Radiomics and Residual Neural Networks for Pretherapeutic MRI Differentiation of *BRAF* Status in Pediatric Low-Grade Gliomas

by

Justin Xu

Supervisor: Farzad Khalvati

September 2022

A thesis submitted in conformity with the requirements
for the degree of Bachelor of Applied Science

Division of Engineering Science
University of Toronto

Radiomics and Residual Neural Networks for Pretherapeutic MRI Differentiation of *BRAF* Status in Pediatric Low-Grade Gliomas

Justin Xu
Bachelor of Applied Science

Division of Engineering Science
University of Toronto
2022

Abstract

Pediatric Low-Grade Gliomas (pLGGs) are the most common set of heterogeneous tumours occurring in children and its variability presents a clinical challenge. The *B-Raf proto-oncogene, serine/threonine kinase* (*BRAF*) status of pLGGs have important prognostic implications. Currently, *BRAF* status differentiation relies on biopsy.

Advances in radiomics and computer vision have established these methods as effective tools for medical image analysis. This project aims to implement and evaluate several methods, including fusion approaches between radiomics and deep learning, to binarily classify *BRAF* fusion and *BRAF* V600E mutation.

FLAIR MR images from a bi-institutional retrospective cohort of 255 pediatric patients acquired between January 2000 and December 2018 were analyzed. Radiomics features were extracted, and MR images were registered and normalized. A radiomics random forest classifier and a 3D residual neural network (ResNet) was implemented. Additionally, appending features and risk scores fusion approaches were also developed. Models were selected using grid search and validation loss. The area under the receiver operating characteristic curve (AUC) was used to evaluate model performance.

A 5-fold cross-validation scheme was executed on an internal training cohort of 161 patients. The radiomics random forest classifier predicted *BRAF* status of the internal testing cohort with an AUC of 0.877 (95% CI, 0.865-0.893). Similarly, the ResNet achieved an AUC of 0.853 (95% CI, 0.842-0.863). The appending features fusion approach achieved an AUC of 0.888 (95% CI, 0.875-0.901), while the risk score approach achieved an AUC of 0.899 (95% CI, 0.891-0.908). The improvement in classification performance of the risk score fusion approach was significant as compared to the radiomics classifier ($p = 0.0429$) and the ResNet ($p < 0.0001$).

Radiomics and machine learning-based differentiation of *BRAF* status in pLGGs appears feasible, and improvements in performances have the potential of enhancing patient prognosis and outcomes.

To Mom, Dad, and Fifi

Acknowledgments

First and foremost, I would like to thank my esteemed supervisor at the Hospital for Sick Children, Dr. Farzad Khalvati, for his invaluable supervision, support, and tutelage during the course of my BASc thesis. My gratitude extends to the Faculty of Applied Science and Engineering and the Division of Engineering Science for shaping me into a professional as I undertake my engineering studies at the University of Toronto.

Additionally, words cannot express my gratitude to my direct advisor, Kareem Kudus, for his treasured patience, expertise, and feedback. His support was tremendously influential in the development of my experimental methods and results.

I would be remiss if I didn't mention my loving parents and baby sister. Their belief in me has kept my spirits and motivation high during the entire process, and I couldn't have done it without them. Lastly, I'd like to also thank the neighbourhood cat for all the entertainment and for acting as my emotional support animal.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Pediatric low-grade gliomas (pLGGs) are one of the most common set of heterogeneous tumours occurring in children [1]. Accounting for nearly 40% of all pediatric cancers of the brain, pLGG's variability in clinical behaviour presents significant therapeutic challenges [2], [3]. The current standard of care for pLGG therapy involves surgical excision wherever possible, and full resection is restorative. Unfortunately, incomplete removal can lead to multiple recurrences in the form of a chronic disorder, bringing about significant reductions in the quality-of-life for the patient [4].

Several researchers have previously illustrated that molecular alterations in genes of mitogen-activated protein kinase pathways are correlated with erratic pLGGs. Notably, different classes of mutations in the *B-Raf proto-oncogene, serine/threonine kinase* (*BRAF*) gene can result in different implications in patient prognosis as illustrated by Lassaletta et al. in 2017 [5]. This finding led to the development of promising therapeutic agents that inhibit associated downstream pathways to target increased risks of disease progression in particular mutation classes [6], [7]. These therapies have the potential to deliver more personalized responses to pLGGs and thus grant more favourable patient outcomes. As such, this pretherapeutic classification of *BRAF* status is critical to determine suitable candidates for targeted treatments.

Currently, this classification relies on biopsies that can be invasive [8], [9]. However, a previous study demonstrated the feasibility of applying radiomics approaches to binarily classify *BRAF* status through the use of FLAIR MR images. The authors trained a random forest classifier to classify outcomes by linking the top 10

quantitative radiomics features extracted from the images to their clinical statuses. Areas under receiver operating characteristic curves (AUCs), which are representative of model performance, were computed [9]. The model yielded an internal testing AUC of 0.75 (95% CI, 0.62–0.89) with a small dataset. Hence, improvements in AUCs and testing on larger datasets are desired for potential model adoption in clinical pathways.

Recent advances in deep learning and computer vision have shown great potential in medical imaging, suggesting the predictive information available in convolutional neural network (CNN) features. Theoretical analysis and preliminary results obtained also indicate that CNNs could be a viable alternative in *BRAF* status classification, yielding similar AUCs as the random forest classifier. Studies suggest that by integrating both CNN and radiomics models, a higher AUC can be achieved [10]. For instance, naïve methods of combining both approaches, such as training models separately and then averaging outputs, would produce a fusion model that achieves a slightly better AUC than either model would individually.

1.2 Objectives

The purpose of this project is to develop improved methods for pretherapeutic differentiation of *BRAF* status using radiomics and machine learning. We hope to explore sophisticated methods of producing a fusion classification model that improves upon the AUC yielded by previous standalone models. Such a model would expand the current understanding of the correlation between molecular markers and imaging features and would enable future machine learning studies of pLGG MRI. The higher AUCs we hope to achieve is necessary for implementation of the algorithm to predict *BRAF* status in clinical settings to enhance patient outcomes.

The objective of this project is three-fold:

(1) *To successfully develop a radiomics random forest classifier as previously seen in Wagner et al. and generate similar results on the larger target dataset;*

(2) *To implement a CNN architecture that classifies* BRAF *status at similar performance as compared to the radiomics random forest classifier;*

(3) *To investigate sophisticated approaches to fuse the two above models for improved classification AUCs.*

After development of the radiomics random forest classifier, a combination of convolutional, pooling, and linear layers will be surveyed to optimize a CNN for increases in AUC. To avoid vanishing/exploding gradients that are characteristic of deep networks, ResNet [11] and related architectures will be explored. Finally, following the implementation of the CNN, sophisticated approaches of fusing these models, such as using feature appending or risk scoring [12], will be investigated to combine features and results for a singular optimized model. Additional approaches, such as feature mapping [13], evidential reasoning [14], and/or soft voting [15] will also be introduced for future implementation. We propose that by implementing a complex approach to merge radiomics and CNN models, a higher AUC can be attained and thus result in a more optimized classification of *BRAF* status in pLGGs.

Chapter 2

Background

2.1 Pediatric Low-Grade Gliomas

Pediatric low-grade gliomas (pLGGs) are one of the most common central nervous system tumours that affect children of all ages. Although pLGG patients have a high survival-rate when undergoing radiation and standard chemotherapy, comorbidities relating to their cancer, such as vision loss, epilepsy, and dysfunctions in endocrine, motor, and neurocognitive functions devastate their quality of life [16]. The World Health Organization categorizes pLGGs as grades I or II, which include juvenile pilocytic astrocytoma (JPA), ganglioglioma, dysembryoplastic neuroepithelial tumor, pleomorphic xanthoastrocytoma, and diffuse low-grade glioma [17].

While an effective strategy to identify and treat all poor-risk groups has yet to be developed due to pLGG's heterogenous nature, precision medicine approaches involving small molecular inhibitors has garnered much attention in the scientific community. pLGGs often exhibit molecular alterations in genes that control associated cell signalling pathways, causing cells to become cancerous.

2.1.1 *BRAF* Status

The *B-Raf proto-oncogene, serine/threonine kinase* (*BRAF*) gene is a key target of pLGG molecular therapies due to its role in encoding instructions for proteins that oversee chemical signal transmission across the cytosol from the extracellular environment to the nucleus. The oncogene plays a role in the Ras/MAPK pathway, which regulates cell proliferation, differentiation, migration, and apoptosis.

Fusions and mutations in the *BRAF* gene lead to significant alterations in molecular signatures and cause deregulation of important pathways. The 2 molecular al-

terations of the *BRAF* gene that is of concern is the *BRAF* fusion, which consists of the duplication of a *BRAF* oncogene leading to loss of an autoregulatory domain through insertion into fusion targets[1] [1], and the *BRAF* V600E, which consists of a point mutation leading to deregulation and constantly active protein kinase pathways[2] [18].

2.1.2 Biopsies

The current standard of differentiation of *BRAF* status involves biopsies. However, there are several disadvantages, including the generally invasive nature of the procedure. Due to the disease site and location, poor accessibility is ever-present, making biopsies complex [16]. Due to the heterogenous nature of the tumours, biopsies may not necessarily always capture an accurate depiction of the molecular character of the entire pLGG. Additionally, biopsy procedures can often lead to infection or hemorrhage if not completed up to standard, and in rare cases, there is the possibility of tumour seeding causing the cancer to spread [17]. All in all, biopsies are undesirable.

2.1.3 FLAIR Magnetic Resonance Imaging

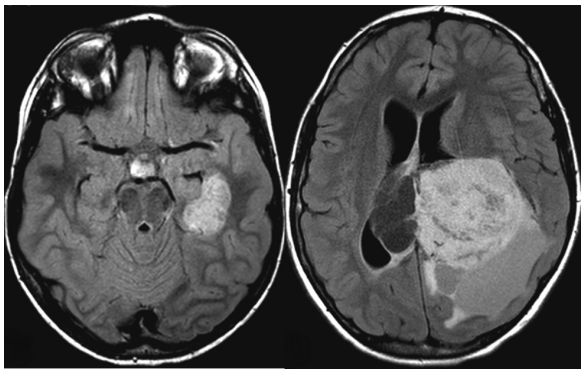

Figure 2.1: Representative FLAIR MR images from the target dataset. White regions in the image correspond to tumorous regions of the brain and contribute to the patient's LGG.

