers, with each concept being defined in terms of simpler concepts, and more abstract representations constructed in terms of less abstracted ones (Ian Goodfellow, Yoshua Bengio, 2016; Lecun, Bengio, & Hinton, 2015; Schmidhuber, 2015). Through learning the automatic discriminative features that are essential for detection or classification, deep learning can significantly outperform systems that rely on features that are designed by hand or supplied by domain experts (Lecun et al., 2015).

Since its inception, the deep learning technology has been garnered remarkable success in solving a wide variety of complex problems such as drug discovery and genomics (Chang et al., 2018; Jiménez, Škalič, Martínez-Rosell, & De Fabritiis, 2018), image analysis (Chen, Tian, & He, 2020; Farabet, Couprie, Najman, & Lecun, 2013), reconstructing brain circuits (Helmstaedter et al., 2013; K. Lee et al., 2019), speech recognition and synthesis (Oord et al., 2016; Xiong et al., 2018), natural language processing (Howard & Ruder, 2018; Peters et al., 2018b, 2018a), and in the analysis of unstructured, tabular-type data using entity

embeddings (De Brébisson, Simon, Auvolat, Vincent, & Bengio, 2015; Guo & Berkhahn, 2016). Moreover, it has been growing rapidly and has also become part of our everyday lives as it is being used by various big technology companies including IBM, Google, Apple, Facebook, NVIDIA, Netflix, Microsoft, and Adobe. OpenAI and DeepMind are also some of the recent deep learning research centers that research and build safe AI technologies that learn how to solve problems and advance scientific discovery for all.

Different categories of deep learning techniques have also been developed and applied by researchers in solving various types of problems. These categories include Convolutional Neural Network (CNN or ConvNet), Auto-Encoder (AE), Deep Boltzmann Machine (DBM), Deep Belief Network (DBN), Capsule Neural Network (CapsNet), Recurrent Neural Network (RNN), Long-Short Term Memory (LSTM), and Generative Adversarial Network (GAN).

ConvNet architectures fall under the category of supervised learning approaches, and they make the explicit assumption that the inputs are images. They have brought about breakthroughs in image processing (Hou et al., 2018), video processing (Ballas, Yao, Pal, & Courville, 2016; Mathieu, Couprie, & LeCun, 2016), speech and audio processing (Hou et al., 2018), and solving graph problems (Niepert, Ahmad, & Kutzkov, 2016). They have also recently been the most preferred choice, over state-of-the-art handcrafted feature detectors, by the computer-vision community for solving image classification (Al-Saffar, Tao, & Talab, 2017; Copur, Melisozyildirim, & Ibrikci, 2018; Pak & Kim, 2018) and object recognition (Eriş & Çevik, 2019; Girshick, Donahue, Darrell, & Malik, 2014; Sultana, Sufian, & Dutta, 2020; Ulku & Akagunduz, 2019).

ConvNets are one of the most researched and high-performing architectures in the field of artificial intelligence. Nonetheless, they have their limits and fundamental drawbacks, which are that they do not take into account spatial relationships, orientation, perspective, and size of features. To tackle these limitations, Geoffrey Hinton and his teammates introduced a modified version of

ConvNets named capsule network (CapsNet) based on the so-called capsules (Sabour, Frosst, & Hinton, 2017). Section 3.2.2 presents more detail about ConvNets.

Recurrent neural networks (RNN) are popular deep learning architectures that have shone a light on many natural language processing (NLP) and speech recognition tasks (Alom et al., 2019). RNNs are suitable to process time-series, DNA sequences, and sequential data (such as text and speech) with no predetermined limit on the input and output size vectors (Lecun et al., 2015). They map variable-length input sequences to variable-length output sequences. Due to this capability, RNNs can also be applied in health care for modeling sequences when huge amounts of sequential data, such as patients' medical history, are available to process (Khan & Yairi, 2018). Since training the conventional RNNs has proved to be problematic due to the vanishing gradient problem, more advanced variants of RNNs, including Long short term memory units (LSTMs) and gated recurrent units (GRUs), were also proposed in the literature (Cho et al., 2014; Hochreiter & Schmidhuber, 1997). Researches combining CNN and RNN architectures for automated diagnosis in the medical field are also rapidly increasing (Abdi et al., 2017; Mao & Yin, 2016; Qin et al., 2019).

Restricted Boltzmann Machine (RBM), Deep Belief Networks (DBN), Discriminative Deep Belief Networks (DDBN), and Deep Boltzmann Machines (DBM) are deep learning models that belong to the "Boltzmann family," in the sense that they utilize the RBM as a learning module. RBM is a generative stochastic neural network containing a layer of observable variables and a single layer of latent variables. DBNs are built with RBM and have multiple latent layers. Unlike CNN algorithms, the DBM, DBN and DDBN algorithms do not rely on labeled datasets and can work in an unsupervised manner (Karabulut & Ibrikci, 2017). Similarly, an autoencoder is a deep neural network approach used for unsupervised feature learning with efficient data encoding and decoding, typically for data compression and dimensionality reduction.

Generative adversarial networks (GAN), which are developed by Ian Goodfellow and his colleagues in 2014 (Goodfellow et al., 2014), are able to generate new data with the same statistics as the training data set. Hence, they have been recently used to produce photorealistic images and “deep fake” videos for application range from medicine to visualization of industrial designs (Vondrick, Pirsiavash, & Torralba, 2016; Yi, Walia, & Babyn, 2019).

Recently, deep learning algorithms have been applied in almost all application domains, and these algorithms are consequently called a universal learning approach. The three major catalysts that contributed to the recent abundant successes of deep learning algorithms have been the empowerment of High-Performance Computing (HPC) systems, the appearance of publicly available high-quality huge data sets, and open-source frameworks.

The development of faster computational devices such as graphics processing units (GPUs), which is a type of processor that was originally designed for gaming and made parallel processing ever faster by a factor of more than 50, cheaper, and more powerful, allowed for significant acceleration of training deep learning models that are big enough to take advantage of the availability huge datasets.

The increased availability of high-quality, enormous publicly available labeled datasets such as ImageNet (Deng et al., 2010), which is a large database of nearly 15 million manually annotated images, contributed to the huge boost of deep learning models. Similarly, the open-source deep learning frameworks, such as TensorFlow (Abadi et al., 2016), Theano (Bergstra et al., 2010), PyTorch (Paszke et al., 2019), Keras (Chollet, 2015), and Caffe (Jia et al., 2014), which allow researchers for faster prototyping through implement deep learning methods in a high-level language rather than implementing low-level GPU operations, have also played their role in the recent success of deep learning models.

1.2. Artificial Intelligence in Medicine

Since the inception of AI, researchers have tried to explore the potential application of AI techniques in every field of medicine from the diagnosis to the prediction that later functioned as the groundwork. Methods such as the Bayesian classifier, knowledge-based expert systems, and artificial neural networks were the potential AI techniques explored in the various subfields of medicine (Kulikowski, 2019). However, medicine is considered to be a perennial latecomer to the technology adoption of AI that had emerged in the early 1970s (Amisha, Malik, Pathania, & Rathaur, 2019; Kaul, Enslin, & Gross, 2020; Kulikowski, 2019; Ramesh et al., 2004).

In 1973, the SUMEX-AIM project was created by Stanford University to explore the technological and clinical opportunities for the emerging field of AI in medicine, such as infectious disease diagnosis, cancer drug development, and work on diabetic retinopathy image (Freiherr, 1980). Similarly, the application of AI technology in the field of surgery for the possibility of a computer-based diagnosis of acute abdominal pain was investigated in 1976 (Gunn, 1976). In addition, a rule-based expert system (EMYCIN) that recommend antibiotic treatment options adjusted appropriately for a patient's body weight (London, 1998), and a decision support system (DXplain) that help to explain the clinical manifestation of more than 500 diseases were also explored in the 1980s (Shortliffe, 1977).

In the 1990's, ANNs were heavily used for a different type of medical diagnosis problems including diagnosis of back pain disorders (Bounds, Lloyd, & Mathew, 1990), analysis of ophthalmological data from glaucoma patients (Henson, Spenceley, & Bull, 1997), predicting survival in colorectal cancer patients (Bottaci et al., 1997), prostate cancer classification (Babaian et al., 1998), diagnosis of acute abdominal pain (Pesonen, Ohmann, Eskelinen, & Juhola, 1998), predict the presence of suspected common bile duct stones in patients at high risk of having duct stones (Golub, Cantu, & Tan, 1998).

AI in medicine has evolved swiftly over the past two decades enabled by the three main trends that are converging and holding the potential to transform healthcare into all areas of medicine. First, the advancements of ML algorithms especially deep learning technologies. Second, the significantly enhanced computing powers, such as GPU, Tensor Processing Unit (TPU), and cloud computing and storage, available to developers and researchers. The third trend is the generation of publicly available huge quantities of medical data from sources such as electronic medical records (EMRs), high-resolution medical images, genome sequencing, and biosensors with a continuous output of physiologic metrics.

The current limitations, including bias, privacy and security, and lack of transparency along with limits on analysis of such huge data by physicians alone have also necessitated computer scientists and AI specialists to improve medical care via developing AI-based automated computer-aided diagnosis systems (CAD). IBM's Watson Health (IBM, 2013) and Stanford University's PAC (Pusiol et al., 2016) AI-systems to efficiently detect heart failure based on EMRs and behavioral changes in elderly people living alone using multi-modal visual data are two of the recent prime examples on the role of AI in medicine.

With an increasing life span, cancer (including brain tumor) has been a major health problem and is now the second leading cause of death worldwide (Cancer - WHO, 2018). Hence, working on techniques to prevent, detect, and cure these diseases is an excellent path towards increasing life expectancy and end-of-life treatments. Thus, this study was initiated within the scope of these issues through developing fully automated computer-aided diagnostic models for brain tumor diagnosis using AI and patients' pre-operative medical imaging scan results. To achieve this, we have developed novel deep learning architectures, and various traditional machine learning methods have also been used in our study.

1.3. Brain Tumor

Biologists define a cell as the smallest structural and functional unit of all known organisms. These cells contain the fundamental molecules of life and of which all living things are composed. Cells in our bodies grow in a controlled fashion as new cells replace old or damaged ones. However, for one or more reasons, some cells may reproduce uncontrollably and crowd out normal cells.

According to the American cancer society (ACS), the term cancer is an umbrella term for more than 100 diseases caused by the uncontrolled growth of abnormal cells that can start anywhere in the body, go beyond their usual boundaries to invade adjoining parts of the body, and/or spread to other organs. The cancer had not been considered a serious human disease for thousands of years. However, during the past century, major advances in the field of modern biology have significantly improved understanding of what cancer really is and how it progresses, which may eventually lead to the patient's death if the spread is not controlled at an early stage.

As reported in the National Cancer Institute statistics, the overall incidence and death rate of cancer had increased highly until the early 1990s. However, due to the decline in death rates for breast, colorectal, lung, and prostate cancers, the death rate of the overall cancer decrease continuously through 2017. Nevertheless, cancer continues to be a major public health problem across worldwide of all income levels. According to recent studies from the World Health Organization (WHO), cancer is the second leading cause of death both in the USA alone and globally, where every year over 9 million patients are dying in the world (Siegel, Miller, & Jemal, 2020).

There were an estimated 14.1 million new cancer cases and 8.2 million cancer deaths worldwide in 2012 (Torre et al., 2015). Similarly, according to the statistics obtained between 2015 and 2017, the cause of more than one in four of all deaths in the UK was cancer (Cancer Research UK, 2017). Like most of the middle-income countries, cancer incidence and mortality rates in Turkey have also

been increasing. In the year 2000, above 13% of all deaths in Turkey were due to cancer (Yilmaz et al., 2011), and around 210,537 people were newly diagnosed with cancer, and among them, more than 55% have died in 2018 (WHO, 2018).

Because of the rapid population growth, aging and increased prevalence of risk factors, unfortunately, it is observed that a higher proportion of cancer-related deaths are gradually increasing in economically undeveloped countries, which possess only 5% of the resources allotted for cancer worldwide. For instance, it is estimated that there were about 715,000 new cancer cases and 542,000 cancer deaths in Africa in 2008, and these numbers are expected to double over the next 20 years (Jemal et al., 2012).

The number of new cases of cancer patients is projected to be increasing every year throughout the world. In the US alone, there will be an estimated 1.8 million new cancer cases diagnosed and 606,520 cancer-related deaths in 2020. In addition, in 2030, cancer incidence rates will almost double (20 million) and are expected to kill around 11.8 million people, and around 70% of these deaths are also expected to occur in low- and middle-income countries (Siegel et al., 2020).

The human brain that weighs around 3lb (1.4kg) and controls all our body's functions is considered as the most complex and important organ of our body. The primary task of the brain is to help keep the entire body in an optimal condition in the environment in order to maximize the chances of survival. In addition to this, it produces the emotions, perceptions, and thoughts that guide our behavior (Carter, Aldridge, Page, & Parker, 2019). It is also considered as one of the most sensitive parts of the human body especially in terms of diseases that affect its functionality. Moreover, it is a soft jelly-like mass of tissue that contains a staggering one hundred billion neurons and up to 5 trillion glial cells. The neurons are responsible for sensing change and communicating with other neurons. Glial cells have four main functions: (1) to surround neurons and hold them in place; (2) to insulate one neuron from another; (3) to supply nutrients and oxygen to neurons; (4) to destroy pathogens and break down dead neurons.

Brain cancers, which are commonly referred to as brain tumors, are not a very common disease. However, they are one of the most common deadly types of cancers faced by modern medicine. Brain tumors are characterized by the uncontrolled growth and spread of abnormal cells in the central nervous system (brain and spinal cord) that can disrupt proper brain function, and are an important cause of morbidity and mortality in both adults and children, often generating severe disabilities (PDQ, 2020; Stewart & Wild, 2014; Kaba & Kyritsis, 1997).

The causes of most brain tumors are unclear yet. But so far, oncologists have identified that exposure to ionizing radiation, exposure to certain chemicals, work in the nuclear industry, race (Caucasians get more often than people of other races), and a family history of a brain tumor or inherited syndromes such as tuberculosis, neurofibromatosis, and von Hippel-Lindau diseases may increase the risk of brain tumor (J Strong & Garces, 2016; Ohgaki, 2009; Rees, Smirniotopoulos, Jones, & Wong, 1996). See Table 1.1 for more potential risk factors associated with brain tumors.

As defined by the WHO, there are more than 120 types of brain tumors that differ in the tumor cell origin, location, size, cancerous, and characteristics of the tumor tissues. However, expert oncologists/pathologists progressively organize them into different classes. According to their origin, they classify the tumor as primary or metastatic, according to the degree of malignancy, they classify them as benign or malignant, and according to the cell of origin, they classify them as neuroma, meningioma, craniopharyngioma, glioma, pituitary adenoma, etc.

Brain tumors that grow from Schwann cells are named neuroma vestibular schwannoma. Meningiomas are the most common primary brain tumor that grows from dura cells. Pituitary adenoma and craniopharyngioma tumors are grown from pituitary gland cells. Similarly, gliomas are brain tumors that grow from the glial cells, where around half of all primary brain tumors grow (Laws & Thapar, 1993). Fig 1.2 illustrates some of the brain tumor types and their origin in the brain.

Table 1.1. Epidemiologic risk factors that may increase a person's possibility of developing a brain tumor.

Risk Factors	**Examples**			
Radiation	Ionizing	Electromagnetic radiation	Cell phones and radiofrequency radiation	
Exposure to Infections and Viruses	Epstein-Barr virus (EBV)	Polyomaviruses	Cytomegalo virus (CMV)	Toxoplasma infection
Diet and Vitamins supplementation	N-nitroso compounds	Fat intake	low in vitamin C-fruits and vegetables	Aspartame ingestion
Head injury and Seizures	Serious head trauma		A history of seizures	
Alcohol and Tobacco				
Chemicals	Hair dyes and sprays	Traffic-related air pollution		
Genetics	Neurofibromatosis type I & II	Von Hippel-Lindau syndrome	Li-Fraumeni syndrome	Turcot syndrome
Home and Work Exposure	Electrical workers and electromagnetic fields	Agriculture workers exposed to pesticides, herbicides, and fungicides	Other industries (vinyl chloride, oil products, and rubber industries)	

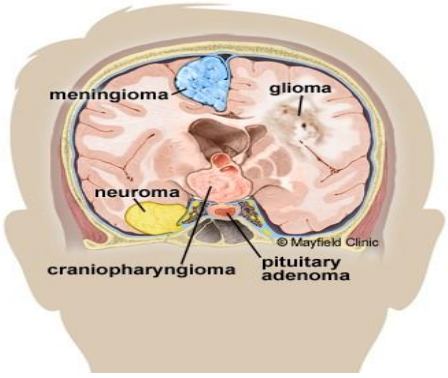

Figure 1.2. Examples of brain tumor types according to the cell of origin (image is taken from https://mayfieldclinic.com)

A brain tumor is considered as primary if there is an abnormal growth of cells that starts in the central nervous system (brain and spinal cord) and usually does not spread to other parts of the body. However, the metastatic (secondary)

brain tumors begin as cancer elsewhere in the body (lung, breast, prostate, etc.) and spread to the brain through the bloodstream or lymphatic vessels. Brain tumors caused by slowly enlarging growths and do not contain cancer cells are labeled as benign. They typically have clear borders that do not spread into other tissue. On the other hand, brain tumors that grow more rapidly and develop from enlarging growths are described as malignant (cancerous). These malignant tumors do not have clear borders and may even spread to other regions or return after surgical removal. Brain tumor differs from other tumors in that both malignant and benign tumors can result in significant morbidity and mortality due to their intracranial location (McKinney, 2004).

A brain tumor is responsible for substantial morbidity and mortality across the globe, and the overall incidence is increasing every year. In 2016, there were 330 000 incident cases of brain cancer and 227 000 deaths globally. In the same year, China, USA, and India were the top three countries with the highest number of incident cases (Patel et al., 2019). According to the report by CBTRUS, a total of 405,740 Americans of all primary brain tumors were diagnosed during the year 2012-2016. During the same year, the overall annual average age-adjusted incidence rate was 23.41; the aggregate total number of observed deaths was 79,718, and an annual average age-adjusted mortality rate of 4.42 per 100,000 populations. For 2017 and 218, the total number of new cases of primary brain and other CNS tumors for all the District of Columbia and the other 50 states was projected to be 78,370 and 79,870, respectively (Ostrom et al., 2019).

According to the report by the American Society of Clinical Oncology (ASCO), a brain tumor is the 10th leading cause of death in the USA, and this year, an estimated 3,540 children under the age of 15, and 13,590 men and 10,300 women adult patients is expected to be diagnosed with primary cancerous tumors of the brain and spinal cord. It is also projected that among the 23,890 adults primary cancerous brain and CNS tumor patients, 18,020 of them will die this year (ASCO, 2020).

Gliomas are a broad category of brain and spinal cord tumors that originate from the precursor or gluey supportive cells (glial cells), including astrocytes, ependymal, and oligodendrocytes cells. Any of these three types of normal glial cells mentioned may produce a specific type of glioma tumor that spans a broad spectrum of biological aggressiveness. Gliomas produced from astrocyte cells are commonly referred to as astrocytoma (including glioblastoma). Similarly, ependymal cells may produce ependymoma tumors, and oligodendroglioma tumors come from oligodendrocyte cells. Astrocyte cells are star-shaped dominant macroglial cells in the brain and spinal cord that outnumber neurons by over fivefold. Astrocytes involve in the repair process of the brain and spinal cord tissue, provision of nutrients to the nervous tissue, neuron maintenance, and in the formation of the blood-brain barrier.

Ependymal cells are the thin neuroepithelial lining of a region of the brain known as the ventricular system, where cerebrospinal fluid is created and circulates, and the central canal of the spinal cord. The ependymal cells are an extremely important component of the developing and mature brain by playing an important role in the production and regulation of cerebrospinal fluid, which has many functions, including nutrient transfer to and from the brain and protection against shock injuries.

Similarly, oligodendrocyte cells are the myelin-forming cells and the most vulnerable cells of the central nervous system. They play an important role in the uptake and inactivation of chemical neurotransmitters released by neurons, formation of the myelin sheath of neurons and mechanical support of neurons in the central nervous system (Bradl & Lassmann, 2010; Johns, 2014; Sofroniew & Vinters, 2010; Wolburg, Wolburg-Buchholz, Reichenbach, & Mack, 2015). An illustration of neurons and glial cells is depicted in Figure 1.3.

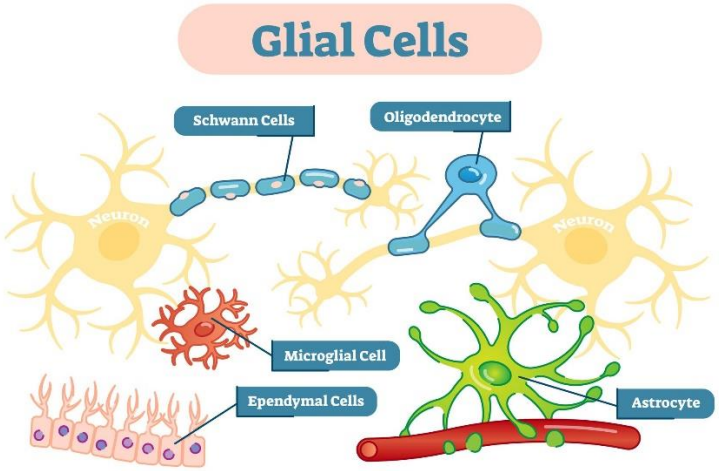

Figure 1.3. Illustration of glial cells (ie, astrocytes, oligodendrocytes, and ependymal cells) (image taken from verywellhealth.com)

Gliomas are also primary tumors that represent about 80 percent of all malignant brain tumors, 30 percent of all brain and central nervous system tumors, and the number of incidences is increasing every year (Goodenberger & Jenkins, 2012; He, Mitteer, Mou, & Fan, 2016). Gliomas can strike at various ages, depending on the subtype, but are more likely to occur in the elderly, where their incidence seems to be increasing. In addition, gliomas affect both sexes and all human races. But gliomas are slightly more likely to occur in men than in women and more common in Caucasians than in African Americans (Desmeules, Mikkelsen, & Mao, 1992; Ostrom et al., 2019). Despite some gliomas that do not show any symptoms, the most common symptoms are; headaches, vision loss, numbness, seizures, problems with speech, personality changes, nausea and vomiting, weakness in some parts of our body, and cranial nerve disorders as a result of increased intracranial pressure.

1.4. Management of Glioma

1.4.1. Classification

Gliomas are a pretty heterogeneous group of brain tumors with multiple histologic types and malignancy grades in which many of them have significantly different biologic properties, treatment options, and prognoses (Louis, Schiff, Batchelor, & Wen, 2017). The WHO classification of tumors of the CNS, which is revised four times since the first published in the year 1979 and the most recent revision published in 2016, is the standard and universally accepted diagnostic system for the grading and classification of brain tumors.

Gliomas range from WHO grade I that are slow-growing to WHO grade IV that are the most malignant and almost incurable. Historically, the slower-growing lesions, corresponding to WHO grade I and grade II, have been commonly referred to as low-grade gliomas (LGGs), while the more rapidly progressive tumors corresponding to WHO grade III and grade IV are commonly referred to as malignant gliomas or high-grade gliomas (HGGs) (Bauer, Wiest, Nolte, & Reyes, 2013; Louis et al., 2007). Further classification of glioma is also possible depending on the underlying histopathological and molecular parameters of the tumor along with the fourth edition of the WHO (2016) CNS Tumors classification guidelines, in which gliomas coarsely classified as diffuse glioma and non-diffuse glioma (Louis et al., 2016; Wesseling & Capper, 2018). Table 1.2 depicts diffuse and non-diffuse glioma subtypes and their corresponding malignancy grade.

Table 1.2. Glioma types and malignancy grading.

WHO grade I	WHO grade II	WHO grade III	WHO grade IV
Pilocytic astrocytoma	Diffuse astrocytoma	Anaplastic astrocytoma	Glioblastoma multiforme
Myxopapillary ependymoma	Ependymoma	Anaplastic Oligodendroglioma	Diffuse midline glioma
Angiocentric glioma	Astroblastoma	Oligoastrocytoma	Diffuse astrocytic glioma
Ganglioglioma	Fibrillary astrocytomas	Anaplastic ganglioglioma	Gliosarcoma

The diffuse gliomas are the most common glioma tumors characterized by extensive infiltrative growth into the surrounding CNS parenchyma, and generally affect the cerebral hemispheres of adults. Oligodendrogliomas, oligoastrocytomas, diffuse midline glioma, and glioblastoma are examples of diffuse glioma types. In addition, based on the absence/presence of histological features such as necrosis, marked mitotic activity, and lorid microvascular proliferation, glioma types within diffuse glioma subgroups have a malignancy grade between WHO grade II and WHO grade IV. However, the diffuse low-grade gliomas represent only 15% of gliomas, and unfortunately, over 70% will transform to high-grade gliomas (Jooma, Waqas, & Khan, 2019).

The diffuse midline gliomas, which are newly introduced in the 2016 revision of the WHO classification of tumors of the CNS, occur in children and young adults, and they are most frequently located in the brain stem, thalamus, and/or spinal cord. In addition, recent studies demonstrate the clinical significance of considering diffuse midline glioma as a universally fatal prognosis and a distinct entity corresponding to WHO grade IV (Karremann et al., 2018). Unlike diffuse gliomas, non-diffuse gliomas are generally much more circumscribed. Different variants of pilocytic astrocytoma and ependymoma are examples of this category (Kristensen, Priesterbach-Ackley, Petersen, & Wesseling, 2019).

1.4.2. Prognosis

Prognosis is a branch of medicine that encompasses many aspects but classically specializes in predicting the likely or expected development of a disease, such as informing patients the likelihood of survival with a certain illness, forecasting the risk of a certain illness recurrence after therapy, predicting the 5-year certain illness risk of an individual, and pre-operative grading/stage of cancer. As such, the prognosis may vary depending on the seriousness of the injury, disease type, race, sex, age, and implemented treatment types (NIH, 2020.).

Knowledge of the patients' prognosis is a key element because it does not only assist physicians in deciding on appropriate treatment options but also in discussing the opinions of patients or relatives concerning management in palliative care settings if the life expectancy is minimal. Appropriate use of treatment options or patient palliative care management will decrease the economic burden and provide optimal life-extending therapies or end-of-life care (Hansebout, Cornacchi, Haines, & Goldsmith, 2009; Kwok, Pang, Salvo, & Chow, 2011). In clinical practice, for instance, the prediction of the 10-year risk of heart attack is applied to decide whether a patient should get drugs to reduce the risk (Babatunde et al., 2020; Benner et al., 2008). The six-month mortality risk for any disease is also used for patients with seriously ill and incurable conditions to decide who should receive end-of-life treatment (D. H. Lee, 2002).

The advanced glioma patients represent a heterogeneous group of people. Thus, prognostication in glioma remains a difficult topic for health care professionals, patients, and patient's families to make important clinical decisions. The prognosis of gliomas is given in relation to the cellular subtype and grade of the tumor the patient presents with. Typically, any patient with more malignant than WHO grade I tumor will have a prognosis resulting in eventual death, varying from years to months (Maher et al., 2001; Sanai, Chang, & Berger, 2011). Hence, prognosis models in glioma are also useful for health professionals in guiding when choosing the right treatment options for each patient, end-of-life planning for seriously ill patients, and avoid unnecessary treatments.

1.4.2.1. Survival Times

Overall survival time (OS) and progression-free survival time (PFS) are two important medical terms in the oncology field often used during the management of glioma patients. PFS is the duration of time from treatment initiation to post-treatment of a disease state, in which a patient lives with the disease condition showing no signs of progression or worsening. Whereas, OS, a

universally-accepted direct measure of clinical benefit, is the duration of time from the date of diagnosis or commencement of treatment of a particular disease to date of death or date of last follow up if the patient is still alive (Hess, Brnabic, Mason, Lee, & Barker, 2019).

1.4.2.2. Low grade gliomas

Lower grade gliomas are typically less aggressive and the prognosis is optimistic than patients with high-grade glioma. One study at the Eskisehir Osmangazi University Hospital Department of Neurosurgery Department reported that low grade gliomas patients diagnosed in the years 1980 - 2006 have a median survival of 11.75+/-1.24 years (Durmaz et al., 2008); another study on low-grade oligodendroglioma patients reported a median survival of 16.7 years (Olson, Riedel, & DeAngelis, 2000). Moreover, children and adolescents patients with low grade glioma have a better prognosis with more than 90 % of all patients survive long term (Yiallouros & Tallen, 2007).

1.4.2.3. High grade gliomas

The median overall survival time of the high grade glioma patients is generally poor compared to the low grade glioma patients. For instance, anaplastic astrocytoma high grade gliomas have a poor median overall survival of 3.5 years but the anaplastic oligodendroglioma patients have a median survival rate of above 10 years. However, the other high grade glioblastoma multiform has a survival of only 15 months, and only 27% survive 2 years (Stupp, Brada, van den Bent, Tonn, & Pentheroudakis, 2014; Bleeker, Molenaar, & Leenstra, 2012). The overall survival time of patients with high grade glioma is often categorized into two (long-term vs. short-term) or three (short-term, medium-term and long-term) groups of survival for valuable insights in advancing surgical and treatment plans (Adeberg et al., 2014; Gandía-González et al., 2019; Lacroix et al., 2001; Macyszyn et al., 2016).

1.4.3. Diagnosis and Treatment

Glioma is a highly heterogeneous and complex type of brain tumor with considerable variability in age of onset, histological features, grade of severity, and survival time. As a result, there is a lack of effective diagnostic strategies for glioma management, and accordingly places substantial psychosocial and financial burdens on patients and their families. In the current trends of glioma management, if any patient admitted at a hospital is suspected of a glioma tumor, after receiving appropriate personal and family medical history interrogation and physical clinical evaluation at the initial visit, the patient should be immediately referred to oncologists who are doctors specialized in cancer. The patient may also be referred to doctors who specialized in brain disorders (neurologists) or to doctors who use radiation to treat cancer (radiation oncologists). A collaboration between various specialists including oncologists, radiation oncologists, neurosurgeons, neuropathologists, and neuro-radiologists may also be required for counseling and to find out whether the patient suffers from glioma.

If any patient is suffered from glioma, a diagnosis will be initiated that includes determining the exact locations of the tumor (localization or segmentation), glioma subtype, malignancy of the disease (grading), and patient's survival time. Obtaining the diagnosis results of these details is unequivocally essential for optimal management and treatment planning of the disease. Medical imaging techniques, such as Magnetic Resonance Imaging (MRI), Magnetic Resonance Spectroscopy (MRS), Computed Tomography (CT), and Positron Emission Tomography (PET), which provide a very detailed image of the brain and tumor, play a central role for non-invasive and preoperative diagnosis, and support during the evaluation of glioma. Moreover, these medical imaging scans show a brain tumor in almost all cases and also may give an idea about the type of tumor based on how it looks on the scan and where it is in the brain (Laws & Thapar, 1993).

For making a definitive diagnosis of glioma, a neurosurgeon may also take a sample of the tumor using surgery/biopsy with subsequent pathological evaluation using a microscope by pathologists. A biopsy, which is performed using a small needle to remove a sample of tissues from the tumor, is usually performed without a larger surgery. This approach is preferred if the tumor is located within a critical area of the brain or if the patient's overall health condition doesn't permit surgery. According to their appearance under the microscope and by using certain molecular and genetic markers, the pathologists will then determine the type and grade of the tumor and the survival time of the patient. For instance, when a pathologist observes a low proliferative potential and infiltrative morphological features, they categorize the malignancy of the tumor as low grade glioma, with patient survival from 5 to 15 years. Similarly, high grade gliomas are exhibited by pathologists through observing the presence of nuclear atypia, brisk mitotic activity, and microvascular proliferation and pseudopalisading necrosis (Ertosun & Rubin, 2015)

Once a glioma is diagnosed, treatment options for glioma encompassing from surgical resection to chemotherapy or radiotherapy, and other therapies and experimental clinical trials are used (Buckner et al., 2007). They may also be used either separately or in combination. However, patients follow different treatment plans according to the grade of the tumor, the extent of tumor resection, biomarker profile, patient's age, and survival time prediction. In terms of current standard treatments, grade I and grade II low-grade glioma tumors are grouped together and treated similarly. Likewise, grade III and grade IV high-grade glioma patients are grouped together and treated similarly (Alifieris & Trafalis, 2015; Altwairgi et al., 2017).

The goal of surgery is to remove as much tumor tissue as possible without causing too much damage to nearby organs or tissues that may disturb neurological functions. Considering that the entire tumor may not be removed using surgery/biopsy alone, radiation therapy may be applied after surgery/biopsy using

high-energy x-rays in several sessions to kill any remaining tumor cancer cells while sparing the healthy tissues in the brain. Similarly, the objective of chemotherapy is to stop or slow the growth of cancer cells by applying medicines. When used in combination with radiation therapy and surgery, chemotherapy may improve survival and quality of life in some patients with high-grade gliomas (Buckner et al., 2007). Other treatments such as corticosteroids and anticonvulsant therapy can also be used in the patients to alleviate symptoms and improve the quality of life (Timmons, 2012). Moreover, for high-grade glioma patients where a short survival time is predicted, hospice care that includes the physical (e.g., pain relief), psychological, social, and spiritual aspects of treatment is recommended, since the disease may not be cured (Giammalva et al., 2018).

1.5. Magnetic Resonance Imaging in Glioma

1.5.1. MRI Techniques

Medical imaging is the technique and process of creating visual representations of the internal of the human body using medical scanners and a computer, with medical imaging required for 30 percent of all medical conditions (DeStigter et al., 2019). It is being used by hospitals and healthcare organizations to aid physicians in the diagnosis and management of patients, and therefore plays an important role in initiatives to improve public health for all population groups.

Magnetic Resonance Imaging (MRI) is an advanced medical imaging scanner of modern healthcare where most commonly used in radiology for imaging the human brain, spinal cord muscles, joints, and other structures in studying anatomical details and pathological abnormalities (Liang & Lauterbur, 2000). In 2003, the Nobel Prize was awarded jointly to Lauterbur (Lauterbur, 1973) and Mansfield and (Mansfield, 1977) to give recognition for their discoveries concerning MRI.

Computer Tomography (CT) and X-Rays are the other medical imaging scanners commonly used in radiology (Drevelegas & Papanikolaou, 2011). X-rays

are most commonly used to detect broken bones and certain cancer types through creating 2D images of the patients' body parts. However, CT and MRI are commonly used to diagnose conditions in patients' organs and soft tissues by creating high-quality cross-sectional three-dimensional (3D) images taken from multiple angles. Moreover, CT and MRI are multi-planar, which means that images can be reconstructed in the axial, coronal and sagittal multiple body planes without changing positions, allowing for better appreciation of anatomical location and for more accurate longitudinal assessment, which provides detailed information to the physicians to detect any abnormality behavior (Reuter et al., 2014).

A major difference between an MRI and a CT scan is that the MRI uses radio waves emitted by atoms subjected to a magnetic field to create images of detailed internal structures of the human body, while CT uses multiple X-rays which provide physicians with another way to see structures of the internal body parts. As a result, MRI is non-ionized and non-invasive, unlike that of X-ray and CT scanners (Liang & Lauterbur, 2000). Moreover, since it provides high soft-tissue contrast resolution compared to the other medical imaging modalities, MRI is the most frequently used brain imaging technique to aid physicians in diagnosing gliomas accurately and choosing the right treatment options (Watanabe, Tanaka, & Takeda, 1992; Wen et al., 2010).

Gliomas have different degrees of aggressiveness, variable prognosis, and several heterogeneous histological sub-regions. As a result, they are often diagnosed and distinguished from healthy tissues by using more than one MRI technique aka multimodality. Each modality reflects diverse tumor biological properties with different image contrasts that offer insights to confirm the presence, location, and type of tumor that may be present (Ellingson, Wen, Van Den Bent, & Cloughesy, 2014). Ordinarily, physicians use the four conventional MRI modalities to diagnose glioma pre-operatively; T1-weighted (T1), Contrast-Enhanced T1-weighted (T1Ce), T2-weighted (T2) and Fluid-Attenuated Inversion Recovery Imaging (FLAIR) (Zelenak, Viera, & Hubert, 2013). Radiographers generate these

four MRI sequences by adjusting the Time to Echo (TE) and Repetition Time (TR) of the MRI scanner parameters. TE is defined as a duration between the series of successive pulses applied to the same slide where TR is defined as the time interval between two continuous radiofrequency pulses used in the same sequence (Brown & Semelka, 2011).

When the radiographers want to generate T1 MRI images, they de-emphasize the longitudinal relaxation by using a short TR between 300-600ms and a TE of 10-30ms. Similarly, T2 MRI images are generated through de-emphasizing the transverse relaxation based on longer TE and TR times than T1 images with TR >2000ms and TE-90-140ms. Both T1 and T2 images can be used to provide a detailed representation of the boundaries between gray matter, white matter, and cerebrospinal fluid (CSF). In addition, pathological lesions are usually high signal on T1 and low signal T2 images (Zimny, Neska-Matuszewska, Bladowska, & Sąsiadek, 2015).

Another common MRI sequence is the FLAIR modality which has longer TR and TE than the T1 and T2 which has the advantage to highlight abnormalities. FLAIR images are considered highly effective scans to distinguish the edema glioma sub-region from the CSF (Liu et al., 2014). In T1Ce, a paramagnetic contrast agent (usually gadolinium-based) is administered and has the effect of reducing the T1 relaxation time and thereby increasing the signal intensity to improve the contrast of the borders of the area that are affected by hemorrhage. The internal architecture seen on T1Ce is determined by the presence of necrosis core, cysts, and tumor boundaries (Mittal, Wu, Neelavalli, & Haacke, 2009). Figure 1.4 depicts examples of the four MRI modalities of a glioma patient in the axial plane.

