ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 110

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 110

PREPROCESSING TECHNIQUES FOR BRAIN MRI SCANS: A COMPARATIVE ANALYSIS FOR RADIOGENOMICS APPLICATIONS

MARIYA MITEVA AND MARIA NISHEVA-PAVLOVA

In this study, we aim to investigate the use of preprocessing techniques on brain magnetic resonance imaging (MRI) scans for the prediction of Methylguanine-DNA methyltransferase methylation (MGMT) status in glioma patients. MGMT methylation is a biomarker that has been linked to treatment response and prognosis in glioma. We review several studies that have applied preprocessing techniques to brain MRI scans, along with molecular genetic information, for this purpose. The preprocessing techniques include but are not limited to image registration, normalization, brain extraction, and tumor segmentation. We compare the effectiveness of the techniques used in these studies and evaluate the performance of each technique in terms of accuracy, computational efficiency and other parameters. Our goal is to identify the most effective preprocessing techniques for radiogenomics applications and to determine the potential of these techniques for improving the accuracy of predictions in brain MRI scans by combining different types of data. The results of this study have the potential to serve as a basis for the development of more accurate and efficient imaging-based diagnostic tools for glioma patients, and to improve the understanding of the relationship between imaging and genomics in glioma.

Keywords: radiogenomics, glioblastoma, MGMT promoter, MRI scans, medical imaging, machine learning, deep learning

CCS Concepts:

• Computing methodologies∼Artificial intelligence∼Computer vision∼Computer vision representations∼Image representations;

• Computing methodologies∼Artificial intelligence∼Computer vision∼Computer vision problems∼Image segmentation;

• Computing methodologies∼Machine learning∼Machine learning approaches∼Neural networks

DOI: 10.60063/GSU.FMI.110.111-125

1. Introduction

Magnetic resonance imaging (MRI) is a widely used modality for non-invasive imaging of the brain and other organs. It provides high-resolution, multi-dimensional images of the tissue anatomy and function, which can be used for various clinical and research purposes. One of the emerging areas of research is radiogenomics, which aims to integrate imaging data with other types of genomic and clinical data to improve the diagnosis, prognosis, and treatment of diseases [18]. For example, the prediction of Methylguanine-DNA methyltransferase methylation (MGMT) status in glioblastoma, a type of brain cancer, has been shown to be a promising marker for personalized medicine and targeted therapy [20].

MGMT is a Deoxyribonucleic acid (DNA) repair enzyme that removes alkyl groups from DNA, and it has been shown to be involved in the resistance of cancer cells to chemotherapy [11]. MGMT methylation status has been shown to be associated with the response to chemotherapy and the prognosis of several types of cancer, including glioblastoma, the most common and aggressive primary brain cancer [11]. There are several methods available to evaluate MGMT promoter methylation, including methylation-specific polymerase chain reaction (MSP), multiplex ligation-dependent probe amplification (MLPA), pyrosequencing (PSQ), quantitative Real-Time PCR, and immunohistochemistry (IHC) to assess protein expression [19].

Medical imaging, such as brain MRI, can provide valuable information about the brain tissue and its structural and functional characteristics, which can be useful for predicting MGMT methylation status and other genomic and clinical phenotypes [16]. However, medical images are often affected by noise, artifact, and intensity variations, which can degrade the quality and accuracy of the images, and hinder the performance of the predictive models [15]. Preprocessing techniques aim to improve the quality and accuracy of the medical images, by removing the noise, artifact, and intensity variations, and by enhancing the tissue characteristics [17].

In addition to improving the quality and accuracy of images, preprocessing techniques also play an important role in the success of Computer-aided detection and diagnosis (CAD) schemes. CAD schemes have become an increasingly important tool in medical imaging to help clinicians read images more efficiently and make diagnoses more accurately and objectively. The use of CAD schemes in medical imaging is not a new concept, with early CAD schemes being developed in the 1970s. However, the development of CAD schemes has accelerated in recent years, particularly since the 1990s [8], due to the integration of more advanced machine learning methods. Conventional CAD schemes typically involve three steps: target segmentation, feature computation, and disease classification. Target segmentation is the process of identifying and isolating the region of interest (ROI) in the image, such as a tumor or lesion. Feature computation involves quantifying the characteristics of the ROI in terms of size, morphology, margin geometry, texture, and so on. Finally, disease classification involves using a classification model, such as linear discrimination analysis (LDA), to identify the malignancy of the ROI.

Deep learning-based models have become increasingly popular in recent years and have shown promising results in CAD schemes. These models involve a hierarchical architecture that can learn important features hidden in the raw image in a self-taught manner, eliminating the need for manual feature development [14]. In deep learning-based models, a neural network is trained to identify and amplify important features related to the specific task, while filtering out irrelevant features. This process is done in a progressive manner, with the neural network gradually recognizing and learning more complex features as the number of layers increases [4].