Fluid-attenuated inversion recovery (FLAIR) is an advanced MRI sequence used in most brain imaging protocols that allows the detection of superficial lesions. The long inversion time of the sequence suppresses cerebrospinal fluid in the

[1]The most common fusion target is the K1AA1549 gene.

[2]The *BRAF* V600E point mutation is associated with increased risk of progression and transformation.

resulting images, where white matter appears darker than grey matter [19]. Radiologists use these FLAIR images to detect pLGGs and to annotate segmentations that show cancerous regions. The dataset used in this project consists of unregistered raw FLAIR images acquired from various sites. All FLAIR images will be registered to an atlas in the preprocessing stage to ensure conformity in coordinate systems prior to model training.

2.2 Radiomics

The growing field of radiomics involves the extraction of quantitative data or features from medical images that are potentially characteristic of tumour behaviour. By implementing machine learning algorithms to analyze a large number of features, radiomics is able to improve prognosis and diagnosis through precision medicine approaches [20].

2.2.1 Radiomics Feature Extraction

Radiomics features are extracted through mathematical approaches and captures textural information, such as spatial distributions of signal intensities or pixel interrelationships [21]. The hundreds of features selected in the development of the radiomics random forest classifier are the same as the pre-extracted features provided by Wagner et al. The PyRadiomics Python package will be employed to extract the features used to construct the models of this project. While the package full documentation is available at "Amadasun and King, PyRadiomics, Readthedocs.io, 2022" [22], feature classes and example features that can be extracted from the image are summarized below.

First Order Statistics

> First-order statistics describe voxel intensity distributions within a region-of-interest (ROI), or segmentation, on the MR image. Several commonly used metrics are grouped into this class. Features of note include voxel energy and entropy.
>
> Energy is a measure of magnitude. Larger values imply greater sums of

the squares of these voxels. It is defined as:

$$energy = \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2 \tag{2.1}$$

where $\mathbf{X}$ is the set of N_p voxels included in the ROI and c represents a value needed to shift intensities to prevent negative values in $\mathbf{X}$.

Entropy specifies the uncertainty when encoding image values. It is defined as:

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 (p(i) + \epsilon) \tag{2.2}$$

where $P(i)$ is the first order histogram with N_g discrete intensity levels and ϵ is an arbitrarily small positive number.

Shape Features

Shape features are descriptors of the image space occupied by the segmentation. Examples of 3D shape features include voxel volume, surface area, and sphericity. Similarly, pixel surface, perimeter, and elongation values are examples of 2D shape features.

Gray Level Matrix (GLM) Features

Gray level matrix features are further subdivided into co-occurrence, run length, size zone, and dependence matrices. These features are established in the field of radiomics and are used for texture analysis and for descriptions of voxel spatial relationships. Specific details are available in the PyRadiomics documentation [23].

Wavelet Derived Features

In addition to features extracted from the original image, features may be extracted from image derivatives after application of filters. One filter of interest is the wavelet filter, which yields 8 image derivatives per level (i.e., for each of the 3 dimension of the image, derivatives of all possible combinations of 'high pass' and 'low pass' filters are generated). The features extracted from these derivatives were also included.

2.2.2 Naïve Feature Selection

Given the number of features extracted, feature selection may be required to limit undesirable outcomes where irrelevant or unimportant features hold substantial weight. A naïve method of feature selection used in this project is based on the fitted `feature_importances_` attribute available in Scikit-learn random forest models. This attribute outlines the importance of input features based on the mean and standard deviation of accumulation of the impurity decrease of each tree in the random forest [24]. For the purposes of this project, the top n number of features as measured by its importance will be extracted and used to train and evaluate a separate random forest classifier to determine the effect of feature selection on performance. While this simple method of feature selection will be used, several other statistical approaches, such as chi-squared or ANOVA tests, are available for future evaluation.

2.2.3 Random Forest Classifiers

Random forest classifiers, made up of decision tree predictors, are models for ensemble learning that can be used for classification tasks. During training, multiple layers of decision trees are constructed based on randomly sampled vectors of input features [25]. Similar to majority voting, the final output of the random forest is the class chosen by the greatest number of trees. Random forests usually outperform traditional decision trees, which often overfit to their training dataset. Avoiding overfitting in *BRAF* differentiation is crucial as the model needs to be generalizable to unseen internal images or data from external institutions for clinical adoption.

2.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a type of neural network primarily used in image related tasks, such as classification, object detection, or segmentations. CNNs have convolutional and pooling layers in addition to the fully connected layers in artificial neural networks. The convolutional layers create feature maps of the input images by iterating a kernel over the image and extracting features, while pooling layers down-sample the convolution output [26]. The final flattening step transforms the data into a one-dimensional array, which is ultimately fed through fully connected layers for classification.

2.3.1 Adam Stochastic Optimization

The Adam optimizer is an established variation of the stochastic gradient descent algorithm used in many deep learning models. Developed at the University of Toronto, the algorithm updates weights of the network iteratively by combining advantages of the Adaptive Gradient algorithm, which sets a learning rate for each parameter, and the Root Mean Square Propagation algorithm, which adapts these learning rates based on the average of previous gradient magnitudes [27]. Adam optimization is proven to empirically converge faster with better results than preceding stochastic methods [27] and will be used to construct the CNN model in this project. For detailed mathematical theory, please consult Kingma and Ba, 2014 [27].

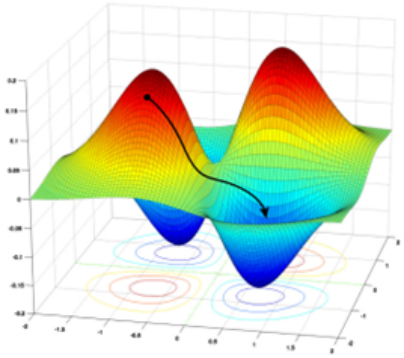

Figure 2.2: Visualization of the "loss" landscape travelled by the Adam stochastic optimization algorithm. Adam seeks to minimize loss by traversing the network and adjusting weights. Reproduced from Jin et al., 2015 [28].

2.3.2 Loss Criterion

In optimization, a loss criterion, or "error function", is a function that maps observations of variables to a real number representing "cost". Machine learning models seek to minimize this "cost" value. The criterion used in the neural networks of this project will be BCEWithLogitsLoss, a form of binary cross entropy loss measured between the input probabilities and target labels with an implicit sigmoid activation layer. The unreduced version is defined as:

$$\ell(x, y) = L = \{l_1, \ldots, l_N\}^\top, \quad l_n = -w_n \left[y_n \cdot \log \sigma(x_n) + (1 - y_n) \cdot \log(1 - \sigma(x_n)) \right], \tag{2.3}$$

where N is the batch size. More information can be found in the PyTorch documentation [29].

2.3.3 Deep CNNs & Limitations

Deep CNNs gained prominence in 2012 where additional layers of convolution were thought to progressively learn more complex features and thus perform better image learning [30]. However, as CNN architectures increased in depth, parameters that were thought to reduce overfitting, such as regularization parameters or dropout, began to be ineffective. Studies show that this reduction in performance could be attributed to the optimizer used or vanishing/exploding gradients [31]. Residual learning was hence developed to combat limitations of deep CNNs.

2.3.4 3D ResNet

Residual neural networks, or "ResNets", are deep neural networks that employ residual blocks or skip connections to bypass certain layers of the network. The layer skips address the issue of vanishing/exploding gradients and contain nonlinearities such as ReLU and batch normalization [11]. By implementing these skips in our deep CNN, any network layers that are disadvantageous to the performance of the architecture will be skipped by the parameter tuning process. An establish 3D version of the ResNet would be able to capture spatial information that exists in the slices of each MR image in the dataset and will be implemented for this project. For more information on the architecture, please refer to Section 3.5.1. Additionally, the architecture used in this project is based on Hara et al., 2017, whose work affirmed the 3D ResNet's ability to extract spatiotemporal features for action recognition [32].

In neural networks, activation functions are responsible for transforming the weighted inputs into outputs of a particular node. Two activation functions are of note for this project, the rectified linear activation function (ReLU) and the sigmoid activation function.

ReLU is a piecewise linear function that outputs the input directly for all positive inputs and 0 for all non-positive inputs [33]. The sigmoid activation function is defined as:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{2.4}$$

Both functions are used in a variety of neural networks and have respective ad-

vantages and disadvantages. The sigmoid function struggles in networks with many layers because of the vanishing gradient problem; however, due to it's mathematical simplicity and compatibility with binary classification, it will be used as a final activation function [33], [34]. In comparison, ReLU is able to overcome the vanishing gradient problem and is suitable for networks with many layers, such as the ResNet.

2.4 Model Implementation & Evaluation

2.4.1 Scikit-learn & PyTorch Packages

Scikit-learn is a Python package that integrates a wide variety of machine learning algorithms and models for various tasks [35]. This project will use Scikit-learn to implement and fit the radiomics random forest classifier, as well as to compute model performance. PyTorch is an open source dynamic machine learning framework that allows easy implementation of CNNs [36]. This project will use PyTorch to implement the 3D ResNet, to train the neural network, and to evaluate its performance through AUCs.

2.4.2 Overfitting

Overfitting is a frequent concern of many deep learning models as it prevents the generalization of the model to unseen data. In medicine, this generalization is crucial for clinical adoption as medical data is often varied across institutions. To prevent overfitting, dropout layers will be implemented for regularization [37]. Additionally, training will cease when the model validation loss fails to decrease for 3 epochs in a row.