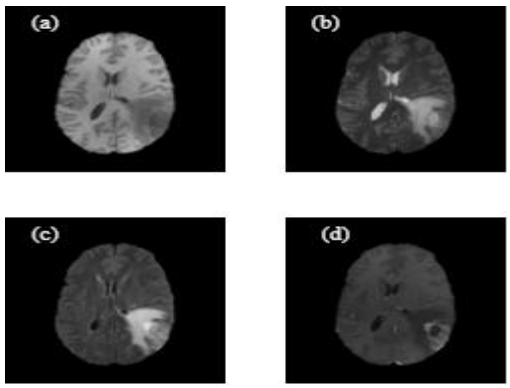

Figure 1.4. MRI images of the brain of glioma patient in the axial plane with different types of image contrasts: (a) T1; (b)T2; (c) FLAIR; and (d) T1Ce

1.5.2. Automated Diagnosis of Glioma using MRI

During the past decades, glioma management based on either removing a section of the tumor during surgery or based on a biopsy through taking a sample of the tumor by inserting a needle through the skull into the brain, then let pathologists examine the tumor tissue under a microscope were the domain of oncologists. However, these approaches are not only invasive but also prone to deformity, tissue trauma and sampling error leading to neurological function impairment or even life-treating effects (Coons, Johnson, Scheithauer, Yates, & Pearl, 1997; Jackson et al., 2001; Shankar, Balaj, Stott, Nahed, & Carter, 2017). Moreover, they are time-consuming, inaccurate and subjective tasks even for experienced pathologists due to the complex mechanical properties of the tissues, which vary from patient to patient (Christakis & Lamont, 2000; Peleg & Biderman, 2000).

Thus, it was indispensable to come up with a non-invasive approach that can overcome the above-listed issues, and the comprehensive MRI imaging techniques that can be applied in glioma management, including prognosis models for classifying glioma grades and survival time rapidly and accurately, has become a continuing trend allowing an effective pre-operative diagnosis. Consequently,

radiologists have increasingly taken over that diagnostic role, and the management of gliomas has improved considerably (Reardon & Wen, 2015; Zelenak et al., 2013).

During the past decades, an increase in the number of patients diagnosed with brain-related diseases, including glioma, has been observed. As a result, demand for MRI exams has been increasing across a range of countries in which around 2 billion brain MRIs are scanned each year and almost 40 million MRI scans worth nearly $50 billion conducted each year in the US, out of which $1 billion worth is spent for brain scans (Callaghan, Kerber, Pace, Skolarus, & Burke, 2014; GE Healthcare Partners, 2018simonyan).

During the pre-operative diagnosis of glioma, a radiologist who is trained in the interpretation of brain MRI, visualize the smaller structures and abnormalities in the brain through visual inspection of the images that may provide valuable insights if the patient has a tumor concerning glioma location and other conditions such as glioma type and grade (Gillies, Kinahan, & Hricak, 2016; Legaz-Aparicio et al., 2017; Smits & Van Den Bent, 2017). However, due to the complex features of glioma morphology and the immense variations of glioma types, even the best-experienced radiologists can fail to accurately localize the glioma region and subsequently finding the other conditions by visual inspection of the images (Upadhyay & Waldman, 2011). Another big problem with MRI screening through visual inspection of the MRI images by the radiologists is that since MRI comes at the cost of generating a bunch number of cross-sectional images per patients, which must be analyzed one by one by the radiologists, it does not appear to be feasible that the radiologist himself has to visually inspect discriminative MRI features of the glioma region in the whole brain MRI images, calculate a bunch of numbers, makes some advanced statistics and then combines the numbers into a meaningful diagnosis, which is both expensive and time-consuming.

Moreover, there is a stagnating number of trained radiologists available to interpret the images despite the growing number of patients, particularly in low- and middle-income countries (LMICs), struggling to meet the ongoing growth in demand in diagnostic imaging (DeStigter et al., 2019; GE Healthcare Partners, 2018; Welling et al., 2011). Thus, the use of a robust automatic computer-aided volumetric image analysis system for brain tumor diagnosis becomes very essential to overcome those shortcomings, due to its ability to quantitatively and reproducibly measure MRI images on a large-scale. Ultimately, the aim of this thesis is also to develop automated brain tumor prognosis models for glioma grade and survival time predictions based on the presurgical multi-modal MRI scans with the goal to significantly automate and reduce the workload of the radiologists.

1.6. Thesis Description and Goal

This PhD thesis focused on the development of fully automated and computerized prognosis models for patients with a brain tumor, in particular glioma, based on multi-modal MRI images acquired pre-surgically. More specifically, pre-operative glioma grading and survival time prediction models have been aimed to develop. For this purpose, novel end-to-end deep learning architectures, in particular convolutional neural networks, have been developed. The multi-modal volumetric MRI images obtained from BraTS'17 challenging data set were used as data in this study. The developed model can have a contribution to an early diagnosis of patients with a brain tumor and most importantly it should provide valuable insights for physicians in advancing treatment planning using only MRI at the time of diagnosis.

Transfer learning techniques consisting of a fine-tuned deep learning models based on natural images and classical data augmentation techniques that increase the available training data sets were used to improve the performance of the proposed deep learning models. During pre-processing, prior to feeding images to the deep learning architectures, the volumetric MRI images were reconstructed

to 2D stacked images to tackle the GPU memory constraint and reduce the training time. Moreover, another novel pre-processing technique, which is a concatenation of two images, was used to leverage pre-trained deep learning models trained on natural images without enlarging them to benefit from the fine-grained details in each slice of the coronal and sagittal projections.

How the incorporation of tissues surrounding the brain tumor to the input during training a deep learning architecture affects the performance of prognosis models was also investigated through providing input images varying from the masked tumor regions to the entire square bounding box around the tumor region. For this, a well-established existing brain tumor segmentation model developed based on the convolutional neural network has been used. Moreover, the individual classification performance of the different MRI modalities and corresponding 2D projection images while developing prognosis models were also investigated.

The performances of single learners are generally unsatisfactory in the small data regime, and ensemble learning is one of the prospective areas of research in CNNs for small data regimes. Hence, an ensemble learning strategy that combines the output of more than one model to improve generalization when compared to a single learner has also been explored in this study to improve the performance of the prognosis models. Thus, besides voting and averaging, different classical machine learning algorithms have been used as an ensembling strategy in our experiments. Ultimately, the proposed approaches combine a highly efficient imaging scanner with artificial intelligence algorithms to allow a reliable, rapid, non-invasive, robust, and automated brain tumor diagnosis.

2. RELATED WORKS

2.1. CNN Architectures for Volumetric Image Recognitions

This section deals with the work related to the various CNN architecture improvements reported from 2012 to 2020 for the volumetric image recognition tasks such as classification, detection, localization and segmentation. This section also gives an insight into data augmentation and transfer learning techniques used during training CNNs for small medical image recognition tasks.

In 1998, the first multilayered CNN architecture, named LeNet-5, was introduced for a handwritten character recognition task (LeCun, Bottou, Bengio, & Haffner, 1998). The LeNet-5 produced promising performances in recognizing characters from 2D images. Hence, the architecture was essentially a 2D-CNN which made a groundwork for the current 2D-CNN architectures. However, it is after 2012 where CNNs have been extensively explored and proved their supremacy for 2D image recognition tasks.

Several different versions of deep 2D-CNN architectures have been proposed and shown substantial performances for 2D medical image analysis tasks. For instance, the AlexNet (Krizhevsky et al., 2012) and GoogLeNet (Szegedy et al., 2015) architectures have won the 2012 and 2014 ImageNet challenges with a large margin, respectively. Similarly, VGGNet (Simonyan & Zisserman, 2015) who has the same but deeper architecture as that of AlexNet secured the first position for the localization task and the second position for the classification task in the 2014 ImageNet Challenge. SqueezeNet (Iandola et al., 2016), ResNet (K. He, Zhang, Ren, & Sun, 2016) and DenseNet (Huang, Liu, Van Der Maaten, & Weinberger, 2017), MobileNet (Howard et al., 2017) and MnasNet (Tan et al., 2018) are also some of the other popular state-of-the-art deep 2D-CNN architectures in the literature who have shown substantial performances in different image recognition tasks.

Nowadays, there is a rapid increase in new CNN architecture innovation and applications. However, most of these new CNN architectures are built upon the principles introduced in the above popular state-of-the-art 2D-CNN architectures. Medical imaging analysis is not an exception, as 2D-CNNs have accomplished expert-level performances in numerous medical fields based on 2D images of a patient's organ. For instance, U-Net, one of the most popular 2D-CNN architecture in the literature, was the winner of the 2015 ISBI cell tracking challenge for the segmentation of neuronal structures in 512 × 512 biomedical microscopy images. U-Net is also applicable for many biomedical image segmentation tasks, such as liver image segmentation (X. Li et al., 2018) and brain image segmentation (Menze et al., 2015). Similarly, Cireşan et al. (2013), Gulshan et al. (2016) and Esteva et al. (2017) demonstrated the potential of 2D-CNN for the mitosis detection in breast cancer using histology images, diabetic retinopathy detection using retinal fundus photographs, and classification of skin cancer directly form images, respectively.

Following their success in 2D images, conventional 2D-CNN architectures have been easily extended to various CNN architectures to exploit volumetric images while considering their challenging properties. However, it is not obvious what architecture will produce a substantial performance in volumetric datasets. Improvements made in the conventional 2D-CNN architectures in order to apply it to volumetric image recognition tasks can be coarsely divided into three-dimensional CNN (3D-CNN) and view-based CNN architectures.

Many researchers who used 3D-CNNs for volumetric image recognition tasks basically apply the conventional 2D-CNN architectures by replacing the 2D convolutional kernels with 3D kernels. Among prior researches on this trend can be observed in the work by Maturana et al. (2015). The authors proposed VoxNet, a basic 3D-CNN architecture developed using 3D kernels, which can be applied in any 3D object recognition task. In a similar fashion, Brock et al. (2016) proposed Voxception-ResNet (VRN), a 3D-CNN that adopts CNN with 3D kernels and

achieved a significant improvement by 51.5% in the ModelNet benchmark for 3D object recognition.

Several researchers have contributed to the use of 3D-CNN in medical image classification and segmentation tasks. For instance, in the work by Yang et al. (2018), the classification of Alzheimer's disease was performed using brain MRI scans based on the 3D counterparts of some well-known state-of-the-art 2D-CNN architectures, specifically ResNet and VGGNet. In a similar fashion, a 3D-CNN was proposed for Alzheimer's disease classification using volumetric CT head scans (Gao, Hui, & Tian, 2017). The authors also compared their proposed results with state-of-the-art hand-crafted-based features and showed that the proposed 3D CNN approach gives around four percent higher classification accuracy.

In regard to medical image segmentation, a 3D-CNN architecture named Deep Medic has been developed for automatic segmentation of brain tumors, which results a high performance on 3D volumetric brain scans (Kamnitsas et al., 2016). A modified version of the original 2D-U-Net i.e., 3D U-Net for various volumetric medical image segmentation tasks has also been used by different researchers (Barıs¸ kayalıbay, Jensen, & Van Der Smagt, 2017; S. Chen et al., 2020; W. Chen, Liu, Peng, Sun, & Qiao, 2019; Isensee, Kickingereder, Wick, Bendszus, & Maier-Hein, 2018; Zeng et al., 2017).

While 3D-CNNs are easily and directly applicable to volumetric images, they have three main challenges: lack of a large number of public volumetric datasets, high computational complexity, and lack of available pre-trained 3D-CNN models (Singh et al., 2020). Most importantly, 3D images may be presented as a combination of multiple 2D images captured for the 3D object from different projections or viewpoints. Learning 3D images from the rendered 2D multi-view images of the same object aims to learn a function modeling each view separately and then jointly optimize all the functions to represent the whole 3D image. To this end, different view-based CNNs that comprise multiple 2D views of a 3D object

are mentioned in the literature to overcome the above three challenges. Most importantly, CNN's categorized under the view-based 2D-CNNs generally perform slightly better than 3D-CNNs in a range of volumetric benchmark datasets (Griffiths & Boehm, 2019; Gupta et al., 2020; Kim et al., 2020).

One of the first attempts on view-based 2D-CNNs for 3D object recognition tasks through combining multi-angle 2D projections of a 3D object in a single end-to-end architecture was presented by Su et al. (2015). The architecture took 12 rendered 2D image views of 3D objects in its simplest pipeline, sent through respective independent 2D-CNNs, and concatenated at the view pooling layer, which is subsequently sent to the second aggregated 2D-CNN architecture. An accuracy of 90.1% was reported by the authors using the proposed view-based 2D-CNNs on the ModelNet40 dataset (Wu et al., 2015), which is composed of objects such as, tables, chairs, desks, etc.

A considerable amount of literature has been published on view-based 2D-CNN for volumetric image recognition tasks. Setio et al. (2016) carried out a study on a view-based 2D-CNN architecture for pulmonary nodule detection and false-positive reduction in volumetric CT scans. The authors first decomposed the volumetric CT scans into multiple 2D views in 9 different fixed planes. Thereafter, the proposed architecture which comprises multiple streams of 2D-CNNs processed the extracted view specific 2D patches. The outputs of each stream 2D-CNN were then concatenated using various fusion methods to get the final classification results using the softmax activation function on the top of the fully connected layer. On the publicly available LIDC-IDRI dataset (Armato et al., 2011), the proposed model achieved a sensitivity of 90.1% and 85.4% at 4 and 1 false positives per scan, respectively.

Wang et al. (2017) used a view-based 2D-CNN architecture with three branches, named MV-CNN, for automated lung nodule segmentation in volumetric CT images. The authors extracted detailed texture and nodule shape information using axial, sagittal, and coronal 2D projection of the volumetric CT images as the

input to the three branches of the MV-CNN. After training on 393 nodules obtained from the publicly available LIDC-IDRI dataset, the MV-CNN obtained encouraging segmentation results with an average dice similarity coefficient (DSC) of 77.67%.

Geras et al. (2017) proposed a multi-view 2D-CNN architecture for breast cancer screening using a set of high-resolution mammographic images of a patient's breast obtained in cranial-caudal and mediolateral oblique views. The proposed method obtained performance comparable to a group of four experienced doctors when examined with the same data. Similarly, Bien et al. (2018) proposed a fully automated view-based 2D-CNN architecture based on AlexNet for feature extraction in knee diagnosis using an MRI examination of patients. The proposed architecture was named MRNet. The inputs to the MRNet were a stack of 2D slices in one of the three views (axial, coronal, and sagittal). The out probability by each view was then combined using logistic regression. The authors achieved AUC values of 0.937, 0.965, 0.847 in detecting abnormalities, ACL tears, and meniscal tears, respectively. Moreover, the authors found no significant difference between the performance of MRNet and that of expert radiologists in detecting abnormalities.

Zhang et al. (2019) proposed a novel view-based 2D-CNN architecture for automated volumetric digital breast tomosynthesis classification that simultaneously handles the variable number of slices and retains slice-to-slice variation. The proposed approach leverage AlexNet pre-trained on ImgeNet and achieved 28.8% higher than approaches based on 3D-CNNs on the same data. Gupta et al. (2020) investigated the incremental impact of volumetric data on 2D reformation combined with novel data augmentation and transfer learning on the performance of view-based 2D CNNs using small coronary CTA volumetric images. The results indicated that a positive impact could be achieved on 2D-CNNs for binary classification in CCTA screening using the proposed 3D to 2D

reformatting based on multiple mosaicked projection views, transfer learning from ImageNet weights, and data augmentation by MPV techniques.

The training process of CNNs from the scratch requires a huge number of annotated datasets, and several studies have revealed that the larger the amount of the data, the better the performances of CNN models (Sun, Shrivastava, Singh, & Gupta, 2017). In natural images, researchers have been creating an abundant amount of datasets for decades, such as MSCOCO (Lin et al., 2014) and ImageNet (Russakovsky et al., 2015), which are large databases including millions of images that have been labeled using a crowd-sourcing effort. However, creating large medical image datasets as comprehensively annotated as ImageNet or MSCOCO remains a challenging task due to the time and labor cost involved in creating expertly annotated medical datasets. As a result, data augmentation and transfer learning techniques have been employed in the literature as an alternative to improve the performances of CNN models for various medical image recognition tasks (Gupta et al., 2020; Zhang et al., 2019).

Data augmentation techniques aim at artificially enlarging the training dataset from existing data using various techniques, such as, translation, random rotations, horizontal and vertical flipping, random cropping, and adding noises (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014). Transfer learning is the process of transferring pre-trained weighted models from the large labeled natural image datasets to reuse and apply to the given task of interest with smaller datasets through fine-tuning the final layer of the CNN while weights of the initial layers are held unchanged (Shin et al., 2016). The fundamental hypothesis of transfer learning is that generic features learned on huge datasets can be shared between apparently different datasets. This portability of learned generic features is a distinctive advantage of deep learning technology, which makes it beneficial in various domain tasks with very small datasets. Currently, many ImageNet pre-trained models, including AlexNet, GoogLeNet, ResNet and VGG, are freely accessible to the public and readily accessible, along with their learned parameters.

2.2. Automated Prognostic Models

Since revealing the potential of multi-modal MRI scans in designing pre-operative prognostic models for patients with a brain tumor, and the weakness of the human-centered healthcare systems, there have been several studies trying to develop automated pre-operative glioma grade and survival time prediction prognostic models using MRI scans obtained pre-surgically. These automated prognostic models can be coarsely classified as handcrafted-based, deep learning-based, and hybrid-based approaches.

The pipeline of handcrafted-based approaches in designing prognostic models for brain tumor require first to identify the region of interest (ROI) from the MRI image, either manually by expert radiologists or alternatively via segmentation algorithms. Subsequently, a large amount of high-throughput quantitative image features is extracted from the ROI. These handcrafted features include tumor volumetric parameters, tumor intensity, histogram, and wavelet textures. The extracted image features are then subjected to a feature selection technique. Finally, the traditional machine learning algorithms using the selected features, and metadata if available, are used to predict glioma grade and survival time.

Automated prognostic models based on handcrafted-based approaches have been studied by many researchers. With the endpoint being OS stratified as short or long survival time, Upadhaya et al. (2015) developed a prognostic model based on handcrafted features extracted from the segmented tumor regions of T1, T2, T1Ce, and FLAIR scans of 40 glioma patients. The authors applied SVM for selecting, ranking, and combining optimal features. K. Wang et al. (2015) investigated a univariate and multivariate survival analysis on handcrafted radiological features, clinical and genetic characteristics to predict the survival time of patients with glioblastoma. The authors demonstrated that IDH1 status combined with handcrafted radiological features, including, peritumoral edema,

multi-enhancing foci, and tumor contrast enhancements were found to predict survival time with more accuracy.

Kickingereder et al. (2016) carried out a study on 119 patients with a newly diagnosed brain tumor to investigate whether handcrafted–based features from MRI allow prediction of survival time with better accuracy compared with that of conventional clinical and radiologic models. The authors extracted 12190 features, including first-order, volume, shape, and texture features, from the multiparametric and multiregional volumetric MRI scans of the patients, out of which 11 features were found to predict survival time with improved accuracy compared with that of conventional clinical and radiologic models.

Li et al. (2017) presented a fully-automatic but handcrafted-based approach that eliminates human factors while predicting the survival time of glioma patients from multi-modal MRI images. In this work, 864 first-order and 44928 high-order image features were extracted from automatically segmented 6 tumor sub-regions. Random forest and multivariate Cox regression were used for automatic segmentation and OS prediction, respectively. The result showed that stratification of glioblastoma patients into short and high survival time was achieved with better performance compared to the conventional factors such as age, tumor volume and Karnofsky Performance Score.

Intending to develop an automated preoperative glioma grading model using multi-modal MRI images of 120 glioma patients, X. Zhang et al. (2017) investigated the influence of 8 different feature selection techniques and 25 commonly used ML methods. Histogram and image texture were extracted from the perfusion, diffusion and permeability parametric maps of MRI tumor images obtained pre-surgically. The authors found that the RFE feature selection technique when combined with SVM exhibited superior performance to the other ML methods. Similarly, H. Cho et al. (2017) extracted handcrafted features from manually segmented tumor regions in multi-modal MRI images and used LASSO for selecting optimal features for categorizing glioma into low-grade and high-

grade. In another work based on a handcrafted-based approach for glioma grading, a two-step SVM-based feature selection technique was also implemented (Tian et al., 2018).

Sanghani et al. (2018) extracted a total of 2200 clinical, volumetric, tumor shape, and textures features from manually segmented tumor regions of multimodal MRI images of brain tumor patients. The authors selected 150 optimum number of features from the 2200 handcrafted features using SVM based recursive feature elimination technique. Finally, SVM classifiers were trained for the OS prediction of patients. Similarly, Shboul et al. (2019) extracted 74 handcrafted features from automatically segmented tumor regions and subsequently combined them with the age of patients to train XGBoost ML classifier for survival time prediction.

Xu Bi et al. (2019) developed a prognostic model for pre-operative automatic classification of glioma grade by extracting 483 first-order statistical, shape and size-based, and texture-based handcrafted features in automatically segmented tumor regions of multi-modal brain tumor patients obtained pre-surgically. LASSO and SVM were used for optimal feature extraction and glioma classification, respectively. In another work for automatic grading of glioma (Reza, Samad, Shboul, Jones, & Iftekharuddin, 2019), the authors used molecular data in addition to the handcrafted features extracted from the automatically segmented 3D tumor region of a magnetic resonance image scanned pre-surgically.

In the literature, designing of brain tumor prognostic models have also been applied through a hybrid approach that comprises CNNs as feature extraction mechanism and traditional ML algorithms to fuse handcrafted features and CNN learned features and give the final prediction result (Chato & Latifi, 2017; Lao et al., 2017; L. Sun, Zhang, Chen, & Luo, 2019). Furthermore, Lao et al. reconstructed the 3D MRIs into 2D slices in the axial plane, and selected three slices that have the largest tumor core, necrosis, and whole tumor area, respectively. The three tumor subregions were also segmented manually by the two

neurosurgeons. In the end, only the selected three slices were used to fine the pre-trained CNN model, and the deep features were extracted subsequently from this finetuned CNN model to combine with handcrafted features.

Handcrafted-based approaches in designing brain tumor prognostic models introduced unavoidably a big challenge during optimizing feature extraction and selection from multi-modal MRI images. Even for a single MRI modality, it is still unobvious which features contribute most to prognostic, those from commonly used histogram parameters or image texture attributes (X. Zhang et al., 2017). Most noticeably, handcrafted features are usually bound to the current limited knowledge of a specific field which often prevents the ability to take full advantage of all the information provided by the multi-modal MRI images from being fully utilized in OS prediction (Nie, Zhang, Adeli, Liu, & Shen, 2016). To tackle such problems, some deep learning algorithms have been explored in the literature for brain tumor prognostic model designing, as they eliminate the need for handcrafted feature extraction and selection.

Nie et al. (2016) proposed a deep learning-based pipeline that predicts the survival time of brain glioma patients in three steps. In the first step, novel 3D-CNN architectures were used for extracting automatic high-level features from fMRI, DTI and T1W MRI scans of brain tumor patients. The inputs to the 3D-CNN architectures were patches from manually delineated extended tumor regions of brain images. In the second step, a feature selection procedure was conducted on the deep features using Principal Component Analysis (PCA) and Sparse Representation (SR). The authors subsequently trained SVM for distinguishing between long and short OS. Moreover, the authors showed that an approximate 30% higher specificity and sensitivity were achieved on the deep learning-based approach compared to the handcrafted-based approach.

There have also been attempts to accurately predict brain tumor patients' glioma grades from volumetric MRI images, using various CNN algorithms. On one hand, the researchers modified the state-of-the-art CNN architectures, which

have been proved to have high accuracy in natural image classification or leverage them via transfer learning. For instance, Khawaldeh et al. (2017) reconstructed a 3D FLAIR modality in the axial plane, and selected slices that contain lesions to develop a 2D CNN model, which was based on AlexNet architecture, for glioma grading.

Similarly, Y. Yang et al. (2018) investigated AlexNet and GoogLeNet CNN architectures, through training both from scratch and trained on natural images, for stratifying glioma grade into LGG and HGG by analyzing on axial T1Ce MRI images of 113 glioma patients obtained pre-surgically. During pre-processing, slices with 80% of the tumor were selected and segmented with a rectangular region of interest (ROI) by two experienced neuroradiologists. In addition, the training data were augmented using random rotation, histogram equalization, flipping, zooming, and adding noise. The result showed that transfer learning improved the result compared with the models trained from scratch, and GoogLeNet performed better than AlexNet.

On the other hand, some researchers have developed their own CNN architectures. For instance, Pereira et al. (2018) developed a 3D CNN architecture to classify glioma grades from automatically defined tumor regions. Similarly, Y. Li et al. (2018) developed an end-to-end CNN-based framework, SPNet, that takes four multi-modal MRI images of brain tumor patients, deep learning-based automatically segmented tumor region and age of the patient as input, and finally predicts the survival time of patients in days.

Banerjee et al. (2019) developed three novel CNN architectures for low/high-grade glioma classification problems. using multi-modal MRI images. The three CNN architectures are based on MRI patches containing the tumor, MRI slices, and multiplanar volumetric slices, respectively. Results demonstrate that the performance of VGGNet and ResNet pre-trained models showed the same results, and the proposed ConvNets were observed to achieve better accuracy when trained on the multi-planar volumetric dataset.

In the most recent work (Zhuge et al., 2020), the authors proposed a fully automated glioma grading on multimodal MRI images using both a novel 3D-CNN architecture (3DConvNet) and a state-of-the-art 2D-CNN model (2D mask R-CNN). The authors first segmented the volumetric tumor region using the pre-existing segmentation algorithm, U-Net. In the first approach (2D Mask R-CNN), the 2D slice with the largest segmented tumor area is chosen to train the mask R-CNN model. To improve the performance of the proposed grading model, a 2D data augmentation technique, which increases both the amount and the diversity of the training images, has been used. In the second approach (3DConvNet), the 3D cropped regions of the tumor were used to train the proposed 3D-CNN using ResNet as a backbone. The experimental result showed that the 3DConvNet technique performs slightly better than the 2D Mask R-CNN technique.

To the best of our knowledge, little work has been done on the construction and evaluation of fully automated prognostic models for brain tumor patients using CNN and multi-modal MRI images. In addition, it can be noticed from the abovementioned studies that there are at least four common issues associated with the existing CNN architectures developed for automated prognosis models for glioma patients. Primarily, most of the existing CNN architectures are 3D CNNs that do not consider the insufficient training available datasets. For example, the well-known BraTS 2017–2018 public datasets contain only 210 training samples. These numbers are far from enough to reach the full potential of 3D CNNs.

Secondly, without considering the prognostic features of slices in coronal and sagittal projections, most of the existing 2D-CNNs are based on slices in the axial projections only. Most importantly, combining the information properly from different modalities and across the three planes may enhance the diagnostic performance. However, when developing prognostic models for glioma patients using these multimodal MRI images, which modalities and 2D-projections are the most effective, and whether combinations of multiple modalities can improve the performance are still unclear and not investigated very well (Y. Liu et al., 2018).

Thirdly, most of the existing CNN networks are based on manually segmented tumor regions neglecting the tumor surrounding tissue that may contain tumor cell infiltrates that contribute to poor prognosis (Choi et al., 2019). In addition, manual tumor segmentation is a very challenging task that requires more time hindering their application for time-constrained largescale medical imaging tasks, like preoperative OS prediction.

The last but not the least common issue in the existing automated prognosis models is that most existing approaches make predictions based on a single classifier. However, single classifiers sometimes present poor results for small and imbalanced datasets. Moreover, previous studies in other medical image analysis problems have shown that ensemble learning of multiple CNN models may yield better stratification accuracy than individual CNN models (Winzeck et al., 2019; Xiao, Wu, Lin, & Zhao, 2018; B. Zhang et al., 2019). Therefore, this thesis aimed in addressing the above-discussed issues and consequently develop a fully automated brain tumor prognostic model using deep learning and classical machine learning algorithms.

3. MATERIALS AND METHODS

3.1. Dataset and Image Pre-processing

In this study, two groups of brain tumor datasets were used in which both were collected from BraTS'17 public dataset (Bakas et al., 2017; Menze et al., 2015). The first dataset consists of multi-modal 3D MRI exams of 75 LGG and 210 HGG glioma patients who underwent the examination before surgery and obtained at different hospitals with a pathologically confirmed diagnosis. Similarly, the second dataset contains pre-operative multi-modal 3D MRI exams, age and OS in days for 163 glioma patients. We used the first and second datasets for glioma grade and survival time classification problems, respectively. In this thesis, two groups of survival were considered, i.e. long-term survivors (OS > 10 months) and short-term survivors (OS <= 10 months). We split the second dataset into two groups such that 75% for training and 25% for testing. We used stratified random sampling for balancing OS distribution between the two groups. Thus, there was no significant difference observed in the demographics and clinical characteristics between the training and testing sets. Detailed patients' information can be found in Table 3.1.

The multi-modal MRI images included for each patient in both datasets were of type T1 MRI, T1Ce MRI, T2 MRI, and FLAIR MRI. Before publicly released the dataset for researchers, the images in the dataset have been skull-striped, co-registered and resampled to a uniform isotropic resolution ($1mm^3$) by the organizer. Moreover, All the glioma subregions in the datasets have been segmented manually, and their annotations were approved by experienced neuroradiologists. These annotations contain GD-enhancing tumor, peritumoral edema, non-enhancing tumor and necrotic core glioma sub-regions. An example of the modalities provided for each patient from the two datasets using an ITK-SNAP visualization toolkit is depicted in Figure 3.1.

Table 3.1 Demographic and clinical characteristics of patients in the training and testing dataset.

Parameters		**Training Dataset**	**Testing Dataset**	**Overall Dataset**
No. of Patients		121	42	163
Age (year)	Range	19 – 86	30 – 85	19 – 86
	Mean	60.181	60.774	60.334
	StdDev	12.379	11.319	12.083
OS (days)	Range	5 - 1767	30 - 1731	5 - 1767
	Mean	424.645	417.881	422.963
	StdDev	346.214	363.787	349.684
	Long (%)	72(59.5)	26(61.9)	98
	Short (%)	49(40.5)	16(38.1)	65
No of Slices	axial	75,020	26,040	101,060
	coronal	116,160	40,320	156,480
	sagittal	116,160	40,320	156,480
No of Slices with no brain tissue	Axial (%)	7,832 (73.2)	2,872 (26.8)	10,704
	Coronal (%)	34,059(73.7)	12,174 (26.3)	46,233
	Sagittal (%)	49,234 (73.5)	17,751 (26.5)	66,985
	Total (%)	91,125 (73.5)	32,797 (26.5)	123,922
No of slices with no brain tumor	Axial (%)	42,166 (73.2)	15,438 (26.8)	57,604
	Coronal (%)	75,705 (73.7)	27,015 (26.3)	102,720
	Sagittal (%)	86,363 (73.5)	31,137 (26.5)	117,500
	Total	94,567 (73.5)	34,043 (26.5)	128,610

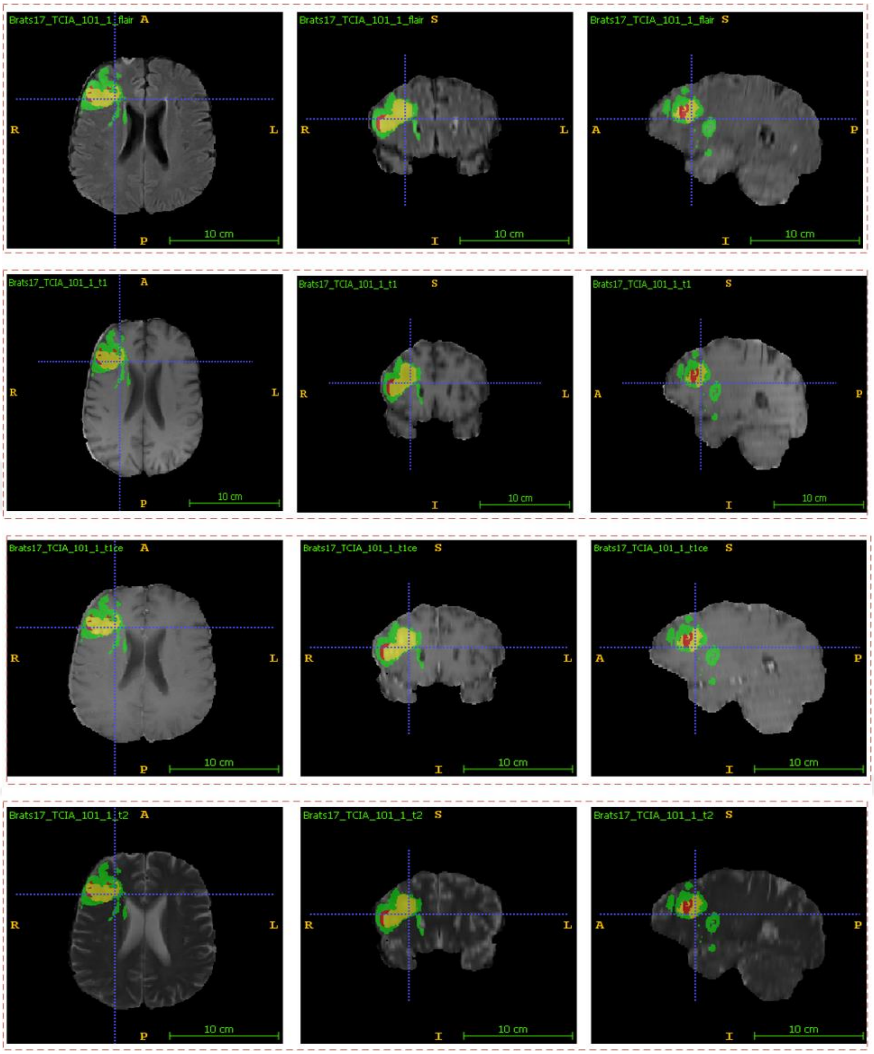

Figure 3.1. Example of MRI sequences of a patient with pathologically proven LGG glioma. From Left to right: Slices of a volumetric MRI scan after represented in the Axial (XY-Plane), Coronal (YZ-Plane) and Sagittal (XZ-Plane) projections. From top to bottom: FLAIR, T1, T1Ce, and T2 MRI modalities. Ground truth for tumor segmentation is shown with colors of Green for edema, Red for non-enhancing tumor, and Yellow for enhancing tumor glioma sub-regions. The axes are shown with blue lines.

The images in the datasets were also stored in the NIFTI medical image file format with nii.gz file extension. These file formats are represented as 3D arrays containing raw voxel intensities in Hounsfield units (HU) along with the corresponding image metadata such as slice duration, image dimensions, orientation, slice thickness, voxel size in physical units, etc. Thus, we used the Python framework with the NiBabel package to interact with these images (Matthew Brett, Michael Hanke, MARC-ALEXANDRE CÔTÉ, Paul McCarthy, Chris Cheng, 2017).

All the images in the dataset were size-adapted uniformly to 240×240×155. In addition, the input sizes of all the images should be the same and should match the input size of the CNN model architecture. Accordingly, as we planned to leverage 2D-CNN models pre-trained on natural images of a fixed image resolution, we reconstructed each 3D MRI scan to a group of 2D stacked slices in the axial, coronal and sagittal planes. These reconstructions of 3D-to-2D lead all the axial slices of every MRI modality to have a shape size of 240×240 with 155 slices per exam. Similarly, all the sagittal and coronal slices lead to have the same shape size of 155×240 with 240 slices per exam.

After converting from 3D to 2D, each modality in the first dataset have 44,175, 68,400 and 68,400 slices in the axial, coronal and sagittal planes, respectively. Similarly, each modality in the second dataset has 25,265, 39,120 and 39,120 slices in the axial, coronal and sagittal planes, respectively. Subsequently, some slices obtained at the beginning and end of each 2D projection, which did not contain any brain tissue intensity value, were also discarded to avoid processing slices that do not add useful information. Moreover, it aids to reduce GPU memory requirement and processing time.

In the first dataset, 4,826, 20,155 and 29,138 slices that did not contain any brain tissue were discarded from the axial, coronal and sagittal projections of the T1 modality, respectively. Similarly, 4,806, 20,159 and 29,153 slices from the axial, coronal and sagittal projections of the T1Ce modality, respectively, 4,853,

20,158 and 29,160 slices from the axial, coronal and sagittal projections of the T2 modality, respectively, and 4,824, 20,492 and 29,135 slices from the axial, coronal and sagittal projections of the FLAIR modality, respectively, were discarded from the 285 patients found in the first dataset. These led to the exclusion of 54,119, 54,118, 54,171 and 54,451 slices from T1, T1Ce, T2 and FLAIR MRI exams of the 285 patients found in the first dataset.

In the second dataset, 2,673, 11,556 and 16,740 slices that did not contain any brain tissue were discarded from the axial, coronal and sagittal projections of the T1 modality, respectively. Similarly, 2,672, 11,557 and 16,741 slices from the axial, coronal and sagittal projections of the T1Ce modality, respectively, 2,684, 11,556 and 16,764 slices from the axial, coronal and sagittal projections of the T2 modality, respectively, and 2,675, 11,564 and 16,740 slices from the axial, coronal and sagittal projections of the FLAIR modality, respectively, were discarded from the 163 patients of the second dataset. These led to the exclusion of 30,969, 30,970, 31,004 and 30,979 slices from T1, T1Ce, T2 and FLAIR MRI exams of 163 patients of the second dataset.

In some of our experiments, slices that do not contain any tumor tissue have also been eliminated before training our CNN architecture. For instance, 25,179, 44,746 and 48,361 slices with no brain tumor tissue were discarded in the first dataset from the axial, coronal and sagittal projections of T1 modality, respectively. These led to the exclusion of 118,286 slices with no brain tumor from T1 MRI exam of the 285 patients found in the first dataset. Similarly, 14,401, 25,680 and 29,375 slices with no brain tumor tissue were discarded in the second dataset from the axial, coronal and sagittal projections of T1 modality, respectively. These led to the exclusion of 69,456 slices with no brain tumor from T1 MRI exam of the 163 patients found in the second dataset. The number of used and discarded axial, coronal and sagittal slices in all of the four MRI modalities for the first and second datasets are shown in Table 3.2 and 3.3, respectively.

Table 3.2. The number of reconstructed and discarded slices from each modality and projections in the first dataset.