A number of studies have compared the performance of different preprocessing techniques for medical images, but most of them have focused on specific types of images and applications, such as CT, PET, or ultrasound [4,9]. It can be concluded that there is a lack of comprehensive and comparative studies on the performance of preprocessing techniques for brain MRI scans in the context of radiogenomics applications, such as predicting MGMT methylation status [1,5,6,12]. This paper aims to fill this gap by providing a comparative analysis of various preprocessing techniques for brain MRI scans, using a radiogenomics dataset and assesses the impact of the preprocessing techniques on the results obtained from deep and machine learning algorithms.

2. Contemporary research methodologies in predicting MGMT methylation status

2.1. Findings and study

Several studies have employed various preprocessing techniques on brain magnetic resonance imaging (MRI) scans to enhance the predictive capabilities of determining methylation status of the O6-methylguanine-DNA methyltransferase gene in brain tumors. The selection of the six studies involved in the comparative analysis of contemporary research methodologies in predicting MGMT methylation status is based on several criteria related to their relevance, scientific rigor, and diversity of methodologies employed. One key criterion for inclusion is the scientific quality and significance of the studies. The selected studies are likely to have undergone rigorous peer-review processes and have been published in high-impact journals, ensuring the reliability and validity of their findings. Another important criterion is the diversity of methodologies employed in the studies. The selected studies are likely to have employed a range of different imaging techniques and data preprocessing methods to enhance the predictive capabilities of determining MGMT methylation status. By including studies with different methodologies, the comparative analysis can provide a more comprehensive overview of the current state of research in the field, identify the strengths and limitations of different approaches, and highlight potential areas for further improvement. The techniques discussed aim to improve the quality and information content of the imaging data for more accurate and efficient assessment of MGMT methylation status. For example, the authors in [2] employed a series of image preprocessing steps to prepare the imaging data for analysis. The primary

goal of the preprocessing was to align and standardize the imaging data across different modalities for each patient. First, the authors used the FMRIB Linear Image Registration Tool (FLIRT) for coregistration of the imaging data across different modalities. The FLIRT algorithm is based on linear affine transformation and uses a mutual-information cost function. The reference volume for coregistration was the highest resolution sequence, most commonly the postcontrast T1-weighted acquisition. On average, coregistration of a single volume took approximately 1 minute.

In this study, MR imaging data and corresponding molecular genetic information were retrospectively obtained from the Cancer Imaging Archives and the Cancer Genome Atlas for patients with low- or high-grade gliomas. Only patients with full preoperative MR imaging, including T2, FLAIR, and T1-weighted pre- and postcontrast acquisitions, were included in the analysis. The molecular information for each patient was obtained, including IDH1 status, 1p/19q codeletion, and MGMT promoter methylation.

Subsequently, each input image was independently normalized using z-score normalization. This step is commonly used to standardize the data, and to ensure that the mean of the data is zero and the standard deviation is one. The authors then employed a custom in-house fully automated whole-brain extraction tool, based on 3D convolutional neural network, to remove extracranial structures. This step aimed to improve computational efficiency and focus the analysis on relevant regions.

Finally, the authors used a fully automated brain tumor segmentation tool to identify lesion margins. The algorithm used in this step was the top-performing tool as evaluated in the international 2016 Multimodal Brain Tumor Segmentation Challenge. It is based on a serial fully convolutional neural network architecture with residual connections, and it performs whole-tumor segmentation in approximately 1 second. These segmentations were used to generate cropped slice-by-slice images of the tumor on all modalities, which were subsequently resized to a $32 \times 32 \times 4$ input.

It is worth noting that the use of convolutional neural network (CNN) based algorithm was developed and tested for the purpose of predicting genetic mutations and methylation status in glioma patients. The algorithm was found to have high accuracy in predicting IDH1 mutation (mean 94%), 1p/19q codeletion (mean 92%), and MGMT promoter methylation (mean 83%), as well as good performance in terms of the area under the curve for these predictions. The CNN was trained for 25,000 iterations before convergence, and the entire imaging workflow takes approximately 5.12 s per patient, which includes time for detection, preprocessing, and classification. This approach demonstrated promising results and may have potential clinical applications.

In a subsequent experiment, discussed in [3], S. Chen et al. obtained approval from the local Institutional Review Board and recruited 111 patients diagnosed with WHO grade 2-4 glioma who had undergone surgical resection and received plain and enhanced scans from the Affiliated Drum Tower Hospital of Nanjing University Medical School between 2018 and 2020. The patients had not received any prior treatment such as radiotherapy, chemotherapy or antitumor drugs before surgery and those with incomplete or poor-quality images were excluded. The data were

divided into a training group and a validation group with a ratio of 8:2, and MRI images acquired using 3.0 T MRI scanners and different protocols were used in the analysis. The image brightness variations due to scanning process were eliminated and deviation field correction was performed on all images before the analysis.

In addition to the previously described methodology, the authors also employed a method for extracting radiomics features from the tumor edema and tumor core area of four sequences, including T1WI, T2WI, T1CE, and ADC. They used a combination of manual segmentation by two radiologists, and an open-source platform called PyRadiomics to extract a total of 688 features from each patient. These features were then normalized by the Z-score and used as input in the deep learning model. The image sequences were registered to the same physical space in order to match the same patient ROIs across each sequence. After completing the preprocessing steps, the study subjects were randomly assigned to either the training or testing group in an 80:20 ratio.