2.4.3 Hyperparameter Tuning with GridSearchCV

GridSearchCV is a Scikit-learn functionality that allows for easy hyperparameter tuning of various deep learning models. Via a parameter grid defined by the user, GridSearchCV will implement, train, and evaluate all combinations of the grid to delineate the combination of hyperparameters that produces the model with the greatest performance [38]. The downfall is the time-consuming nature of the process; however, GridSearchCV will likely find a better performing model than if only a few parameter combinations were tested.

2.4.4 Nested Cross-Validation

k-fold cross-validation is often used to estimate the true performance of a machine learning model when making classification predictions. This can be useful when attempting to optimize the hyperparameters of a model or when comparing analogous models on the same classification task. By resampling the testing dataset into k folds and holding out one fold for validation in each iteration, the model is trained and evaluated with different combinations of samples, thus avoiding overestimates of model skill [39]. To prevent biases when the testing dataset is too hard or too easy, nested cross-validation will be used for an accurate representation of accuracy [40].

2.4.5 Widening Factor

The widening factor of neural networks control the width of the network architecture and can expand or reduce the number of channels in the CNN. A widening factor of 0.5 was used in this project that effectively speeds up the training of the ResNet by a factor of 2 without affecting the ultimate performance.

2.4.6 Area Under Receiver Operating Characteristic Curve

The receiver operating characteristic (ROC) curve is a measure of usefulness for a particular binary classification model. The curve accounts for biases caused by the dataset distribution between the two labels and is created by plotting the true positive rate (TPR) against the false positive rate (FPR) [41]. By computing the area under this resulting ROC curve (AUC), an estimation of the model performance can be inferred.

2.4.7 Mann-Whitney U Test

Comparing many models and methods is common in applied machine learning. As multiple methods are developed in this project, a statistical test is needed to compare significance in differences of model skill. The k-fold cross-validation procedure generates observations that are dependent, which violates a key assumption of the commonly used paired Student's t-test [42]. Additionally, other assumptions of the distribution of the AUCs, such as its Gaussian nature, cannot be made. Thus, an unpaired, nonparametric test is needed. The Mann-Whitney U test, which tests the

null hypothesis that the distribution underlying the two sample groups are equal, meets these requirements [42].

2.5 Radiomics & ResNet Fusion

2.5.1 Appending Features Approach

One possible approach of integrating radiomics and the ResNet model is by appending the radiomics features within the ResNet. Radiomics features extracted via PyRadiomics and ResNet image features extracted from the convolutional layers may work in harmony to achieve a higher AUC. The radiomics features would be implemented and fed into the fully connected layers of the ResNet during the feedforward segment.

2.5.2 Risk Score Approach

Another possible fusion method is referred to as the "risk score" approach. Zhang et al., 2021 developed an analytic pipeline combining radiomics and transfer learning features to predict overall survival in pancreatic ductal adenocarcinoma (PDAC) patients [12]. Zhang et al. fitted 2 separate random forest classifiers using two feature banks, one derived from PyRadiomics and the other extracted from an 8-layer CNN. The probabilities of death (output) from these two models were then referred to as risk scores and used to fit a final fusion random forest prognosis model. A 40% increase in AUC was observed (yielded 0.84), indicating the method to be a tremendous success in PDAC survival prediction [12].

A similar approach may be applied to this project by treating the output labels of the radiomics random forest classifier and the ResNet as risk scores. These outputs will then be inputted into a final fusion random forest classifier as features and evaluated, whereby a higher AUC may be achieved.

2.5.3 Feature Mapping

Feature mapping is the process of representing image features along with its relevancy or score on a graph. This method could be used to create a map of radiomics features which is then fed into a deep CNN for convolutional classification. Zhang and Qi et al. demonstrated the effectiveness of this method at increasing model accuracy when predicting EGFR gene mutation status of lung adenocarcinoma using

CT radiomics features [43]. The method could potentially be translated for MRI radiomics features in pLGGs.

2.5.4 Analytic Evidential Reasoning

Analytic evidential reasoning can be used to fuse the outputs of the radiomics and CNN models. This approach was used to combine many-objective radiomics and a 3D CNN to predict lymph node metastasis in head and neck cancer with successful results [14]. The original analytical evidential reasoning recursive algorithm is outlined by Wang et al., 2006 [44].

2.5.5 Soft Voting

Whitney et al, 2019 demonstrated an approach of using soft voting to construct a fusion classifier combining radiomics based methods and CNN based methods [15]. Soft voting involves summing up the weighted probability (based on the importance of the model that was used to generate the probability) of a particular input for each target class. The label corresponding to the greatest sum of weighted probabilities becomes the classification output. Whitney et al. observed a statistically significant improvement in performance when using this approach to classify breast tumour using MRI [15].

Chapter 3

Methodology

3.1 Data

3.1.1 Patient Cohort Criteria

The target dataset contained patients diagnosed with pLGG in 2 children's hospitals between January 2000 and December 2018. It is more than double the size of datasets used in previous studies. These patients have received a FLAIR MRI segmented by a licensed radiologist. The Hospital for Sick Children dataset contained a total of 397 patients, while the Stanford University dataset contained 41 patients, for a bi-institutional dataset totalling 438 patients. Of these, only patients who have a molecular signature identified to be *BRAF* fusion or *BRAF* V600E point mutation were included in this project.

3.1.2 The Hospital for Sick Children

The compiled dataset of The Hospital for Sick Children in Toronto consisted of 397 patients with tumours of varying pathological diagnoses. 143 patients exhibited a *BRAF* fusion molecular character, while 71 exhibited a *BRAF* V600E point mutation.

3.1.3 Stanford University

The compiled dataset of The Lucile Packard Children's Hospital at Stanford University consisted of 41 patients of the target molecular character. 32 patients were identified to have a *BRAF* fusion alteration while 9 patients had a *BRAF* V600E point mutation molecular character.

Table 3.1: Final number of patients included in dataset by gene alteration classes and institution.

	BRAF Fusion	*BRAF* V600E Mutation
The Hospital for Sick Children	143	71
Stanford University	32	9

3.1.4 MRI Sanity Check & Additional Patient Exclusions

MR images underwent validation where segmentations by radiologists were verified. Images with visual artefacts, such as areas of the brain that is unclear or cut off, were excluded from model training. Additionally, images with missing or unclear segmentations, or any form of corrupted data, were also excluded from this project.

3.1.5 Final Dataset Summary

MR images from a bi-institutional cohort of 255 pediatric patients were included. The cohort consisted of 136 boys and 119 girls, and has a mean age of 8.62 years. pLGGs had varying pathological diagnoses and clinical parameters, such as tumour location in the brain.

3.2 Data Preprocessing

3.2.1 Clinical Data

The dataset contained clinical information of all patients, including gender, age, tumour location, such as supratentorial, infratentorial, or trans-tentorial, and whether the tumour exists in multiple locations. These characteristics could assist with classification; however, to simplify testing of fusion models, they were excluded at this stage. To prepare for future use, genders were binarily encoded, ages were reformatted to datetime format, and tumour locations were one-hot encoded.

3.2.2 Label Encoding

To aid in classification, the ground truth labels, *BRAF* fusion and *BRAF* V600E, were encoded binarily in the dataset. A ‘0’ indicates *BRAF* fusion, whereas ‘1’ indicates *BRAF* V600E point mutation.

3.2.3 Image Registration & Normalization

Raw 3-Dimensional MRI of each patient were of varying sizes. As such, all images were registered to the SR124 atlas to ensure that all images being fed into models were on the same coordinate system. Bias correction and normalization between 0 and 1 via the NumPy Python package was also applied to increase computation efficiency and time, as well as to remove inconsistencies and avoid biases in the data from pixel intensity variations across FLAIR sequences from different sources.

3.2.4 Radiomics Features

Radiomics features were extracted from preprocessed MRI using PyRadiomics. The set of extracted radiomics features from the original image (i.e. no filters applied) are available in Appendix B (total of 107).

Extracted features required additional unsupervised filtering to account for various biases and potential for errors. Groups of highly correlated features will increase the complexity of the model without bringing substantial additional predictive information. As this increases the risk of errors, highly correlated features with a correlation factor greater than 0.95 were removed. A baseline method of feature selection known as variance thresholding was also applied to remove all zero-variance and low-variance ($correlation < 0.05$) features, such as features that exhibit the same value in all samples across the dataset. Figure 3.1 depicts the feature correlation maps before and after filtering. Removed features were set to 0 and spliced out prior to model fitting.

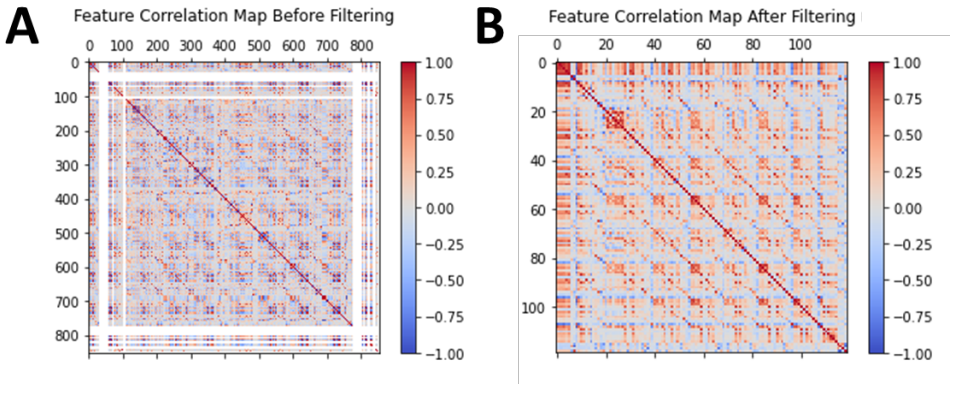

Figure 3.1: Representative correlation map of radiomics features. a) Correlation map of 851 radiomics features before preprocessing. b) Correlation map of radiomics features after preprocessing where highly correlated and zero-variance or low-variance features are removed.

Additionally, naïve feature selection was implemented, which selects for the top $n = 50$ important features according to the `feature_importances_` attribute available in random forest models. The effect of this method was evaluated, and results are shown in Section 4.1.1.