Modalities	Slices	Axial	Coronal	Sagittal	Total
T1	All	44,175	68,400	68,400	180,975
	No brain tissue	4,826	20,155	29,138	54,119
	No brain tumor	25,179	44,746	48,361	118,286
T1Ce	All	44,175	68,400	68,400	180,975
	No brain tissue	4,806	20,159	29,153	54,118
	No brain tumor	25,179	44,746	48,361	118,286
T2	All	44,175	68,400	68,400	180,975
	No brain tissue	4,853	20,158	29,160	54,171
	No brain tumor	25,179	44,746	48,361	118,286
FLAIR	All	44,175	68,400	68,400	180,975
	No brain tissue	4,824	20,492	29,135	54,451
	No brain tumor	25,179	44,746	48,361	118,286
Total	All	176,700	273,600	273,600	723,900
	No brain tissue	19,309	80,964	116,586	216,859
	No brain tumor	100,716	178,984	193,444	473,144

Table 3.3. The number of reconstructed and discarded slices from each modality and projections in the second dataset

Modalities	Slices	Axial	Coronal	Sagittal	Total
T1	All	25,265	39,120	39,120	103,505
	No brain tissue	2,673	11,556	16,740	30,969
	No brain tumor	14,401	25,680	29,375	69,456
T1Ce	All	25,265	39,120	39,120	103,505
	No brain tissue	2,672	11,557	16,741	30,970
	No brain tumor	14,401	25,680	29,375	69,456
T2	All	25,265	39,120	39,120	103,505
	No brain tissue	2,684	11,556	16,764	31,004
	No brain tumor	14,401	25,680	29,375	69,456
FLAIR	All	25,265	39,120	39,120	103,505
	No brain tissue	2,675	11,564	16,740	30,979
	No brain tumor	14,401	25,680	29,375	69,456
Total	All	101,060	156,480	156,480	414,020
	No brain tissue	10,704	46,233	66,985	123,922
	No brain tumor	57,604	102,720	117,500	277,824

Another operation that we conducted during the pre-processing stage was cropping each slice to a fixed size of 224×224 pixels for axial, and 112×224 pixels for coronal and sagittal views, respectively. To avoid the effect of image enlarging on the classification performance, two consecutive coronal slices were concatenated and reshaped to 224×224 pixels. Similarly, two consecutive sagittal slices were concatenated and reshaped to 224×224 pixel sizes. The example of reconstructed, cropped and concatenated 2D images of a sample patient MRI exam is depicted in Figure 3.2. A similar approach was used and reported promising result in our published works where each consecutive coronal and sagittal slices at the same position were concatenated before fed to the CNN model in deep learning-based automated Tuberculosis diagnosis using chest CT scans (Mossa, Eriş, & Çevik, 2020; Mossa, Yibre, & Çevik, 2019).

In the experiments for the survival time prediction, since the size of the dataset is very small and a large portion of the MRI slices regions are non-tumor tissues, the CNN models may focus on unnecessary and redundant information, which may lead to degraded performance. Thus, in order for the proposed CNN models to focus only on learning notable features related to the tumor region and its surrounding tissues, rather than feeding the whole slide image while training the models, it may be essential to feed them only the rectangular box containing the tumor regions and it‘s surrounding tissues. Thus, in this study, three different types of slice inputs with different sizes of extended tumor regions (ROI-15, ROI-30 and ROI-45) were extracted in the pre-processing stage. ROI-15 represents the rectangular bounding box containing tumor regions and surrounding tissues up to 15 pixels from the tumor border. Similarly, ROI-30 and ROI-45 represent the rectangular bounding box containing tumor regions and surrounding tissues up to 30 and 45 pixels from the tumor border, respectively.

At the end of pre-processing, we rescaled the intensity values of all the slices in the (0,255) range, convert them to PNG format, and normalized them to have zero mean and unit variance. This is a common technique in deep image

processing since standardization makes it much easier for the network to learn. In addition, it facilitates model training by standardizing the input distribution. Then, all the reconstructed PNG images of an MRI in the axial, coronal and sagittal views were stacked together and saved it in serialized form with a pickle toolbox, respectively. Therefore, our input shape turned to be (n, 3,224,224). The values can be interpreted such that the first value holds for the number of stacked images in the axial, coronal or sagittal views, respectively. The second value for the number of color channels, third and fourth values for width and height of PNG images. The first value varies across different views and modalities. The axial slices extraction procedure is detailed in Algorithm 1. The coronal and sagittal slices extraction and concatenation algorithms are detailed in Algorithm2. Similarly, Algorithm 3 detailed the extraction of the extended ROI with a rectangular bounding box that includes the surrounding tumor tissues of the glioma region. These algorithms were implemented using the Python programming language and the NiBabel package.

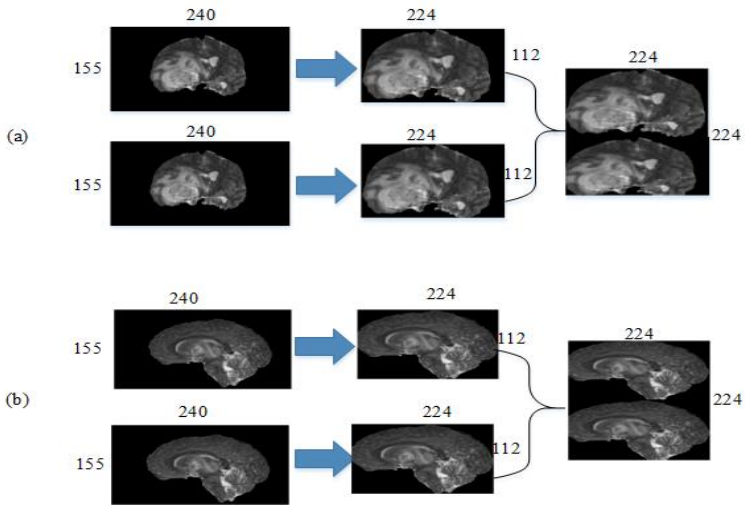

Figure 3.2. Concatenation. (a) two consecutive coronal slices having the same shape size (155 × 240) were sliced from the 3D MRI scans, center cropped and then concatenated, (b) two consecutive sagittal slices having the same shape size (155 × 240) were sliced from 3D MRI scans, center cropped and then concatenated.

Algorithm 1: Algorithm for Axial Slices Extraction from the volumetric MRI Images.

```
Input: volumetric MRI image
Dimension = [240,240,155]
 [row intensity, column intensity, number of slices] = dimensions of MRI image in the axial plane
 [row intensity, number of slices, column intensity] = dimensions of MRI image in the coronal plane
 [ number of slices, row intensity, column intensity] = dimensions of MRI image in the sagittal plane
 MAX_PIXEL_VAL = 255

 axialslices = []
 for k ← 1 to number of slices do
        axial_slice = Input[:,:,k]
        if IsNot_empty (slice) then
            axial_slice = resize(axial_slice,[224,224])
            axial_slice = Standardize_Normalize(axial_slice)
            axialslices.append(axial_slice)
        k = k + 1
axialslices = numpy.stack((axialslices,)*3, axis=1)
axialslices = Standardize_Normalize(axialslices)

IsNot_empty(slice)
 {
        index_xy = numpy.where(slice != 0)
        if index_xy[0].shape[0] != 0 then
                return true
 }
Standardize_Normalize(slice)
{
 slice = (slice - numpy.min(slice)) / (np.max(slice) - np.min(slice)) * MAX_PIXEL_VAL
 slice = (slice - numpy.mean(slice)) / numpyp.std(slice)
}
```

Algorithm 2: Algorithm for Coronal and Sagittal Slices Extraction and Concatenation.

```
Input: volumetric MRI image
Dimension = [240,240,155]
 [row intensity, column intensity, number of slices] = dimensions of MRI image in the axial plane
 [row intensity, number of slices, column intensity] = dimensions of MRI image in the coronal plane
 [ number of slices, row intensity, column intensity] = dimensions of MRI image in the sagittal plane
 MAX_PIXEL_VAL = 255

 coronalslices = []
 for k⟵ 1 to number of slices do
        coronal_slice1 = Input[:,k,:]
        coronal_slice2 = Input[:,k+1,:]
        if IsNot_empty (coronal_slice) then
            coronal_slice1 = resize(coronal_slice1,[112,224])
            coronal_slice1 = Standardize_Normalize(coronal_slice1)
            coronal_slice2 = resize(coronal_slice2,[112,224])
            coronal_slice2 = Standardize_Normalize(coronal_slice1)
            coronal_slice = Concatenate(coronal_slice, coronal_slice2)
            coronalslices.append(coronal_slice)
        k = k + 2
coronalslices = numpy.stack((coronalslices,)*3, axis=1)
coronalslices = Standardize_Normalize(coronalslices)

sagittalslices = []
 for k⟵ 1 to number of slices do
        sagittal_slices1 = Input[k,:,:]
        sagittal_slices2 = Input[k+1,:,:]
        if IsNot_empty (slice) then
            sagittal_slice1 = resize(coronal_slice1,[112,224])
            sagittal_slice1 = Standardize_Normalize(sagittal_slice1)
            sagittal_slice2 = resize(sagittal_slice2,[112,224])
            sagittal_slice2 = Standardize_Normalize(sagittal_slice1)
            sagittal slice = Concatenate(coronal_slice, sagittal_slice2)
            sagittalslices.append(sagittal_slice)
        k = k + 2
sagittalslices = numpy.stack((sagittalslices,)*3, axis=1)
sagittalslices = Standardize_Normalize(sagittalslices)

IsNot_empty(slice)
 {
        index_xy = numpy.where(slice != 0)
        if index_xy[0].shape[0] != 0 then
                return true
 }
Standardize_Normalize(slice)
{
 slice = (slice - numpy.min(slice)) / (numpy.max(slice) - numpy.min(slice)) *  MAX_PIXEL_VAL
 slice = (slice - numpy.mean(slice)) / numpy.std(slice)
}
```

Algorithm 3: Algorithm for ROI and Extended-ROI Slices Extraction from MRI Images.

```
Input: volumetric MRI image
Dimension = [240,240,155]
 [row intensity, column intensity, number of slices] = dimensions of MRI image in the axial plane
 MAX_PIXEL_VAL = 255
 mask = NiftyNet_Segment(Input)
 axialslices = []
 for k⟵ 1 to number of slices do
        mask_axial = mask[:,:,k]
        axial_slice = Input[:,:,k]
        d = {15,30,45}
        if IsNot_empty (mask_axial) then
            mask_axial_xy = np.where(mask_axial==1)
            imag_axial_xy = np.where(axial_slice!=0)
            Eroi = EROI(image_axial_xy,mask_axial_xy,d)
            axial_slice = resize(Eroi,[224,224])
            axial_slice = Standardize_Normalize(axial_slice)
            axialslices.append(axial_slice)
        k = k + 1
axialslices = numpy.stack((axialslices,)*3, axis=1)
axialslices = Standardize_Normalize(axialslices)

IsNot_empty(slice)
{
        index_xy = numpy.where(slice != 0)
        if index_xy[0].shape[0] != 0 then
                return true
}

Standardize_Normalize(slice)
{
 slice = (slice - numpy.min(slice)) / (np.max(slice) - np.min(slice)) *  MAX_PIXEL_VAL
 slice = (slice - numpy.mean(slice)) / numpyp.std(slice)
}

EROI()
 {
 if(max(image_axial_xy[0]) - max(mask_axial_xy[0]) >= d) then
                max_EROI_xy_0 = max(msk_axial_xy[0]) + d
            else
                max_EROI_xy_0 = max(img_axial_xy[0])

            if(max(img_axial_xy[1]) - max(msk_axial_xy[1]) >= d) then
                max_EROI_xy_1 = max(msk_axial_xy[1]) +d
            else
                max_EROI_xy_1 = max(img_axial_xy[1])

            if(min(msk_axial_xy[0]) - min(img_axial_xy[0]) >= d) then
                min_EROI_xy_0 = min(msk_axial_xy[0]) – d
             else
                min_EROI_xy_0 = min(img_axial_xy[0])

            if(min(msk_axial_xy[1] - min(img_axial_xy[1])) >= d) then
                min_EROI_xy_1 = min(msk_axial_xy[1]) – d
            else
                min_EROI_xy_1 = min(img_axial_xy[1])

            center_x = (max_EROI_xy_0 + min_EROI_xy_0) / 2
            center_y = (max_EROI_xy_1 + min_EROI_xy_1) / 2

             length_EROI = max_EROI_xy_0 - min_EROI_xy_0
             width_EROI = max_EROI_xy_1 - min_EROI_xy_1

            EROI = np.zeros(length,widthl)
            EROI[:,:] = image_axial[int(center_x - length/2) : int(center_x + length/2), int(center_y - width/2) : int(center_y + width_EROI/2)]

           return EROI
 }
```

3.2. Deep Learning

3.2.1. Artificial Neural Network (ANN)

An artificial Neural Network, usually simply called a neural network, is a type of learning algorithm inspired by the early models of sensory processing in the brain. It has emerged as a powerful technique for modeling the general relationship between input and output by applying algorithms that mimic the processes of biological neurons. Most importantly, it is now the basis for most deep learning algorithms.

Essentially, ANNs are comprised of several interconnected computational units, named neurons, arranged in layers. Neurons are the basic operating unit in a neural network. Each neuron takes several inputs from other neurons or external sources and produces a single output that may be passed to other neurons through a connection link. Each connection link is associated with a weight that determines the strength of the interconnection and thus the contribution of that interconnection to the following neurons. Moreover, each neuron has an internal state, which is called an activation function.

Using Equation 3.1, the output of a neuron, z, is computed through combining the input values x = {x_1, x_2, …. x_n} and learnable parameters Θ = {W, b}, where W is a set of weights {w_1, w_2, …. w_n} and b represent bias, and subsequently, pass them through an activation function, f(.). A single neuron model is referred to as perceptron and its function is illustrated in Fig 3.3. The output, y, of the perceptron, can be computed using Equation 3.2.

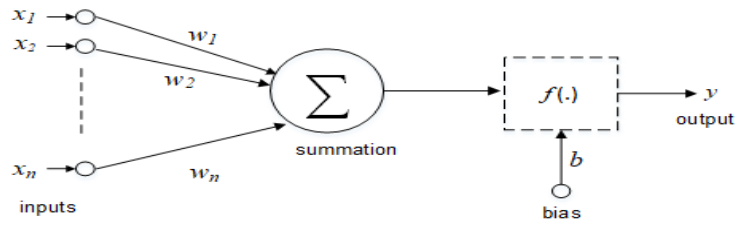

Figure 3.3. Graphical representation of a neuron model or perceptron

$$z = \sum_{i=1}^{n} w_i x_i + b \tag{3.1}$$

$$y = f(z) \tag{3.2}$$

Typically, activation functions are non-linear functions that allow the neural network to combine the inputs in a more complex way and in turn provide a richer capability in the functions they can model. The most commonly used activation functions for the traditional neural networks include the binary threshold, bipolar threshold, rectified linear unit, and sigmoid functions. The output y of a neuron with the binary threshold activation function is 1 if $z \geq 0$ in Equation 1, i.e., if $(\sum_{i=1}^{n} w_i x_i + b) \geq 0$. The output y of a neuron with the bipolar threshold activation function is similar to that of the binary threshold, but 0 is represented as -1 when $z < 0$. The rectified linear unit (ReLU) computes a linear weighted sum of their inputs, and the output is a non-linear function of the total input computed using the max operation, f(x) = max (0, x). The sigmoid function transforms a real-valued output to the value in the range between 0 and 1, which is a smooth and bounded function of their total input. Typically, sigmoid activation functions use the logistic function. The binary threshold, bipolar threshold, ReLU, and sigmoid activation functions are represented in Equation 3.3, Equation 3.4, Equation 3.5, and Equation 3.6, respectively.

$$f(\mathrm{x};\ \Theta) = \begin{cases} 0, & z < 0 \\ 1, & z \geq 0 \end{cases} \tag{3.3}$$

$$f(\mathrm{x};\ \Theta) = \begin{cases} -1, & z < 0 \\ 1, & z \geq 0 \end{cases} \tag{3.4}$$

$$f(\mathrm{x};\ \Theta) = \begin{cases} 0, & z \leq 0 \\ z, & z > 0 \end{cases} \tag{3.5}$$

$$f(\mathrm{x};\ \Theta) = \frac{1}{1+e^{-z}} \tag{3.6}$$

A multilayer neural network is essentially built using three layers: input, hidden, and output layers. The input layer is used to interface with the real world for receiving the input and passes them to the next layer, often referred to as a hidden layer. Last comes the output layer that produces the neural network's predictions. If the network comprises more than one hidden layer, it is typically considered as a deep neural network (DNN), hence the term ‘deep learning’.

As depicted in Figure 3.4, the multi-layered perceptron (MLP) is one of the most well-known of the traditional Feed-Forward DNN characterized by several layers of nodes connected as a directed graph between the input and output layers, which means that the signals that pass through the nodes go only one way. Each layer of an MLP has a weight (W) vector and a bias vector (b) to be assigned or learned. Suppose that the input pattern $x = (x_1, \ldots, x_n)$ is to induce the desired output y at the final layer (L), and Wn is a matrix comprising of rows w^k, associated with activation function k in the output, where the symbol n indicates the number of the current layer, then, the output of an MLP network at the final layer is computed using Equation 3.7, which is obtained through applying Equation 2 for each neuron of the network.

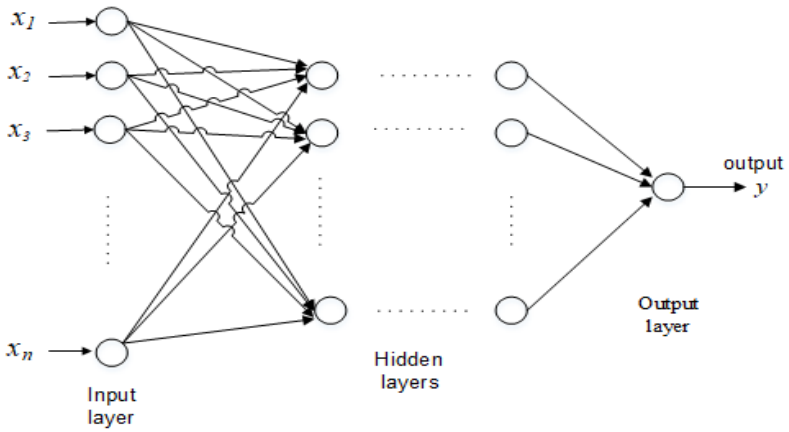

Figure 3.4. Examples of Multi Layer Neural Network Architecture

$$y = f(W^L f(W^{L-1} \ldots f(W^0 x + b^0) + \ldots + b^{L-1}) + b^L) \quad (3.7)$$

For the binary classification problem, the final layer of a neural network may have a single output neuron with a sigmoid activation function to predict a value between 0 and 1 that represents the probability of predicting a value for class 1. This can be turned into a crisp class value by using a threshold. If the total input is above a threshold, the output of the neuron is interpreted as class 1; otherwise, it is class 0.

Neural Networks can be trained to perform a particular task by adjusting the values of weights and biases parameters in such a way that the performance of the networks is improved through reducing the network loss (cost function), which measures the compatibility between the actual prediction of the network and the desired output as given on a training dataset. The choice of the proper cost function is an important aspect of a neural network algorithm. The binary cross-entropy loss for two-class problems and the categorical cross-entropy loss for the multi-class problems that are associated with the Softmax functions are the most preferred techniques in the deep learning field for computing the cost function. The cross-entropy equation for n classes using the Softmax function to compute the data loss

for a single training example is shown in Equation 3.8, where t_i is the truth label and p_i the Softmax probability for the i^{th}class scores of that example.

$$\boldsymbol{L = -\sum_{i=1}^{n} t_i \log(p_i)} \tag{3.8}$$

During training the neural network, the performance of the network under particular weight and bias parameters is evaluated by the cost function during the forward propagation stage. Subsequently, the gradient descent algorithm, the classical and still the most preferred parameter optimization technique is used to iteratively update the parameters until the cost function is lower than a predetermined threshold value or the loss no longer changes after a certain number of iterations by using the gradient of the loss, which is computed using the back-propagation algorithm (Rumelhart, Hinton, & Williams, 1986), often simply called backprop. Modern optimization techniques are improvements and refinements on the idea of this gradient descent algorithm. The gradient is, mathematically, a partial derivative of the loss with respect to each learnable parameter, and a single update of a weight and bias parameters are formulated as using Equation 3.9 and Equation 3.10, respectively:

$$w \coloneqq w - \alpha * \frac{\partial L}{\partial w} \tag{3.9}$$

$$b \coloneqq b - \alpha * \frac{\partial L}{\partial b} \tag{3.10}$$

where *w* and *b* stand for the biases and weights learnable parameter, L stands for the cost function and α stands for the learning rate, a positive scalar determining the step size. There are also some parameters, such as momentum and weight decay that need tuning while training the neural network. In the end, when the network has learned to solve problems based on the training datasets, the new

dataset from the same knowledge domain can then be put into the trained neural network to estimate the generalization performance of the network.

3.2.2. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks also called Convolutional Networks or ConvNets are a class of ANN techniques that have shown exemplary performances in various computer vision tasks. CNNs make the explicit assumption that the inputs are images, and they are designed to automatically and adaptively learn hierarchies of features from input images by building high-level attributes from low-level ones. CNNs were built as early as 1989 ultimately for the purpose of automatic 2D image classification tasks (Y. LeCun et al., 1989) and were trained to recognize images of numbers from handwritten ZIP-codes collected by the US post office (LeCun and Cortes, 2005).

Several pioneering ideas have been explored in the literature to bring advancements in CNNs, such as the use of different architectural innovations, various optimization algorithms, activation function, loss function and regularization techniques. However, as depicted in Figure 3.5, the majority of any typical CNN architecture is based on a sequence of layers that transform one volume of activation to another through a differentiable function. These typical CNN architectures consist of repetitions of a stack of several convolutions and pooling layers, followed by a classification module (i.e. MLP) at the end. The first two convolutions and pooling layers perform feature extraction, whereas the third, an MLP, maps the extracted features by the convolution and pooling layers into the final output, such as classification.

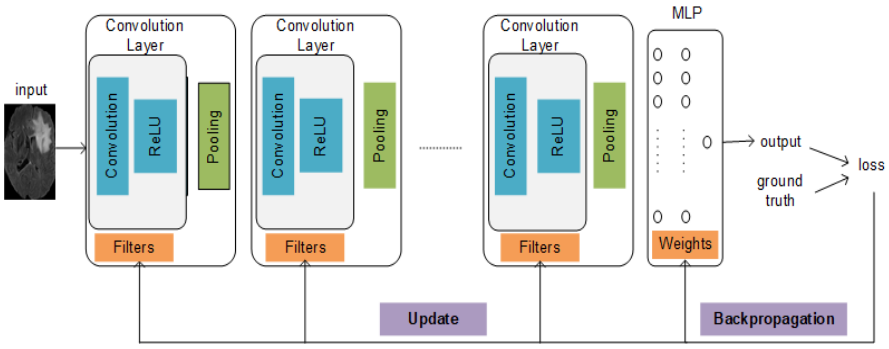

Figure 3.5. Building blocks of a typical CNN architecture

The convolution layer is the most essential component of any CNN architecture that performs feature extraction using a stack of convolution operations and activation functions by dividing the input into small parts. This stack of convolutions and activation functions will make the extracted features become more abstract with the increase of depth. Moreover, since digital images are stored in a 2D grid of pixels, the input and output of each convolution layer for the case of image recognition are sets of a multi-dimensional array of data called feature maps. For example, if the input is a colour image, each feature map would be a 2D array containing a colour channel of the input image. In addition, the output of the first convolution layer can be used as an input for the second convolution layer, which can be used as an input for the third convolution layer, and so on. Each output feature map of any convolutional layer represents a particular feature extracted at many spatial locations on the input. As a result, to obtain feature maps that represent many kinds of features, a small grid of multiple trainable parameters called filters (kernels), which are adapted by the learning algorithm, are applied at every location on the input. Thus, different filters can be considered as different feature extractors, and they are shared among all units of the same feature maps. As a result, convolutional layers have fewer parameters to train than MLP layers, making CNN easier to train and less prone to overfitting.

The step where input data are transformed into output through the network is called forward propagation. During this forward propagation, an element-wise product between each element of the filter and the input feature map is calculated at each location of the input and summed to obtain the output value in the corresponding position of the output feature map. This special linear type of mathematical operation is called convolution, which is typically denoted with an asterisk. In the convolution layer, the output feature maps obtained after the convolution operator are also passed through a non-linear activation function, responsible for non-linearly transforming the data. The most commonly used nonlinear activation function implemented in the convolution layers is the very simple ReLU activation function, which is defined in Equation 5.

Two key hyperparameters that define the convolution operation are the size and number of filters. The most common filter size in the literature is 3×3, but 5×5 or 7×7 are also rarely used in the literature. The number of features determines the depth of output feature maps. For instance, suppose an f1 × f2 size filter k connects a 2D input feature map x of size n1 × n2 to output the feature map, y, then the convolution operation to get the desired feature map can be expressed using Equation 3.11. With stride s and amount of padding p used on the border, the height and width dimensions of the output feature map, y, can also be expressed using Equation 3.12 and Equation 3.13, respectively.

$$y(i,j) = \sum_m \sum_n x(i+m, j+n)k(m,n) \tag{3.11}$$

$$\frac{n1+2p-f1}{2s} + 1 \tag{3.12}$$

$$\frac{n2+2p-f2}{2s} + 1 \tag{3.13}$$

The pooling layer provides a downsampling operation that reduces the dimensionality of the feature maps in order to speed up the computation and make the features detected by the convolution layers more robust. The pooling operator operates using the average function (average-function) or the max function (max-pooling) using values in each feature map within a rectangular neighbourhood with a stride larger than 1, and subsequently, makes the representation invariant to small translations and distortions of the input. Although several subsampling operations are in use, the widely used ones are 2×2 max or average-pooling with a stride of 2, which reduces the dimension of the feature maps by a factor of 2. The 2×2 max pooling operation computes the maximum values of the four neighbourhood values in each feature map and discards all the other values. Similarly, the average pooling computes the mean of the 4 neighbourhood values in each feature map.

The outputs of the final pooling layers of any ConvNet architecture are typically flattened into a one-dimensional (1D) array of numbers, and connected to a traditional ANN, such as MLP, in which every input is connected to every output by a learnable weight. Just like the regular neural network, the final fully connected layer of an MLP in a CNN has output nodes followed by a nonlinear activation function that needs to be chosen according to the task, as discussed in section 3.2. Similarly, supervised training of the network is performed using a backpropagation approach and an optimization technique to reduce the discrepancy between the ground truth labels and the actual outputs of the ConvNet as for the standard ANN. All the learnable parameters such as filter weights and biases in all the building blocks of the ConvNet are updated concurrently by the learning procedure. On the other hand, the type of loss function, filter size, number of filters, padding, learning rate, and stride size are some of the hyperparameters that need to be chosen before the training process starts.

3.2.3. State-of-the-art CNN Models

In medical image analysis tasks, the standard of finding the optimal CNN architecture is through trying out multiple architectures and see which one achieved an enhanced performance for a particular task. As a result, some of the famous state-of-the-art CNN architectures such as AlexNet, GoogLeNet, VGG, ResNet, UNet, etc have been customized and applied to various image analysis problems. In this thesis, fifteen most common state-of-the-art ImageNet CNN architectures were investigated in finding an optimal feature extraction layer for the proposed CNN architectures and subsequently, leverage weights learned from natural images. A UNet based state-of-the-art CNN model developed by Wang et al. (2018) was also utilized in this thesis for the automatic segmentation of the ROI tumour region. All of these state-of-the-art CNN architectures vary in their architectural design as discussed in the following paragraphs.

AlexNet is considered as the first deep CNN architecture that launched the current deep learning thrive by winning the 2012 ILSVRC competition by a huge margin for image classification and recognition tasks. The basic architectural design of AlexNet is shown in Figure 3.6. It consists of five convolution layers, three pooling layers and three fully connected layers. Input RGB images with a size of 227×227 go through the convolutional and pooling layers, and subsequently to the fully connected layers.

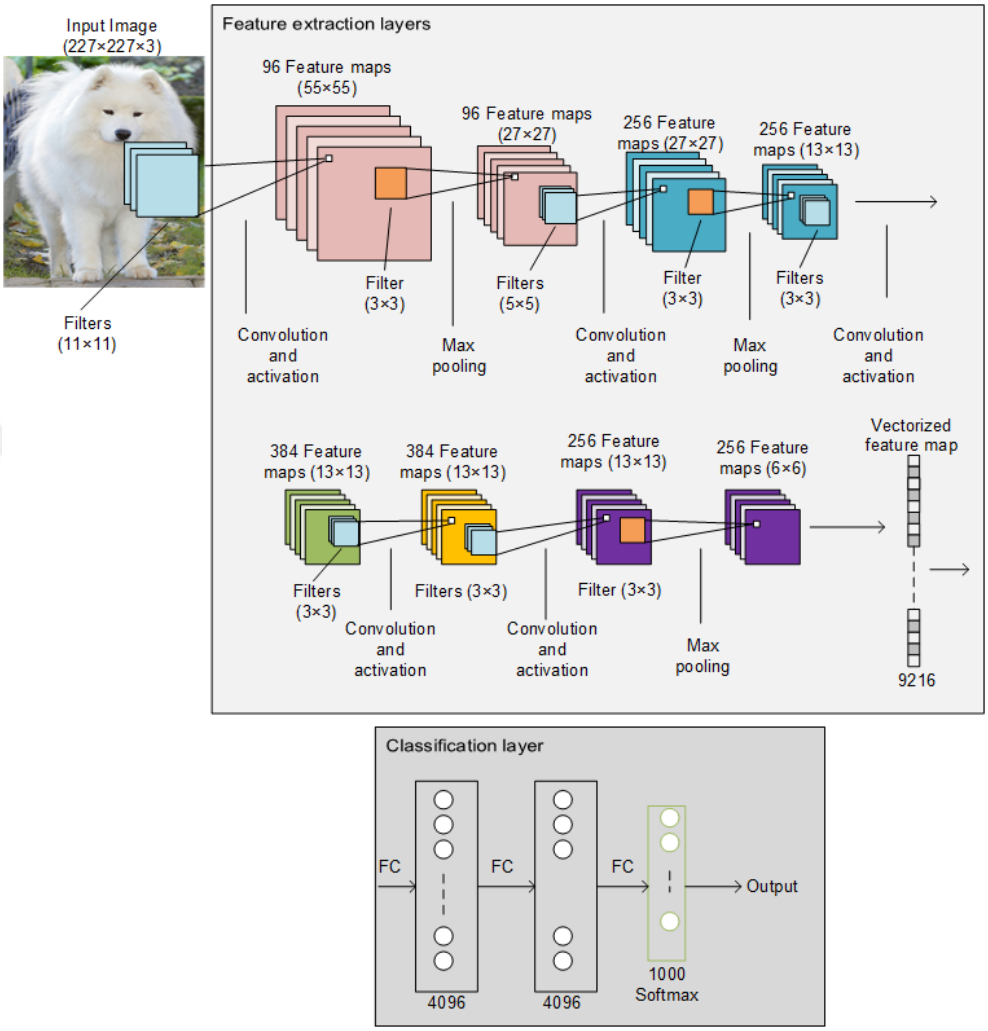

Figure 3.6. Building blocks of the AlexNet architecture

The input images in the AlexNet architecture are convolved with 96 filters of size 11×11, with 256 filters of size 5×5, with 384 filters of size 3×3, with 384 filters of size 3×3, with 256 filters of size 3×3, in the first, second, third, fourth and fifth convolution layers, respectively. In the architecture, the immediate feature maps obtained after the first convolutional layer are used as input to the second convolutional layer. Subsequently, the max-pooling operation with 3×3 filter size is performed on the output of each convolutional layer where the output feature

maps are used as input to the immediate consecutive convolutional layer in which similar max-pooling is applied to reduce the learnable parameters. At the end of the convolution layer, the max-pooling with a 3×3 filter size and a stride of 2 is performed on the output. In addition to the convolutional and pooling layers, three fully connected layers are also constructed at the end of the last pooling layer in the AlexNet architecture. Moreover, notable features including the use of the ReLU as a non-saturating activation function to improve the convergence rate by alleviating the problem of vanishing gradient to some extent, training on multiple GPUs to overcome the shortcoming of the hardware were used in the architecture.

With the goal of achieving high accuracy with a very deep CNN architecture while maintaining minimal computational time, GoogLeNet architecture, also called Inception-v1, was built by a Google research team using a new concept called inception. The architecture contains 22 layers based on conventional convolutional and pooling layer and stacked inception modules with different filter sizes of 5×5, 3×3, and 1×1 to parallelly filter the layers and subsequently concatenate their results and fed to the subsequent inception module (Szegedy et al. 2015). It was much deeper than AlexNet and was the winner of the ILSVRC14 competition. GoogLeNet also brought the popularized idea of not using fully-connected layers at the end, but rather global average pooling, significantly reduce the number of parameters from 138 million to 4 million parameters which improve the computational efficiency. Other regulatory factors applied were batch normalization and the use of RmsProp as an optimizer (Dauphin, De Vries, & Bengio, 2015).

VGG was proposed with a deeper CNN architecture through improving AlexNet architecture by replacing the 11×11 and 5×5 large-sized filters with multiple 3×3 small-sized filters stacked one after another. The use of the small-size filters provides an additional benefit of low computational complexity by reducing the number of parameters. For the tuning of the network, max-pooling is placed after the convolutional layer, while padding was performed to maintain the spatial

resolution (Huang et al. 2007). VGG showed good results both for image classification and localization problems and was in 2nd place in the 2014-ILSVRC competition. VGG11, VGG13, VGG16 and VGG19 are variants of the VGG architecture with 11,13,16 and 19 layers, respectively.

The main limitation associated with VGG was the use of very deep architecture with 138 million parameters, which make it computationally expensive and difficult to train because of the vanishing and exploding gradient type of problems. Moreover, while training deep CNN architectures with a large number of layers, the layers do not increase performance efficiency beyond a certain point. In addition, in such a scenario depth can also increase the model error. As a result, skip connections that allow taking the activation from one layer and suddenly feed to another layer even to the much deeper in the network, created an efficient methodology for the training such kind of deep CNN architectures. Proposed by He et al. (2015), ResNet is built using the concept of skip connections that allows the network to copy the activation from one ResNet block to another ResNet block, preserving information as the data goes through the layers. As a result, it provides a solution to the problem of the vanishing gradient and the learning speed of the model is increased and still achieves compelling performance. ResNet18, ResNet34 and ResNet50 are some of the variants of the ResNet architecture with 18, 34 and 50 layers, respectively.

DenseNet is a network architecture in which each layer is connected to the other in a feed-forward manner. These connections allow the model to reuse features and avoid the problem of vanishing gradient. DenseNet121, DenseNet169 and DensNet201 are some of the variants of the DensNet architecture with 121, 169 and 201 layers, respectively. In this thesis, MansNet and MobileNet architectures, which belong to the family of the second generation computer vision networks and designed to perform classification tasks at a very low computational cost, were also investigated.

The objective of image segmentation problems is to label every pixel in the image with the class of its enclosing object or region. The ML-based automatic segmentation of brain tumours from MRI images is extremely challenging because the method should have a strong classification and localization properties at the same time. Recently, considerable advancement in the area of automatic medical image segmentation was achieved after state-of-the-art methods based on convolutional neural networks were developed. Among this, the UNet architecture, which was first proposed and implemented by Ronneberger et al. (2015) for the segmentation of neuronal structures in 512 × 512 biomedical microscopy images and won the cell tracking Challenge proposed during the 2015 ISBI, has gained wide popularity within the research community implementing for many biomedical image segmentation tasks with satisfactory segmentation results, including for brain tumor segmentation challenges (Dong, Yang, Liu, Mo, & Guo, 2017; G. Wang, Li, Ourselin, & Vercauteren, 2018).

As depicted in Figure 3.7, the UNet CNN architecture, which has a symmetrical letter "U" shape, is comprised of three core parts: (1) encoding/contracting/downsampling; (2) bottleneck; and (3) decoding/expanding/upsampling. The encoding part is composed of multiple blocks comprising of convolutional and pooling layers, where each block is formed by a number of convolutional layers (each followed by a ReLU activation function) downsampled by the max pooling layer of stride 2. Through this path, the network learns the important features (context) but not their locations. The decoding part has a reversed effect as it works to progressively propagate context information to a higher resolution until it reaches its original size. For this, it is composed of a number of transpose convolution (deconvolution) layers with a stride of 2. The bottleneck appears between the contracting and expanding paths, and it is built with two convolutional layers and a dropout layer. In addition, based on the ideas of ResNet, there are skip connections that concatenates features from the

downsampling to the upsampling paths to ensure the precise localization of context at each stage of the expanding path.

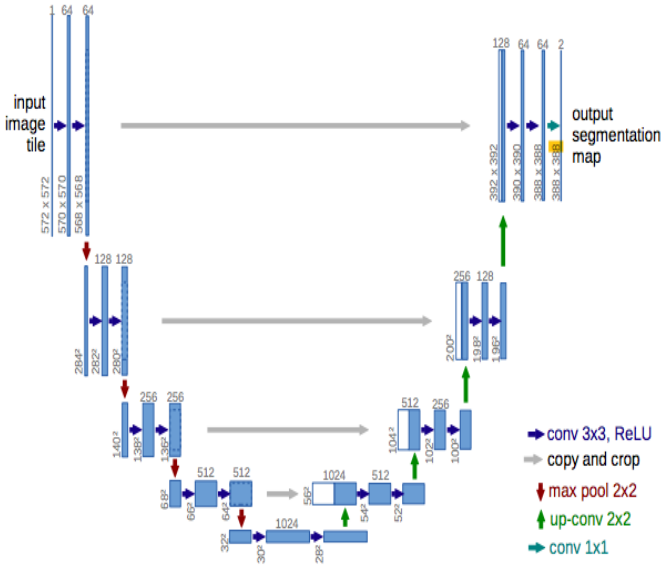

Figure 3.7. The Original U-Net architecture (Ronneberger et al., 2015)

3.2.4. Multi-View CNN (Mv-CNN)

This thesis aimed to tackle a binary classification problem on volumetric medical images through leveraging the knowledge learned from state-of-the-art 2D-CNN architectures pre-trained over large natural image datasets (ImageNet). Thus, the Mv-CNN architecture that we proposed is similar to the original Multi-View CNN architecture developed for 3D shape recognition but unlike the original Multi-View CNN, our proposed Mv-CNN used the feature extraction layers of the pre-trained ImageNet architecture as a backend. The basic concept of the proposed Mv-CNN architecture is that during the training process we can provide the model a series of 2D slices sampled from the volumetric MRI scan of a patient, which can be processed using 2D-CNNs for the classification.