3.3 Data Allocation

Figure 3.2 summarizes the breakdown for the complete compiled dataset undergoing 5-fold nested cross-validation. The Hospital for Sick Children "internal" dataset was randomly split for each iteration n, such that 75% of that dataset was referred to as the "development" set. The remaining 25% was known as the "evaluation" set. The development dataset was used for training and validation (ie. 5-folds dataset split during nested cross-validation), while the evaluation dataset contained images that were unseen to the model and was used to evaluate model skill. Patients in the Stanford "external" dataset is allocated to be used for external testing only. As external images were not fully available at the time of this project, they will be evaluated by developed models in the future. Thus, only the complete set of internal data from The Hospital for Sick Children was used for training and evaluation, with results summarized in the subsequent sections.

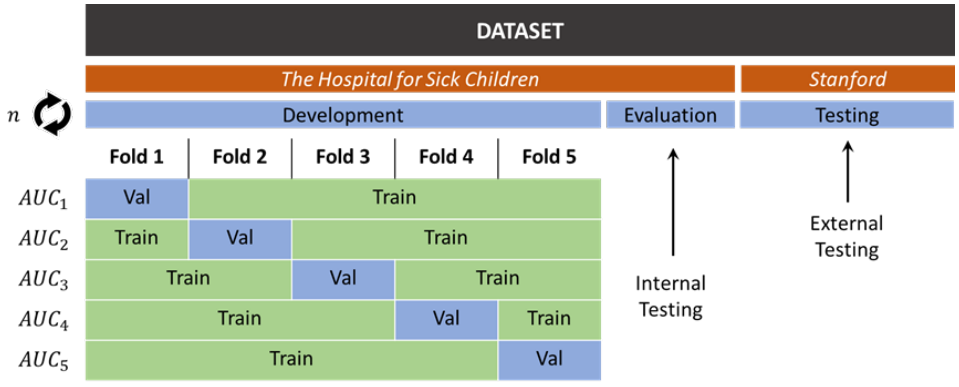

Figure 3.2: Dataset split for development, internal testing, and external testing. 75% of the data from The Hospital for Sick Children was used as the development dataset where 5-folds cross-validation was implemented. The remaining 25% is known as internal testing, whereas external testing refers to the unseen testing set of data from Stanford. A random seed is set to split the data for each of the n iterations and noted for reproducibility.

The mean AUC may be calculated by:

$$Mean\ AUC = \frac{1}{5}\sum_{k=1}^{5} AUC_k \tag{3.1}$$

3.4 Radiomics Random Forest Classifier

A similar radiomics model illustrated by Wagner et al., 2021 [9] was developed using Scikit-learn. Fitting and prediction were completed on remote workstations and High-Performance Computing (HPC) clusters available at The Hospital for Sick Children.

3.4.1 Model Fitting & Evaluation

The development dataset, containing a random 75% of the internal patients (161 patients), was used for model fitting. The evaluation dataset, containing the remaining 25% of the internal patients (53 patients), was reserved for internal testing.

5-fold nested cross-validation was conducted. GridSearchCV was applied over the set of model parameters outlined in Table 3.2. The model with the best performance as outlined by Scikit-learn's `best_estimator_` attribute was selected and used to evaluate the internal testing dataset. This process was repeated twice, once with naïve feature selection and once without.

Table 3.2: GridSearchCV parameter grid for the radiomics random forest classifier.

Hyperparameters	*Candidate Values*
n_estimators	25, 50, 100
min_samples_leaf	2, 4, 8
max_depth	1, 2, 4
max_features	'auto', None
max_samples	0.5, 0.75, 1

3.5 3D ResNet

A 3D ResNet model was implemented using PyTorch. Training and evaluation were completed on remote workstations and HPC clusters available at The Hospital for Sick Children.

3.5.1 Architecture

A 3D ResNet architecture with skip connections was implemented according to Hara et al., 2017 [32]. Normalized tensors of MR images were fed through a combination of residual blocks and dropout functions (see Figure 3.3). Each residual block contains 2 sequences of a 3D convolution followed by batch normalization and ReLU. Disadvantageous network layers are skipped during parameter tuning, addressing the vanishing gradient problem.

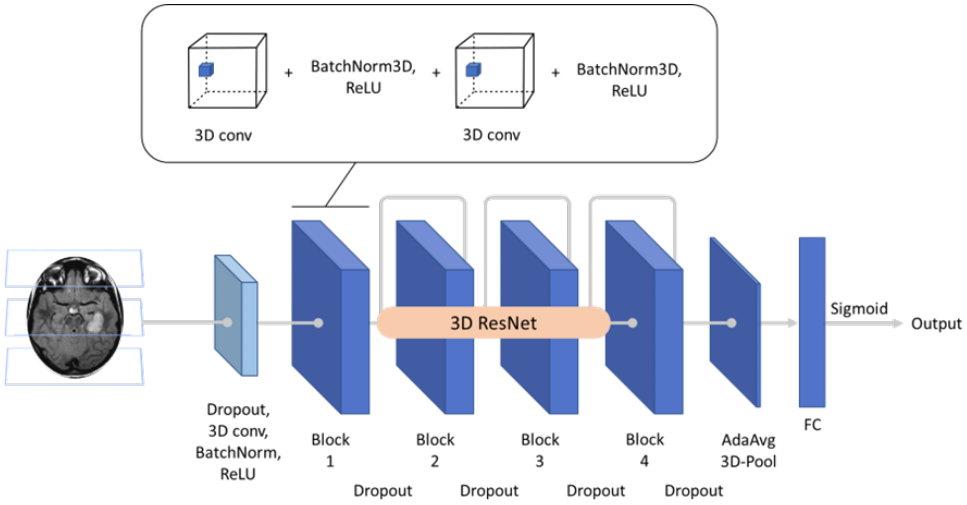

Figure 3.3: Implemented 3D ResNet architecture. Convolutional, pooling, and fully connected layers are depicted along with the breakdown for each residual block in the ResNet architecture.

The loss function used was BCEWithLogitsLoss. Adaptive average 3D pooling was used prior to the fully connected layers and the sigmoid activation function was used to produce the output. A widening factor and dropout rate of 0.5 was also incorporated.

3.5.2 Model Training & Evaluation

The 3D ResNet was trained with an optimizer learning rate of 0.01 and batch size of 8. The model was allowed to be trained for 15 epochs; however, to increase efficiency and to prevent overfitting, training ceased when the validation loss failed to decrease for 3 epochs in a row. Checkpointing was done for each epoch such that the model performance was determined by testing with the evaluation dataset.

3.6 Fusion Approaches

3.6.1 Appending Features

The appending features approach involves appending radiomics features to the fully connected layers of the 3D ResNet in the hopes of providing additional predictive information for the classifier to use. The model was developed using Scikit-learn and training/evaluation were completed on remote workstations available at The Hospital for Sick Children.

The 3D ResNet was implemented and trained using methods and parameters outlined in Section 3.5. The post-processed PyRadiomics-extracted features of each respective patient was appended to the first fully connected layer (Figure 3.4). An additional fully connected layer was also added with dropout prior to classification output.

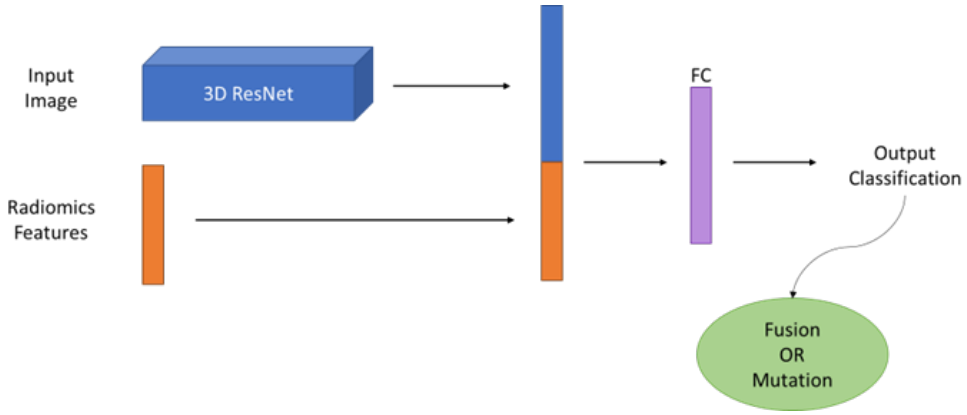

Figure 3.4: Schematic of the appending features fusion approach. Radiomics features are appended to the fully connected layer of the neural network as seen.

3.6.2 Risk Score

As seen in Figure 3.5, the risk score approach was implemented as described by Zhang et al., 2021, with the radiomics random forest model and 3D ResNet implemented and trained using methods and parameters outlined in the above sections [12].

Conforming with the nested cross-validation data allocation scheme, patients in the development dataset was split into 5 folds. Throughout the training process, each of these folds acted as the validation set at some point for both the radiomics random forest classifier and the 3D ResNet, where outputs were generated by re-

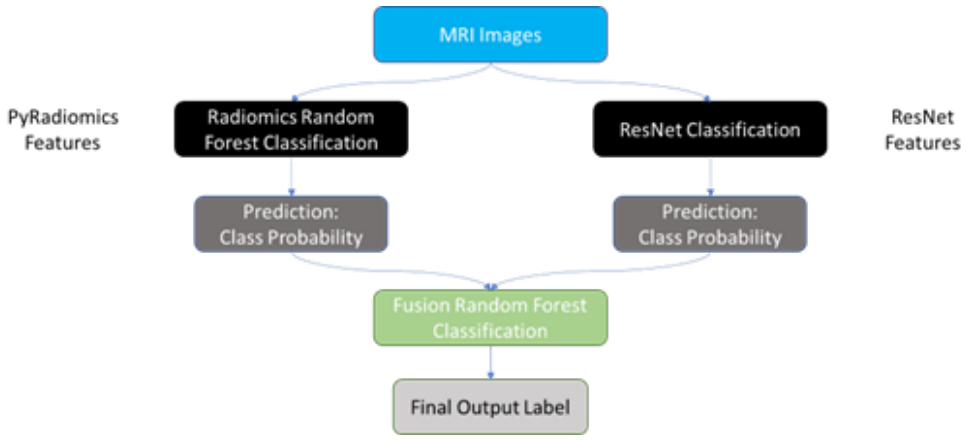

Figure 3.5: Schematic of the pipeline for the risk score fusion approach. The Radiomics random forest classifier and the 3D ResNet are trained separately. Outputs are then fused as inputs into the final fusion random forest model.

spective models trained using data from the other 4 folds. By doing so, a prediction from the radiomics random forest classifier and a prediction from the 3D ResNet is available for each patient in the development dataset. These predictions acted as the input features of a final fusion random forest model that classifies *BRAF* status, where another 5-folds cross-validation and GridSearchCV was conducted (hyperparameters outlined in Table 3.3). This same pipeline was used on the evaluation dataset to evaluate whole model skill.