The proposed Mv-CNN architectures consist of three core modules: feature extraction layers of pre-trained ImageNet architecture, pooling layers (global average and max pooling) and Fully connected layers. Figure 3.8 depicts an example for the AlexNet based Mv-CNN architecture. Similar architectures were developed using the other 14 state-of-the-art ImageNet architectures which were discussed in Section 3.2.3. The first two modules of the Mv-CNN architectures operate as feature extraction layers and the third as a classification layer. Details of each module are briefly discussed in the following paragraphs.

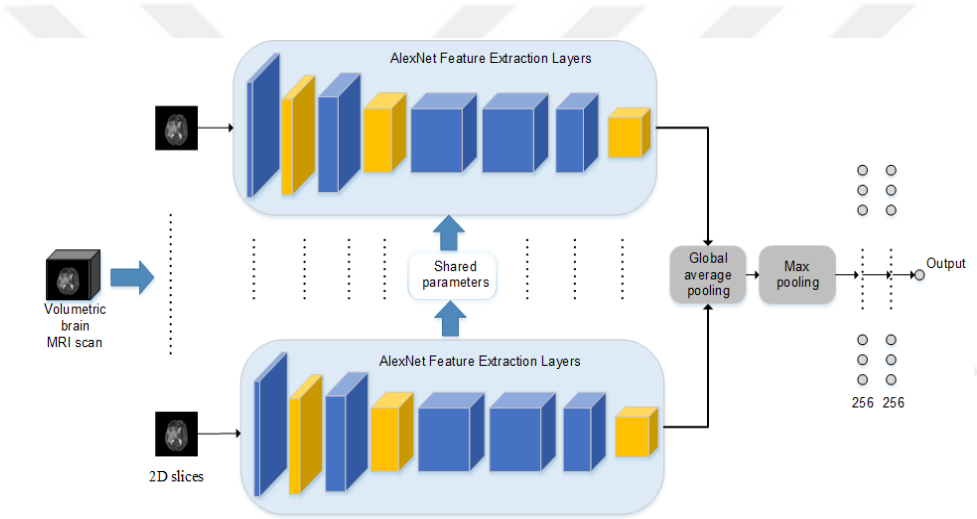

Figure 3.8. The architecture of the proposed Mv-CNN architecture for volumetric image classification using AlexNet as a backend.

The feature extraction layers of the pre-trained AlexNet, which contains five convolution and three pooling layers, are used to exploit the hierarchical features uniformly from each slice. Given a volumetric MRI scan represented by k slices in one of the three planes obtained after the pre-processing stage (discussed in Section 3.1), the size of the training data inputs to the network will be fixed as $k \times 3 \times 224 \times 224$ stacked slices, where 3 indicates the number colour channels. For each slice, the output feature map after applying the AlexNet feature extraction

layers will be $Y_j \in \mathbb{R}^{256\times6\times6}$, and for the k slices of the volumetric MRI, we will obtain varying feature map shapes $\mathbb{R}^{k\times256\times6\times6}$, where $s = \{y_i\}_{i=1}^{k}$ is the set of feature maps of all the slices of a volumetric MRI. Most importantly, each slice in the AlexNet feature extraction layers is treated fairly by sharing the learnable parameters, which makes it easier to train and less prone to overfitting.

The global average pooling computes the average value of all elements in the feature map (M. Lin, Chen, & Yan, 2014). Thus, the Global average pooling and max pooling layers were incorporated on the top of the feature extraction layers of the ImageNet architectures to reduce the generated feature maps obtained in the previous module and increase the compactness of the models. The global average pooling layer applied across the spatial dimensions, reduce the generated features maps in the AlexNet based Mv-CNN from $k \times 256 \times 6 \times 6$ to $k \times 256$. On the other hand, the max pooling layer applied across slices on top of the global average pooling layer reduces the generated feature maps from $k \times 256$ to reduce to a 1D 256-dimensional vector, as a total feature representation. The description of the feature extraction layers of the AlexNet based Mv-CNN with dimensions of feature maps at each layer is detailed in Figure 3.9.

The last module that is incorporated in the Mv-CNN architecture is the three fully connected layers with 256 nodes on each of the first two layers and a node with a sigmoid activation function on the last year that map the computed hidden representation to the output probability prediction in the 0 to 1 range indicating whether an instance belongs to the given classes or not. In the experiments performed in this thesis, class labels both for HGG and short-term OS patient exams were set to 1 (positive), and class labels both for LGG and long-term OS patient exams were set to 0 (negative).

Layer	Kernel size	Stride	#Features
max pooling			256
global average pooling			k x 256
max pooling	3 x 3	2 x 2	k x 256 x 6 x 6
convolutions	3 x 3	1 x 1	k x 256 x 13 x 13
convolutions	3 x 3	1 x 1	k x 256 x 13 x 13
convolutions	3 x 3	1 x 1	k x 384 x 13 x 13
max pooling	3 x 3	2 x 2	k x 192 x 13 x 13
convolutions	5 x 5	1 x 1	k x 192 x 27 x 27
max pooling	3 x 3	2 x 2	k x 64 x 27 x 27
convolution	11 x 11	4 x 4	k x 64 x 55 x 55
input (k x 3 x 224 x 224)			

Figure 3.9. Description of the AlexNet based Mv-CNN feature extraction layers, which transform the input stacked slices with k × 3× 224 × 224 into a 256-dimensional vector

The Mv-CNN architectures were trained using an adaptive moment estimation (Adam) optimization algorithm that is an extension of the stochastic gradient descent (SGD) that has recently seen broader adoption for deep learning applications in computer vision and natural language processing. The Adam optimization technique has many advantages including: computationally efficient, invariant to diagonal rescale of the gradient, appropriate for problems with very noisy gradients. The optimizer is responsible for updating the learnable parameters via backpropagation The parameters of the AlexNet portion of the Mv-CNN architecture were leveraged from the AlexNet trained on the ImageNet dataset, then fine-tuned these parameters using our dataset.

The proposed Mv-CNN architecture consists of a large number of trainable parameters. Moreover, the available training datasets are small and unbalanced class sizes which may lead the Mv-CNNs to converge towards the majority class sizes. Thus, apart from transfer learning, conventional data augmentation techniques, such as Gaussian noise, horizontal and vertical flipping, random

rotation, and elastic deformation, were used during training the proposed CNN models to increase the size of the training data and make the models more robust by learning more features. This helps increase the diversity of the general MRI images, which may improve the generalization of the proposed Mv-CNN on the test dataset samples. In addition, since we are dealing with the binary classification problem with imbalanced class sizes, the weighted binary cross-entropy loss function through modifying the loss function which is defined in Equation (8), to weight the two classes differently based on the class frequencies, has been applied for all the tasks while training the models. With this setting, we can have all of the examples for the loss contribution from both the negative and positive classes to be the same.

The Mv-CNN architectures were trained twelve times using each pre-processed 2D images of the four modalities projected in the three planes. With such trained twelve models, patient-level twelve predicted probabilities were generated. Then, we fused these twelve predicted probabilities using different ensemble methods, which is discussed in Section 3.3. The idea is that a more accurate stratification of OS patients with glioma may be obtained when the individual learners' outputs, which are trained separately on 2D projected images of multimodal MRI scans, are combined.

3.2.5. Multi-Column CNN (Mc-CNN)

To investigate and compare the performances of each modality and 2D sliced images, and investigate the learning capability of different Multi-View CNN architectures for the brain tumor diagnostic system, we developed two Multi-Column CNN (Mc-CNN) architectures that can combine information from multiple 2D-projection images using multiple base learners within a single CNN architecture by sharing a substantial amount of the learning parameters among the base learners, which is inspired by prior work of Nan Wu et al. (2019). In this

thesis, the two proposed Mc-CNN architectures are named TriPlanar-CNN and QuadPlanar-CNN according to their architectural design style.

The TriPlanar-CNN architecture is a three-column CNN architecture that takes as input the 2D stacked slices of the volumetric MRI scan of the three views (axial, coronal and sagittal), which takes advantage of the multi-view property of the volumetric MRI. As shown in Figure 3.10, the model computes the output probability of a patient being HGG or LGG and short-term OS or long-term OS in three stages. In the first stage, each input is processed independently by the three columns aiming to extract features from each view-based slice. In the second stage, the feature outputs by the three columns are consolidated into a fixed-size feature map by using concatenation. In the third stage, Concatenated features are processed together and classification is performed using a fully connected layer and a sigmoid activation on top of it. Finally, the individual classification performance of each modality was compared and the probability predictions of the four modalities were fused to obtain the final more accurate classification result at the patient level.

The feature extractor module in each column of the TriPlanar-CNN architecture adopts the feature extraction layer of the Mv-CNN architecture, which is explained in the previous section. Similar to the Mv-CNN architecture, transfer learning from AlexNet pre-trained on natural images was applied on the feature extractor module and fine-tuned each column with the pre-processed axial, coronal and sagittal slices, respectively. Similarly, each of the three columns outputs a fixed-dimension hidden representation for each of the three views. Subsequently, a fusion layer is used to integrate the output of the three columns vector into a 768-dimension vector. In the end, a fully connected layer with a sigmoid activation function was used to obtain the classification result in the 0 to 1 range.

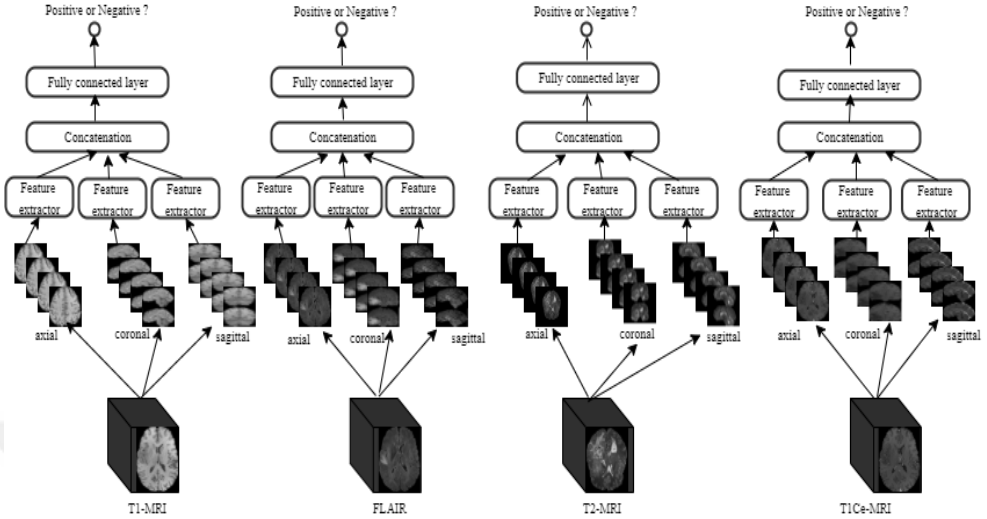

Figure 3.10. A schematic representation of the Triplanar-CNN architecture pipeline. Each 3D MRI modality is represented by a sequence of 2D slices in the axial, coronal and sagittal planes, giving rise to our Triplanar-CNN architecture. Consecutive axial, coronal, and sagittal slices, respectively, are stacked as image input to the feature extractor. The feature extraction layers are employed to extract view-specific features from each modality. The concatenation layer integrates features obtained from each column, followed by a fully connected layer with a sigmoid activation function that maps the 3D MRI scan to probability prediction in the 0 to 1 range.

The TriPlanar-CNN architecture was trained four times using each preprocessed 2D images of the four modalities. Similarly, The QuadPlanar-CNN architecture was trained three times using each preprocessed 2D images sliced in the axial, coronal, and sagittal planes. With such trained 7 models, patient-level seven predicted probabilities were generated. Then, we fused these seven predicted probabilities using different ensemble methods, which were discussed in Section 3.3.

The network architecture of the QuadPlanar-CNN is similar to that of the TriPlanar-CNN architecture. It consists of four feature extractor sub-networks that aim to learn view-specific features of the four modalities within the same

architecture, followed by a common fusion and fully connected sub-networks that compute the output (see Figure 3.11). The feature extractor modules in each column adopt the feature extraction layer of the Mv-CNN architecture, and similar to the Mv-CNN architecture, transfer learning from AlexNet pre-trained on natural images was applied on the feature extractor modules and fine-tuned each column with the pre-processed 2D stacked slices. Each of the four columns outputs the same 256-dimensional hidden representation for each 2D stacked slices projected in the same plane from the four modalities.

The fusion layer in the QuadPlanar-CNN was added to integrate the outputs of the four-column vectors into a 1024-dimension vector through concatenation. The QuadPlanar architecture also further incorporated a fully connected layer and a sigmoid activation to map the computed hidden representations to the binary classification results in the 0 to 1 range. Using the same training fashion to that of the Mv-CNN, we trained the QuadPlanar-CNN architecture three times, one for each 2D projections of the four modalities, resulting in three different binary probability estimates for each patient's prognostic prediction. In the end, probability predictions by the TriPlanar-CNN and QuadPlanar-CNN models were fused to improve the classification performance and compare the result with the Mv-CNN approach.

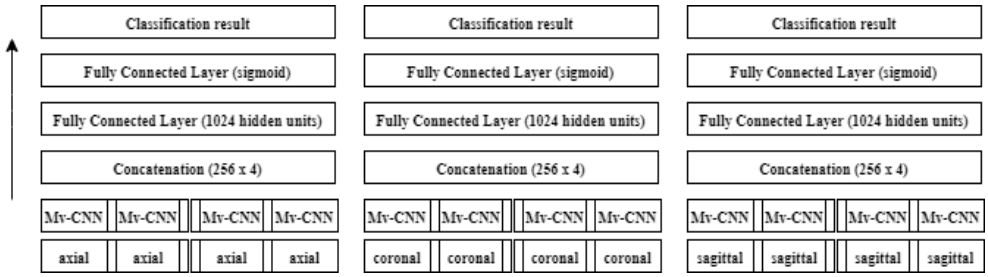

Figure 3.11. The Mc-CNN architecture pipeline. Mv-CNN refers to the feature extraction layers of the Mv-CNN architecture. The arrow indicates the direction of the information flow.

3.3. Ensemble Learning

Ensemble learning is the process by which multiple independent learning algorithms, often called the base learners, are trained separately either in parallel or in a cascading manner, and the decisions made by each base learners are subsequently combined with an optimal fusion strategy. Thus, even if a weak classifier got a wrong prediction, the whole ensemble classifiers could correct the error back. In general, the main purpose of ensemble learning is to obtain an overall better predictive performance than that could be achieved from any of the constituent base learning algorithms or to reduce the variance and bias of the base learner model predictions (Zhou, Wu, & Tang, 2002).

In this thesis, we have performed an ensembling learning experiment by combining the outputs of multiple deep learning-based base learner models to effectively improve the performance and generalization of our automated prognostic models. Hence, the individual Mv-CNN base learner prediction results after trained in each 2D stacked image types sampled from the four volumetric MRI modalities in the axial, coronal, and sagittal planes (12 in total) were fused using different fusing strategies in order to reach the final better conclusion. Similarly, the individual TriPlanar-CNN and QuadPlanar-CNN prediction results after trained in each 2D stacked slices (7 in total) were combined with different fusing strategies. In our experiments, we used the majority voting (VOT), weighted VOT, averaging (AVE), weighted AVE, and stacking ensembling strategies.

Unlike that of VOT and AVE, stacking involves training a learning algorithm using the available data, then a combiner algorithm is subsequently trained to make a final prediction using all the predictions of the other algorithms as additional inputs. In this study, traditional ML algorithms including logistic regression, MLP, Naïve Bayes, SVM, Random Forest (RF), and Random Tree (RT) were used to fuse the prediction of the individual DL models and correspondingly generate 8 ensemble learners. Finally, the comparison of results on

the effects of different combining strategies is compared. The conceptual diagram of the ensemble approach used in this study is illustrated in Figure 3.12.

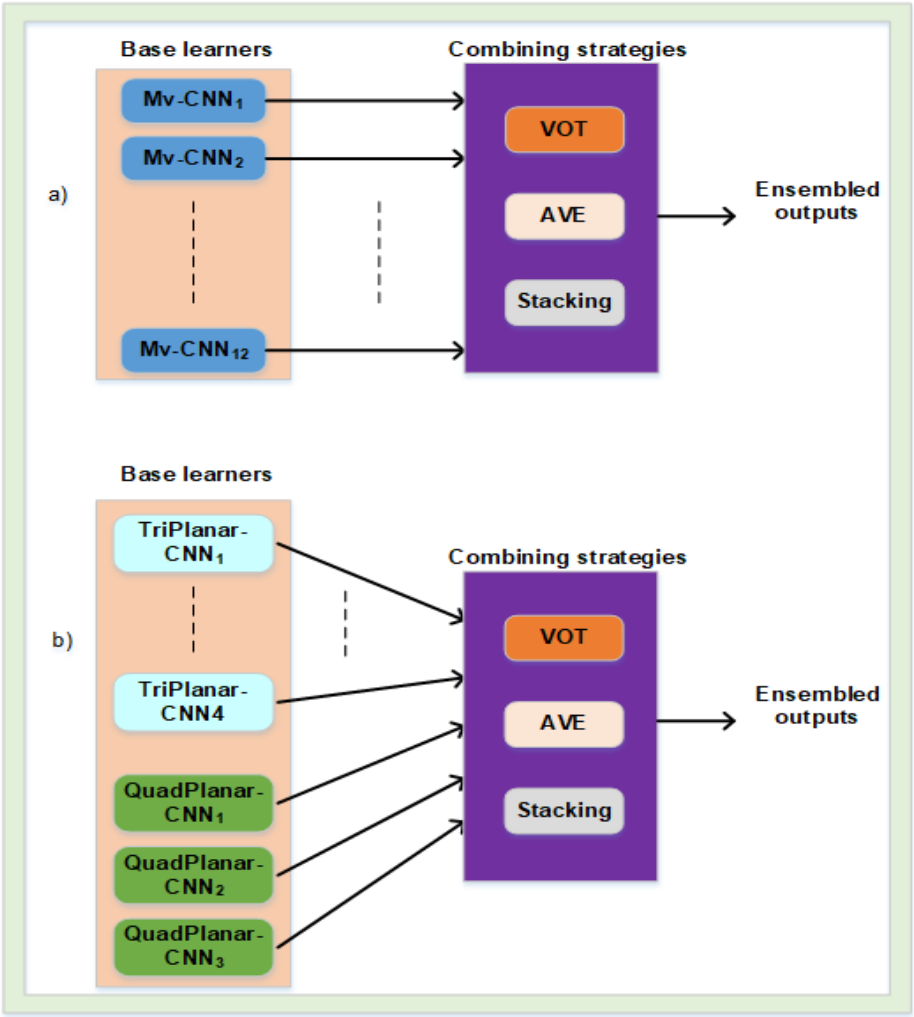

Figure 3.12. Deep learning-based individual classifiers combined using different fusion strategies. a) The individual Mv-CNN models outputs combined using the VOT, AVE, and Stacking ensembling strategies. b) The individual Multi-Column CNN models outputs combined using VOT, AVE, and Stacking ensembling strategies

3.3.1. Majority Voting (VOT)

In majority voting ensemble strategies, the predicted class label for a particular sample is the class label that represents the majority of the class labels predicted by each individual deep learning model. Given that our prognostic task is a binary classification problem, VOT can be evaluated using Equation 3.14,

$$P(x_n) = \frac{\sum_{i=1}^{k} Y_{DL_i}(x_n)}{k} \tag{3.14}$$

where $P(x_n)$ is the final prediction result for a patient x_n using the VOT method. $Y_{DL_i}(x_n)$ represents the classification result for each DL model of the patient x_n. If $P\left(Y_{DL_i}(x_n)\right) > 0.5$, $Y_{DL_i}(x_n)$ is defined as 1, otherwise $Y_{DL_i}(x_n)$ is 0. k represents the number of base learners, i.e. k is 12 if the base learners are Mv-CNN otherwise it is 7, which represents the number of Multi-Column CNN models as base learners.

3.3.2. Averaging (AVE)

Similar to the VOT, AVE can also be evaluated using the following Equation 3.15,

$$P(x_n) = \frac{\sum_{i=1}^{k} P(Y_{DL_i}(x_n))}{k} \tag{3.15}$$

where $P(x_n)$ is the final prediction result for a patient x_n using the AVE method. $P\left(Y_{DL_i}(x_n)\right)$ is the predicted result for each DL model of the patient x_n.

3.3.3. Logistic Regression

Logistic regression is one of the most popular supervised ML classification techniques for modeling a categorical outcome variable (or output), Y, from a set of features (or inputs), X, using the sigmoid function to model a binary dependent variable and the maximum likelihood estimation approach to optimize the model. The output prediction by the logistic regression is based on a linear combination of more than one feature value as an argument to the sigmoid function, resulting in output between 0 and 1; a probability. The logistic regression can be evaluated based on the Equation 3.16,

$$Y = P(y = 1|x) = \frac{1}{1+e^{W^T X+ b}} \tag{3.16}$$

where Y is the predicted probability that the output is 1 given its input features, b is the bias, and W is the learnable parameter vectors for the input feature vectors X. For the binary classification, the middle predicted value of Equation 14 is considered as a threshold that establishes what belongs to class 1, and to class 0. In particular, input having features that produce an outcome of less than 0.5 is considered to belong to class 0. Otherwise, the corresponding inputs are classified as belonging to class 1. In our experiments, we trained the logistic regression to weight the probability prediction from the twelve Mv-CNN base learners and generate a single output for each patient. We also trained the logistic regression to weight the probability prediction from the seven Multi-Column CNN base learners and generate a single output for each patient.

3.3.4. Naïve Bayes

Despite its seemingly simple nature, the Naïve Bayes classifier is one of the most efficient and well-known statistical learning methods (Minsky, 1961). The Naïve Bayes classifier model predicts the probabilities of each class member based

on applying Bayes' theorem in the decision rule with strong assumptions that each feature makes an independent and equal contribution to the outcome and that is why it is named 'Naïve'. Bayes theorem is used for calculating conditional probabilities, which is the probability of an event occurring given the probability of another event has already occurred. If A and B are two events, the conditional probability of A in such a condition that B has already occurred, which is represented as P(A|B), can be equated using Equation (3.17), and the Bayes' theorem can be derived from the following mathematical equations from our probability course basics:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{3.17}$$

$$P(A \cap B) = P(A|B)P(B) \tag{3.18}$$

$$P(A \cap B) = P(B|A)P(A) \tag{3.19}$$

Then, when the right sides of Equation 3.18 and Equation 3.19 are equated the Bayes' theorem is obtained as shown in the following Equation 3.20

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{3.20}$$

The Naïve Bayes algorithm predicts the class probabilities of a given observation and selects the class with maximum probability value as the decision. It assumes that each feature value of an observation is independent of other the feature values, but the class value of the observation affects all the other features. This assumption, also known as "the conditional independence", simplifies the probability calculation and works very efficiently for problems in medical fields,

probably connected to the fact that the chosen symptoms are independent to some degree (Sierra et al., 2001).

In statistic and computer science literature, Naïve Bayes classifier is also simply called Bayesian classifiers, and it works as follows:

1. Assume that a training data sample X in the dataset has features = {x1, x2, … , xn}, and there are m classes represented as C1, C2, …, Cm. Using Bayes' theorem defined in Equation 20, it is possible to determine the class of sample X from each of the possible classes as depicted in Equation 3.21:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{3.21}$$

2. In the Naïve Bayes-based classification, it is assumed that xi feature values are conditionally independent. As a result, in order to calculate P ($X|C_i$) Equation 3.22 can be used (Han and Kamber, 2000):

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{3.22}$$

3. Since the P(X) value will be the same in all the class probability calculations, and also since the maximum class probability is going to be selected as a class, we can discard P(X) from calculations as shown in Equation 3.23.

$$\arg\max\{P(X|C_i)P(C_i)\} \tag{3.23}$$

3.3.5. Support Vector Machine

Support Vector Machine (SVM) is a supervised ML algorithm, which can be generalized for classification and regression problems. Given a set of training features and their corresponding classes, the SVM aims to find an optimal hyperplane that can split new features into two classes for the case of binary classification problems. A support vector is a data point that is nearest to the hyperplane. They are accepted as the critical elements of the training dataset. The term margin is defined as the distance between the support vector and hyperplane. There can be more than one hyperplane solution in determining the optimal hyperplane but SVM picks the best hyperplane having the largest marginal distance (maximal margin) to any of the candidate hyperplanes for the generalization performance of the classifier. This optimal hyperplane is usually named the optimal separating hyperplane (OSH). However, the optimal hyperplane of any SVM is unique (Theodoridis & Koutroumbas, 2010). Figure 3.13. illustrates the binary classification problem with the SVM classifier.

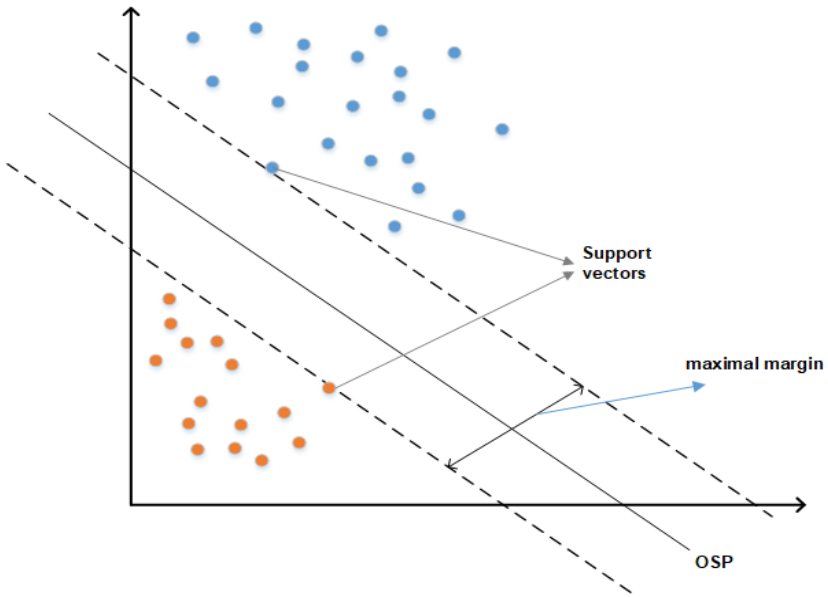

Figure 3.13. The Support Vector Machine Classifier

3.3.6. Random Forest

A decision tree can be defined as a decision support tool that uses a tree-like graph or model of decisions and their possible outcomes. Each branch of a decision tree represents the outcomes of a test, and each leaf node represents a class label. Introduced by Breiman (2001), Random Forest (RF) is one of the most effective supervised machine learning models that perform well, especially when few training data examples with many features are available (Witten, Frank, Hall, & Pal, 2011). RF operates by creating an ensemble of a large number of independent decision trees, and every individual decision tree, which is built by a random selection of Ni features, spits out a class prediction. Each tree being a weak classifier, the RF approach inspires the Bagging idea in the opinion that if a group of weak learners is ensemble they may result in a strong classifier outperforming any of the individual classifiers. The final class prediction of a random forest is based on the majority voting, which is the class with the most votes. The hyperparameters to be tuned while training an RF model include a number of trees, max depth of the forest, and class weights. Figure 3.14 is the representation of a random forest classifier. For each feature, there is one decision tree.

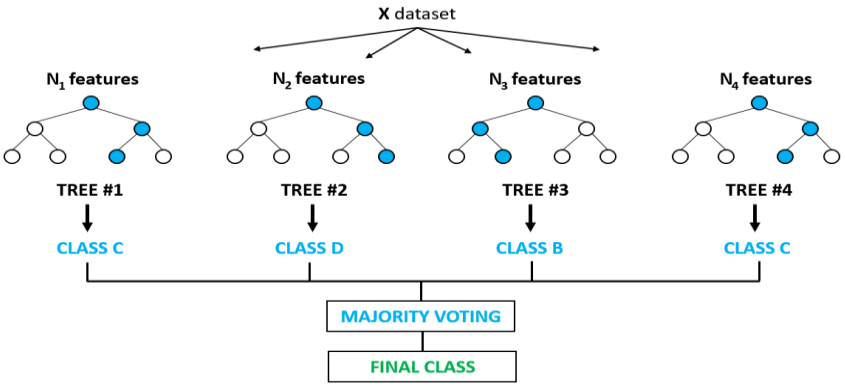

Figure 3.14. The Random Forest Classifier (Taken from the Global Software Support))

3.3.7. K-Nearest-Neighbour

The K-Nearest-Neighbour (KNN) algorithm is one of the simplest supervised ML algorithms that can be used to solve both regression and classification tasks. It works based on the assumption that similar examples exist in close proximity in the feature space, i.e. similar examples are near to each other. Therefore, for KNN distance or closeness is the metric for similarity. It calculates the distance between points on a graph. That way, when a new sample is introduced to the system, the classifier checks what the labels of the nearest data points are, and assigns the label that occurs the most to this new sample. Figure 3.15 is the representation of KNN classifier. For each feature, there is one decision tree. Significant hyperparameters include a number of neighbors, and class weights.

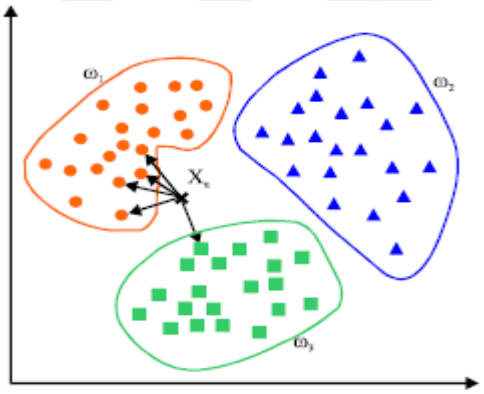

Figure 3.15. KNN Classifier (Adapted from mathwork))

3.4. Baseline Approach

Radiomics, the prediction of disease characteristics using quantitative image features from the segmented tumor regions has also been used in this study as a baseline to the proposed fully automated approach. The age of patients, and the 12 standard radiomics (imaging) features depicted in Table 3.4 were extracted from the Multi-Modal MRI images of each patient using ITK-SNAP toolkit. After

extraction, each feature was subsequently normalized as shown in the table. Finally, the combined features were used for the two tasks to train the six standard ML algorithms discussed in the previous section. The age of patients was available only for survival time prediction tasks, but not for glioma grading prediction. Hence, the age feature was not used for the glioma grading task.

Table 3.4. Description of imaging features, and normalization techniques

Features	**Normalized Features**
Age	Age/100
Extent of NCR & NET tumor	Number of Voxels of Necrotic and non-enhancing tumor/Total number of Voxels
Extent of Edema tumor	Number of voxels of Peritumoral Edema/Total number of Voxels
Extent of Enhancing tumor	Number of voxels of GD-Enhancing tumor/Total Number of Voxels
Extent of Other	Number of voxels of any region outside the whole tumor/Total Number of Voxels
Mean-NCR and NET	(The raw voxel mean image intensity of NCR and NET – mean of all patients' raw voxel mean image intensity of NCR and NET)/SDV
Mean-Edema	(The raw voxel mean image intensity of Edema – mean of all patients' raw voxel mean image intensity of Edema)/SDV
Mean-Enhancing tumor	(The raw voxel mean image intensity of Enhancing tumor – mean of all patients' raw voxel mean image intensity of Enhancing)/SDV
Mean-Other	(The raw voxel mean intensity of any region outside the whole tumor – mean of all patients' raw voxel mean intensity of any region outside the whole tumor)/SDV

STDV- NCR and NET	(Standard deviation raw voxel mean image intensity of NCR and NET – mean of all patients' raw voxel mean image intensity of NCR and NET)/SDV
STDV-Edema	(Standard deviation of raw voxel image intensity of Edema – mean of all patients' raw voxel mean image intensity of Edema)/SDV
STDV - Enhancing tumor	(Standard deviation of raw voxel image intensity of Enhancing – mean of all patients' raw voxel mean image intensity of Enhancing)/SDV
STDV - Other	(Standard deviation of raw voxel image intensity of any region outside the whole tumor – mean of all patients' raw voxel mean image intensity of any region outside the whole tumor)/SDV

3.5. Hardware and Frameworks

3.5.1. Hardware

The proposed deep learning implementations were developed using a remote machine, and desktop computers that have the following features:

1. Desktop computer 1: Intel Corei7 2.8GHz, 1GB RAM, NVIDIA GeForce GTX 1050 2GB, Windows 10 Operating System.
2. Desktop computer 2: AMD Phenom Quad-Core 2.5GHz, 8GB RAM, NVIDIA GeForce RTX 2070 8GB, Windows 10 Operating System.
3. Remote Machine (Google's free cloud service for AI developers): Google Colab with free Tesla K80 GPU

3.5.2. PyTorch

PyTorch is an optimized tensor open-source framework for the implementation of DL models using GPUs and CPUs. It is primarily developed by Facebook's AI Research lab group (FAIR), and open-sourced on GitHub in 2017. It offers a platform to research the computer vision and natural language processing fields to implement new DL architectures or state-of-the-art DL architectures with trained models for achieving state-of-the-art results quickly and efficiently, regardless of resource constraints. In general, it has a reputation for simplicity, ease of use, flexibility, debugging capabilities, and efficient memory usage. In our study, the proposed deep learning architectures, including the Mv-CNN and Multi-Column CNN architectures were implemented using this PyTorch library.

3.5.3. NiftyNet

NiftyNet is an open-source deep learning platform built on the TensorFlow framework to do research in both 2D and 3D medical image analysis applications, including segmentation and classification (Gibson et al., 2018). It is developed by a consortium of research groups, including the School of Biomedical Engineering and Imaging Sciences at King's College London (BMEIS), Wellcome Centre for Medical Engineering (CME), and the High-dimensional Imaging Group (HIG) at the UCL Institute of Neurology. NiftyNet’s modular structure is designed for sharing networks and pre-trained models to accelerate and simplify the development of DL models. In summary, using NiftyNet’s modular structure it is possible to get started with established pre-trained models, adapt existing state-of-the-art architectures to your imaging data, and rapidly develop new solutions to new medical image analysis problems. In our study, NiftyNet were used to do the automatic segmentation of tumor regions from the multi-modal MRI images using the well-established existing model, which is based on the paper entitled “Automatic Brain Tumour Segmentation Using Cascaded Anisotropic Convolutional Neural Network” (Wang, Li, Ourselin, & Vercauteren, 2018).

3.5.4. ITK-SNAP

ITK-SNAP is a free, open-source, and multi-platform powerful toolkit for 3D medical image segmentation and visualization through overlaying the segmentation on the image (Yushkevich, Gao, & Gerig, 2016). It supports different volumetric medical image formats, including NIFTI, and DICOM. It is the product of a decade-long collaboration between Paul Yushkevich, Ph.D., of the Penn Image Computing and Science Laboratory (PICSL) at the University of Pennsylvania, and Guido Gerig, Ph.D., of the Scientific Computing and Imaging Institute (SCI) at the University of Utah, whose vision was to create a tool that would be dedicated to a specific function, segmentation, and would be easy to use and learn. In order to evaluate the baseline approach, all the radiomics features discussed in Section 3.4 were extracted using the ITK-SNAP software.

3.5.5. Weka

Weka stands for Waikato Environment for Knowledge Analysis, was developed in New Zealand, at the University of Waikato for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions (Holmes, Donkin, & Witten, 1994). It is free a software licensed under the GNU General Public License, and the companion software to the book "Data Mining: Practical Machine Learning Tools and Techniques" (Witten et al., 2011). Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (numeric or nominal attributes). In this study, all the standard ML algorithms that are discussed in Section 3.3 were implemented using the Weka software.

4. RESULTS

4.1. Evaluation Metrics

In order to evaluate the proposed models' performance, the results predicted by the models have been compared with the ground truths using different metrics, including the Accuracy, Area Under the Receiver Operating Curve (ROC-AUC), and F1-Score (FSC). In addition, we used the Specificity (SPE), the Sensitivity (SEN), and precision (PRE) evaluation metrics, which are key ingredients in evaluating models in medical settings. These evaluation metrics are calculated based on the outcome of a model interpreted with four possible outcomes represented by True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), which are elements of a confusion matrix as shown in Table 4.1.

Table 4.1. A Confusion Matrix for a binary classification system

	Model Output	
Ground Truth	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

The confusion matrix can show the performance of a classification model in the form of a table by showing how many positive or negative cases are predicted truly, or falsely. The rows in the table correspond to the ground truth, and the columns correspond to the model predictions. The cells in the table correspond to the number of elements corresponding to each ground-truth model prediction combination.

The True Positive (TP) in the confusion matrix refers to the number of positive events correctly predicted as positive. Similarly, the False Positive (FP) refers to the number of events predicted as positive when in fact they are negative.