Table 3.3: GridSearchCV parameter grid for the final fusion random forest model.

Hyperparameters	*Candidate Values*
n_estimators	5, 10, 25
min_samples_leaf	1, 2, 4
max_depth	1, 2
max_features	None
max_samples	0.5, 0.75, 1

Chapter 4

Results

4.1 Radiomics Random Forest Classifier

4.1.1 Naïve Feature Selection

Table 4.1 and Figure 4.1 summarizes the AUCs achieved by the radiomics random forest classifier over 30 trials and the effect of naïve feature selection. AUCs were determined using Scikit-learn's `roc_auc_score` and applied to predictions from the radiomics random forest classifier and labels of the respective datasets. 95% confidence intervals are indicated in brackets. Naïve feature selection showed minimal, non-significant improvement in AUCs across the datasets. An internal testing AUC of 0.877 was achieved with a 95% confidence interval of 0.865 to 0.893.

Table 4.1: Summary of average AUCs achieved by the radiomics random forest classifier over 30 trials with and without naïve feature selection.

	Without Naive Feature Selection	With Naive Feature Selection
Training	**0.959** (95% CI, 0.949-0.968)	**0.963** (95% CI, 0.947-0.978)
Internal Testing	**0.876** (95% CI, 0.863-0.890)	**0.877** (95% CI, 0.865-0.893)

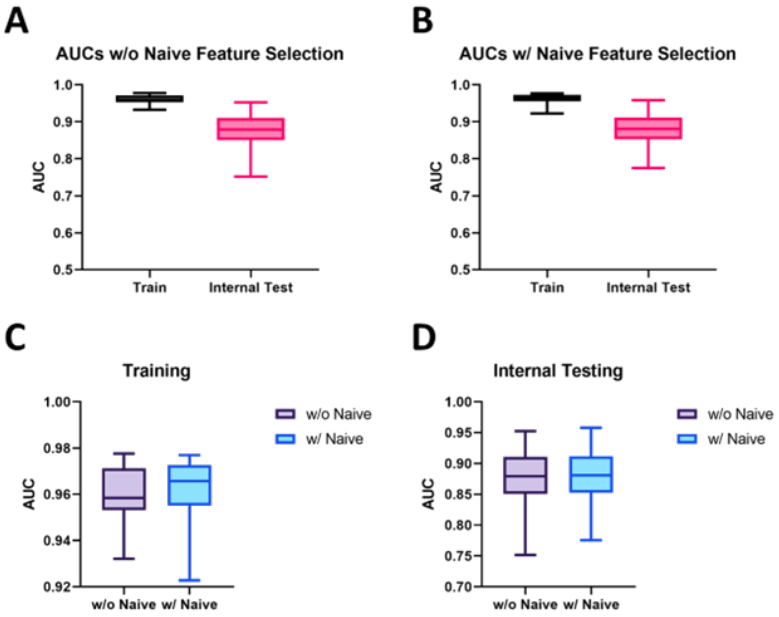

Figure 4.1: Radiomics random forest classifier results over 30 trials. a) AUC of training and internal testing without naïve feature selection. b) AUC of training and internal testing with naive feature selection. c) Training AUC without and with naïve feature selection. d) Internal testing AUC without and with naïve feature selection. Naïve feature selection showed minimal increase in AUC; however, the change is not significant. Results in boxes are presented as means and interquartile range.

4.1.2 Summary of AUC

Table 4.2: Average training and internal testing AUCs achieved by the radiomics random forest classifier over 30 trials.

	AUC
Training	**0.963** (95% CI, 0.947-0.978)
Internal Testing	**0.877** (95% CI, 0.865-0.893)

4.2 3D ResNet

4.2.1 Training Curves & Loss

Figure 4.2 illustrates representative training and internal testing error/loss curves over epochs. AUCs were determined using Scikit-learn's `roc_auc_score` and ap-

plied to predictions from the 3D ResNet and labels of the respective datasets. Loss value was measured using the network criterion. On average, training ceased around epoch 12, where validation loss often sees an increase for 3 epochs in a row. The optimal epoch is between the range of 8-10 on average and the model corresponding to this epoch is used to determine model performance using the evaluation dataset (internal testing).

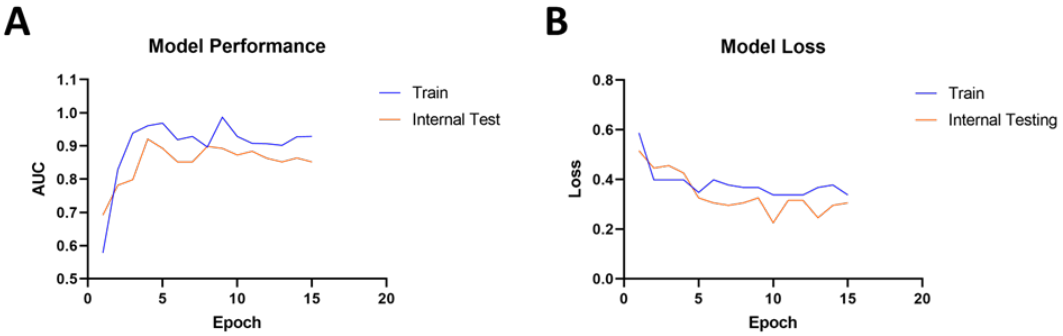

Figure 4.2: Representative training curves of the 3D ResNet classifier. a) Model performance of training and internal testing datasets as measured by AUC over 15 epochs. b) Model loss of training and internal testing datasets over 15 epochs.

Table 4.3 summarizes average loss measured by the 3D ResNet criterion over 15 trials. 95% confidence intervals are indicated in brackets.

Table 4.3: Average loss measured by the 3D ResNet classifier over 15 trials.

	Loss
Training	**0.382** (95% CI, 0.368-0.396)
Internal Testing	**0.341** (95% CI, 0.293-0.389)

4.2.2 Summary of AUC

Table 4.4: Average training and internal testing AUCs achieved by the 3D ResNet classifier over 15 trials.

	AUC
Training	**0.903** (95% CI, 0.886-0.919)
Internal Testing	**0.853** (95% CI, 0.842-0.863)

4.3 Appending Features Fusion Approach

AUCs achieved by the appending features fusion approach were measured over 15 trials. Likewise to the 3D ResNet, AUC was determined using Scikit-learn's `roc_auc_score` and applied to predictions from the 3D ResNet post-feature appending and labels of the respective datasets. 95% confidence intervals are indicated in brackets. An internal testing AUC of 0.888 was achieved with a 95% confidence interval of 0.875 to 0.901.

4.3.1 Summary of AUC

Table 4.5: Average training and internal testing AUCs achieved by the appending features fusion approach model over 15 trials.

	AUC
Training	**0.900** (95% CI, 0.872-0.929)
Internal Testing	**0.888** (95% CI, 0.875-0.901)

4.4 Risk Score Fusion Approach

AUCs achieved by the risk score fusion approach were measured over 15 trials. AUC was determined using Scikit-learn's `roc_auc_score` and applied to predictions of the final fusion classifier and labels of the respective datasets. 95% confidence intervals are indicated in brackets. An internal testing AUC of 0.899 was achieved with a 95% confidence interval of 0.891 to 0.908.

4.4.1 Summary of AUC

Table 4.6: Average training and internal testing AUCs achieved by the risk score fusion approach model over 15 trials.

	AUC
Training	**0.918** (95% CI, 0.902-0.934)
Internal Testing	**0.899** (95% CI, 0.891-0.908)

4.5 Overview of Methods

Table 4.7 provides an overview of the training and internal testing AUCs achieved by each of the methods outlined in this project. The highest internal testing validation was achieved by the risk score fusion method at 0.899.

Table 4.7: Overview of internal testing AUCs by method. 95% confidence intervals are indicated in brackets.

	Risk Score	Append	ResNet	Radiomics
Training	**0.918** (95% CI, 0.902-0.934)	**0.900** (95% CI, 0.872-0.929)	**0.903** (95% CI, 0.886-0.919)	**0.963** (95% CI, 0.947-0.978)
Internal Testing	**0.899** (95% CI, 0.891-0.908)	**0.888** (95% CI, 0.875-0.901)	**0.853** (95% CI, 0.842-0.863)	**0.877** (95% CI, 0.865-0.893)

4.5.1 Statistical Evaluation

Figure 4.3 shows the statistical comparison between the internal testing AUCs achieved by the outlined methods. The Mann-Whitney U test was used. The risk score fusion approach showed a significant improvement in performance as compared to the standalone 3D ResNet ($p < 0.0001$), and the standalone radiomics random forest classifier ($p = 0.0429$). The appending features fusion approach showed a significant improvement in performance as compared to the standalone 3D ResNet ($p = 0.0007$), but not the standalone radiomics random forest classifier. The differences between all other unmentioned comparisons were also not significant.