The True Negative (TN) in the confusion matrix refers to the number of negative samples correctly predicted as negative. Similarly, the False Negative (FN) refers to the number of samples predicted as negative when in fact they are positive. Equation 4.1 to Equation 4.9 shows the definitions, and mathematical descriptions of the parameters, which are used to evaluate the proposed models.

$$TP = Number\ of\ HGG\ patients\ correctly\ predicted\ as\ HGG\ or\ Number\ of\ short-term\ OS\ patients\ correctly\ predicted\ as\ short-term\ OS \quad (4.1.)$$

$$FP = Number\ of\ LGG\ patients\ incorrectly\ predicted\ as\ HGG\ or\ Number\ of\ long-term\ OS\ patients\ incorrectly\ predicted\ as\ short-term\ OS \quad (4.2.)$$

$$TN = Number\ of\ LGG\ patients\ correctly\ predicted\ as\ LGG\ or\ Number\ of\ long-term\ OS\ patients\ correctly\ predicted\ as\ long-term\ OS \quad (4.3.)$$

$$FN = Number\ of\ HGG\ patients\ incorrectly\ predicted\ as\ LGG\ or\ Number\ of\ short-term\ OS\ patients\ incorrectly\ predicted\ as\ long-term\ OS \quad (4.4.)$$

$$\mathrm{ACC} = \frac{TP+TN}{(TP+FP+TN+FN)} \times 100 \quad (4.5.)$$

$$SEN = Recall = \frac{TP}{TP+FN} \quad (4.6.)$$

$$SPE = \frac{TN}{(TN+FP)} \quad (4.7.)$$

$$PRE = \frac{TP}{(TP+FP)} \quad (4.8.)$$

$$FSC = 2 \times \frac{PRE \times Recall}{PRE + Recall} \quad (4.9.)$$

4.2. Parameter Optimization

It is not intuitive what the most efficient parameters would be for training CNN algorithms. It requires extensive testing and artistry to choose the optimal values among the possible optimal parameters values. Thus, for all the proposed deep learning models, we sought to find some of the optimal training parameter settings, including learning rate, and optimizing algorithm. Hence, our initial experiment revolved around training the designed Mv-CNN architecture with AlexNet as a backend using the different parameter settings. The training was performed for the glioma grading prediction problem using FLAIR axial slices as input. Thus, we perform an ablation study on the three most commonly used optimizers in the literature, including Adam, RMSProp, and SGD. In addition, since a learning rate that is too small leads to a slow convergence, and a learning rate that is too large can limit convergence, and make the loss function fluctuate around the minimum, we investigated the six most commonly used learning rates in the literature, ranging from 10^{-4} to 10^{-6}. However, we used the same weighted binary cross-entropy loss metric for the loss function, and hyperparameters values such as 0.9 for the momentum and 0.01 for the weight decay across all the experiments.

The dataset was split into 80% for training, and 20 % for testing at the patient level to avoid any overlap in data between the sets. Epoch is the number of times for the learning algorithm to pass through the training set, and the maximum number of the epoch was set to 500 for all the training. However, if there was no further improvement in the validation loss after 20 epochs, the training was automatically stopped and the model with the least validation loss was used. This approach is helpful because the last epoch may not give the best validation performance to be updated in the weight. In the end, the final chosen optimized parameter settings (learning rate and optimizer) were subsequently used across all the deep learning models used in this thesis.

The results of the Mv-CNN model when Adam, RMSProp, and SGD optimizers were adapted using the six different learning rates with its corresponding best model epochs number are illustrated in Table 4.2, Table 4.3, and Table 4.4, respectively. The best results of the Mv-CNN model with Adam optimizer were obtained when a learning rate of 10^{-5} was used. It converged after 12 epochs of training. The loss of the training set and that of the test set decreased to 0.0597, and 0.1642, respectively.

Table 4.2. Evaluation of Mv-CNN using Adam optimizer

Learning Rate	Training			Validation			Epoch
	AUC	ACC	Loss	AUC	ACC	Loss	
5×10^{-4}	0.5189	56.1	0.2756	0.6968	75.4	0.2645	4
10^{-4}	0.8687	79.4	0.1776	0.8238	61.4	0.2046	8
5×10^{-5}	0.8677	77.2	0.1808	0.7937	63.2	0.1964	2
10^{-5}	**0.9987**	**98.9**	**0.0597**	**0.8397**	**77.2**	**0.1642**	12
5×10^{-6}	0.9984	96.1	0.0640	0.8349	73.7	0.1862	19
10^{-6}	0.993	93	0.0921	0.827	68.4	0.1866	56

Table 4.3. Evaluation of Mv-CNN using RMSProp optimizer

Learning Rate	Training			Validation			Epoch
	AUC	ACC	Loss	AUC	ACC	Loss	
5×10^{-4}	0.6293	66.7	0.2657	0.8175	73.7	0.2639	93
10^{-4}	0.5029	62.3	3.6177	0.6508	71.9	0.2551	1
5×10^{-5}	0.8505	0.75	0.1785	0.754	0.5789	0.2152	43
10^{-5}	0.8145	73.7	0.2002	0.7698	54.4	0.2252	3
5×10^{-6}	0.9562	89.5	0.1070	0.8095	75.4	0.2015	14
10^{-6}	**1.0**	**99.6**	**0.0417**	**0.8381**	**75.4**	**0.1682**	16

Table 4.4. Evaluation of Mv-CNN using SGD optimizer

Learning Rate	Training			Validation			Epoch
	AUC	ACC	Loss	AUC	ACC	Loss	
5×10^{-4}	0.873	78.95	0.1705	0.8063	71.93	0.2096	3
10^{-4}	0.9488	83.8	0.1170	**0.8143**	**77.2**	**0.1734**	10
5×10^{-5}	0.9408	85.1	0.1313	0.8048	68.4	0.1959	12
10^{-5}	**0.9777**	**89.5**	**0.1056**	0.8095	70.2	0.1965	74
5×10^{-6}	0.9698	87.7	0.1131	0.8048	68.4	0.1974	176
10^{-6}	0.9077	75.9	0.1665	0.7984	68.4	0.2034	225

As depicted in Figure 4.1, after 12 epochs, while the training loss was decreasing the testing loss was not decreasing. The same trend was observed for the AUC values as well. Similarly, the best model for RMSprop optimizer was obtained with a 10^{-6} learning rate and epoch number 16, and 10^{-4} learning rate, and epoch number 10 for SGD optimizer. Figure 4.1 indicates that the best performing model with higher validation AUC and lesser validation loss was obtained when Adam optimizer was adapted with a 10^{-5} learning rate than the SGD and RMSProp optimizers. For this reason, we chose Adam optimizer with a learning rate of 10^{-5} in all deep learning experiments implemented in this thesis.

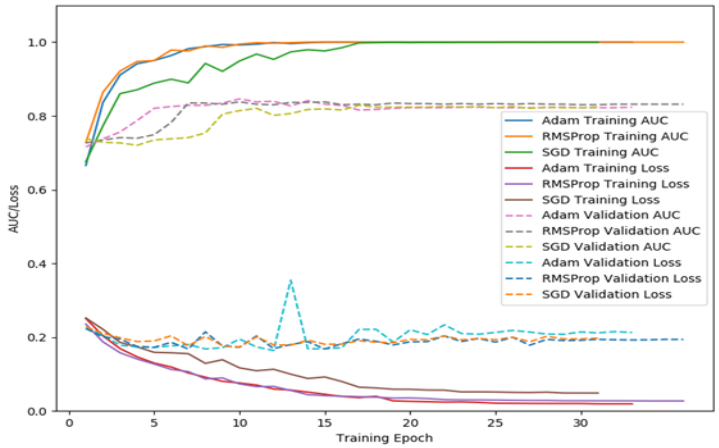

Figure 4.1. Comparison of optimization algorithms.

4.3. Choice of Feature Extractor

After we did the first experiment for choosing the optimal training parameter settings, our second experiment was sought to determine the optimal feature extractor layers for our proposed deep learning architectures. Hence, the second ablation study was revolved around training the designed Mv-CNN architecture fifteen times using the feature extractor layers of the 15 most common state-of-the-art CNN models developed for ImageNet as a backend. The

experiments were performed using the same training strategy on the glioma grading prediction problem based on the T1 axial slices as input. The dataset was split into 80% for training, and 20 % for testing at the patient level to avoid any overlap in the data between the sets.

We evaluate all the 15 models using different evaluation metrics, and show the result in Table 4.5. In general, our results show that the Mv-CNN architecture using the VGG-19 feature extractor layers as a backend achieves the best performance with an AUC of 0.981, ACC of 0.947, FSE of 0.964, and 0.0627 loss on the test dataset. GoogLeNet was the second best with an AUC of 0.981, ACC of 0.93, FSE of 0.95, and 0.0677 loss on the test dataset, while, DenseNet121 got the worst performance with an AUC of 0.752 AUC, ACC of 0.79, FSE of 0.867, and 0.2426 loss on the test set. GoogleLeNet has a similar AUC to that of VGG-19. However, the VGG-19 model has higher ACC, and FSC and smaller validation loss than GoogLeNet, and the other models. In addition, we know that the higher the SPE, the lower the chance of misclassifying the LGG patients, and the higher the SEN, the lower the chance of misclassifying the HGG patients. But, regardless of the class imbalance on the dataset, the VGG-19 model stratifies the LGG and HGG patients with 0.952 SEN and 0.933 SPE, which shows the robustness of the feature extractor compared to the other 14 models. As a result, we chose VGG-19 as a feature extractor for all the proposed CNN architectures used for glioma grading and AlexNet for survival time prediction problems used in this thesis.

Table 4.5. Performance comparison of state-of-the-art CNN pre-trained models

Model	AUC	ACC	SEN	SPE	PRE	FSE	v.losss
AlexNet	0.962	87.7	0.881	0.867	0.949	0.914	0.0923
DenseNet121	0.752	79	0.929	0.4	0.812	0.867	0.2426
DenseNet169	0.929	87.7	**0.952**	0.667	0.889	0.92	0.2100
DenseNet201	0.937	86	0.857	0.867	0.947	0.9	0.1535
GoogLeNet	**0.981**	93	0.905	**1.0**	**1.0**	0.95	0.0677
MobileNetV2	0.97	93	0.929	0.933	0.975	0.951	0.0858
MnasNet	0.967	93	0.905	**1.0**	**1.0**	0.95	0.0964
ResNet18	0.944	87.7	0.881	0.867	0.949	0.914	0.1200
ResNet34	0.97	89.5	0.857	**1.0**	**1.0**	0.923	0.0785
ResNet50	0.94	89.5	0.857	**1.0**	**1.0**	0.923	0.0881
SqeezNet1	0.933	89.5	0.905	0.867	0.895	0.927	0.1226
VGG11	0.968	91.2	0.905	0.933	0.974	0.938	0.0893
VGG13	0.956	87.7	0.881	0.867	0.949	0.914	0.0981
VGG16	0.962	94.7	**0.952**	0.933	0.976	**0.964**	0.1061
VGG19	**0.981**	**94.7**	**0.952**	0.933	0.976	**0.964**	**0.0627**

While training from scratch, we knew that large CNN architectures performed great for big datasets, and for small datasets, the performance decreased when the architecture becomes bigger. The possible reason is that when the layers go deeper, more abstract information can be extracted, but this comes at the cost of training more parameters which requires a larger training dataset, and larger training time. Thus, for small medical image datasets, like the proposed brain tumor diagnosis in this thesis, instead of training large CNN architectures from scratch, transfer learning can solve the diagnosis problem quite well. From our experimental results depicted in Table 4.5, we can also observe that 93.3% and 60% of the pre-trained models achieved AUC values of above 0.9, and 0.95, respectively, in classifying glioma into two, while the training time of each model is less than 5 hours which normally takes more than a day while training without fine-tuning. Even, light architectures like AlexNet, and MobileNet achieved an AUC value above 0.95 while the training time is less than 3 hours.

Moreover, we examined whether larger pre-trained architectures performed better than smaller architectures on glioma grade prediction. We display this relationship in Table 4.6, and Figure 4.2. Our experimental results show that, with ImageNet pre-training, there is no linear association between the number of

parameters and the glioma grading classification's AUC, with Spearman correlation coefficient (ρ) = -0.174 as shown in Figure 4.2. These results demonstrate that bigger does not necessarily guarantee better performance in pre-trained models for small medical datasets. Similarly, smaller architecture does not also guarantee better performance in pre-trained models for small medical datasets. For example, the AlexNet based Mv-CNN architecture, a very light architecture with approximately 2.6 million parameters, achieved an AUC value 2.5% larger than the ResNet50 based Mv-CNN architecture which has approximately 50 million parameters. Similarly, the VGG-19 based Mv-CNN architecture which has approximately 20.2 million parameters achieved an AUC value of 1.9% larger than the AlexNet based Mv-CNN architecture, but is 7.77x larger. In general, based on our experimental investigations, we can conclude that transfer learning parameter efficient ImageNet models, such as AlexNet, can be effectively used for biomedical image classification problems in small resource environments without degrading the performance.

Table 4.6. Number of parameters for ImageNet models and Mv-CNN performance for glioma grading

Model	AUC	#Total Params	#Feature extractor Params	Mv-CNN	
				#Params	Log_{10}
AlexNet	0.962	61,100,840	2,469,696	2,601,537	6.415
DenseNet121	**0.752**	7,978,856	6,953,856	7,282,305	6.862
DenseNet169	0.944	14,149,480	12,484,480	12,976,769	7.113
DenseNet201	0.937	20,013,928	18,092,928	18,650,753	7.271
GoogLeNet	**0.981**	6,624,904	5,599,904	5,928,353	6.773
MobileNetV2	0.97	3,504,872	2,223,872	2,617,857	6.418
MnasNet	0.967	4,383,825	3,102,312	3,497,321	6.544
ResNet18	0.944	11,689,512	11,176,512	11,373,889	7.056
ResNet34	0.938	21,797,672	21,284,672	21,482,049	7.331
ResNet50	0.937	25,557,032	23,508,032	24,098,625	**7.382**
SqeezNet1	0.933	1,248,424	735,424	932,801	**5.970**
VGG11	0.968	132,863,336	9,220,480	9,417,857	6.974
VGG13	0.956	133,047,848	9,404,992	9,602,369	6.982
VGG16	0.962	138,357,544	14,714,688	14,912,065	7.174
VGG19	**0.981**	143,667,240	20,024,384	20,221,761	7.306

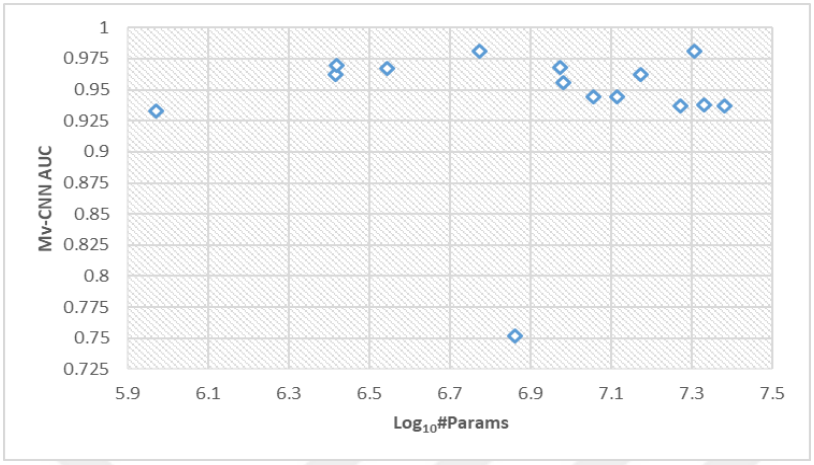

Figure 4.2. Mv-CNN AUC vs. Model Size. The plot shows the logarithm of the model size has no linear relationship with model performance (ρ=-0.174).

Besides, we have also tried to examine whether higher performance on ImageNet guarantees higher performance for small medical imaging tasks like the glioma grading task. We depicted the relationship between ImageNet accuracy, and the glioma grading AUC in Figure 4.3. We notice that for small medical datasets like glioma grading, we find no monotonic relationship with ImageNet accuracy, with Spearman correlation coefficient (ρ) = -0.237. Our findings suggest that pre-trained models that perform better on ImageNet do not necessarily guarantee better performances than the other pre-trained models on small medical image task datasets, at least on the dataset we examine. However, contrary to our experimental result, a strong monotonic correlation between performances on natural image tasks and ImageNet has been reported in the work by Kornblith et al (2019), supporting the implicit assumption that CNN architectures that perform better on ImageNet necessarily perform better on other image analysis tasks. Thus, a further systematic study was required across pre-trained CNN architectures for other small medical image tasks to explain the weak monotonic relationship obtained in our experiment.

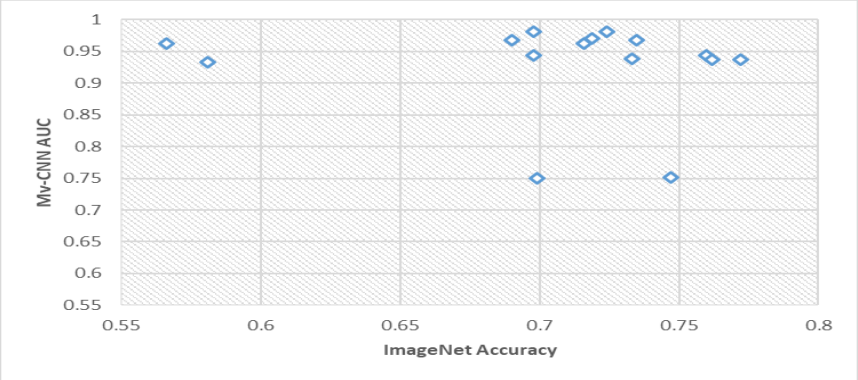

Figure 4.3. Mv-CNN AUC vs. ImageNet Accuracy. The plot shows performances of state-of-the-art CNN models trained on the ImageNet dataset have no linear relationship with its customized CNN models on a small medical image dataset (ρ= -0.237).

4.4. Comparisons of different inputs

4.4.1. Modalities

We examined the individual classification performances of all the four MRI modalities using multiple evaluation metrics obtained after training the proposed TriPlanar-CNN architecture both for the glioma grade, and survival time classifications problems, which are presented in Table 4.7, and Table 4.8, respectively.

Table 4.7. Performance comparison of MRI Modalities on glioma grading

Modality	Evaluation Metrics					
	AUC	ACC	SEN	SPE	PRE	FSC
FLAIR	0.89	83.2	0.84	0.8	0.92	0.88
T1	**0.983**	**94.7**	**0.95**	**0.95**	**0.98**	**0.96**
T1Ce	0.96	92.3	0.92	0.93	0.97	0.94
T2	0.957	90.2	0.89	**0.95**	**0.98**	0.93

Table 4.8. Performance comparison of MRI modalities on survival time prediction

Modality	Evaluation Metrics					
	AUC	ACC	SEN	SPE	PRE	FSC
FLAIR	**0.863**	**78.6**	**0.769**	0.812	0.87	**0.816**
T1	0.755	71.4	0.615	**0.875**	**0.889**	0.727
T1Ce	0.683	64.3	**0.769**	0.438	0.69	0.727
T2	0.714	61.9	0.577	0.688	0.75	0.652

For glioma grading, the average AUC of the five-fold cross-validation was 0.89, 0.957, 0.96, and 0.983 for FLAIR, T2, T1Ce and T2 modalities, respectively. For the overall survival time prediction, the AUC performances on the test dataset were 0.683, 0.714, 0.755, and 0.863 for T1Ce, T2, T1 and FLAIR modalities, respectively. The results indicate that brain images from different MRI modalities show different sensitivity to a brain tumor. Moreover, the performance of each modality for glioma grading classification has an AUC, and Accuracy values greater than 95%, except the FLAIR modality which has 89% AUC, and 83% ACC, demonstrating that each of the four MRI modalities is more or less capable of distinguishing between HGGs, and LGGs. However, the classification performances of each modality on the survival time are obviously dropped compared with the performances on the glioma grading, and it is believed to be reasonable because the survival time dataset is extremely smaller. Thus, we can conclude that survival time prediction solely based on a single modality is a difficult task when the dataset is extremely small. This necessitates the need to extract and fuse features from the multimodal MRI images using different strategies to improve the OS classification accuracy.

As shown in Figure 4.4, our experiments demonstrate that performances of the four MRI modalities based on the AUC are in the following order FLAIR > T1 > T2>T1Ce for overall survival time prediction whereas for glioma grade prediction T1>T1Ce>T2>FLAIR. These results indicate that the T1 modality has more discriminative features for glioma grading prediction compared to the other MRI modalities. Similarly, the FLAIR modality has more discriminative

features for survival time prediction compared to the other MRI modalities. However, in contrary to previous studies where FLAIR modality was reported to be superior to T1Ce images (Gurwara et al., 2010), and FLAIR to be more superior to T2 (Ge, Gu, Jakola, & Yang, 2018) in providing a more complete characterization of glioma, our findings show that T1 modality is more predictive among the four MRI modalities, and FLAIR gets the least performance. However, our finding corroborates Ge et al. (2018) which shows individual glioma grading classification performance of T1Ce is more superior to FLAIR, and T2. However, in their experiment, they did not include the classification performance of the T1 modality. In summary, based on our findings we suggest that the T1, and FLAIR modalities should be the preferred image sequence in clinical trials for glioma grading, and survival time prediction, respectively.

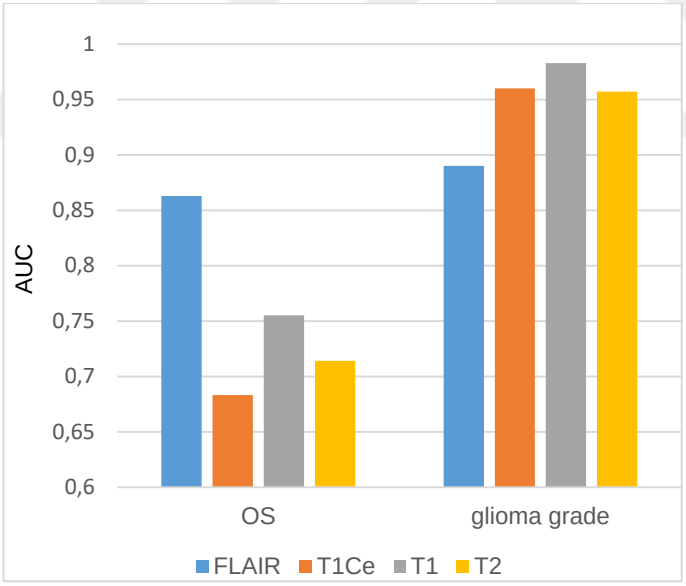

Figure 4.4. Performance comparison of the four MRI modalities on the OS and glioma grade predictions

4.4.2. 2D-projections

It is obvious that the vast amount of the brain tumor regions are found in the axial plane of the volumetric MRI images, and reasonably outperform the corresponding coronal, and sagittal 2D images when used as input while training the deep learning-based prognostic models. However, in this study, we hypothesized that through concatenating the two consecutive coronal slices, and sagittal slices, respectively, we may improve the performances of deep learning models for glioma grading, and survival time classifications. Thus, in this section, we examined the classification performance of all the twelve 2D projected images obtained from the four MRI modalities based on the glioma grade, and overall survival time classification problems through training the Mv-CNN architecture. Figure 4.5 presents comparisons of the glioma grading, and survival time classifications of the Mv-CNN model with the different twelve 2D projection images as input. As shown in the Figure, it is noted that the axial projection of the FLAIR modality, and the coronal projection of the T1Ce modality perform the best than the other inputs for survival time and glioma predictions, respectively. Most importantly, when we observe the performances of the 2D-projected images of each modality, the results are comparable and arbitrary, which complements our hypothesis where concatenation may enhance the performance of deep learning models. For instance, the sagittal, and coronal images of the FLAIR modality outperform the corresponding axial images in the glioma grade prediction. Similarly, the sagittal images of the T1Ce image outperform the corresponding axial images in the OS prediction.

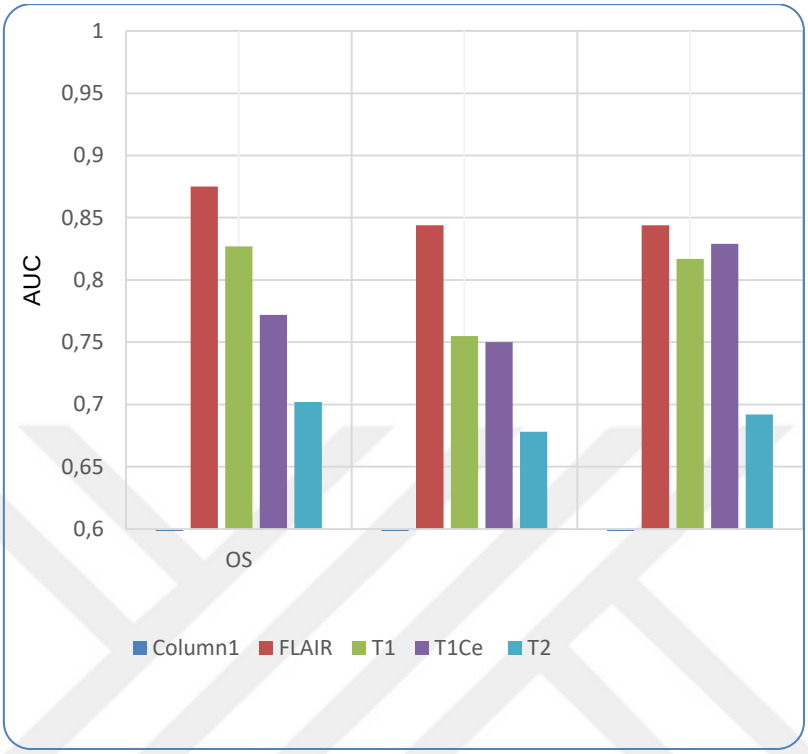

Figure 4.5. Performance comparison of all the twelve 2D-projections of the four MRI modalities on the OS and glioma grade prediction.

Furthermore, to further prove our hypothesis on concatenated and resized sagittal, and coronal slices, we did an experiment on all of the four modalities in coronal and sagittal planes using the Mv-CNN model. As shown in Figure 4.6, for all the MRI modalities (except the T2 modality), the concatenation of two consecutive slices has improved the performance of the Mv-CNN model by more than 4.6%.

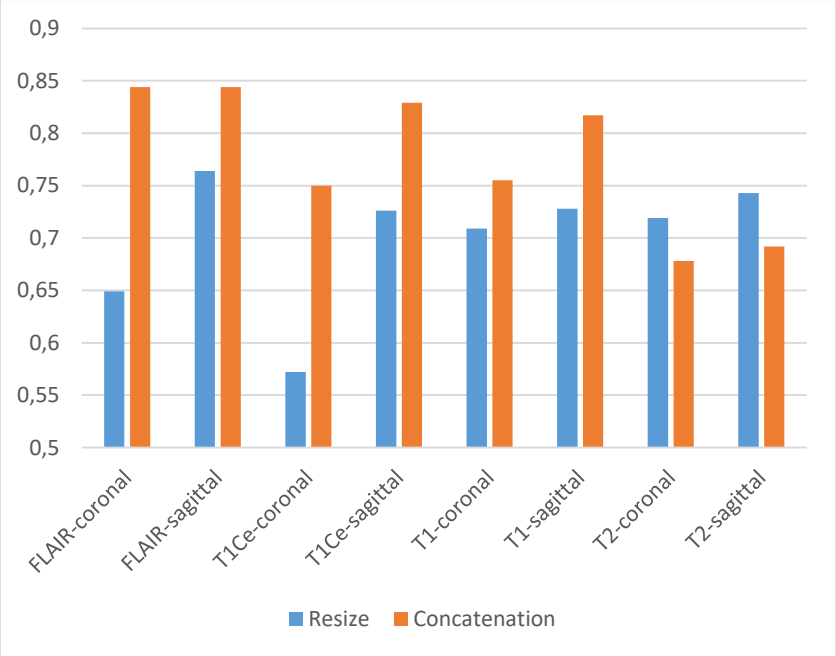

Figure 4.6. Comparisons of the survival time prediction performance of the Mv-CNN model with resized and concatenated image as inputs

4.4.3. Extended-ROI

Table 4.9 presents the comparison of the survival time prediction performance of the Mv-CNN model with five different inputs: segmented tumor region only (ROI), ROI-15 - extended tumor regions through incorporating up to 15 pixels of the surrounding tumor tissues of the ROI, ROI-30 - extended tumor regions through incorporating up to 30 pixels of the surrounding the tumor tissues of the ROI, ROI-45 - extended tumor region through incorporating up to 45 pixels of the surrounding tumor tissues of the ROI, and WSI - the whole slide image that contains the whole brain tissue. It is noted that including the surrounding tumor tissues as input while training the Mv-CNN model effectively improves the performance of survival time classification. Nevertheless, ROI-30 performs the best, indicating that adding excessive surrounding tumor tissues or discarding surrounding tumor tissues can significantly decrease the performance of the deep

learning models. Figure 4.7 also depicts the ROC curves for the Mv-CNN model with the five different inputs.

Table 4.9. Comparison of the survival time prediction performance of the Mv-CNN model with different inputs

Metrics	ROI	ROI-15	ROI-30	ROI-45	WSI
AUC	0.832	0.839	**0.875**	0.776	0.81
ACC	71.4	**73.8**	**73.8**	66.7	71.4
SEN	0.692	**0.731**	0.654	0.654	0.692
SPE	0.75	0.75	**0.875**	0.688	0.75
PRE	0.818	0.826	**0.895**	0.773	0.818
F1SC	0.75	**0.776**	0.756	0.708	0.75
LOSS	0.2417	0.1745	**0.1652**	0.2701	0.2486

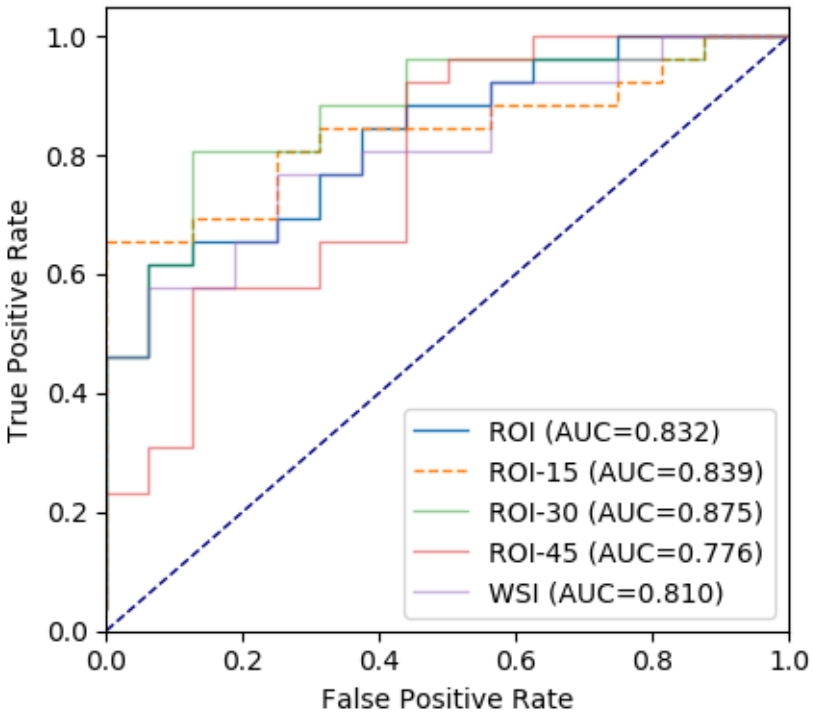

Figure 4.7. ROC curves of Mv-CNN with different inputs

4.5. Glioma grade prediction

Cross-validation is one of the most commonly used approaches in the ML field to evaluate the generalization ability of trained models. It is carried out based on splitting the whole dataset into n disjoint groups of the same size, then n-1 subgroups are used for training and the remaining subgroup for testing. Thus, we used the five-fold stratified cross-validation in our experiment upon the 285 cases, wherefrom each class, 80% of all the image dataset were used as the training set, and the rest 20% as the testing set to assess the effectiveness of the proposed models in differentiating glioma between HGG, and LGG. That is, for each testing fold, the remaining other folds were used to train all the models. This operation was reiterated five times. Only the seventy-five patients (26.3%) in fact had low-grade gliomas, whereas the rest 210 cases (73.7%) had high-grade gliomas. The results obtained for the classification of glioma grade using the Mv-CNN architecture are reported in the following sections.

4.5.1. Performances of Mv-CNN Models

The same training process of the Mv-CNN architecture was applied to each of the twelve 2D-projection images of the multi-modal MRI images, resulting in a total of 60 training operations. The five-fold cross-validation results of the Mv-CNN model for the axial, coronal, and sagittal projections of the FLAIR modality are presented in Table 4.10, 4.11, and 4.12, respectively. The AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the FLAIR axial images, when measured averaged over all the five-folds, were 83.5%, 71.9%, 69%, 80%, 90.7%, and 78.3%, respectively. Similarly, the AUC, ACC, SEN, SPE, PRE and FSC evaluation metrics of the Mv-CNN model based on the FLAIR coronal images when measured averaged over all the five folds were 85.9%, 78.3%, 76.7%, 82.7%, 92.7%, and 83.6%, respectively, when using the FLAIR sagittal images were 88.8%, 77.2%, 74.8%, 84%, 93.2%, and 82.2, respectively.

Table 4.10. Evaluation results of Mv-CNN architecture when axial projections of the FLAIR modalities are used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	0.819	66.7	0.619	0.8	0.897	0.732	0.2021
2	0.91	79	0.762	0.867	0.941	0.842	0.1405
3	0.921	77.2	0.714	0.933	0.968	0.822	0.1549
4	0.938	80.7	0.762	0.933	0.97	0.853	0.1191
5	0.589	56.1	0.595	0.467	0.758	0.667	0.2561
Average	0.835	71.9	0.690	0.8	0.907	0.783	0.1745

Table 4.11. Evaluation results of Mv-CNN architecture when coronal projections of the FLAIR modalities are used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	0.843	75.44	0.81	0.6	0.85	0.829	0.2020
2	0.954	91.23	0.881	1.0	1.0	0.937	0.084
3	0.841	78.95	0.786	0.8	0.917	0.846	0.2071
4	0.905	77.19	0.738	0.867	0.939	0.827	0.1488
5	0.754	68.42	0.619	0.867	0.929	0.743	0.2099
Average	0.859	78.25	0.767	0.827	0.927	0.836	0.1704

Table 4.12. Evaluation results of Mv-CNN architecture when sagittal projections of the FLAIR modalities are used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	0.84	73.7	0.714	0.8	0.909	0.8	0.1799
2	0.968	87.7	0.929	0.733	0.907	0.918	0.1062
3	0.941	80.7	0.786	0.867	0.943	0.857	0.1212
4	0.925	80.7	0.786	0.867	0.943	0.857	0.1374
5	0.768	63.2	0.524	0.933	0.957	0.677	0.1990
Average	0.888	77.2	0.748	0.84	0.932	0.822	0.1487

The five-fold cross-validation results of the Mv-CNN model for the axial, coronal, and sagittal projections of the T1 modality are presented in Table 4.13, 4.14, and 4.15, respectively. The AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the T1 axial images, when measured averaged over all the five-folds, were 93.3%, 88%, 85.7%, 94.6%, 97.7%, and 90.5%, respectively. Similarly, the AUC, ACC, SEN, SPE, PRE, and FSC

evaluation metrics of the Mv-CNN model based on the T1 coronal images when measured averaged over all the five folds were 81.8%, 73.3%, 74.3%, 70.7%, 88.4%, and 79.7%, respectively, when using the T1 sagittal images were 84.6%, 73.3%, 69.1%, 85.3%, 93.5%, and 78.3%, respectively.

Table 4.13. Evaluation results of Mv-CNN architecture when axial projections of the T1 modality is used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	0.989	94.7	0.952	0.933	0.976	0.964	0.0541
2	0.986	91.2	0.905	0.933	0.974	0.938	0.0507
3	0.981	94.7	0.952	0.933	0.976	0.964	0.0627
4	0.967	94.7	0.929	1.0	1.0	0.963	0.0952
5	0.741	64.9	0.548	0.933	0.958	0.697	0.2487
Average	0.933	88	0.857	0.946	0.977	0.905	0.1023

Table 4.14. Evaluation results of Mv-CNN architecture when coronal projections of the T1 modality is used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	0.684	63.16	0.69	0.467	0.784	0.734	0.2396
2	0.918	80.7	0.857	0.667	0.878	0.867	0.1551
3	0.838	73.68	0.667	0.933	0.966	0.789	0.1883
4	0.884	85.96	0.952	0.6	0.87	0.909	0.1997
5	0.764	63.16	0.548	0.867	0.92	0.687	0.2202
Average	0.818	73.3	0.743	0.707	0.884	0.797	0.2006

Table 4.15. Evaluation results of Mv-CNN architecture when sagittal projections of the T1 modality is used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	0.767	61.4	0.476	1.0	1.0	0.645	0.2095
2	0.948	80.7	0.762	0.933	0.97	0.853	0.1188
3	0.771	64.9	0.619	0.733	0.867	0.722	0.1972
4	0.946	87.7	0.929	0.733	0.907	0.918	0.1183
5	0.8	71.9	0.667	0.867	0.933	0.778	0.1964
Average	0.846	73.3	0.691	0.853	0.935	0.783	0.1680

The five-fold cross-validation results of the Mv-CNN model for the axial, coronal, and sagittal projections of the T1Ce modality are presented in Table 4.16, 4.17, and 4.18, respectively. The AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the T1Ce axial images, when measured averaged over all the five-folds, were 97.9%, 95.1%, 94.1%, 98.7%, 99.4%, and 96.3%, respectively. Similarly, the AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the T1Ce coronal images when measured averaged over all the five folds were 98.3%, 93.7%, 92.4%, 97.3%, 99%, and 95.4%, respectively, when using the T1Ce sagittal images were 97.2%, 91.9%, 92.4%, 90.7%, 96.6%, and 94.4%.

Table 4.16. Evaluation results of Mv-CNN architecture when axial projections of the T1Ce modality is used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	1.0	98.3	0.976	1.0	1.0	0.988	0.0116
2	1.0	98.3	0.976	1.0	1.0	0.988	0.0103
3	1.0	100	1.0	1.0	1.0	1.0	0.0047
4	0.997	98.3	0.976	1.0	1.0	0.988	0.0300
5	0.898	80.7	0.7762	0.933	0.97	0.853	0.1535
Average	0.979	95.1	0.941	0.987	0.994	0.963	0.042

Table 4.17. Evaluation results of Mv-CNN architecture when coronal projections of the T1Ce modality is used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	1.0	100	1.0	1.0	1.0	1.0	0.0156
2	1.0	98.25	0.976	1.0	1.0	0.988	0.0169
3	0.989	94.74	0.952	0.933	0.976	0.964	0.0654
4	0.963	91.23	0.905	0.933	0.974	0.938	0.1040
5	0.962	84.21	0.786	1.0	1.0	0.88	0.1121
Average	0.983	93.7	0.924	0.973	0.99	0.954	0.0628

Table 4.18. Evaluation results of Mv-CNN architecture when sagittal projections of the T1Ce modality is used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	0.973	93	0.952	0.867	0.952	0.952	0.1091
2	0.968	91.2	0.905	0.933	0.974	0.938	0.0766
3	0.983	93	0.905	1.0	1.0	0.95	0.0718
4	0.987	93	0.929	0.933	0.975	0.951	0.0678
5	0.949	89.5	0.929	0.8	0.929	0.929	0.1159
Average	0.972	91.9	0.924	0.907	0.966	0.944	0.0882

The five-fold cross-validation results of the Mv-CNN model for the axial, coronal, and sagittal projections of the T2 modality are presented in Table 4.19, 4.20, and 4.21, respectively. The AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the T2 axial images, when measured averaged over all the five-folds, were 95.9%, 90.2%, 89%, 93.3%, 97.3%, and 92.3%, respectively. Similarly, the AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the T2 coronal images when measured averaged over all the five folds were 82.6%, 70.2%, 63.8%, 88%, 95.1%, and 72.8%, respectively, when using the T2 sagittal images were 80.6%, 69.8%, 70.5%, 68%, 86.6%, and 77.1%.

Table 4.19. Evaluation results of Mv-CNN architecture when axial projections of the T2 modality is used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	0.998	96.5	0.976	0.933	0.976	0.976	0.0316
2	1.0	98.3	0.976	1.0	1.0	0.988	0.0113
3	0.998	96.5	0.976	0.933	0.976	0.976	0.0276
4	0.983	91.2	0.929	0.867	0.951	0.94	0.0704
5	0.817	68.4	0.595	0.933	0.962	0.735	0.2266
Average	0.959	90.2	0.89	0.933	0.973	0.923	0.074

Table 4. 20. Evaluation results of Mv-CNN architecture when coronal projections of the T1Ce modality is used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	0.786	59.6	0.476	0.933	0.952	0.635	0.2129
2	0.965	87.7	0.857	0.933	0.973	0.911	0.0855
3	0.838	80.7	0.881	0.6	0.86	0.871	0.2133
4	0.91	78.9	0.738	0.933	0.969	0.838	0.151
5	0.633	43.8	0.238	1.0	1.0	0.385	0.2497
Average	0.826	70.2	0.638	0.88	0.951	0.728	0.1825

Table 4. 21. Evaluation results of Mv-CNN architecture when sagittal projections of the T1Ce modality is used as an input

Fold	Evaluation Metrics						
	AUC	ACC	SEN	SPE	PRE	F1SC	loss
1	0.659	52.6	0.595	0.333	0.714	0.649	0.2317
2	0.914	78.9	0.762	0.867	0.941	0.842	0.1402
3	0.816	66.7	0.571	0.933	0.96	0.716	0.1996
4	0.867	80.7	0.881	0.6	0.86	0.871	0.1819
5	0.776	70.2	0.714	0.667	0.857	0.779	0.2138
Average	0.806	69.8	0.705	0.68	0.866	0.771	0.1934

In summary, regardless of the implemented MRI modality, and its corresponding 2D-projection images, we could find that the proposed Mv-CNN architecture using the 2D images sliced from a volumetric MRI as a training input has impressive glioma grade classification performances where the minimum, max, and average AUCs achieved across the twelve Mv-CNN models were 0.806, 0.983, and 0.892, respectively. It is also observed that for most of the twelve models the FSC can reach above 80%. These results indicate that, in clinical trials where only a single MRI modality is available, the Mv-CNN could be effectively used to stratify brain tumor patients into high, and low-grade gliomas preoperatively from the MRI images.

Moreover, we know that the higher the SPE, the lower the chance of misclassifying the LGG patients, and the higher the SEN, the lower the chance of misclassifying the HGG patients. Nevertheless, regardless of the class imbalance

on the dataset, we could find that the SEN, and SPE across the twelve models were quite good where the average SEN, and SPE were 0.794 and 0.861, respectively. In addition, we could find that the Mv-CNN model with the T1Ce axial slices as input stratifies the HGG, and LGG patients with the highest specificity and sensitivity values of 0.987, and 0.941, respectively, which has a very important clinical impact.

4.5.1. Comparison of Different Ensemble Learning Strategies

After performing the cross-validation experiments using the twelve Mv-CNN models, their respective probability predictions are combined in ten different ways yielding ten different prediction results. In this section, we analyzed and compared the results between the ten different combining strategies, including AVE, W.AVE, VOT, W.VOT, and stacking (KNN, LR, MLP, NB, RF, and SVM).

The stacking-based ensemble methods were implemented and optimized using the default settings of the Weka toolkit with modification of some of the parameters. For instance, while training the RF classifier the number of iterations parameter was changed to 10,000, for the KNN the distanceWeighting parameter was changed to weight by 1/distance, for the Naïve Bayes the useSupervisedDiscretization parameter was changed to True, for the logistic regression the useConjugateGradientDescent parameter was set True, for the MLP a learning rate of 0.001, a momentum value of 0.9, and epoch number 500 were used. However, the SVM classifier was optimized using the default settings of all the parameters. The unweighted voting (VOT), and unweighted average (AVE) ensemble methods were implemented using the Python software based on the formulas defined in Equation 12, and Equation 13, respectively. For the weighted voting (W.VOT), and weighted average (W.AVE) methods, we used double weights on the three Mv-CNN models which have the highest AUC in the single classifier experiment. Thus, because the Mv-CNN models with T1Ce axial, T1Ce

coronal, and T1Ce sagittal were the top three performers in the single classifier experiments, we set a weight value of 2 for each of them.

The overall classification performances of each of the twelve Mv-CNN models, and the ten ensemble methods with multiple evaluation metrics are presented in Table 4.22. The individual classification performances of each of the twelve Mv-CNN models were elaborated on in the previous section. According to the results, the RF method yielded the highest AUC (0.997), and SEN (0.986) values, and LR yielded an SEN of 0.952 with the least AUC value (0.961). However, from the classification results achieved by each of the ten ensemble methods, it is observed that seven of them achieved AUC, and ACC values of more than 99%, and 95%, respectively, while only four of the twelve Mv-CNN models could achieve an AUC, and ACC values of more than 90%. Whereas the SEN, SPE, PRE, and FSC values yielded by each of the ten ensemble methods range from 93.3% to 98.6%, 92% to 100%, 97.1% to 100%, and 97.1% to 98.4%, respectively. Therefore, the developed ensemble method can achieve impressive glioma grade predicting performance as compared with a single CNN-based system.

Table 4.22. The individual classification Performances of the proposed Mv-CNN models and for the ten different ensemble methods for the pre-operative glioma grading.

Models		Evaluation Metrics					
		AUC	**ACC**	**SEN**	**SPE**	**PRE**	**FSC**
Mv-CNN	FLAIR-axial	0.835	71.9	0.690	0.8	0.907	0.783
	FLAIR-coronal	0.859	78.25	0.767	0.827	0.927	0.836
	FLAIR-sagittal	0.888	77.2	0.748	0.84	0.932	0.822
	T1-axial	0.933	88	0.857	0.946	0.977	0.905
	T1-coronal	0.818	73.3	0.743	0.707	0.884	0.797
	T1-sagittal	0.846	73.3	0.691	0.853	0.935	0.783
	T1Ce-axial	0.979	**95.1**	**0.941**	**0.987**	**0.994**	**0.963**
	T1Ce-coronal	**0.983**	93.7	0.924	0.973	0.99	0.954
	T1Ce-sagittal	0.972	91.9	0.924	0.907	0.966	0.944
	T2-axial	0.959	90.2	0.89	0.933	0.973	0.923
	T2-coronal	0.826	70.18	0.638	0.88	0.951	0.728
	T2-sagittal	0.806	69.8	0.705	0.68	0.866	0.771
Ensemble Learning	AVE	0.992	95.1	0.933	**1.0**	**1.0**	0.963
	WAVE	0.996	95.4	0.938	**1.0**	**1.0**	0.968
	VOT	0.978	90.9	0.886	0.973	0.989	0.935
	WVOT	0.988	93.3	0.91	**1.0**	**1.0**	0.953
	NaiveBayes	0.99	95.8	0.952	0.973	0.99	0.971
	KNN	0.968	96.1	0.971	0.933	0.976	0.974
	LR	0.961	94.7	0.957	0.92	0.971	0.964
	MLP	0.995	97.2	0.981	0.947	0.981	0.981
	RF	**0.997**	**97.5**	**0.986**	0.947	0.981	**0.984**
	SVM	0.995	96.8	0.981	0.933	0.976	0.979

For the ten types of ensemble method strategies, the goal is to find the best combination of the twelve Mv-CNN models. Figure 4.8 illustrates the ROC curves, and AUC values yielded by each of the ten combining strategies. These curves were generated by plotting the true positive rates versus the false-positive rates at different threshold settings. The ROC curves for each ensemble strategy are plotted to the same coordinates to visually examine the cons, and pros of the ensemble method. As can be seen on the ROC curves, most of the implemented ensemble strategies achieve the highest accuracy as they are closest to the upper left of the corner of the curve. When comparing all the ensemble strategies, the difference is not large for most of the curves. However, it can be observed from the figure indicate that the RF classifier performs better than the other ensemble strategies.

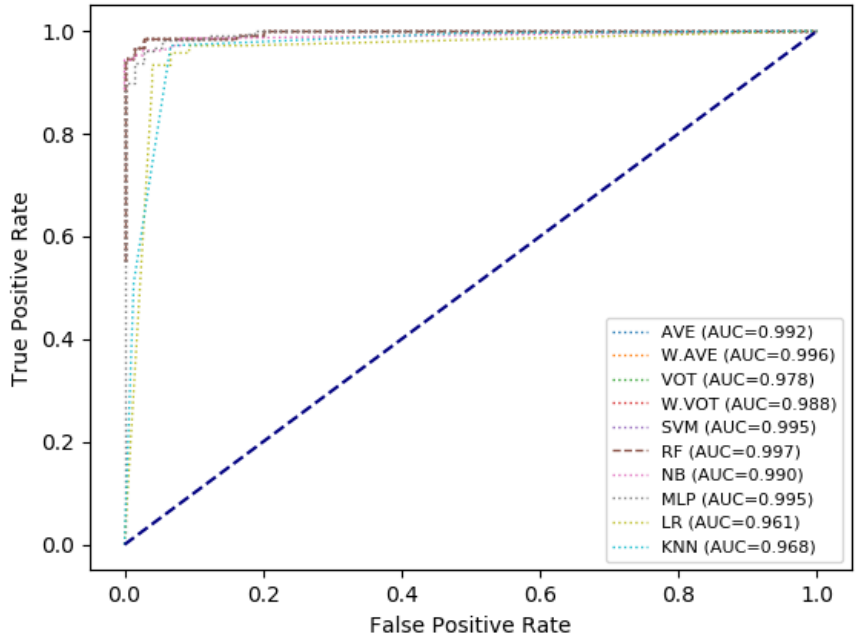

Figure 4.8. The ROC Curves yielded by the all ensemble strategies for the preoperative grading of glioma. The random forest (RF) has the highest AUC value (0.997) than the other ensembling strategies.

For a more detailed analysis, confusion matrices of the ten ensemble strategies are also presented in Table 4.23. The confusion matrices show the performances of all the ten ensemble strategies in terms of the number of glioma patients that have been correctly classified as well as misclassification cases for each class. One can see that the RF classifier was able to correctly classify 207 patients out of 210 HGG patients, yielding significantly the lowest false-negative rate as compared with the other ensemble methods. Meanwhile, the AVE, W.AVE, and W.VOT methods, interestingly, were able to classify all the LGG patients correctly, yielding significantly the lowest false positive rate. Thus, it is reasonable that the combination of multiple MRI modalities using an ensemble of multiple CNN models produces more features about the brain tumor; which leads to a more effective and accurate prediction of glioma grade pre-operatively solely from MRI scans.

Table 4.23. Confusion matrices of the ten ensembling strategies

	Predicted	
Actual	**HGG**	**LGG**
KNN		
HGG	204 (97.1%)	6 (2.9%)
LGG	5 (6.7%)	70 (93.3%)
LR		
HGG	201 (95.7%)	9 (4.3%)
LGG	6 (8%)	69 (92%)
MLP		
HGG	206 (98.1%)	4 (1.9%)
LGG	4 (5.3%)	71 (94.7%)
NB		
HGG	200 (95.2%)	10 (4.8%)
LGG	2 (2.7%)	73 (97.3%)
RF		
HGG	207 (98.6%)	3 (1.4%)
LGG	4 (5.3%)	71 (94.7%)

	Predicted	
Actual	**HGG**	**LGG**
SVM		
HGG	206 (98.1%)	4 (1.9%)
LGG	5 (6.7%)	70 (93.3%)
AVE		
HGG	196 (93.3%)	14 (6.7%)
LGG	0 (0%)	75(100%)
W.AVE		
HGG	197 (93.8)	13 (6.2%)
LGG	0 (0%)	75 (100%)
VOT		
HGG	186 (88.6%)	24 (11.4%)
LGG	2 (2.7%)	73 (97.3%)
W.VOT		
HGG	191 (91%)	19 (9%)
LGG	0 (0%)	75 (100%)

4.6. Survival Time Prediction

We split the dataset for the overall survival time (OS) classification task into training, and testing set, where the multimodal MRI scans of the 122 patients (75%) were used for training, and the remaining multimodal MRI scans of the 41 patients (25%) were used for testing. We used stratified random sampling for balancing the overall survival time distribution between the training and testing datasets, and assess the effectiveness of the prosed models in differentiating between long-term OS, and short-term OS. Detailed patients' information for the two groups can be found in Table 3.1. In the following sections, we present a more detailed result analysis for the classification of overall survival time using the Mv-CNN architecture, Multi-Column CNN architecture, and the ensemble learning strategies.

4.6.1. Performance of Mv-CNN and Multi-Column CNN Models

The same training process of the Mv-CNN architecture was applied to each of the twelve 2D-projection images of the multi-modal MRI images, resulting in a total of 12 different training operations, and results. The results of the twelve Mv-CNN models using multiple evaluation metrics are reported in Table 4.24.

Table 4.24. The individual classification performances of the Mv-CNN architecture for the OS binary classification

Mv-CNN	Evaluation Metrics					
	AUC	ACC	SEN	SPE	PRE	FSC
FLAIR-axial	**0.875**	73.8	0.654	0.875	0.895	0.756
FLAIR-coronal	0.844	**81**	0.731	**0.938**	**0.95**	**0.826**
FLAIR-sagittal	0.844	76.2	0.769	0.75	0.833	0.8
T1-axial	0.827	64.3	0.731	0.5	0.704	0.717
T1-coronal	0.755	64.3	0.654	0.625	0.739	0.694
T1-sagittal	0.817	76.2	**0.885**	0.562	0.767	0.821
T1Ce-axial	0.772	66.7	0.769	0.5	0.714	0.741
T1Ce-coronal	0.75	66.7	0.654	0.688	0.773	0.708
T1Ce-sagittal	0.829	73.8	0.692	0.812	0.857	0.766
T2-axial	0.702	61.9	0.615	0.625	0.727	0.667
T2-coronal	0.678	57.1	0.423	0.812	0.786	0.55
T2-sagittal	0.692	59.5	0.731	0.375	0.655	0.691

The AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the FLAIR axial images on the test dataset were 87.5%, 73.8%, 65.4%, 87.5%, 89.5%, and 75.6%, respectively. Similarly, the AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the FLAIR coronal images on the test dataset were 84.4%, 81%, 73.1%, 93.8%, 95%, and 82.6%, respectively, when using the FLAIR sagittal images were 84.4%, 76.2%, 76.9%, 75%, 83.3%, and 80%.

The AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the T1 axial images on the test dataset were 82.7%, 64.3%, 73.1%, 50%, 70.4%, and 71.7%, respectively. Similarly, the AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the T1 coronal images on the test dataset were 75.5%, 64.3%, 65.4%, 62.5%, 73.9%, and

69.4%, respectively, when using the T1 sagittal images were 81.7%, 76.2%, 88.5%, 56.2%, 76.7% and 82.1%, respectively.

The AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the T1Ce axial images on the test dataset were 77.2%, 66.7%, 76.9%, 50%, 71.4%, and 74.1%, respectively. Similarly, the AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the T1Ce coronal images on the test dataset were 75%, 66.7%, 65.4%, 68.8%, 77.3%, and 70.8%, respectively, when using the T1Ce sagittal images were 82.9%, 73.8%, 69.2%, 81.2%, 85.7%, and 76.6%, respectively.

The AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the T2 axial images on the test dataset were 70.2%, 61.9%, 61.5%, 62.5%, 72.7%, and 66.7%, respectively. Similarly, the AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics of the Mv-CNN model based on the T2 coronal images on the test dataset were 67.8%, 57.1%, 42.3%, 81.2%, 55%, and 70.8%, respectively, when using the T2 sagittal images were 69.2%, 59.5%, 73.1%, 37.5%, 65.5%, and 69.1%, respectively.

Regardless of the implemented MRI modality, and its corresponding 2D-projection slices, we could find that the proposed Mv-CNN architecture using the 2D images sliced from a volumetric MRI as a training input has good OS classification performances where the maximum, minimum, and average AUCs achieved across the twelve Mv-CNN models were 0.875, 0.678, and 0.782, respectively. We observed that half of the twelve models obtained an AUC value greater than 0.8 on the test dataset.

Moreover, we know that the higher the SPE, the lower the chance of misclassifying the long OS patients, and the higher the SEN, the lower the chance of misclassifying short OS patients. Thus, another observation we made is that, the SEN, and SPE across the twelve models were poor compared to the glioma grade prediction where the average SEN and SPE were 0.69, and 0.67, respectively. However, the average SEN, and SPE for the glioma grade classification using the

same architecture were 0.794, and 0.864, respectively. We assume that the reason for these poor results is due to the small number of patients in the OS classification training dataset, and the cases are also not balanced.

The OS stratification performances of each of the seven multicolumn Mv-CNN models have also been evaluated in terms of multiple metrics as shown in Table 4.25. Their AUC, ACC, SEN, SPE, PRE, and FSC range from 68.3% to 86.3%, 57.7% to 80.8%, 43.8% to 87.5%, 69% to 91.3%, 65.2% to 85.7% and 61.9% to 83.3%, respectively. We observed that there were less significant differences between the average performances of the twelve Mv-CNN models, and the average performances of the seven Multi-Column CNN models. However, the Multi-Column CNN models have increased the average SPE by 5.1% to that of Mv-CNN models. Similar to the Mv-CNN models, we assume that the reason for these poor results by the majority of the Multi-Column is due to the small number of patients in the OS classification training dataset, and the cases are also not balanced.

Table 4.25. The individual classification performances of the Multi-Column CNN architecture for the OS binary classification

Multi-Column CNN	Evaluation Metrics					
	AUC	ACC	SEN	SPE	PRE	FSC
axial	**0.82**	**83.3**	**0.808**	**0.875**	**0.913**	**0.857**
coronal	0.779	73.8	0.654	0.75	0.81	0.723
sagittal	0.69	64.3	0.654	0.625	0.739	0.694
FLAIR	**0.863**	78.6	0.769	0.812	0.87	0.816
T1	0.755	71.4	0.615	0.875	0.889	0.727
T1Ce	0.683	64.3	0.769	0.438	0.69	0.727
T2	0.714	61.9	0.577	0.688	0.75	0.652

4.6.2. Comparison of Different Ensemble Learning Strategies

After performing the experiments for the OS classification using the twelve Mv-CNN models, and seven Multi-Column CNN models, their respective probability predictions are combined in ten different ways yielding ten different

prediction results for both scenarios. In this section, we analyzed and compared the results between the ten different combining strategies, including AVE, W.AVE, VOT, W.VOT, and stacking (KNN, LR, MLP, NB, RF, and SVM).

The implementation of the stacking-based ensemble methods was implemented and optimized using the default settings of the Weka toolkit with modification of some of the parameters. For instance, while training the RF classifier the number of iterations parameter were changed to 1000, for the KNN the distanceWeighting parameter was changed to weight by 1/distance, for the Naïve Bayes the useSupervisedDiscretization parameter was changed to True, for the logistic regression the useConjugateGradientDescent parameter was set True, for the MLP a learning rate of 0.02, a momentum value of 0.3, and an epoch number of 500 were used. However, the SVM classifier was optimized using the default settings of all the parameters. The unweighted voting (VOT) and unweighted average (AVE) ensemble methods were implemented using the Python software based on the formulas defined in Equation 12, and Equation 13, respectively. For the weighted voting (W.VOT), and weighted average (W.AVE) methods, we used double weights on the three Mv-CNN, and Multi-Column CNN models which have the highest AUC in the single classifier experiments. Thus, because the Mv-CNN models with FLAIR axial, FLAIR coronal, and FLAIR sagittal were the top three performers in the single classifier experiments, we set a weight value of 2 for each of them. Similarly, the Multi-Column models with the FLAIR modality, the Triplanar-CNN with axial, and coronal slices were the top three performers in the single classifier experiments, thus we set a weight value of 2 for each of them.

The overall survival time classification results obtained when the Mv-CNN, and the Multi-Column CNN models were fused using the ten different ensemble learning strategies are presented in Table 4.26. In general, almost all the ensemble strategies improved the OS classification performance when compared with the average of the individual classification performances of the base learners.

Moreover, the ensemble of the Mv-CNN models worked significantly better than the ensemble of the Multi-Column CNN models on all the ensemble strategies with regard to almost all evaluation metrics. Another observation we made is that 70% of the ensemble strategies were achieved AUC, and ACC values of more than 90% and 85%, respectively, while none of the twelve Mv-CNN models could achieve these results.

Moreover, we observed all of the ensemble strategies based on the Mv-CNN base learners yielded quite impressive classification performances on every evaluation metrics, ranging from 88.2% to 93%, 81% to 92.9%, 76.9% to 92.3%, 75% to 93.8%, 85.7% to 96%, and 83.3% to 94.1% in AUC, ACC, SEN, SPE, PRE, and FSC evaluation metrics, respectively. However, The SVM-based ensemble strategy yielded the highest AUC (0.93), ACC (92.9%), and FSC (94.1%), and VOT and W.VOT yielded the least AUC (0.886), and ACC (81%), respectively. In addition, the SVM fusing strategy improved the AUC, ACC, and FSC OS classification performance of the base learners by 5%, 11.9%, and 11.5%, respectively. Therefore, the developed ensemble method can achieve impressive survival time classification performance as compared with a single CNN-based system, and could be considered as a promising strategy for any other medical classification problems with limited, and unbalanced volumetric medical image datasets.

Table 4.26. Evaluation measures of the proposed Mv-CNN and Multi-Column CNN models for survival time classification when fused using ten different ensemble learning strategies

	Ensemble Strategy	**Evaluation Metrics**					
		AUC	**ACC**	**SEN**	**SPE**	**PRE**	**FSC**
Ensemble of Mv-CNN models	AVE	0.925	85.7	0.808	0.938	0.955	0.875
	W.AVE	0.928	83.3	0.769	0.938	0.952	0.851
	VOT	0.886	83.3	0.846	0.812	0.88	0.863
	W.VOT	0.899	81	0.769	0.875	0.909	0.833
	KNN	0.913	88.1	0.885	0.875	0.92	0.882
	LR	0.882	88.1	**0.923**	0.813	0.889	0.880
	MLP	0.923	90.5	**0.923**	0.875	0.923	0.905
	NB	0.908	85.7	**0.923**	0.75	0.857	0.855
	RF	0.909	85.7	0.846	0.875	0.917	0.858
	SVM	**0.93**	**92.9**	**0.923**	**0.938**	**0.96**	**0.941**
Ensemble of Multi-Column CNN Models	AVE	0.837	76.2	0.731	**0.812**	**0.864**	0.792
	WAVE	0.856	76.2	0.731	**0.812**	**0.864**	0.792
	VOT	0.85	76.2	0.731	**0.812**	**0.864**	0.792
	WVOT	**0.87**	**78.6**	0.769	**0.812**	0.87	**0.816**
	KNN	0.861	73.8	0.625	0.808	0.667	0.645
	LR	0.809	69.1	**0.813**	0.615	0.565	0.667
	MLP	0.846	76.2	**0.813**	0.731	0.65	0.722
	NB	0.862	**78.6**	**0.813**	0.769	0.684	0.743
	RF	0.849	78.6	0.75	0.808	0.706	0.727
	SVM	0.832	81	0.938	0.731	0.682	0.789

5. DISCUSSION

Using a prognostic model for pre-operative prediction of glioma grade, and survival time is one of the most commonly used critical steps in brain tumor diagnosis as it can help to choose the appropriate treatment planning strategy, and MRI is usually the preferred noninvasive imaging technique in the pre-operative assessment of patients with a brain tumor. Hence, studies have been conducted in developing prognostic models for glioma grading, and survival time prediction tasks based on the handcrafted features extracted from MRI scans of patients. However, the conventional grading of glioma, and survival time predictions based on the classical machine learning methods that rely on handcrafted features is time-consuming, and subjective. Moreover, this approach may not fully capture the useful information from every cross-sectional MRI and subsequently could lead to an inaccurate prognostic model. In this work, we substantially advanced this body of knowledge by using classical machine learning, and deep learning techniques that are no longer limited to handcrafted features and do predictions directly from the images.

Deep learning, and classical machine learning techniques alone and altogether have been used in medical image processing. Appropriate application of these two techniques has the potential to revolutionize radiology and open up a new field in medical imaging. To accomplish this, the right combination of classical machine learning, and deep learning methods is very important. Moreover, the pretreatment and number of the input images affect the classification findings of deep learning models. Thus, throughout this study, to develop an effective end-to-end, powerful, and fully automated CAD system for a correct diagnosis of brain tumor diseases, we proposed and explored different image pre-processing techniques, novel CNN architectures, and ensemble learning strategies via classical machine learning techniques. Rather than training CNN models from

scratch using random initialization, we exploited the concept of transfer learning by leveraging the pre-trained weights learned from large image datasets (ImageNet). Moreover, we examined the proper tuning of optimization parameters, and loss functions that guarantee the finest network learning for better classification performances. Thus, it was possible to fulfill the initial goal of fully automating, and improving the accuracy of prognostic models from pre-operative multi-modal MRI scans to assist physicians in clinical diagnosis and treatment. Besides, considering the different approaches, and a large number of implemented experiments, it is possible to make some assumptions about the achieved results. Moreover, the main findings of our study are briefly discussed in the following sections.

5.1 Automated Glioma Grading

In this work, we developed and applied new approaches based on deep learning, and classical machine learning techniques to the problem of automated prediction of glioma grade from volumetric multi-modal MRI images. We found that our best model, which is based on ensemble multiple CNN models using random forest classifier, achieved impressive classification performances in multiple evaluation metrics (AUC = 99.7%, Accuracy = 97.5%, SEN = 98.6, SPE = 94.7%, PRE = 98.1%, and FSC = 98.4%) for distinguishing high-grade, and low-grade gliomas using a small dataset of less than 300 patients.

For the types of fifteen pre-trained ImageNet architectures that were customized and trained for glioma grade classification, we found that VGG-19 had the best performance and DenseNet-121 had the worst performance. We found that small ImageNet pre-trained architectures might perform better than deep ImageNet pre-trained architectures in small medical image classification problems. Moreover, we found that using state-of-the-art CNN models no statistically significant relationship was observed between glioma grading accuracy, and ImageNet accuracy. This finding extends the work by Kornblith et al. (2019) - which supports

the implicit assumption that CNN architectures that perform better on ImageNet necessarily perform better on typical image classification tasks - to the medical setting of glioma grade classification from MRI images. The contradiction in the finding could be attributed to the difference in the number of classes, and the unique characteristics medical images have. However, supporting our assumptions, the same authors showed that even among natural image datasets fine-tuning did not provide a substantial benefit over training from random initialization on small fine-grained classification datasets.

For the types of multi-modal MRI images that were used to train the deep learning model, the results indicate that brain images from different MRI modalities show different sensitivity to a brain tumor. Moreover, our findings are consistent with some previous reports that classify glioma using solely on single and multiple MRI modalities, in that using multiple modalities improves classification performance (Ali, Gu, & Jakola, 2019; Ye, Pu, Wang, Li, & Zha, 2017). However, in contrast to earlier findings where FLAIR modality was reported to be more superior to T1Ce images (Gurwara et al., 2010), and T2 images (Ali et al., 2019; Ge, Gu, et al., 2018) in providing a more complete characterization of glioma, our findings showed that T1 modality was more predictive than the other three modalities, and the individual classification performance of FLAIR modality was the lowest. Besides, based on our findings there is no significant difference between the individual classification performances of the T1, and T1Ce modalities. Moreover, our finding corroborates Ge et al. (2018), and Ali et al. (2019) where the individual glioma grading classification performance of T1Ce is more superior to that of the FLAIR, and T2 MRI modalities. However, in their experiment, they did not include the classification performance of the T1 modality. In summary, based on our findings we suggest that the T1, and T1Ce modalities should be the preferred image sequences than the others in clinical trials for glioma grading. Though in our experiment, the individual classification performance of the four modalities, in

multiple performance measurements, were consistent across different 2D architectures and sub-datasets, more investigation has to be done on comparison of the individual classification performance of various modalities for glioma grading.

In this study, we investigated ten fusion strategies for obtaining effective glioma grade classification results through ensembling the probability predictions of multiple CNN models. Our finding corroborates other recent works in medical image classification tasks using small, and unbalanced datasets, which shows ensemble models guarantee performance over individual CNN models if the right fusion strategy is used (Monkam et al., 2019; Mossa, Eriş, & Çevik, 2020; B. Zhang et al., 2019). Moreover, in terms of performance comparison among different ensembling strategies, it is noticed that random forest outperformed the other implemented fusion strategies in multiple evaluation metrics for glioma grading. However, it is the average, weighted average, and voting ensembling strategies that significantly reduce the false positive rate to zero for the glioma grade classification task.

To demonstrate that our proposed approach for the glioma grade classification is original and valuable, a comparison of our results with recently reported algorithms for the same glioma classification problem using the same or different datasets is summarized in Table 5.1. It is clear that the proposed ensemble learning framework based on the Mv-CNN architecture gives the best glioma grade classification result compared to other previous studies for the same problem which demonstrates the reliability of the proposed system. Besides, the proposed Mv-CNN method is a fully automated, segmentation-free approach as we train using the whole slide image in an end-to-end fashion to get the corresponding class directly. In contrast, authors in (Bi, Liu, & Cao, 2019; W. Chen, Liu, Peng, Sun, & Qiao, 2018; H.-H. Cho & Park, 2017; H. ho Cho, Lee, Kim, & Park, 2018; Hsieh, Lo, & Hsiao, 2017; Reza, Samad, Shboul, Jones, & Iftekharuddin, 2019; X. Zhang et al., 2017) used the hand-crafted feature approach that mainly relies on the extraction of highly discriminating features using the traditional feature extraction

mechanisms and then reduces their dimensions before applying a powerful classifier to distinguish between HGG and LGG accurately. Using this approach an accuracy that ranges from 88% to 94.5% was obtained by these different authors.

Using methods based on the deep learning approach, Banerjee et al. (2019) proposed a novel CNN architecture, and trained through scratch using three different types of inputs: slices, patches, and multiplanar volumetric slices of MRIs. In addition, the authors examined the pre-trained VGGNet, and ResNet models by fine-tuning the last few layers of the architectures. In distinguishing between LGG, and HGG from pre-operative multi-modal MRI images, the performance of the proposed CNN with volumetric slices achieved the highest accuracy (95%) compared to the other architectures. Moreover, the proposed CNN achieved better performance compared to the fine-tuned ResNet, and VGG pre-trained models. However, our experimental investigation showed that both ResNet, and VGG pre-trained models outperformed the results obtained by the authors' proposed CNN architecture. Moreover, in the work by Y. Yang et al. (2018b), with fine-tuning and transfer learning, better results were obtained compared with CNN trained from scratch to pre-operative glioma grading. This indicates that if transfer learning is properly utilized for initializing weights then it can outperform CNN models trained from scratch with random weight initialization. The authors implemented only T1Ce-axial slices, and similar data augmentation techniques used in our study. With five-fold cross-validation on a patient-level split, they obtained 90.9%, and 93.9% values in accuracy and AUC evaluation metrics, respectively. However, using similar MRI Modality (T1Ce-axial), and five-fold-validation on patient-level, we obtained a 95.1% ACC, and 97.9% AUC throughout our experiments indicating the effectiveness of our proposed Mv-CNN architecture.

Table 5.1. Results Comparison between Our Proposed Method and Recently Reported Algorithms for the Glioma Grade Classification. * *indicates authors used the same cohort dataset which used in this thesis.*

Study	Approaches	Dataset Source	MRI Modalities	# of Subjects	AUC	ACC (%)
Ye et al. (2017)	3D-CNN	BraTS'2015	T1,T2, and FLAIR	274	-	82.1
Hsieh et al. (2017)	Handcrafted Features	TCGA	T1Ce-axial	107	0.89	88
H.-H. Cho et al. (2017)	Handcrafted features	BraTS'2015	T1, T1Ce,T2 and FLAIR	108	0.887	89.8
Ge et al. (2017)	2D-CNN	BraTS'2017	T1,T2, and FLAIR	285	-	90.9
X. Zhang et al., (2017)	Handcrafted Features	Private	T1, T1Ce, DCE, FLAIR, DWI, and ASL	120	0.945	94.5
Cho et al. (2018)	Handcrafted Feature	BraTS'2017	T1, T1Ce,T2 and FLAIR	*285	0.921	88.8
Ge et al. (2018)	3D CNN	BraTS'2017	T1, T1Ce,T2 and FLAIR	*285	-	89.5
Y. Yang et al. (2018)	2D-CNNs	Private	T1Ce-axial	113	0.939	90.9
Khawaldeh et al. (2018)	2D-CNN	TCIA	FLAIR	130	-	91.2
Chen et al. (2018)	Handcrafted Features	BraTS'2015	T1, T1Ce,T2 and FLAIR	274	0.96	91.3
Pereira et al. (2018)	3D-CNN	BraTS'2017	T1, T1Ce,T2 and FLAIR	*285	0.984	93
Ge et al. (2018)	2D-CNN	BraTS'2017	T1, T2 and FLAIR	*285	-	90.9
Zhu et al. (2019)	2D-CNN	private	T1, T2 and T1Ce	181	0.82	-
Reza et al. (2019)	Handcrafted Features	BraTS'2017	T1, T1Ce,T2 and FLAIR	*285	0.88	88
Bi et al. (2019)	Handcrafted Features	BITE	T1, T2 and FLAIR	60	0.862	91
Ali et al. (2019)	2D	BraTS'2017	T1Ce, T2 and FLAIR	*285	-	92
Banerjee et al. (2019)	3D-CNN	BraTS'2017	same	*285	-	95
Mossa et al. (2020)	2D-CNN	BraTS'2017	T1, T1Ce,T2 and FLAIR	*285	0.985	95.8
Zhuge et al. (2020)	3D-CNN	TCGA and BraTS'2018	T1, T1Ce,T2 and FLAIR	315	-	97.1
Çinarer et al. (2020)	Handcrafted and DNN	TCIA	T2 and FLAIR	121	0.987	96.1
Proposed	Ensemble Learning	BraTS'2017	T1, T1Ce,T2 and FLAIR	*285	**0.997**	**97.5**

Another method based on the deep learning approach, Ge et al. (2017) proposed a novel Multi-stream 2D CNN architecture for pre-operative glioma grade classification on multi-modal MRI scans (T1-MRI, T2-MRI, and FLAIR). In the proposed architecture, generated features from each modality were aggregated before fed to the FC layer for classification, and achieved a 90.9% accuracy. Using similar Multi-stream CNN architecture based on autoencoder and data augmentation using GAN, Ali et al. (2019) achieved an accuracy of 92% for the classification of LGG, and HGG. However, our proposed Triplanar-CNN architecture, which is similar to the Multi-stream architecture, achieved an accuracy of 94.7% solely on the T1 MRI modality. Moreover, the proposed ensemble of multiple DL models has increased the accuracy by 6% over the Multi-stream CNN, complementing the effectiveness of our proposed approaches over recently developed algorithms for glioma grade classification.

Zhuge et al. (2020) proposed two novel and effective methods for automated, non-invasively distinguishing LGG, and HGG gliomas on conventional MRI images by using 2D R-CNN and 3D CNNs (3DConvNet) without manual segmentation of the tumor regions. While using the 2D R-CNN model, the slice with the largest area of the tumor is determined first, and then the mask R-CNN model is trained for tumor grading. In the 3DConvNet, a 3D volumetric CNNs is applied directly on bounding image regions of the segmented tumor for classification. In terms of accuracy, the results of the proposed 3D CNN-based method (97.1%) outperformed the 2D CNN method by a small margin (0.8%). This finding is not unexpected because the inputs to the 2D-CNN model were only slices from the three modalities combined, and set as RGB channel. All the T1 slices and all the slices in the other three modalities not having the largest tumor area were discarded. We assume this decreases the performance of the 2D CNN model. The 3DConvNet has nearly the same accuracy as our proposed approach based on ensemble learning. However, unlike that of the 3DConNet, our proposed approach was based on training 2D-CNN in an end-to-end fashion without the need

for segmenting the tumor regions, which reduces training time and GPU memory requirement. Therefore, it is possible to hypothesize that if proper 2D-CNN architectures are designed and trained with all the MRI slices we can get an accuracy that can outperform 3D-CNN models in distinguishing between HGG, and LGG.

In summary, our method is superior to the other existing algorithms in multiple evaluation metrics including AUC, and accuracy. However, direct comparisons between our study, and some previous studies based on private data are not possible without having access to MRI data of the patient cohort. Nevertheless, we show the state-of-the-art studies on glioma grading that have used pre-operative MRI modalities. Overall, the results reported in this paper are seemed to be quite impressive and innovative, and can be considered as the state-of-the-art method for stratifying glioma grades into HGG, and LGG.