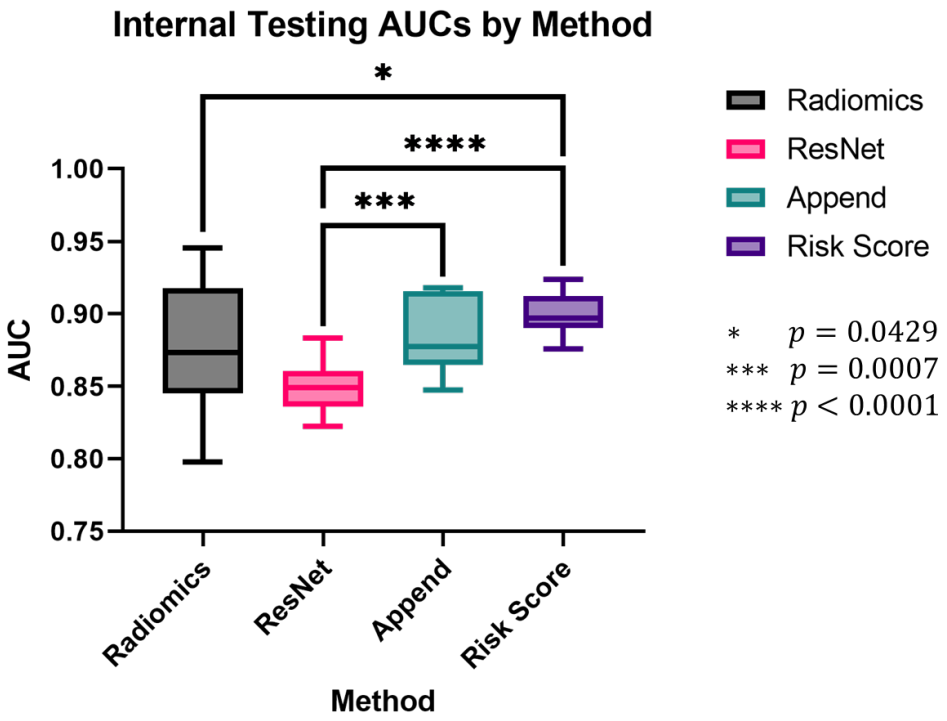

Figure 4.3: Overview of internal testing AUCs achieved by method. Risk score fusion approach showed a significant difference in performance as compared the to standalone radiomics random forest classifier or the 3D ResNet. Results in boxes are presented as means and interquartile range.

Chapter 5

Discussion & Future Directions

Several methods of *BRAF* status differentiation have been illustrated in this project. While each approach varies in their algorithm complexity and structure, their performances indicate predictive signatures and biomarkers of pLGG *BRAF* status were successfully extracted from MR images.

Molecular signatures of pLGG are currently assessed through tumor tissue analysis via biopsies wherever possible. Patients with nonresectable tumors are submitted to surgical procedures instead of targeted precision medicine therapy. The heterogenous nature of pLGG tumours was observed on the FLAIR sequences used in this project, where qualitative factors of the tumour, such as its volume, shape, and location, varied as expected. However, MR image analysis is better equipped to accommodate the heterogeneity as opposed to biopsies. As such, the analysis of MR images is pivotal in identifying pLGG molecular character for improved prognostication.

5.1 External Dataset Evaluation

As stated in Section 3.3, external testing images from Stanford University were not yet available at the time of experiments. As such, the models' generalizability to data from external institutions was not evaluated, which is desirable for clinical adoption. An immediate future step is to complete the compilation and preprocessing of external institutional data, and complete preparation such that they will be ready for evaluation by each of the illustrated methods.

5.2 Radiomics Feature Selection

As expected, based on previous relevant studies, radiomics features were predictors of mutational status. However, further investigation of additional feature selection methods would be advantageous. While naive feature selection was implemented in this project, it did not have a significant effect on model performance. As feature selection is regarded as a crucial step in applied machine learning, other methods of selection, such as using Recursive Feature Elimination (RFE) [45], should be explored. Statistical methods of measuring feature correlation, such as via chi-squared or ANOVA tests can also be explored for feature scoring [46].

5.3 Random Forest Classifier Optimization

In addition to evaluating the effect of various feature selection methods to optimize the random forest classifier, additional hyperparameter tuning is warranted. Continued use of GridSearchCV will be needed to arrive at an optimized model capable of yielding maximum AUC. Other criteria, such as the Gini index, can also be explored as a replacement for the entropy criterion [47].

5.4 3D ResNet Optimization

Due to the complex nature of ResNet models, many adjustments are possible to optimize performance. Elementary training hyperparameters, such as batch size, learning rate, and epochs can be further tuned to pinpoint the optimal combination. Additionally, architectural modifications, either to model depth or convolutional layer dimensions, can be explored for improvements in performance. Further testing is also mandated to elucidate the true performance of the 3D ResNet. More trials should be executed to arrive at larger sample sizes of AUCs for greater confidence in performance.

5.5 Fusion Optimization

Based on the results of the 2 fusion models, approaches to combining radiomics and convolutional neural networks seem to be feasible pathways to improve classification performance for *BRAF* differentiation. Notably, the risk score approach generates results having lower bounds that are greater or on par with the averages of other

methods. However, it is at a disadvantage in terms of the training process due to its model pipeline complexity. Specifically, the risk score model was much more computationally expensive and time consuming to train and evaluate. As such, there is a need to evaluate the trade-off between performance and model pipeline complexity. Perhaps it is viable to reduce the number of hyperparameters or channels/layers involved without dramatically affecting performance.

5.6 Integration of Clinical Features

As clinical features were set aside for these experiments, the effect on performance by the presence of clinical features cannot be concluded. Radiological features, such as clinical tumour location, and patient demographic information, such as age, may hold predictive power and should be analyzed. Additionally, pathological diagnoses of these pLGGs were not examined. The skewed nature of the external testing dataset in terms of pathological diagnoses warrants further investigation on its predictability.

A restrictive approach may be adopted for future studies such that only a particular pathological diagnosis or anatomical location is analyzed and tested for predictability of *BRAF* status [9]. However, due to the limited sample size and the low prevalence of certain diagnoses, these studies will likely be limited to multi-institutional collaborations.

5.7 MRI Sequences

The retrospective nature of the dataset contained acquired images from varying MRI vendors and parameters, leading to heterogeneity in the FLAIR MRI. The technical variation in images is representative of real-world clinical practice and was included in the attempt to develop robust models. To that end, model predictability to classify other MRI sequences, such as T2WI, DWI, and contrast enhanced T1WI, could be explored [9].

5.8 Multi-class Classification

While the two most common pLGG molecular alterations were the focus of this project, other characters relating to the NF1 (neurofibromatosis 1), CDKN2A (cyclin-dependent kinase inhibitor 2A), and FGFR (fibroblast growth factor receptor) genes

are also present [48]–[50]. As such, multi-class classification may be investigated to incorporate the other molecular signatures of pLGGs that were currently excluded.

Chapter 6

Conclusion

This project presents radiomics and residual neural network methods for the prediction of *BRAF* status in pLGGs using FLAIR MR images. The optimal model developed was the risk score fusion approach which achieved an AUC of 0.899 on the internal testing cohort. This fusion classification model will expand the current understanding of the correlation between molecular markers and imaging features and will enable future machine learning studies of pLGG MRI. With continued improvements in performance and increases in AUCs, radiomics and machine learning-based differentiation may be adopted for prediction of *BRAF* status in clinical pathways and has the potential of enhancing patient prognosis and outcomes.

Appendix A

Code Repository

The code repository corresponding to this project is available on Github at https://github.com/IMICSLab/pLGG_Predict_Status.

Appendix B

Original Image Radiomics Features

#	**Pyradiomics Class**	**Feature Name**
1	Shape	VoxelVolume
2	Shape	Maximum3DDiameter
3	Shape	MeshVolume
4	Shape	MajorAxisLength
5	Shape	Sphericity
6	Shape	LeastAxisLength
7	Shape	Elongation
8	Shape	SurfaceVolumeRatio
9	Shape	Maximum2DDiameterSlice
10	Shape	Flatness
11	Shape	SurfaceArea
12	Shape	MinorAxisLength
13	Shape	Maximum2DDiameterColumn
14	Shape	Maximum2DDiameterRow
15	Gray Level Difference Matrix	GrayLevelVariance
16	Gray Level Difference Matrix	HighGrayLevelEmphasis
17	Gray Level Difference Matrix	DependenceEntropy
18	Gray Level Difference Matrix	DependenceNonUniformity
19	Gray Level Difference Matrix	GrayLevelNonUniformity
20	Gray Level Difference Matrix	SmallDependenceEmphasis
21	Gray Level Difference Matrix	SmallDependenceHighGrayLevelEmphasis
22	Gray Level Difference Matrix	DependenceNonUniformityNormalized
23	Gray Level Difference Matrix	LargeDependenceEmphasis
24	Gray Level Difference Matrix	LargeDependenceLowGrayLevelEmphasis
25	Gray Level Difference Matrix	DependenceVariance
26	Gray Level Difference Matrix	LargeDependenceHighGrayLevelEmphasis
27	Gray Level Difference Matrix	SmallDependenceLowGrayLevelEmphasis

Table B.1 continued from previous page

#	Pyradiomics Class	Feature Name
28	Gray Level Difference Matrix	LowGrayLevelEmphasis
29	Gray Level Co-Occurrence Matrix	JointAverage
30	Gray Level Co-Occurrence Matrix	SumAverage
31	Gray Level Co-Occurrence Matrix	JointEntropy
32	Gray Level Co-Occurrence Matrix	ClusterShade
33	Gray Level Co-Occurrence Matrix	MaximumProbability
34	Gray Level Co-Occurrence Matrix	Idmn
35	Gray Level Co-Occurrence Matrix	JointEnergy
36	Gray Level Co-Occurrence Matrix	Contrast
37	Gray Level Co-Occurrence Matrix	DifferenceEntropy
38	Gray Level Co-Occurrence Matrix	InverseVariance
39	Gray Level Co-Occurrence Matrix	DifferenceVariance
40	Gray Level Co-Occurrence Matrix	Idn
41	Gray Level Co-Occurrence Matrix	Idm
42	Gray Level Co-Occurrence Matrix	Correlation
43	Gray Level Co-Occurrence Matrix	Autocorrelation
44	Gray Level Co-Occurrence Matrix	SumEntropy
45	Gray Level Co-Occurrence Matrix	MCC
46	Gray Level Co-Occurrence Matrix	SumSquares
47	Gray Level Co-Occurrence Matrix	ClusterProminence
48	Gray Level Co-Occurrence Matrix	Imc2
49	Gray Level Co-Occurrence Matrix	Imc1
50	Gray Level Co-Occurrence Matrix	DifferenceAverage
51	Gray Level Co-Occurrence Matrix	Id
52	Gray Level Co-Occurrence Matrix	ClusterTendency
53	Histogram	InterquartileRange
54	Histogram	Skewness
55	Histogram	Uniformity
56	Histogram	Median
57	Histogram	Energy
58	Histogram	RobustMeanAbsoluteDeviation
59	Histogram	MeanAbsoluteDeviation
60	Histogram	TotalEnergy
61	Histogram	Maximum
62	Histogram	RootMeanSquared
63	Histogram	90Percentile
64	Histogram	Minimum
65	Histogram	Entropy
66	Histogram	Range
67	Histogram	Variance
68	Histogram	10Percentile