We believe that the substantial improvement in glioma grade classification performance yield by our model compared to the other published works in the literature is attributed to the following four-fold unique features implemented in our proposed framework. First, the virtual increment of images per patient from four volumetric MRI scans to twelve 2D images through projection in the axial, coronal, and sagittal planes, which allowed us to leverage state-of-the-art CNN models trained on the ImageNet dataset. This conclusion is consistent with the previous studies for other small, and unbalanced medical image classification tasks (Gupta et al., 2020; Mormont, Geurts, & Maree, 2018).

Slices in the axial plane have less noise, and higher resolution in comparison to slices in the coronal and sagittal planes. Moreover, the majority of the tumor regions are obtained in the axial slices. As a result, it can be seen from studies in the literature that using axial slices only for tumor detection and classification is a common practice (Khawaldeh et al., 2018; Mormont, Geurts, & Maree, 2018). Therefore, concatenation of two consecutive coronal slices, and two consecutive sagittal slices before fed to the deep learning models could also be our

second potential reason for the impressive glioma grade classification performance. To prove our hypothesis, we examined the performance of CNN models using the two kinds of inputs (concatenation, and enlarging). Throughout this work, we have shown that concatenated slices as input to the model yield greatly improved AUC, ACC, and other evaluation metrics than enlarging the coronal, and sagittal slices to fit the fixed input size of CNN models. A similar concatenation strategy where each coronal, and sagittal slices at the same position were concatenated before fed to the deep learning model was also used in our previous work for Tuberculosis diagnosis, which achieved promising results (Mossa et al., 2020; Mossa, Yibre, & Çevik, 2019).

Third, the proposed Mv-CNN architecture could also be one of the other potential reasons for the impressive glioma grade classification performance. Since volumetric medical image datasets are always small and unbalanced, we needed to design a proper CNN architecture for the small and unbalanced data. The proposed Mv-CNN architecture can achieve a relatively good performance in such kind of unbalanced and small medical datasets. This conclusion is consistent with the previous studies for 3D object recognition, and medical image analysis tasks. However, until now the concept of Mv-CNN architecture has not been used for glioma grading. Fourth, we also assume the proposed and implemented ensemble learning strategies could also be the other potential reason. This conclusion is also consistent with the previous studies for other small and unbalanced medical image classification problems (Mormont et al., 2018; Mossa et al., 2020; Winzeck et al., 2019; Xiao et al., 2018; B. Zhang et al., 2019).

5.2. Automated Survival Time Prediction

Automated computerized image analysis can gain far more experience in a significantly shorter amount of time than human subjects can acquire in their lifetime. Differentiating between short-term and long-term overall survival time is a critical step in the management of brain tumor patients. In this study, we

developed and applied new approaches based on deep learning and classical machine learning techniques to the problem of automated classification of pre-operative survival time from volumetric multi-modal MRI images. We found that our best model, which is based on ensembling multiple CNN models using an SVM classifier, achieved state-of-the-art results in multiple evaluation metrics (AUC = 93%, Accuracy = 92.9%, SEN = 92.3%, SPE = 93.8%, PRE = 96%, and FSC = 94.1%) in distinguishing survival time between long-term and short-term overall survival time of brain tumor patients using a small pre-operative MRI dataset of 163 patients.

In order to demonstrate our proposed approach for the pre-operative survival time classification task is original and valuable, a comparison of our results with recently reported algorithms for the same survival time classification problem on the same or different datasets is discussed in the following paragraphs. To further validate the effectiveness of the proposed approach, we also presented a reference baseline (discussed in Section 3.4) based on limited demographic, and tumor volume-related features, and trained with an SVM classifier. The baseline approach achieved a performance of 0.65, and 60% in AUC and accuracy, respectively. It is clear that the proposed ensemble learning framework based on the Mv-CNN architecture gives the best classification performance when it is compared to the baseline approach and other previous studies for the same problem, which demonstrates the reliability and validity of the proposed system (Mossa & ÇEVİK, 2021).

Deep learning algorithms work well often when large annotated datasets are available. As a result, the handcrafted-based approaches have been used in the literature more often than deep learning algorithms when datasets sizes are limited. Similarly, several studies have considered this handcrafted based approach for predicting the survival of glioma patients from MRI images. However, in contrast to our proposed approach, the majority of these handcrafted based approaches are not fully automatic or mainly relies on the extraction of large amount quantitative

imaging features using the traditional feature extraction mechanisms, and then reduces their dimensions before applying a powerful classifier to the survival time classification, which is time-consuming, and may lead to false positives.

For instance, Shboul et al. (2019) extracted a total of 31,000 texture, histogram, and volume related features among which only 74 features were selected, and subsequently combined with the age of patients to train the XGBoost classifier. Using the same public dataset that we used for this study, they obtained an accuracy of 73% in stratifying survival into three groups. Using the same dataset, Chato et al. (2018) achieved an accuracy of 66.7% using histogram features and SVM classifier, and Sanghani et al. (2018) achieved a 3-class classification accuracy of 87% on stratified 5-fold cross-validation using clinical, volumetric, tumor shape, and texture features trained with ML algorithm.

Likewise, authors extracted gene expression, GLCM, and JIM based imaging texture, and intensity features from manually segmented tumor region on T1CE and FLAIR MRI modalities of 73 patients, and combined into a single radiogenomic feature to train RF classifier. Their method achieved a leave-one-out-cross-validation AUC of 0.776 (Chaddad, Daniel, Desrosiers, Toews, & Abdulkarim, 2019). Similarly, X. Chen et al. (2019) showed that extracted handcrafted features, such as intensity, shape, texture, and wavelet, from manually segmented tumor regions in pre-surgical axial T1Ce modality, and subsequently combined with clinical data allowed stratifying patients' survival into a low- or high-risk group with an AUC of 0.851.

In another work (Macyszyn et al., 2016), using clinical information (age, and gender) and more than 60 features derived from computer-based segmented tumor regions, authors achieved an overall accuracy of 80% in classifying patients into short, medium, and long survival. In (Liu et al., 2018), using SVM classifier and 2D texture features extracted from slices with the largest tumor size that are manually segmented by two experienced radiologists, authors compared the performance of four MRI modalities when used individually, and in combination

for classifying survival into two groups. The result showed that when using only T1Ce, and the four MRI modalities separately, both models achieved nearly equal accuracy, and AUC value of 80.7%, and 0.79, respectively. Likewise, Upadhaya et al. (2015) achieved a classification accuracy of 90% based on heterogeneity textural features extracted from T1Ce, and T1 MRI modalities. However, considering only one MRI modality at a time, the best classification accuracy (82.5%) was obtained using T1Ce, followed by FLAIR. In addition, the experimental result indicated that considering only a single MRI modality at a time, the survival classification performance using only the T1Ce modality is comparable to that of using all the four MRI modalities together. However, inconsistent with the other study (Nie, Zhang, Adeli, Liu, & Shen, 2016), our proposed approach shows that the performances combined from multi-modal MRI were superior to using T1Ce alone. In addition, our experimental results show that considering one MRI sequence at a time, the best stratification performance was obtained using FLAIR MRI modality. Another report supporting our finding also stated that FLAIR sequences were reported to be superior to post-contrast T1Ce images postoperatively in providing more accurate information about tumor volume, and treatment response (Gurwara et al., 2010). However, a further thorough investigation is needed on the discriminative performances of MRI modalities in glioma patients' survival time prediction.

So far, the investigation of fully automated survival time prediction based on deep learning techniques is very limited. In fact, the performances of these limited works are also unsatisfactory. For instance, Chato & Latifi (2017) used pre-trained AlexNet for extracting deep features from specific slices that clearly contained the three regions of HGG glioma, a 91% accuracy was obtained for the Linear Discriminant classifier, and 86.4% for the linear SVM classifier. In other recently published work, Nie et al. (2019) proposed a multi-channel 3D CNN architecture that extracts deep features from multi-modal MRI scans, and subsequently combined deep features, demographic and tumor-related features to

train SVM that classified OS into two groups with an accuracy of 90.7%, and 90.5% on 3-fold and, 10-fold cross-validation, respectively. The paper also justified the importance of using multi-modal MRI in predicting OS, and rs-fMRI resulted in the best classification performance among all the single modalities.

In summary, the performance of our proposed ensemble learning approach regarding classification accuracy, and AUC is superior to several-state-of-the-art results in brain tumor patients' OS stratification. However, direct comparisons between our study, and some previous studies based on private data are not possible without having access to MRI data of the patient cohort. Nevertheless, we show the state-of-the-art studies on survival time classification that have used pre-operative MRI modalities. Overall, the results reported in this paper seem to be quite impressive and innovative, and can be considered as the state-of-the-art method for stratifying survival time into short-term and long-term.

We believe that the substantial improvement in the survival time classification task yield by our model compared to the other published works in the literature is attributed to the following five-fold unique features implemented in our proposed framework. First, the virtual increment of images per patient from four volumetric MRI scans to twelve 2D images through projection in the axial, coronal, and sagittal planes, which allowed us to leverage state-of-the-art CNN models trained on the ImageNet dataset. This conclusion is consistent with the previous studies for other small and unbalanced medical image classification tasks (Gupta et al., 2020; Mormont, Geurts, & Maree, 2018).

Slices in the axial plane have less noise, and higher resolution in comparison to slices in the coronal, and sagittal planes. Moreover, the majority of the tumor regions are obtained in the axial slices. As a result, it can be seen from studies in the literature that using axial slices only for tumor detection, and classification is a common practice (Khawaldeh et al., 2018; Mormont, Geurts, & Maree, 2018). Therefore, concatenation of two consecutive coronal slices, and two consecutive sagittal slices before fed to the deep learning models could also be the

second potential reason for the impressive classification performances of the survival prediction task. To prove our hypothesis, we examined the performance of CNN models using the two kinds of inputs (concatenation, and enlarging). A similar concatenation strategy where each coronal, and sagittal slices at the same position were concatenated before fed to the deep learning model was also used in our previous work for Tuberculosis diagnosis, which achieved promising results (Mossa et al., 2020; Mossa, Yibre, & Çevik, 2019).

In addition, the proposed Mv-CNN architecture could also be one of the other potential reasons for the impressive classification performance for the survival prediction task. Since volumetric medical image datasets are always small and unbalanced, we needed to design a proper CNN architecture that the small and unbalanced data. The proposed Mv-CNN architecture can achieve a relatively good performance in such kind of unbalanced and small medical datasets. This conclusion is consistent with the previous studies for 3D object recognition, and medical image analysis tasks. However, until now the concept of Mv-CNN architecture has not been used for the brain tumor patients' survival prediction.

The above-mentioned deep learning models in the literature for survival time prediction are based on manually or automatically segmented tumor regions as input to the CNN. However, including surrounding tumor regions provides more information on the tumor-surrounding tissues, and therefore also the location of the tumor (W. Yang et al., 2012). In this work, we have also investigated the difference in the performance of the Mv-CNN architecture when trained with different inputs such as the tumor segmented region only (ROI), the extended tumor segmented regions that include the surrounding tumor tissues alone (EROI), and the whole slide image alone (WSI). As shown by our experiments, we found the inclusion of tumor surrounding tissue to the input of a CNN yields superior performance over the ROI region or the WSI image. Moreover, the benefit of including surrounding tissue (through disk-shaped image dilation) based on handcrafted imaging features is shown by other publications for brain tumor

classification (Cheng et al., 2015). Therefore, we believe the inclusion of context information on surrounding brain tumor tissue, and location effects are also one of the main potential reasons for the increased performance over the other results as the location is an important distinguishing marker. Finally, we assume the proposed and implemented ensemble learning strategies could also be the other potential reason. This conclusion is also consistent with the previous studies for other small and unbalanced medical image classification problems (Mormont et al., 2018; Mossa et al., 2020; Winzeck et al., 2019; Xiao et al., 2018; B. Zhang et al., 2019).

6. CONCLUSION AND FUTURE WORKS

6.1. Conclusion

Using a prognostic model for pre-operative prediction of glioma grade, and survival time is one of the most commonly used critical steps in the brain tumor diagnosis field as it can help to choose the appropriate treatment planning strategy, and MRI is usually the preferred noninvasive imaging technique in the pre-operative assessment of patients with a brain tumor. In this thesis, we proposed to develop a machine learning, and deep learning-based fully automated prognostic modes cable of predicting glioma grade, and survival time of patients with brain tumors on pre-operative multi-modal MRI images (FLAIR, T1Ce, T1, and T2).

The proposed system was developed in three main stages: pre-processing, deep learning development, and ensembling model outputs. At each stage, different strategies were proposed, and the strategies used varied in terms of preprocessing techniques, CNN architectures, pre-trained state-of-the-art deep learning models, inputs to the deep learning model (concatenated vs enlarged image, and ROIs vs Extended-ROI vs whole slide image), and ensemble strategies (averaging, voting, and classical machine learning methods). The final goal was to get a high classification accuracy when evaluating the developed prognostic model, which assesses the patient's degree of glioma grade (LGG vs HGG), and overall survival time (short-term vs long-term) before a brain tumor surgery.

In the preprocessing stage, each volumetric MRI image was reconstructed to a stack of 2D slices in the axial, coronal, and sagittal planes, which virtually increased images per patient from four to twelve. Subsequently, some slices obtained at the beginning, and end of each 2D projection, and which did not contain any brain tissue intensity value were also discarded to avoid processing slices that do not add useful information. Moreover, another novel pre-processing technique, which is a concatenation of two images, was used to leverage pre-trained deep learning models trained on natural images without enlarging them to

benefit from the fine-grained details in each slice of the coronal, and sagittal projections. The individual classification performance of the four MRI modalities both for glioma grading and survival time predictions were compared. The experimental results show that T1Ce coronal slices, and FLAIR axial slices have more discriminative deep features for glioma grading, and survival time prediction, respectively. For the survival time prediction, the effecting of using segmented glioma region (ROIs), determined by a deep learning-based segmentation model (Wang et al., 2018), or extended glioma region (EROIs) or whole slide image (without segmentation) as input while training deep learning model were analyzed. The conclusion is that incorporating the surrounding tumor tissues while training the deep learning models for survival time prediction improves the model performance.

When training datasets are scarce, deep learning techniques knowns as transfer learning, and data augmentation have been used in the literature. In the deep learning development stage of this work, since the datasets that we used are small cross-sectional volumetric datasets, proper CNN architectures that suit the data were designed using the feature extractor layers of the state-of-the-art pre-trained ImageNet CNN architectures as a backend. As a result, inspired by the first Multi-View CNN architecture developed for 3D objection recognition (Su et al., 2015), three novel CNN architectures (Mv-CNN, Triplanar-CNN, and Mc-CNN) were proposed in this thesis. In the meanwhile, 15 ImageNet CNN models were examined, and found VGG-19 performs the best and DenseNet-121 performs the worst on the glioma grading task. This thesis has also investigated if pre-trained ImageNet architectures performance could relate with glioma grading performance and found that no statistically significant relationship between glioma grading accuracy, and ImageNet accuracy (ρ= -0.237), and between the number of training parameters and the glioma grading classification's performance (ρ = -0.174) on small medical image datasets.

In the ensembling stage, ten different fusion strategies combining the prediction of twelve Mv-CNN models through averaging, voting, and classical machine learning methods were implemented both for glioma grading, and survival time prediction. In terms of fusion strategies, RF, and SVM achieved an excellent performance than the others for glioma grading and survival time prediction, respectively. Our final prognostic models implemented on the BraTS'17 dataset significantly outperforms state-of-the-art methods achieving an ACC of 97.5%, AUC of 0.997, SEN of 0.986, SPE of 0.947, and FSC of 0.984 for glioma grade classification, and an ACC of 92.9%, AUC of 0.93, SEN of 0.923, SPE of 0.938, and FSC of 0.941 for survival time classification.

6.2. Future works

The work in this thesis is a first step in dealing with designing a fully automated prognostic model, which is one of the next frontiers to be reached by new deep learning techniques in the medical scenario. The results obtained both for glioma grading, and survival time prediction are impressive. However, the works done in this thesis can be advanced with the following contributions: First of all, increasing the number of classes that can be classified by including more sub-OS, and sub-grades of brain tumors will make our work clinically more significant. Second, there were only very few structural MRI scans of brain tumor patients studied in this thesis. Besides, structural MRIs do not provide information about the biological details of brain tumors. Therefore, we believe increasing the dataset and combining it with functional MRI scans, digital pathology images, demographics, and genomic biomarkers of brain tumor patients may improve the performance of the proposed prognostic models.

Third, MRI images are susceptible to noise and do not have a tissue-specific pixel value even within the same modality and subject. Thus, although we applied the simple min-max method to normalize the pixel intensity values, applying various advanced image preprocessing techniques, such as inhomogeneity

correction, and noise removal methods, to the MRI images before fed to the deep learning models may enhance our work. Fourth, for the OS classification task, the clinical data can be combined with the deep features. The combination of clinical data and deep features might give good results for survival time classification in brain tumor patients.

Fifth, one of the critics of deep learning algorithms, especially by non-experts, is for being a black-box approach that does not reveal intuitions into their prediction mechanisms. Moreover, in medical applications, such as the one we addressed in this thesis, it is essential for deep learning models to provide explainable outputs to patients and physicians in order to build trust in the predicted output. Thus, in order to understand and explain what our developed model predicts, our works can be enhanced through the design of an explainable CNN model by adding a class activation maps Grad-CAMs (Selvaraju et al., 2017) or attention modules (Gu et al., 2018) into the proposed framework.

Sixth, the ensembling approach we used was based on homogenous CNN architecture trained with different inputs. We believe that using heterogeneous CNN architecture may enhance the overall classification performance. Last but not the least, the approaches we used in this work are powerful, and can be extended to other classification tasks in volumetric medical images without significantly changing the methodologies.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*.

Abdi, A. H., Luong, C., Tsang, T., Allan, G., Nouranian, S., Jue, J., … Abolmaesumi, P. (2017). Automatic Quality Assessment of Echocardiograms Using Convolutional Neural Networks: Feasibility on the Apical Four-Chamber View. *IEEE Transactions on Medical Imaging*, *36*(6), 1221–1230. https://doi.org/10.1109/TMI.2017.2690836

Adeberg, S., Bostel, T., König, L., Welzel, T., Debus, J., & Combs, S. E. (2014). A comparison of long-term survivors and short-term survivors with glioblastoma, subventricular zone involvement: A predictive factor for survival? *Radiation Oncology*, *9*(1), 95. https://doi.org/10.1186/1748-717X-9-95

Al-Saffar, A. A. M., Tao, H., & Talab, M. A. (2017). Review of deep convolution neural network in image classification. *Proceeding - 2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications, ICRAMET 2017*, *2018-January*, 26–31. https://doi.org/10.1109/ICRAMET.2017.8253139

Ali, M. B., Gu, I. Y. H., & Jakola, A. S. (2019). Multi-stream Convolutional Autoencoder and 2D Generative Adversarial Network for Glioma Classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11678 LNCS*, 234–245. https://doi.org/10.1007/978-3-030-29888-3_19

Alifieris, C., & Trafalis, D. T. (2015). Glioblastoma multiforme: Pathogenesis and treatment. *Pharmacology and Therapeutics*, *152*, 63–82. https://doi.org/10.1016/j.pharmthera.2015.05.005

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*. https://doi.org/10.1007/978-3-030-22475-2_1

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., … Asari, V. K. (2019). A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics*, *8*(3), 292. https://doi.org/10.3390/electronics8030292

Altwairgi, A. K., Raja, S., Manzoor, M., Aldandan, S., Alsaeed, E., Balbaid, A., … Alsharm, A. A. (2017). Management and treatment recommendations for World Health Organization Grade III and IV gliomas. *International Journal of Health Sciences*, *11*(3), 54–62. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/28936153

Amisha, Malik, P., Pathania, M., & Rathaur, V. (2019). Overview of artificial intelligence in medicine. *Journal of Family Medicine and Primary Care*, *8*(7), 2328. https://doi.org/10.4103/jfmpc.jfmpc_440_19

Armato, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., … Clarke, L. P. (2011). The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, *38*(2), 915–931. https://doi.org/10.1118/1.3528204

ASCO, 2020. American Society of Clinical Oncology. Brain Tumor: Statistics | Cancer.Net. Retrieved September 17, 2020, from https://www.cancer.net/cancer-types/brain-tumor/statistics

Babaian, R. J., Fritsche, H. A., Zhang, Z., Zhang, K. H., Madyastha, K. R., & Barnhill, S. D. (1998). Evaluation of ProstAsure index in the detection of prostate cancer: A preliminary report. *Urology*, *51*(1), 132–136. https://doi.org/10.1016/S0090-4295(97)00574-8

Babatunde, O. A., Olarewaju, S. O., Adeomi, A. A., Akande, J. O., Bashorun, A., Umeokonkwo, C. D., & Bamidele, J. O. (2020). 10-year risk for cardiovascular diseases using WHO prediction chart: Findings from the civil servants in South-western Nigeria. *BMC Cardiovascular Disorders*, *20*(1), 154. https://doi.org/10.1186/s12872-020-01438-9

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., … Davatzikos, C. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, *4*, 170117. https://doi.org/10.1038/sdata.2017.117

Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers and Security*, *90*, 101710. https://doi.org/10.1016/j.cose.2019.101710

Ballas, N., Yao, L., Pal, C., & Courville, A. (2016). Delving deeper into convolutional networks for learning video representations. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

Banerjee, S., Mitra, S., Masulli, F., & Rovetta, S. (2019). Deep radiomics for brain tumor detection and classification from multi-sequence mri. *ArXiv*.

Barıs¸kayalıbay, B. B., Jensen, G., & Van Der Smagt, P. (2017). CNN-based Segmentation of Medical Imaging Data. *ArXiv*. Retrieved from https://github.com/BRML/

Bauer, S., Wiest, R., Nolte, L. P., & Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumor studies. *Physics in Medicine and Biology*, *58*(13), R97. https://doi.org/10.1088/0031-9155/58/13/R97

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. https://doi.org/10.1109/72.279181

Benner, J. S., Erhardt, L., Flammer, M., Moller, R. A., Rajicic, N., Changela, K., … Girerd, X. (2008). A novel programme to evaluate and communicate 10-year risk of CHD reduces predicted risk and improves patients' modifiable risk factor profile. *International Journal of Clinical Practice*, *62*(10), 1484–1498. https://doi.org/10.1111/j.1742-1241.2008.01872.x

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., … Bengio, Y. (2010). Theano: A CPU and GPU Math Compiler in Python. *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/majora-92bf1922-003

Bi, X., Liu, J. G., & Cao, Y. S. (2019). Classification of low-grade and high-grade glioma using multiparametric radiomics model. *Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2019*, 574–577. https://doi.org/10.1109/ITNEC.2019.8729245

Bien, N., Rajpurkar, P., Ball, R. L., Irvin, J., Park, A., Jones, E., … Lungren, M. P. (2018). Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Medicine*, *15*(11). https://doi.org/10.1371/journal.pmed.1002699

Bleeker, F. E., Molenaar, R. J., & Leenstra, S. (2012). Recent advances in the molecular understanding of glioblastoma. *Journal of Neuro-Oncology*, *108*(1), 11–27. https://doi.org/10.1007/s11060-011-0793-0

Bottaci, L., Drew, P. J., Hartley, J. E., Hadfield, M. B., Farouk, R., Lee, P. W. R., … Monson, J. R. T. (1997). Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *Lancet*, *350*(9076), 469–472. https://doi.org/10.1016/S0140-6736(96)11196-X

Bounds, D. G., Lloyd, P. J., & Mathew, B. G. (1990). A comparison of neural network and other pattern recognition approaches to the diagnosis of low back disorders. *Neural Networks*, *3*(5), 583–591. https://doi.org/10.1016/0893-6080(90)90008-9

Bradl, M., & Lassmann, H. (2010). Oligodendrocytes: Biology and pathology. *Acta Neuropathologica*, Vol. 119, pp. 37–53. https://doi.org/10.1007/s00401-009-0601-5

Brady, A. P. (2017). Error and discrepancy in radiology: inevitable or avoidable? *Insights into Imaging*, *8*(1), 171–182. https://doi.org/10.1007/s13244-016-0534-1

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brock, A., Lim, T., Ritchie, J. M., & Weston, N. (2016). *Generative and Discriminative Voxel Modeling with Convolutional Neural Networks*. Retrieved from https://www.youtube.com/watch?v=LtpU1yBStlU

Brown, M. A., & Semelka, R. C. (2011). MRI: Basic Principles and Applications. In *Wiley-Liss*. Retrieved from https://books.google.com.tr/books?hl=en&lr=&id=oYOIHi3YkuMC&oi=fnd&pg=PR9&ots=ZoElAzTjCT&sig=FQIJm2_sO0xGjetucnyUD4QvWag&redir_esc=y#v=onepage&q&f=false

Buckner, J. C., Brown, P. D., O'Neill, B. P., Meyer, F. B., Wetmore, C. J., & Uhm, J. H. (2007). Central nervous system tumors. *Mayo Clinic Proceedings*, *82*(10), 1271–1286. https://doi.org/10.4065/82.10.1271

Callaghan, B. C., Kerber, K. A., Pace, R. J., Skolarus, L. E., & Burke, J. F. (2014). Headaches and neuroimaging: High utilization and costs despite guidelines. *JAMA Internal Medicine*, Vol. 174, pp. 819–821. https://doi.org/10.1001/jamainternmed.2014.173

Cancer Research UK, 2017. *Cancer Research UK: Cancer mortality statistics. Retrieved from https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk*

Carter, R., Aldridge, S., Page, M., & Parker, S. (2019). The human brain book: An illustrated guide to its structure, function, and disorders. In *Penguin*. Penguin.

Chaddad, A., Daniel, P., Desrosiers, C., Toews, M., & Abdulkarim, B. (2019). Novel Radiomic Features Based on Joint Intensity Matrices for Predicting Glioblastoma Patient Survival Time. *IEEE Journal of Biomedical and Health Informatics*, *23*(2), 795–804. https://doi.org/10.1109/JBHI.2018.2825027

Chang, Y., Park, H., Yang, H. J., Lee, S., Lee, K. Y., Kim, T. S., … Shin, J. M. (2018). Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. *Scientific Reports*, *8*(1), 8857. https://doi.org/10.1038/s41598-018-27214-6

Chato, L., Chow, E., & Latifi, S. (2018). Wavelet transform to improve accuracy of a prediction model for overall survival time of brain tumor patients based on MRI Images. *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, 441–442. https://doi.org/10.1109/ICHI.2018.00091

Chato, L., & Latifi, S. (2017). Machine Learning and Deep Learning Techniques to Predict Overall Survival of Brain Tumor Patients using MRI Images. *Proceedings - 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering, BIBE 2017*, *2018-Janua*, 9–14. https://doi.org/10.1109/BIBE.2017.00-86

Chen, S., Zhong, X., Hu, S., Dorn, S., Kachelrieß, M., Lell, M., & Maier, A. (2020). Automatic multi-organ segmentation in dual-energy CT (DECT) with dedicated 3D fully convolutional DECT networks. *Medical Physics*, *47*(2), 552–562. https://doi.org/10.1002/mp.13950

Chen, W., Liu, B., Peng, S., Sun, J., & Qiao, X. (2018). Computer-Aided Grading of Gliomas Combining Automatic Segmentation and Radiomics. *International Journal of Biomedical Imaging*, *2018*. https://doi.org/10.1155/2018/2512037

Chen, W., Liu, B., Peng, S., Sun, J., & Qiao, X. (2019). S3D-UNET: Separable 3D U-Net for brain tumor segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11384 LNCS*, 358–368. https://doi.org/10.1007/978-3-030-11726-9_32

Chen, X., Fang, M., Dong, D., Liu, L., Xu, X., Wei, X., … Liu, Z. (2019). Development and Validation of a MRI-Based Radiomics Prognostic Classifier in Patients with Primary Glioblastoma Multiforme. *Academic Radiology*, *26*(10), 1292–1300. https://doi.org/10.1016/j.acra.2018.12.016

Chen, Y., Tian, Y., & He, M. (2020). Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, *192*, 102897. https://doi.org/10.1016/j.cviu.2019.102897

Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., … Feng, Q. (2015). Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS ONE*, *10*(10), e0140381. https://doi.org/10.1371/journal.pone.0140381

Cho, H.-H., & Park, H. (2017). Classification of low-grade and high-grade glioma using multi-modal image radiomics features. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 3081–3084. https://doi.org/10.1109/EMBC.2017.8037508

Cho, H. ho, Lee, S. hak, Kim, J., & Park, H. (2018). Classification of the glioma grading using radiomics analysis. *PeerJ*, *2018*(11). https://doi.org/10.7717/peerj.5982

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1724–1734. https://doi.org/10.3115/v1/d14-1179

Choi, Y., Ahn, K. J., Nam, Y., Jang, J., Shin, N. Y., Choi, H. S., … Kim, B. soo. (2019). Analysis of heterogeneity of peritumoral T2 hyperintensity in patients with pretreatment glioblastoma: Prognostic value of MRI-based radiomics. *European Journal of Radiology*, *120*. https://doi.org/10.1016/j.ejrad.2019.108642

Chollet, F. (2015). Keras: The Python Deep Learning library. *Keras.Io*.

Christakis, N. A., & Lamont, E. B. (2000). Extent and determinants of error in doctors' prognoses in terminally ill patients: Prospective cohort study. *British Medical Journal*, *320*(7233), 469–472. https://doi.org/10.1136/bmj.320.7233.469

Çinarer, G., Gürsel, B., & Haşim, A. (2020). Prediction of glioma grades using deep learning with wavelet radiomic features. *Applied Sciences (Switzerland)*, *10*(18). https://doi.org/10.3390/APP10186296

Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis detection in breast cancer histology images with deep neural networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8150 LNCS*(PART 2), 411–418. https://doi.org/10.1007/978-3-642-40763-5_51

Coons, S. W., Johnson, P. C., Scheithauer, B. W., Yates, A. J., & Pearl, D. K. (1997). Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary gliomas. *Cancer*, *79*(7), 1381–1393. https://doi.org/10.1002/(SICI)1097-0142(19970401)79:7<1381::AID-CNCR16>3.0.CO;2-W

Copur, M., Melisozyildirim, B., & Ibrikci, T. (2018). Image Classification of Aerial Images Using CNN-SVM. *Proceedings - 2018 Innovations in Intelligent Systems and Applications Conference, ASYU 2018*. https://doi.org/10.1109/ASYU.2018.8554008

Dauphin, Y. N., De Vries, H., & Bengio, Y. (2015). Equilibrated adaptive learning rates for non-convex optimization. *Advances in Neural Information Processing Systems, 2015-Janua*, 1504–1512. Retrieved from http://arxiv.org/abs/1502.04390

De Brébisson, A., Simon, É., Auvolat, A., Vincent, P., & Bengio, Y. (2015). Artificial neural networks applied to taxi destination prediction. *CEUR Workshop Proceedings, 1526*. Retrieved from http://arxiv.org/abs/1508.00021

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2010, March 1). *ImageNet: A large-scale hierarchical image database*. 248–255. https://doi.org/10.1109/cvpr.2009.5206848

Desmeules, M., Mikkelsen, T., & Mao, Y. (1992). Increasing incidence of primary malignant brain tumors: Influence of diagnostic methods. *Journal of the National Cancer Institute*, *84*(6), 442–445. https://doi.org/10.1093/jnci/84.6.442

DeStigter, K., Horton, S., Atalabi, O. M., Garcia-Monaco, R. D., Gharbi, H. A., Hlabangana, L. T., … Mendel, J. (2019). Equipment in the Global Radiology Environment: Why We Fail, How We Could Succeed. *Journal of Global Radiology, 5*(1), e1079. https://doi.org/10.7191/jgr.2019.1079

Dong, H., Yang, G., Liu, F., Mo, Y., & Guo, Y. (2017). Automatic brain tumor detection and segmentation using U-net based fully convolutional networks. *Communications in Computer and Information Science, 723*, 506–517. https://doi.org/10.1007/978-3-319-60964-5_44

Drevelegas, A., & Papanikolaou, N. (2011). Imaging modalities in brain tumors. In *Imaging of Brain Tumors with Histological Correlations* (pp. 13–33). https://doi.org/10.1007/978-3-540-87650-2_2

Durmaz, R., Vural, M., Işldi, E., Coşan, E., Özkara, E., Bal, C., … Atasoy, M. (2008). Efficacy of prognostic factors on survival in patients with low grade glioma. *Turkish Neurosurgery, 18*(4), 336–344. Retrieved from http://neurosurgery.dergisi.org/pdf/pdf_JTN_608.pdf

Ellingson, B. M., Wen, P. Y., Van Den Bent, M. J., & Cloughesy, T. F. (2014). Pros and cons of current brain tumor imaging. *Neuro-Oncology*, *16*(Suppl 7), vii2–vii11. https://doi.org/10.1093/neuonc/nou224

Eriş, H., & Çevik, U. (2019). Implementation of Target Tracking Methods on Images Taken from Unmanned Aerial Vehicles. *SAMI 2019 - IEEE 17th World Symposium on Applied Machine Intelligence and Informatics, Proceedings*, 311–316. https://doi.org/10.1109/SAMI.2019.8782768

Ertosun, M. G., & Rubin, D. L. (2015). Automated Grading of Gliomas using Deep Learning in Digital Pathology Images: A modular approach with ensemble of convolutional neural networks. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2015*, 1899–1908. Retrieved from /pmc/articles/PMC4765616/?report=abstract

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Farabet, C., Couprie, C., Najman, L., & Lecun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1915–1929. https://doi.org/10.1109/TPAMI.2012.231

Freiherr, G. (1980). *The Seeds of Artificial Intelligence - SUMEX-AIM. - The Edward A. Feigenbaum Papers - Spotlight at Stanford*. Retrieved from https://exhibits.stanford.edu/feigenbaum/catalog/kd321pr5751

Gandía-González, M. L., Cerdán, S., Barrios, L., López-Larrubia, P., Feijoó, P. G., Palpan Jr., A., … Solivera, J. (2019). Assessment of Overall Survival in Glioma Patients as Predicted by Metabolomic Criteria. *Frontiers in Oncology*, *9*(MAY), 328. https://doi.org/10.3389/fonc.2019.00328

Gao, X. W., Hui, R., & Tian, Z. (2017). Classification of CT brain images based on deep learning networks. *Computer Methods and Programs in Biomedicine*, *138*, 49–56. https://doi.org/10.1016/j.cmpb.2016.10.007

Ge, C., Gu, I. Y.-H., Jakola, A. S., & Yang, J. (2017). BRAIN TUMOR CLASSIFICATION USING SLICE-BASED DEEP LEARNING AND FUSION OF MULTI-MODAL MR IMAGES. *Proceedings of the 40th Annual Int'l Conf of the IEEE Engineering in Medicine and Biology Society (EMBC18)*, 17–21. Retrieved from https://clasp.gu.se/digitalAssets/1696/1696299_ssdl2018_paper_1.pdf

Ge, C., Gu, I. Y. H., Jakola, A. S., & Yang, J. (2018). Deep Learning and Multi-Sensor Fusion for Glioma Classification Using Multistream 2D Convolutional Networks. *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *2018*, 5894–5897. https://doi.org/10.1109/EMBC.2018.8513556

Ge, C., Qu, Q., Gu, I. Y. H., & Store Jakola, A. (2018). 3D multi-scale convolutional networks for glioma grading using MR images. *Proceedings - International Conference on Image Processing, ICIP*, 141–145. https://doi.org/10.1109/ICIP.2018.8451682

GE Healthcare Partners. (2018). *Workforce Review: Radiology Aligning demand and capacity in a changing healthcare environment GE Healthcare Partners*. Retrieved November 17, 2020, from https://emea.gehealthcarepartners.com/images/pdfs/Rapid-Review--Radiology-Workforce-Review-FINAL.pdf

Geras, K. J., Wolfson, S., Shen, Y., Wu, N., Gene Kim, S., Kim, E., … Cho, K. (2017). High-resolution breast cancer screening with multi-view deep convolutional neural networks. *ArXiv*. Retrieved from http://arxiv.org/abs/1703.07047

Giammalva, G. R., Iacopino, D. G., Azzarello, G., Gaggiotti, C., Graziano, F., Gulì, C., … Maugeri, R. (2018). End-of-life care in high-grade glioma patients. The palliative and supportive perspective. *Brain Sciences*, *8*(7). https://doi.org/10.3390/brainsci8070125

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., … Vercauteren, T. (2018). NiftyNet: a deep-learning platform for medical imaging. *Computer Methods and Programs in Biomedicine*, *158*, 113–122. https://doi.org/10.1016/j.cmpb.2018.01.025

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology*, *278*(2), 563–577. https://doi.org/10.1148/radiol.2015151169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 580–587. https://doi.org/10.1109/CVPR.2014.81

Golub, R., Cantu, R., & Tan, M. (1998). The prediction of common bile duct stones using a neural network. *Journal of the American College of Surgeons*, *187*(6), 584–590. https://doi.org/10.1016/S1072-7515(98)00241-5

Goodenberger, M. L., & Jenkins, R. B. (2012). Genetics of adult glioma. *Cancer Genetics*, Vol. 205, pp. 613–621. https://doi.org/10.1016/j.cancergen.2012.10.009

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *3*(January), 2672–2680. Retrieved from http://www.github.com/goodfeli/adversarial

Griffiths, D., & Boehm, J. (2019). A Review on deep learning techniques for 3D sensed data classification. *Remote Sensing*, *11*(12). https://doi.org/10.3390/rs11121499

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., … Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - Journal of the American Medical Association*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

Gunn, A. A. (1976). The diagnosis of acute abdominal pain with computer analysis. *Journal of the Royal College of Surgeons of Edinburgh*, *21*(3), 170–172. Retrieved from https://pubmed.ncbi.nlm.nih.gov/781220/

Guo, C., & Berkhahn, F. (2016). Entity Embeddings of Categorical Variables. *ArXiv:1604.06737*. Retrieved from http://arxiv.org/abs/1604.06737