Table B.1 continued from previous page

#	Pyradiomics Class	Feature Name
69	Histogram	Kurtosis
70	Histogram	Mean
71	Gray Level Run-Length Matrix	ShortRunLowGrayLevelEmphasis
72	Gray Level Run-Length Matrix	GrayLevelVariance
73	Gray Level Run-Length Matrix	LowGrayLevelRunEmphasis
74	Gray Level Run-Length Matrix	GrayLevelNonUniformityNormalized
75	Gray Level Run-Length Matrix	RunVariance
76	Gray Level Run-Length Matrix	GrayLevelNonUniformity
77	Gray Level Run-Length Matrix	LongRunEmphasis
78	Gray Level Run-Length Matrix	ShortRunHighGrayLevelEmphasis
79	Gray Level Run-Length Matrix	RunLengthNonUniformity
80	Gray Level Run-Length Matrix	ShortRunEmphasis
81	Gray Level Run-Length Matrix	LongRunHighGrayLevelEmphasis
82	Gray Level Run-Length Matrix	RunPercentage
83	Gray Level Run-Length Matrix	LongRunLowGrayLevelEmphasis
84	Gray Level Run-Length Matrix	RunEntropy
85	Gray Level Run-Length Matrix	HighGrayLevelRunEmphasis
86	Gray Level Run-Length Matrix	RunLengthNonUniformityNormalized
87	Gray Level Size Zone Matrix	GrayLevelVariance
88	Gray Level Size Zone Matrix	ZoneVariance
89	Gray Level Size Zone Matrix	GrayLevelNonUniformityNormalized
90	Gray Level Size Zone Matrix	SizeZoneNonUniformityNormalized
91	Gray Level Size Zone Matrix	SizeZoneNonUniformity
92	Gray Level Size Zone Matrix	GrayLevelNonUniformity
93	Gray Level Size Zone Matrix	LargeAreaEmphasis
94	Gray Level Size Zone Matrix	SmallAreaHighGrayLevelEmphasis
95	Gray Level Size Zone Matrix	ZonePercentage
96	Gray Level Size Zone Matrix	LargeAreaLowGrayLevelEmphasis
97	Gray Level Size Zone Matrix	LargeAreaHighGrayLevelEmphasis
98	Gray Level Size Zone Matrix	HighGrayLevelZoneEmphasis
99	Gray Level Size Zone Matrix	SmallAreaEmphasis
100	Gray Level Size Zone Matrix	LowGrayLevelZoneEmphasis
101	Gray Level Size Zone Matrix	ZoneEntropy
102	Gray Level Size Zone Matrix	SmallAreaLowGrayLevelEmphasis
103	Neighborhood Gray Tone Difference Matrix	Coarseness
104	Neighborhood Gray Tone Difference Matrix	Complexity
105	Neighborhood Gray Tone Difference Matrix	Strength
106	Neighborhood Gray Tone Difference Matrix	Contrast
107	Neighborhood Gray Tone Difference Matrix	Busyness

References

[1] J. AlRayahi, M. Zapotocky, V. Ramaswamy, *et al.*, "Pediatric brain tumor genetics: What radiologists need to know," *Radiographics: A Review Publication of the Radiological Society of North America, Inc*, vol. 38, no. 7, pp. 2102–2122, Dec. 2018, ISSN: 1527-1323. DOI: 10.1148/rg.2018180109.

[2] Q. T. Ostrom, P. M. de Blank, C. Kruchko, *et al.*, "Alex's lemonade stand foundation infant and childhood primary brain and central nervous system tumors diagnosed in the united states in 2007–2011," *Neuro-Oncology*, vol. 16, pp. x1–x36, suppl_10 Jan. 1, 2015, ISSN: 1522-8517. DOI: 10.1093/neuonc/nou327. [Online]. Available: https://doi.org/10.1093/neuonc/nou327 (visited on 07/15/2022).

[3] S. Ryall, U. Tabori, and C. Hawkins, "Pediatric low-grade glioma in the era of molecular diagnostics," *Acta Neuropathologica Communications*, vol. 8, no. 1, p. 30, Mar. 12, 2020, ISSN: 2051-5960. DOI: 10.1186/s40478-020-00902-z. [Online]. Available: https://doi.org/10.1186/s40478-020-00902-z (visited on 07/15/2022).

[4] G. T. Armstrong, H. M. Conklin, S. Huang, *et al.*, "Survival and long-term health and cognitive outcomes after low-grade glioma," *Neuro-Oncology*, vol. 13, no. 2, pp. 223–234, Feb. 1, 2011, ISSN: 1522-8517. DOI: 10.1093/neuonc/noq178. [Online]. Available: https://doi.org/10.1093/neuonc/noq178 (visited on 07/15/2022).

[5] A. Lassaletta, M. Zapotocky, M. Mistry, *et al.*, "Therapeutic and prognostic implications of BRAF v600e in pediatric low-grade gliomas," *Journal of Clinical Oncology*, vol. 35, no. 25, pp. 2934–2941, Sep. 2017, Publisher: Wolters Kluwer, ISSN: 0732-183X. DOI: 10.1200/JCO.2016.71.8726. [Online]. Available: https://ascopubs.org/doi/full/10.1200/JCO.2016.71.8726 (visited on 07/15/2022).

[6] D. R. Hargrave, E. Bouffet, U. Tabori, *et al.*, "Efficacy and safety of dabrafenib in pediatric patients with BRAF v600 mutation-positive relapsed or refractory low-grade glioma: Results from a phase i/IIa study," *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, vol. 25, no. 24, pp. 7303–7311, Dec. 15, 2019, ISSN: 1557-3265. DOI: 10.1158/1078-0432.CCR-19-2177.

[7] C. Rouleau, P. Bandopadhayay, L. Ramkissoon, and R. Beroukhim, "LGG-06. FUNCTIONAL AND THERAPEUTIC RELEVANCE OF MYB-QKI-INDUCED TRANSCRIPTIONAL EVENTS TO GLIOMAGENESIS," *Neuro-Oncology*, vol. 19, pp. iv34–iv34, Jun. 1, 2017. DOI: 10.1093/neuonc/nox083.139.

[8] S. N. McNulty, K. E. Schwetye, C. Ferguson, *et al.*, "BRAF mutations may identify a clinically distinct subset of glioblastoma," *Scientific Reports*, vol. 11, no. 1, p. 19 999, Oct. 8, 2021, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-021-99278-w. [Online]. Available: https://www.nature.com/articles/s41598-021-99278-w (visited on 07/15/2022).

[9] M. W. Wagner, N. Hainc, F. Khalvati, *et al.*, "Radiomics of pediatric low-grade gliomas: Toward a pretherapeutic differentiation of BRAF-mutated and BRAF-fused tumors," *American Journal of Neuroradiology*, vol. 42, no. 4, pp. 759–765, Apr. 1, 2021, Publisher: American Journal of Neuroradiology Section: Pediatrics, ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A6998. [Online]. Available: http://www.ajnr.org/content/42/4/759 (visited on 07/14/2022).

[10] E. Calabrese, J. D. Rudie, A. M. Rauschecker, *et al.*, "Combining radiomics and deep convolutional neural network features from preoperative MRI for predicting clinically relevant genetic biomarkers in glioblastoma," *Neuro-Oncology Advances*, vol. 4, no. 1, vdac060, Jan. 1, 2022, ISSN: 2632-2498. DOI: 10.1093/noajnl/vdac060. [Online]. Available: https://doi.org/10.1093/noajnl/vdac060 (visited on 09/12/2022).

[11] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, Dec. 10, 2015. DOI: 10.48550/arXiv.1512.03385. arXiv: 1512.03385[cs]. [Online]. Available: http://arxiv.org/abs/1512.03385 (visited on 07/15/2022).

[12] Y. Zhang, E. M. Lobo-Mueller, P. Karanicolas, S. Gallinger, M. A. Haider, and F. Khalvati, "Improving prognostic performance in resectable pancreatic ductal adenocarcinoma using radiomics and deep learning features fusion in CT images," *Scientific Reports*, vol. 11, no. 1, p. 1378, Jan. 14, 2021, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-

021-80998-y. [Online]. Available: https://www.nature.com/articles/s41598-021-80998-y (visited on 09/12/2022).

[13] Z. Baihua, S. Qi, X. Pan, *et al.*, "Deep CNN model using CT radiomics feature mapping recognizes EGFR gene mutation status of lung adenocarcinoma," *Frontiers in Oncology*, vol. 10, Feb. 12, 2021. DOI: 10.3389/fonc.2020.598721.

[14] L. Chen, Z. Zhou, D. Sher, *et al.*, "Combining many-objective radiomics and 3d convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer," *Physics in Medicine and Biology*, vol. 64, no. 7, Mar. 29, 2019, ISSN: 0031-9155. DOI: 10.1088/1361-6560/ab083a. [Online]. Available: http://www.scopus.com/inward/record.url?scp=85063935032&partnerID=8YFLogxK (visited on 07/15/2022).

[15] H. M. Whitney, H. Li, Y. Ji, P. Liu, and M. L. Giger, "Comparison of breast MRI tumor classification using human-engineered radiomics, transfer learning from deep convolutional neural networks, and fusion methods," *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, vol. 108, no. 1, pp. 163–177, Jan. 2020, ISSN: 0018-9219. DOI: 10.1109/jproc.2019.2950187. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8152568/ (visited on 05/21/2022).

[16] J. Fangusaro and P. Bandopadhayay, "The "risk" in pediatric low-grade glioma," *Cancer Cell*, vol. 37, no. 4, pp. 424–425, Apr. 13, 2020, ISSN: 1535-6108. DOI: 10.1016/j.ccell.2020.03.020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1535610820301604 (visited on 07/15/2022).

[17] D. Sturm, S. M. Pfister, and D. T. Jones, "Pediatric gliomas: Current concepts on diagnosis, biology, and clinical management," *Journal of Clinical Oncology*, vol. 35, no. 21, pp. 2370–2377, Jul. 20, 2017, Publisher: Wolters Kluwer, ISSN: 0732-183X. DOI: 10.1200/JCO.2017.73.0242. [Online]. Available: https://ascopubs.org/doi/full/10.1200/JCO.2017.73.0242 (visited on 07/15/2022).

[18] A. Chalil and V. Ramaswamy, "Low grade gliomas in children," *Journal of Child Neurology*, vol. 31, no. 4, pp. 517–522, Mar. 2016, ISSN: 1708-8283. DOI: 10.1177/0883073815599259.

[19] R. Bakshi, S. Ariyaratana, R. H. B. Benedict, and L. Jacobs, "Fluid-attenuated inversion recovery magnetic resonance imaging detects cortical and juxtacortical multiple sclerosis lesions," *Archives of Neurology*, vol. 58, no. 5, pp. 742–

748, May 1, 2001, ISSN: 0003-9942. DOI: 10.1001/archneur.58.5.742. [Online]. Available: https://doi.org/10.1001/archneur.58.5.742 (visited on 07/15/2022).

[20] J. D. Shur, S. J. Doran, S. Kumar, *et al.*, "Radiomics in oncology: A practical guide," *RadioGraphics*, vol. 41, no. 6, pp. 1717–1732, Oct. 2021, Publisher: Radiological Society of North America, ISSN: 0271-5333. DOI: 10.1148/rg.2021210037. [Online]. Available: https://pubs.rsna.org/doi/10.1148/rg.2021210037 (visited on 07/15/2022).

[21] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—"how-to" guide and critical reflection," *Insights into Imaging*, vol. 11, no. 1, p. 91, Aug. 12, 2020, ISSN: 1869-4101. DOI: 10.1186/s13244-020-00887-2. [Online]. Available: https://doi.org/10.1186/s13244-020-00887-2 (visited on 07/15/2022).

[22] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research*, vol. 77, no. 21, e104–e107, Nov. 1, 2017, ISSN: 1538-7445. DOI: 10.1158/0008-5472.CAN-17-0339.

[23] M. Amadasun and R. King, "Textural features corresponding to textural properties," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 5, pp. 1264–1274, Oct. 1989, ISSN: 00189472. DOI: 10.1109/21.44046. [Online]. Available: http://ieeexplore.ieee.org/document/44046/ (visited on 09/12/2022).

[24] "Feature importances with a forest of trees," scikit-learn. (), [Online]. Available: https://scikit-learn/stable/auto_examples/ensemble/plot_forest_importances.html (visited on 09/12/2022).

[25] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 1, 2001, ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. [Online]. Available: https://doi.org/10.1023/A:1010933404324 (visited on 07/15/2022).

[26] F. Sultana, A. Sufian, and P. Dutta, "Advancements in image classification using convolutional neural network," in *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, Nov. 2018, pp. 122–129. DOI: 10.1109/ICRCICN.2018.8718718. arXiv: 1905.03288[cs]. [Online]. Available: http://arxiv.org/abs/1905.03288 (visited on 07/15/2022).

[27] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, Jan. 29, 2017. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980[cs]. [Online]. Available: http://arxiv.org/abs/1412.6980 (visited on 07/14/2022).

[28] C. Jin, "A sequential process monitoring approach using hidden markov model for unobservable process drift," Ph.D. dissertation, Jul. 28, 2015. DOI: 10.13140/RG.2.2.32537.83041.

[29] "BCEWithLogitsLoss — PyTorch 1.12 documentation." (), [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html (visited on 09/14/2022).

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., 2012. [Online]. Available: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (visited on 08/14/2022).

[31] R. Pascanu, T. Mikolov, and Y. Bengio, *On the difficulty of training recurrent neural networks*, Feb. 15, 2013. DOI: 10.48550/arXiv.1211.5063. arXiv: 1211.5063[cs]. [Online]. Available: http://arxiv.org/abs/1211.5063 (visited on 08/14/2022).

[32] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, ISSN: 2473-9944, Oct. 2017, pp. 3154–3160. DOI: 10.1109/ICCVW.2017.373.

[33] J. Brownlee. "A gentle introduction to the rectified linear unit (ReLU)," Machine Learning Mastery. (Jan. 8, 2019), [Online]. Available: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/ (visited on 09/14/2022).

[34] L. Nieradzik, G. Scheuermann, D. Saur, and C. Gillmann, *Effect of the output activation function on the probabilities and errors in medical image segmentation*, Sep. 2, 2021. DOI: 10.48550/arXiv.2109.00903. arXiv: 2109.00903[cs,eess]. [Online]. Available: http://arxiv.org/abs/2109.00903 (visited on 09/14/2022).

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011, ISSN: 1533-7928. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html (visited on 07/15/2022).

[36] A. Paszke, S. Gross, F. Massa, *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html (visited on 07/15/2022).

[37] J. Brownlee. "A gentle introduction to dropout for regularizing deep neural networks," Machine Learning Mastery. (Dec. 2, 2018), [Online]. Available: https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/ (visited on 07/15/2022).

[38] P. Liashchynskyi and P. Liashchynskyi, *Grid search, random search, genetic algorithm: A big comparison for NAS*, Dec. 12, 2019. DOI: 10.48550/arXiv.1912.06059. arXiv: 1912.06059[cs,stat]. [Online]. Available: http://arxiv.org/abs/1912.06059 (visited on 07/15/2022).

[39] J. Brownlee. "A gentle introduction to k-fold cross-validation," Machine Learning Mastery. (May 22, 2018), [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/ (visited on 07/15/2022).

[40] ——, "Nested cross-validation for machine learning with python," Machine Learning Mastery. (Jul. 28, 2020), [Online]. Available: https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/ (visited on 07/14/2022).

[41] N. R. Cook, "Use and misuse of the receiver operating characteristic curve in risk prediction," *Circulation*, vol. 115, no. 7, pp. 928–935, Feb. 20, 2007, Publisher: American Heart Association. DOI: 10.1161/CIRCULATIONAHA.106.672402. [Online]. Available: https://www.ahajournals.org/doi/10.1161/circulationaha.106.672402 (visited on 07/15/2022).

[42] J. Brownlee. "Statistical significance tests for comparing machine learning algorithms," Machine Learning Mastery. (Jun. 19, 2018), [Online]. Available: https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms/ (visited on 09/15/2022).

[43] B. Zhang, S. Qi, X. Pan, *et al.*, "Deep CNN model using CT radiomics feature mapping recognizes EGFR gene mutation status of lung adenocarcinoma," *Frontiers in Oncology*, vol. 10, 2021, ISSN: 2234-943X. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2020.598721 (visited on 07/15/2022).

[44] Y.-M. Wang, J.-B. Yang, and D.-L. Xu, "Environmental impact assessment using the evidential reasoning approach," *European Journal of Operational Research*, vol. 174, no. 3, pp. 1885–1913, Nov. 1, 2006, ISSN: 0377-2217. DOI: 10.1016/j.ejor.2004.09.059. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221705003164 (visited on 07/15/2022).

[45] X. Zeng, Y.-W. Chen, and C. Tao, "Feature selection using recursive feature elimination for handwritten digit recognition," in *2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Sep. 2009, pp. 1205–1208. DOI: 10.1109/IIH-MSP.2009.145.

[46] J. Laborda and S. Ryoo, "Feature selection in a credit scoring model," *Mathematics*, vol. 9, no. 7, p. 746, Jan. 2021, Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2227-7390. DOI: 10.3390/math9070746. [Online]. Available: https://www.mdpi.com/2227-7390/9/7/746 (visited on 09/15/2022).

[47] T. Suryakanthi, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm*," *International Journal of Advanced Computer Science and Applications*, vol. 11, Jan. 1, 2020. DOI: 10.14569/IJACSA.2020.0110277.

[48] A. R. Tang, J. Haizel-Cobbina, P. Paueksakon, *et al.*, "Disseminated craniospinal low-grade glioma in a patient with NF-1 without optic pathway pathology: Illustrative case," *Journal of Neurosurgery: Case Lessons*, vol. 2, no. 18, Nov. 1, 2021, Publisher: American Association of Neurological Surgeons Section: Journal of Neurosurgery: Case Lessons. DOI: 10.3171/CASE21378. [Online]. Available: https://thejns.org/caselessons/view/journals/j-neurosurg-case-lessons/2/18/article-CASE21378.xml (visited on 09/15/2022).

[49] M. Mistry, N. Zhukova, D. Merico, *et al.*, "BRAF mutation and CDKN2a deletion define a clinically distinct subgroup of childhood secondary high-grade glioma," *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 33, no. 9, pp. 1015–1022, Mar. 20, 2015, ISSN: 1527-7755. DOI: 10.1200/JCO.2014.58.3922.

[50] S. Ryall, M. Zapotocky, K. Fukuoka, *et al.*, "Integrated molecular and clinical analysis of 1,000 pediatric low-grade gliomas," *Cancer Cell*, vol. 37, no. 4, 569–583.e5, Apr. 13, 2020, ISSN: 1878-3686. DOI: 10.1016/j.ccell.2020.03.011.