Gupta, V., Demirer, M., Bigelow, M., Little, K. J., Candemir, S., Prevedello, L. M., … Erdal, B. S. (2020). Performance of a Deep Neural Network Algorithm Based on a Small Medical Image Dataset: Incremental Impact of 3D-to-2D Reformation Combined with Novel Data Augmentation, Photometric Conversion, or Transfer Learning. *Journal of Digital Imaging*, *33*(2), 431–438. https://doi.org/10.1007/s10278-019-00267-3

Gurwara, S., Azzawe, A., Jacobs, S., Ravi, A., Hardjasudarma, M., & Toledo, E. G. (2010). Comparing the volume of gliomas in the brain in FLAIR and post-contrast T1-weighted MRI sequences. *The Journal of the Louisiana State Medical Society : Official Organ of the Louisiana State Medical Society*, *162*(5), 265–266. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21141258

Hansebout, R. R., Cornacchi, S. D., Haines, T., & Goldsmith, C. H. (2009). How to use an article about prognosis. *Canadian Journal of Surgery. Journal Canadien de Chirurgie*, *52*(4), 328–336. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19680521

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

He, Z., Mitteer, R. A., Mou, Y., & Fan, Y. (2016). Multimodality Targeting of Glioma Cells. In S. Brem & K. G. Abdullah (Eds.), *Glioblastoma* (pp. 55–72). https://doi.org/10.1016/B978-0-323-47660-7.00005-7

Helmstaedter, M., Briggman, K. L., Turaga, S. C., Jain, V., Seung, H. S., & Denk, W. (2013). Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, *500*(7461), 168–174. https://doi.org/10.1038/nature12346

Henson, D. B., Spenceley, S. E., & Bull, D. R. (1997). Artificial neural network analysis of noisy visual field data in glaucoma. *Artificial Intelligence in Medicine*, *10*(2), 99–113. https://doi.org/10.1016/S0933-3657(97)00388-6

Hess, L. M., Brnabic, A., Mason, O., Lee, P., & Barker, S. (2019). Relationship between progression-free survival and overall survival in randomized clinical trials of targeted and biologic agents in oncology. *Journal of Cancer*, *10*(16), 3717–3727. https://doi.org/10.7150/jca.32205

Hinton, G., Sejnowski, T. J., & Wang, D. (2001). *Book Review Unsupervised Learning-Foundations of Neural Computation* (Vol. 22). The MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holmes, G., Donkin, A., & Witten, I. H. (1994). WEKA: A machine learning workbench. *Australian and New Zealand Conference on Intelligent Information Systems - Proceedings*, 357–361. https://doi.org/10.1109/anziis.1994.396988

Hou, J. C., Wang, S. S., Lai, Y. H., Tsao, Y., Chang, H. W., & Wang, H. M. (2018). Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(2), 117–128. https://doi.org/10.1109/TETCI.2017.2784878

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., … Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*. Retrieved from http://arxiv.org/abs/1704.04861

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, *1*, 328–339. https://doi.org/10.18653/v1/p18-1031

Hsieh, K. L. C., Lo, C. M., & Hsiao, C. J. (2017). Computer-aided grading of gliomas based on local and global MRI features. *Computer Methods and Programs in Biomedicine*, *139*, 31–38. https://doi.org/10.1016/j.cmpb.2016.10.021

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

Ian Goodfellow, Yoshua Bengio, A. C. (2016). *Deep Learning*. Retrieved from http://www.deeplearningbook.org

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). *SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size*. Retrieved from http://arxiv.org/abs/1602.07360

IBM, 2013. IBM News room - 2013-10-09 $2 Million Awarded to Sutter Health, IBM and Geisinger Health System to Study Heart Failure Prediction - United States. Retrieved November 8, 2020, from https://www-03.ibm.com/press/us/en/pressrelease/42156.wss?_ga=2.208765714.1247324728.1604865674-768700670.1604865674

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., & Maier-Hein, K. H. (2018). Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10670 LNCS*, 287–297. https://doi.org/10.1007/978-3-319-75238-9_25

J Strong, M., & Garces, J. (2016). Brain Tumors: Epidemiology and Current Trends in Treatment. *Journal of Brain Tumors & Neurooncology*, *01*(01). https://doi.org/10.4172/2475-3203.1000102

Jackson, R. J., Fuller, G. N., Abi-Said, D., Lang, F. F., Gokaslan, Z. L., Shi, W. M., … Sawaya, R. (2001). Limitations of stereotactic biopsy in the initial management of gliomas. *Neuro-Oncology*, *3*(3), 193–200. https://doi.org/10.1093/neuonc/3.3.193

Jemal, A., Bray, F., Forman, D., O'Brien, M., Ferlay, J., Center, M., & Parkin, D. M. (2012). Cancer burden in Africa and opportunities for prevention. *Cancer*, *118*(18), 4372–4384. https://doi.org/10.1002/cncr.27410

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., … Darrell, T. (2014). *Caffe*. https://doi.org/10.1145/2647868.2654889

Jiménez, J., Škalič, M., Martínez-Rosell, G., & De Fabritiis, G. (2018). KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, *58*(2), 287–296. https://doi.org/10.1021/acs.jcim.7b00650

Johns, P. (2014). Neurons and glial cells. In *Clinical Neuroscience* (pp. 61–69). https://doi.org/10.1016/b978-0-443-10321-6.00005-9

Jooma, R., Waqas, M., & Khan, I. (2019). Diffuse low-grade glioma – Changing concepts in diagnosis and management: A review. *Asian Journal of Neurosurgery*, *14*(2), 356. https://doi.org/10.4103/ajns.ajns_24_18

Kaba, S. E., & Kyritsis, A. P. (1997). Recognition and management of gliomas. *Drugs*, Vol. 53, pp. 235–244. https://doi.org/10.2165/00003495-199753020-00004

Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A. V., Criminisi, A., … Glocker, B. (2016). DeepMedic for brain tumor segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10154 LNCS*, 138–149. https://doi.org/10.1007/978-3-319-55524-9_14

Karabulut, E. M., & Ibrikci, T. (2017). Discriminative deep belief networks for microarray based cancer classification. *Biomedical Research (India)*, *28*(3), 1016–1024. Retrieved from www.biomedres.info

Karremann, M., Gielen, G. H., Hoffmann, M., Wiese, M., Colditz, N., Warmuth-Metz, M., … Kramm, C. M. (2018). Diffuse high-grade gliomas with H3 K27M mutations carry a dismal prognosis independent of tumor location. *Neuro-Oncology*, *20*(1), 123–131. https://doi.org/10.1093/neuonc/nox149

Kaul, V., Enslin, S., & Gross, S. A. (2020). History of artificial intelligence in medicine. *Gastrointestinal Endoscopy*, *92*(4), 807–812. https://doi.org/10.1016/j.gie.2020.06.040

Khan, S., & Yairi, T. (2018). A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, *107*, 241–265. https://doi.org/10.1016/j.ymssp.2017.11.024

Khawaldeh, S., Pervaiz, U., Rafiq, A., & Alkhawaldeh, R. S. (2018). Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks. *Applied Sciences (Switzerland)*, *8*(1), 27. https://doi.org/10.3390/app8010027

Kickingereder, P., Burth, S., Wick, A., Götz, M., Eidel, O., Schlemmer, H. P., … Bonekamp, D. (2016). Radiomic profiling of glioblastoma: Identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology*, *280*(3), 880–889. https://doi.org/10.1148/radiol.2016160845

Kim, E. Y., Shin, S. Y., Lee, S., Lee, K. J., Lee, K. H., & Lee, K. M. (2020). Triplanar convolution with shared 2D kernels for 3D classification and shape retrieval. *Computer Vision and Image Understanding*, *193*, 102901. https://doi.org/10.1016/j.cviu.2019.102901

Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do better ImageNet models transfer better? *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2661–2671.

Kristensen, B. W., Priesterbach-Ackley, L. P., Petersen, J. K., & Wesseling, P. (2019). Molecular pathology of tumors of the central nervous system. *Annals of Oncology*, *30*(8), 1265–1278. https://doi.org/10.1093/annonc/mdz164

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105. https://doi.org/10.1145/3065386

Kulikowski, C. A. (2019). Beginnings of Artificial Intelligence in Medicine (AIM): Computational Artifice Assisting Scientific Inquiry and Clinical Art - with Reflections on Present AIM Challenges. *Yearbook of Medical Informatics*, *28*(1), 249–256. https://doi.org/10.1055/s-0039-1677895

Kwok, S., Pang, J., Salvo, N., & Chow, E. (2011). Prognostic assessment of the cancer patient. In *Supportive Oncology* (1st Editio, pp. 472–484). https://doi.org/10.1016/B978-1-4377-1015-1.00046-1

Lacroix, M., Abi-Said, D., Fourney, D. R., Gokaslan, Z. L., Shi, W., DeMonte, F., … Sawaya, R. (2001). A multivariate analysis of 416 patients with glioblastoma multiforme: Prognosis, extent of resection, and survival. *Journal of Neurosurgery*, *95*(2), 190–198. https://doi.org/10.3171/jns.2001.95.2.0190

Lao, J., Chen, Y., Li, Z. C., Li, Q., Zhang, J., Liu, J., & Zhai, G. (2017). A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-10649-8

Lauterbur, P. C. (1973). Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature*. https://doi.org/10.1038/242190a0

Laws, E. R., & Thapar, K. (1993). Brain tumors. *CA: A Cancer Journal for Clinicians*, *43*(5), 263–271. https://doi.org/10.3322/canjclin.43.5.263

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

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Lecun, Y., & Cortes, C. (2005). THE MNIST DATABASE of Handwritten Digits. *The Courant Institute of Mathematical Sciences*, 1–10. Retrieved from https://ci.nii.ac.jp/naid/10027939599

LeCun, Yann, Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2323. https://doi.org/10.1109/5.726791

Lee, D. H. (2002). Approach to end of life care. *Ochsner Journal*, *4*(2), 98–103.

Lee, K., Turner, N., Macrina, T., Wu, J., Lu, R., & Seung, H. S. (2019, April 1). Convolutional nets for reconstructing neural circuits from brain images acquired by serial section electron microscopy. *Current Opinion in Neurobiology*, Vol. 55, pp. 188–198. https://doi.org/10.1016/j.conb.2019.04.001

Legaz-Aparicio, A. G., Verdú-Monedero, R., Larrey-Ruiz, J., Morales-Sánchez, J., López-Mir, F., Naranjo, V., & Bernabéu, Á. (2017). Efficient variational approach to multimodal registration of anatomical and functional intra-patient tumorous brain data. *International Journal of Neural Systems*, *27*(6). https://doi.org/10.1142/S0129065717500149

Li, Q., Bai, H., Chen, Y., Sun, Q., Liu, L., Zhou, S., … Li, Z. C. (2017). A Fully-Automatic Multiparametric Radiomics Model: Towards Reproducible and Prognostic Imaging Signature for Prediction of Overall Survival in Glioblastoma Multiforme. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-14753-7

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C. W., & Heng, P. A. (2018). H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes. IEEE Transactions on Medical Imaging, 37(12), 2663–2674.

Li, Y., & Shen, L. (2018). Deep learning based multimodal brain tumor diagnosis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10670 LNCS*, 149–158. https://doi.org/10.1007/978-3-319-75238-9_13

Liang, Z.-P., & Lauterbur, P. C. (2000). *Principles of magnetic resonance imaging: a signal processing perspective*. https://doi.org/10.1109/memb.2000.870245

Lin, M., Chen, Q., & Yan, S. (2014). Network in network. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. Retrieved from https://arxiv.org/abs/1312.4400v3

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8693 LNCS*(PART 5), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, J., Li, M., Wang, J., Wu, F., Liu, T., & Pan, Y. (2014). A survey of MRI-based brain tumor segmentation methods. *Tsinghua Science and Technology*, *19*(6), 578–595. https://doi.org/10.1109/TST.2014.6961028

Liu, Y., Zhang, X., Feng, N., Yin, L., He, Y., Xu, X., & Lu, H. (2018). The effect of glioblastoma heterogeneity on survival stratification: a multimodal MR imaging texture analysis. *Acta Radiologica*, *59*(10), 1239–1246. https://doi.org/10.1177/0284185118756951

London, S. (1998). Dxplain™: A web-based diagnostic decision support system for medical students. *Medical Reference Services Quarterly*, *17*(2), 17–28. https://doi.org/10.1300/J115v17n02_02

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*, *114*(2), 97–109. https://doi.org/10.1007/s00401-007-0243-4

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., … Ellison, D. W. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica*, *131*(6), 803–820. https://doi.org/10.1007/s00401-016-1545-1

Louis, D. N., Schiff, D., Batchelor, T., & Wen, P. Y. (2017). *Classification and pathologic diagnosis of gliomas - UpToDate,Waltham, MA: Walters Kluwer Health.* Retrieved from https://www.uptodate.com/contents/classification-and-pathologic-diagnosis-of-gliomas#PATIENT_INFORMATION

Macyszyn, L., Akbari, H., Pisapia, J. M., Da, X., Attiah, M., Pigrish, V., … Davatzikos, C. (2016). Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-Oncology*, *18*(3), 417–425. https://doi.org/10.1093/neuonc/nov127

Maher, E. A., Furnari, F. B., Bachoo, R. M., Rowitch, D. H., Louis, D. N., Cavenee, W. K., & DePinho, R. A. (2001). Malignant glioma: Genetics and biology of a grave matter. *Genes and Development*, *15*(11), 1311–1333. https://doi.org/10.1101/gad.891601

Mansfield, P. (1977). Multi-planar image formation using NMR spin echoes. *Journal of Physics C: Solid State Physics*, *10*(3), L55. https://doi.org/10.1088/0022-3719/10/3/004

Mao, Y., & Yin, Z. (2016). A hierarchical convolutional neural network for mitosis detection in phase-contrast microscopy images. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9901 LNCS*, 685–692. https://doi.org/10.1007/978-3-319-46723-8_79

Mathieu, M., Couprie, C., & LeCun, Y. (2016). Deep multi-scale video prediction beyond mean square error. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

Matthew Brett, Michael Hanke, MARC-ALEXANDRE CÔTÉ, Paul McCarthy, Chris Cheng, B. C. (2017). *NiBabel- Access a cacophony of neuro-imaging file formats*. https://doi.org/10.5281/zenodo.1011207.svg

Maturana, D., & Scherer, S. (2015). VoxNet: A 3D Convolutional Neural Network for real-time object recognition. *IEEE International Conference on Intelligent Robots and Systems*, *2015-Decem*, 922–928. https://doi.org/10.1109/IROS.2015.7353481

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. In *AI Magazine* (Vol. 27). https://doi.org/10.1609/AIMAG.V27I4.1904

McKinney, P. A. (2004). Brain tumours: Incidence, survival, and aetiology. *Neurology in Practice*, *75*(2), 12–17. https://doi.org/10.1136/jnnp.2004.040741

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., … Van Leemput, K. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, *34*(10), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

Minsky, M. (1961). Steps Toward Artificial Intelligence. *Proceedings of the IRE*, Vol. 49, pp. 8–30. https://doi.org/10.1109/JRPROC.1961.287775

Mittal, S., Wu, Z., Neelavalli, J., & Haacke, E. M. (2009). Susceptibility-weighted imaging: Technical aspects and clinical applications, part 2. *American Journal of Neuroradiology*, *30*(2), 232–252. https://doi.org/10.3174/ajnr.A1461

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Monkam, P., Qi, S., Xu, M., Li, H., Han, F., Teng, Y., & Qian, W. (2019). Ensemble Learning of Multiple-View 3D-CNNs Model for Micro-Nodules Identification in CT Images. *IEEE Access*, *7*, 5564–5576. https://doi.org/10.1109/ACCESS.2018.2889350

Mormont, R., Geurts, P., & Maree, R. (2018). Comparison of deep transfer learning strategies for digital pathology. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2018-June*, 2343–2352. https://doi.org/10.1109/CVPRW.2018.00303

Mossa, Abdela A., Yibre, A. M., & Çevik, U. (2019). Multi-view CNN with MLP for diagnosing tuberculosis patients using CT scans and clinically relevant metadata. In H. M. Linda Cappellato, Nicola Ferro, David E. Losada (Ed.), *CEUR Workshop Proceedings* (Vol. 2380). Lugano, Switzerland: CEUR-WS.org 2019.

Mossa, Abdela A, Eriş, H., & Çevik, U. (2020). Ensemble of Deep Learning Models for Automatic Tuberculosis Diagnosis Using Chest CT Scans : Contribution to the ImageCLEF-2020 Challenges. *CEUR Workshop Proceedings*, 22–25. Retrieved from http://www.dei.unipd.it/~ferro/CLEF-WN-Drafts/CLEF2020/paper_64.pdf

Mossa, Abdela Ahmed, & Çevik, U. (2020). TRIPLANAR-CNN FOR AUTOMATED GRADING OF GLIOMAS USING PREOPERATIVE MULTI-MODAL MR IMAGES. *Proc. Of the International E-Conference OnAdvances in Engineering,Technology and Management -ICETM 2020*, 21–27. https://doi.org/10.15224/978-1-63248-188-7-05

Mossa, Abdela Ahmed, & Çevik, U. (2021). Ensemble learning of Multi-View CNN models for survival time prediction of brain tumor patients using multi-modal MRI scans. *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES*, *29*(2), 616–631. https://doi.org/10.3906/elk-2002-175

National Cancer Institute, & PDQ Adult Treatment Editorial Board. (2020). Adult Central Nervous System Tumors Treatment (PDQ®): Health Professional Version.

Nie, D., Lu, J., Zhang, H., Adeli, E., Wang, J., Yu, Z., … Shen, D. (2019). Multi-Channel 3D Deep Feature Learning for Survival Time Prediction of Brain Tumor Patients Using Multi-Modal Neuroimages. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-018-37387-9

Nie, D., Zhang, H., Adeli, E., Liu, L., & Shen, D. (2016). 3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9901 LNCS*, 212–220. https://doi.org/10.1007/978-3-319-46723-8_25

Niepert, M., Ahmad, M., & Kutzkov, K. (2016). Learning convolutional neural networks for graphs. *33rd International Conference on Machine Learning, ICML 2016*, *4*, 2958–2967.

NIH,2020. National Institutes of Health. Genetics Home Reference. What is the prognosis of a genetic condition?. Retrieved September 30, 2020, from https://ghr.nlm.nih.gov/primer/consult/prognosis

Ohgaki, H. (2009). Epidemiology of brain tumors. *Methods in Molecular Biology*, *472*, 323–342. https://doi.org/10.1007/978-1-60327-492-0_14

Olson, J. D., Riedel, E., & DeAngelis, L. M. (2000). Long-term outcome of low-grade oligodendroglioma and mixed glioma. *Neurology*, *54*(7), 1442–1448. https://doi.org/10.1212/WNL.54.7.1442

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., … Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. *ArXive:1609.03499v2*. Retrieved from http://arxiv.org/abs/1609.03499

Ostrom, Q. T., Cioffi, G., Gittleman, H., Patil, N., Waite, K., Kruchko, C., & Barnholtz-Sloan, J. S. (2019). CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2012-2016. *Neuro-Oncology*, Vol. 21, pp. V1–V100. https://doi.org/10.1093/neuonc/noz150

Pak, M., & Kim, S. (2018). A review of deep learning in image recognition. *Proceedings of the 2017 4th International Conference on Computer Applications and Information Processing Technology, CAIPT 2017*, *2018-January*, 1–3. https://doi.org/10.1109/CAIPT.2017.8320684

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*.

Patel, A. P., Fisher, J. L., Nichols, E., Abd-Allah, F., Abdela, J., Abdelalim, A., … Fitzmaurice, C. (2019). Global, regional, and national burden of brain and other CNS cancer, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, *18*(4), 376–393. https://doi.org/10.1016/S1474-4422(18)30468-X

Peleg, R., & Biderman, A. (2000). A memorable patient: The scars of the Jewish holocaust. *BMJ*, *320*(7233), 473–473. https://doi.org/10.1136/bmj.320.7233.473

Pereira, S., Meier, R., Alves, V., Reyes, M., & Silva, C. A. (2018). Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11038 LNCS*, 106–114. https://doi.org/10.1007/978-3-030-02628-8_12

Pesonen, E., Ohmann, C., Eskelinen, M., & Juhola, M. (1998). Diagnosis of acute appendicitis in two databases. Evaluation of different neighborhoods with an LVQ neural network. *Methods of Information in Medicine*, *37*(1), 59–63. https://doi.org/10.1055/s-0038-1634497

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018a). Deep contextualized word representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 2227–2237. https://doi.org/10.18653/v1/n18-1202

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018b). Improving Language Understanding by. *OpenAI*, 1–10. Retrieved from https://gluebenchmark.com/leaderboard

Pouratian, N., & Schiff, D. (2010). Management of low-grade glioma. *Current Neurology and Neuroscience Reports*, *10*(3), 224–231. https://doi.org/10.1007/s11910-010-0105-7

Pusiol, G., Esteva, A., Hall, S. S., Frank, M., Milstein, A., & Li, F. F. (2016). Vision-based classification of developmental disorders using eye-movements. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9901 LNCS*, 317–325. https://doi.org/10.1007/978-3-319-46723-8_37

Qin, C., Schlemper, J., Caballero, J., Price, A. N., Hajnal, J. V., & Rueckert, D. (2019). Convolutional recurrent neural networks for dynamic MR image reconstruction. *IEEE Transactions on Medical Imaging*, *38*(1), 280–290. https://doi.org/10.1109/TMI.2018.2863670

Ramesh, A. N., Kambhampati, C., Monson, J. R. T., & Drew, P. J. (2004). Artificial intelligence in medicine. *Annals of the Royal College of Surgeons of England*, *86*(5), 334–338. https://doi.org/10.1308/147870804290

Reardon, D. A., & Wen, P. Y. (2015). Glioma in 2014: Unravelling tumour heterogeneity - Implications for therapy. *Nature Reviews Clinical Oncology*, *12*(2), 69–70. https://doi.org/10.1038/nrclinonc.2014.223

Rees, J. H., Smirniotopoulos, J. G., Jones, R. V., & Wong, K. (1996). From the Archives of the AFIP - Glioblastoma Multiforme: Radiologic-Pathologic Correlation. *Radiographics*, *16*(6), 1413–1438. https://doi.org/10.1148/radiographics.16.6.8946545

Reuter, M., Gerstner, E. R., Rapalino, O., Batchelor, T. T., Rosen, B., & Fischl, B. (2014). Impact of MRI head placement on glioma response assessment. *Journal of Neuro-Oncology*, *118*(1), 123–129. https://doi.org/10.1007/s11060-014-1403-8

Reza, S. M. S., Samad, M. D., Shboul, Z. A., Jones, K. A., & Iftekharuddin, K. M. (2019). Glioma grading using structural magnetic resonance imaging and molecular data. *Journal of Medical Imaging*, *6*(2). https://doi.org/10.1117/1.JMI.6.2.024501

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9351*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

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

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Russell, S., & Norvig, P. (2002). *Artificial Intelligence: A Modern Approach*.

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, *2017-Decem*, 3857–3867. Retrieved from http://arxiv.org/abs/1710.09829

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *44*(1–2), 207–219. https://doi.org/10.1147/rd.441.0206

Sanai, N., Chang, S., & Berger, M. S. (2011). Low-grade gliomas in adults: A review. *Journal of Neurosurgery*, *115*(5), 948–965. https://doi.org/10.3171/2011.7.JNS101238

Sanghani, P., Ang, B. T., King, N. K. K., & Ren, H. (2018). Overall survival prediction in glioblastoma multiforme patients from volumetric, shape and texture features using machine learning. *Surgical Oncology*, *27*(4), 709–714. https://doi.org/10.1016/j.suronc.2018.09.002

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 618–626. https://doi.org/10.1109/ICCV.2017.74

Setio, A. A. A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., Van Riel, S. J., … Van Ginneken, B. (2016). Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Transactions on Medical Imaging*, *35*(5), 1160–1169. https://doi.org/10.1109/TMI.2016.2536809

Shankar, G. M., Balaj, L., Stott, S. L., Nahed, B., & Carter, B. S. (2017). Liquid biopsy for brain tumors. *Expert Review of Molecular Diagnostics*, *17*(10), 943–947. https://doi.org/10.1080/14737159.2017.1374854

Shboul, Z. A., Alam, M., Vidyaratne, L., Pei, L., Elbakary, M. I., & Iftekharuddin, K. M. (2019). Feature-Guided Deep Radiomics for Glioblastoma Patient Survival Prediction. *Frontiers in Neuroscience*, *13*. https://doi.org/10.3389/fnins.2019.00966

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. https://doi.org/10.1109/TMI.2016.2528162

Shortliffe, E. H. (1977). Mycin: A knowledge-based computer program applied to infectious diseases. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 66. American Medical Informatics Association.

Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(1), 7–30. https://doi.org/10.3322/caac.21590

Sierra, B., Serrano, N., Larrañaga, P., Plasencia, E. J., Inza, I., Jiménez, J. J., … Mora, M. L. (2001). Using Bayesian networks in the construction of a bi-level multi-classifier. A case study using intensive care unit patients data. *Artificial Intelligence in Medicine*, *22*(3), 233–248. https://doi.org/10.1016/S0933-3657(00)00111-1

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., … Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. Retrieved from http://arxiv.org/abs/1409.1556

Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., & Gulyás, B. (2020). 3d deep learning on medical images: A review. *Sensors (Switzerland)*, *20*(18), 1–24. https://doi.org/10.3390/s20185097

Smits, M., & Van Den Bent, M. J. (2017). Imaging correlates of adult glioma genotypes. *Radiology*, *284*(2), 316–331. https://doi.org/10.1148/radiol.2017151930

Sofroniew, M. V., & Vinters, H. V. (2010). Astrocytes: Biology and pathology. *Acta Neuropathologica*, Vol. 119, pp. 7–35. https://doi.org/10.1007/s00401-009-0619-8

Stewart, B. W., & Wild, C. P. (2014). World Cancer Report 2014- IARC WHO. In *World Cancer Report*.

Stupp, R., Brada, M., van den Bent, M. J., Tonn, J. C., & Pentheroudakis, G. (2014). High-grade glioma: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, *25*, 93–101. https://doi.org/10.1093/annonc/mdu050

Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E. (2015). Multi-view convolutional neural networks for 3D shape recognition. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 945–953. https://doi.org/10.1109/ICCV.2015.114

Sultana, F., Sufian, A., & Dutta, P. (2020). Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey. *Knowledge-Based Systems*, *201–202*, 106062. https://doi.org/10.1016/j.knosys.2020.106062

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 843–852. https://doi.org/10.1109/ICCV.2017.97

Sun, L., Zhang, S., Chen, H., & Luo, L. (2019). Brain tumor segmentation and survival prediction using multimodal MRI scans with deep learning. *Frontiers in Neuroscience*, *13*(JUL). https://doi.org/10.3389/fnins.2019.00810

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., & Le, Q. V. (2018). MnasNet: Platform-Aware Neural Architecture Search for Mobile. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 2815–2823. Retrieved from http://arxiv.org/abs/1807.11626

Theodoridis, S., & Koutroumbas, K. (2010). *Pattern recognition* (Fourth Edi). https://doi.org/10.1002/wics.99

Tian, Q., Yan, L. F., Zhang, X., Zhang, X., Hu, Y. C., Han, Y., … Cui, G. Bin. (2018). Radiomics strategy for glioma grading using texture features from multiparametric MRI. *Journal of Magnetic Resonance Imaging*, *48*(6), 1518–1528. https://doi.org/10.1002/jmri.26010

Timmons, J. (2012). Primary Brain Tumours – Everything a Medical Student Needs to Know. *Scottish Universities Medical Journal*, *1*(1), 31–37. Retrieved from http://sumj.dundee.ac.uk/data/uploads/volume1/SUMJVol1-p31-37.pdf

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*, *65*(2), 87–108. https://doi.org/10.3322/caac.21262

Turing, A. (1950). Alan Turing Scrapbook - Turing Test. Retrieved March 15, 2021, from Turing.org website: https://www.turing.org.uk/scrapbook/test.html

Ulku, I., & Akagunduz, E. (2019). *A Survey on Deep Learning-based Architectures for Semantic Segmentation on 2D images*. Retrieved from http://arxiv.org/abs/1912.10230

Upadhaya, T., Morvan, Y., Stindel, E., Le Reste, P. J., & Hatt, M. (2015). A framework for multimodal imaging-based prognostic model building: Preliminary study on multimodal MRI in Glioblastoma Multiforme. *IRBM*, *36*(6), 345–350. https://doi.org/10.1016/j.irbm.2015.08.001

Upadhyay, N., & Waldman, A. D. (2011). Conventional MRI evaluation of gliomas. *British Journal of Radiology*, *84*(SPEC. ISSUE 2), S107. https://doi.org/10.1259/bjr/65711810

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating videos with scene dynamics. *Advances in Neural Information Processing Systems*, 613–621. https://doi.org/10.13016/m26gih-tnyz

Wang, G., Li, W., Ourselin, S., & Vercauteren, T. (2018). Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10670 LNCS*, 178–190. https://doi.org/10.1007/978-3-319-75238-9_16

Wang, K., Wang, Y., Fan, X., Wang, J., Li, G., Ma, J., … Dai, J. (2015). Radiological features combined with IDH1 status for predicting the survival outcome of glioblastoma patients. *Neuro-Oncology*, *18*(4), 589–597. https://doi.org/10.1093/neuonc/nov239

Wang, S., Zhou, M., Gevaert, O., Tang, Z., Dong, D., Liu, Z., & Tian, J. (2017). A multi-view deep convolutional neural networks for lung nodule segmentation. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 1752–1755. https://doi.org/10.1109/EMBC.2017.8037182

Watanabe, M., Tanaka, R., & Takeda, N. (1992). Magnetic resonance imaging and histopathology of cerebral gliomas. *Neuroradiology*, *34*(6), 463–469. https://doi.org/10.1007/BF00598951

Welling, R. D., Azene, E. M., Kalia, V., Pongpirul, K., Starikovsky, A., Sydnor, R., … Mollura, D. J. (2011). White paper report of the 2010 RAD-AID Conference on International Radiology for Developing Countries: Identifying sustainable strategies for imaging services in the developing world. *Journal of the American College of Radiology*, *8*(8), 556–562. https://doi.org/10.1016/j.jacr.2011.01.011

Wen, P. Y., Macdonald, D. R., Reardon, D. A., Cloughesy, T. F., Sorensen, A. G., Galanis, E., … Chang, S. M. (2010). Updated response assessment criteria for high-grade gliomas: Response assessment in neuro-oncology working group. *Journal of Clinical Oncology*, *28*(11), 1963–1972. https://doi.org/10.1200/JCO.2009.26.3541

Wesseling, P., & Capper, D. (2018). WHO 2016 Classification of gliomas. *Neuropathology and Applied Neurobiology*, *44*(2), 139–150. https://doi.org/10.1111/nan.12432

Winzeck, S., Mocking, S. J. T., Bezerra, R., Bouts, M. J. R. J., Mcintosh, E. C., Diwan, I., … Wu, O. (2019). Ensemble of convolutional neural networks improves automated segmentation of acute ischemic lesions using multiparametric diffusion-weighted MRI. *American Journal of Neuroradiology*, *40*(6), 938–945. https://doi.org/10.3174/ajnr.A6077

Witten, I. H., Frank, E., Hall, mark A., & Pal, C. J. (2011). Data Mining: Practical Machine Learning Tools and Techniques. In *Data Mining: Practical Machine Learning Tools and Techniques*. https://doi.org/10.1016/c2009-0-19715-5

Wolburg, H., Wolburg-Buchholz, K., Reichenbach, A., & Mack, A. F. (2015). Ependymal Cells☆. In *Reference Module in Biomedical Sciences*. https://doi.org/10.1016/b978-0-12-801238-3.04586-4

WHO,2018. World Health Organization. *Globocan 2018 - Turkey*. Retrieved from https://gco.iarc.fr/today/data/factsheets/populations/792-turkey-fact-sheets.pdf

WHO,2018. Cancer - World Health Organization. Retrieved from https://www.who.int/health-topics/cancer#tab=tab_1

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., … Geras, K. J. (2019). *Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening*. Retrieved from http://arxiv.org/abs/1903.08297

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., & Xiao, J. (2015). 3D ShapeNets: A deep representation for volumetric shapes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*, 1912–1920. https://doi.org/10.1109/CVPR.2015.7298801

Xiao, Y., Wu, J., Lin, Z., & Zhao, X. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine*, *153*, 1–9. https://doi.org/10.1016/j.cmpb.2017.09.005

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The Microsoft 2017 Conversational Speech Recognition System. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2018-April*, 5934–5938. https://doi.org/10.1109/ICASSP.2018.8461870

Xu, Z., Bagci, U., Kubler, A., Luna, B., Jain, S., Bishai, W. R., & Mollura, D. J. (2013). Computer-aided detection and quantification of cavitary tuberculosis from CT scans. *Medical Physics*, *40*(11). https://doi.org/10.1118/1.4824979

Zhu, Y., Man, C., Gong, L., Dong, D., Yu, X., Wang, S., … Tian, J. (2019). A deep learning radiomics model for preoperative grading in meningioma. *European Journal of Radiology*, *116*, 128–134. https://doi.org/10.1016/j.ejrad.2019.04.022

Yang, C., Rangarajan, A., & Ranka, S. (2018). Visual Explanations From Deep 3D Convolutional Neural Networks for Alzheimer's Disease Classification. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2018*, 1571–1580. Retrieved from /pmc/articles/PMC6371279/?report=abstract

Yang, W., Feng, Q., Yu, M., Lu, Z., Gao, Y., Xu, Y., & Chen, W. (2012). Content-based retrieval of brain tumor in contrast-enhanced MRI images using tumor margin information and learned distance metric. *Medical Physics*, *39*(11), 6929–6942. https://doi.org/10.1118/1.4754305

Yang, Y., Yan, L. F., Zhang, X., Han, Y., Nan, H. Y., Hu, Y. C., … Wang, W. (2018). Glioma grading on conventional MR images: A deep learning study with transfer learning. *Frontiers in Neuroscience*, *12*(NOV). https://doi.org/10.3389/fnins.2018.00804

Ye, F., Pu, J., Wang, J., Li, Y., & Zha, H. (2017). Glioma grading based on 3D multimodal convolutional neural network and privileged learning. *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017, 2017-Janua*, 759–763. https://doi.org/10.1109/BIBM.2017.8217751

Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, *58*, 101552. https://doi.org/10.1016/j.media.2019.101552

Yiallouros, M., & Tallen, G. (2007). Niedrigmaligne Gliome (Kurzinformation). In M. Yiallouros (Ed.), *Niedrigmaligne Gliome (Kurzinformation)*. https://doi.org/10.1591/poh.patinfo.ng.kurz.1.20070627

Yilmaz, H. H., Yazihan, N., Tunca, D., Sevinç, A., Olcayto, E. Ö., Özgül, N., & Tuncer, M. (2011). Cancer trends and incidence and mortality patterns in Turkey. *Japanese Journal of Clinical Oncology*, *41*(1), 10–16. https://doi.org/10.1093/jjco/hyq075

Yushkevich, P. A., Gao, Y., & Gerig, G. (2016). ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2016-Octob*, 3342–3345. https://doi.org/10.1109/EMBC.2016.7591443

Zelenak, K., Viera, C., & Hubert, P. (2013). Radiology Imaging Techniques of Brain Tumours. In *Clinical Management and Evolving Novel Therapeutic Strategies for Patients with Brain Tumors*. https://doi.org/10.5772/53470

Zeng, G., Yang, X., Li, J., Yu, L., Heng, P. A., & Zheng, G. (2017). 3D U-net with multi-level deep supervision: Fully automatic segmentation of proximal femur in 3D MR images. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10541 LNCS*, 274–282. https://doi.org/10.1007/978-3-319-67389-9_32

Zhang, B., Qi, S., Monkam, P., Li, C., Yang, F., Yao, Y.-D., & Qian, W. (2019). Ensemble Learners of Multiple Deep CNNs for Pulmonary Nodules Classification Using CT Images. *IEEE Access*, *7*, 110358–110371. https://doi.org/10.1109/access.2019.2933670

Zhang, X., Yan, L. F., Hu, Y. C., Li, G., Yang, Y., Han, Y., … Cui, G. Bin. (2017). Optimizing a machine learning based glioma grading system using multi-parametric MRI histogram and texture features. *Oncotarget*, *8*(29), 47816–47830. https://doi.org/10.18632/oncotarget.18001

Zhang, Y., Wang, X., Blanton, H., Liang, G., Xing, X., & Jacobs, N. (2019). 2D Convolutional Neural Networks for 3D Digital Breast Tomosynthesis Classification. *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*, 1013–1017. https://doi.org/10.1109/BIBM47256.2019.8983097

Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, *137*(1–2), 239–263. https://doi.org/10.1016/S0004-3702(02)00190-X

Zhuge, Y., Ning, H., Mathen, P., Cheng, J. Y., Krauze, A. V., Camphausen, K., & Miller, R. W. (2020). Automated glioma grading on conventional MRI images using deep convolutional neural networks. *Medical Physics*, *47*(7), 3044–3053. https://doi.org/10.1002/mp.14168

Zimny, A., Neska-Matuszewska, M., Bladowska, J., & Sąsiadek, M. J. (2015). Intracranial lesions with low signal intensity on T2-weighted MR images - review of pathologies. *Polish Journal of Radiology*, *80*(1), 40–50. https://doi.org/10.12659/PJR.892146

AUTOBIOGRAPHY

He attended his elementary and secondary school respectively at Meshentiee Elementary School and Merawi Senior Secondary School. He completed his Bachelor's degree in Computer Science from Bahirdar University in 2008, Ethiopia. Immediately after graduation, he had been working as an assistant lecturer at the University of Gondar (UOG), Ethiopia. In 2013, he obtained his Master's degree in Computer Science from Osmania University, India, and continued working as a lecturer in the same university, UOG, until September 2014 when he began his Ph.D study at the Department of Computer Engineering of Çukurova University after winning an international scholarship. During his studies and thesis work, he worked on the application of deep learning and classical machine learning techniques for fully automated 3D medical image analysis problems under the supervision of Prof. Dr. Ulus Çevik.