In this study, a ResNet deep learning model based on radiomics features was used to predict the MGMT promoter methylation status of gliomas. The study found that among single MRI modalities, the T1CE model based on the Region of Interest (ROI) of the tumor core achieved the highest AUC value of 0.84. When multiple MRI modalities were combined, the T1CE model combined with the ADC model based on the ROI of the tumor core achieved the highest AUC value of 0.90. The final model, which was a combination of T1CE and ADC modalities, based on the ROI of the tumor core, showed the best performance among all the models, with the highest accuracy of 0.91 and AUC of 0.90. Ten features were found to be the most important radiomics features for the prediction. The study suggests that the combination of T1CE and ADC MRI modalities could be superior to other single or multiple MRI sequences in the prediction of MGMT promoter methylation and that a deep learning model based on radiomics features could help in identifying molecular biomarkers from routine medical images, and therefore facilitate treatment planning.

In another work [10], the authors used brain MRI scans from patients diagnosed with glioblastoma multiforme (GBM) in order to analyze the relationship between methylation status and imaging characteristics. GBM is a highly malignant brain tumor with a poor prognosis, and identifying biomarkers that can inform diagnosis and treatment strategies is a crucial area of research.

The MRI scans used in the study were obtained from the Cancer Imaging Archive (TCIA) and consisted of 5,235 scans from 262 patients. Each scan was a 3-dimensional reconstruction of the brain and were provided in a DICOM format, which is a non-proprietary data interchange protocol, digital image format, and file structure for biomedical images and related information. These scans were preprocessed to remove noise by looking at the distribution of Hounsfield Units in the pixels and only retain the slices that contain the tumor. Additionally, all images were resized to 128x128 dimensions for consistency.

The study also utilized methylation data from the Cancer Genome Atlas (TCGA) for 423 unique patients, which were preprocessed to extract methylation sites that are located in minimal promoter and enhancer regions. These regions are known

to have a high level of methylation activity and affect MGMT expression, a DNA repair protein that is commonly inactivated in GBM. Specifically, the study focused on three methylation sites cg02941816, cg12434587, and cg12981137 which had been used in previous studies on MGMT methylation and GBM. A patient was considered to have positive methylation status if any of the three sites had a methylation beta value of at least 0.2.

The study also employed data augmentation techniques to increase the size of the dataset and prevent overfitting in the convolutional recurrent neural network (CRNN) used in the analysis. The data augmentation involved applying image rotation and reversing the MRI scans, so that the methylation status and location of the tumor was preserved. The images were rotated every 4° from −90° to +90°, and were flipped so that the MRI scans were represented from superior to inferior and vice versa. This resulted in a 90-fold increase in the number of MRI scans available for training the CRNN, which is expected to boost the performance and robustness of the network.

The study found that the CRNN obtained modest patient-level accuracy of 0.67 on the validation set and 0.62 on the test set. It was also better in performance when compared to the random forest model. The CRNN also provided a generalizable platform for visualizing the different filters and layers of deep learning architectures for brain MRI scans.

In a subsequent research, P. Korfiatis et al. [13] emphasize the importance of reducing manual steps required by computer-aided diagnosis systems in order to facilitate the translation of such systems into clinical practice. The proposed system uses image normalization and bias corrections as the only preprocessing steps, which are fully automated and computationally efficient, taking less than 2 minutes on a typical desktop computer. The system also focuses on a three-class problem rather than a binary approach, enabling the algorithm to operate without the need for tumor segmentation step, thus reducing the complexity of the process.

This study aims to investigate the relationship between methylation status and imaging characteristics in patients with newly diagnosed GBM using MRI scans. The study was approved as minimal risk by the institution's Internal Review Board and included 155 presurgery MRI examinations from patients treated at Mayo Clinic between 2007 and 2015. The inclusion criteria were age $\geq$ 18 years and preoperative MR scans with T2 and T1 weighted post-contrast images performed at Mayo Clinic with known MGMT methylation status. The images were anonymized and the image processing pipelines were managed with MIRMAID. The study found 66 patients had methylated and 89 patients had unmethylated tumors. For the methylated group, 53 scans were performed on a 1.5T scanner and 13 were performed on a 3T scanner, while for the unmethylated group, 76 scans were performed on a 1.5T scanner and 13 were performed on a 3T scanner. For the purpose of this study, only T2 images were used and N4 was used for bias field correction to eliminate the low-frequency and smooth signal that corrupts MRI images and potentially affect image analysis steps.

The authors evaluated the ability of three different residual deep neural network (ResNet) architectures to predict methylation status of the O6-methylguanine

methyltransferase gene using magnetic resonance imaging without the need for a distinct tumor segmentation step. The study found that the ResNet50 architecture (50 layers) was the best performing model, achieving an accuracy of 94.90% for the test set. This performance was statistically significantly better than both ResNet18 (18 layers) and ResNet34 (34 layers) architectures. This study proposes a method that eliminates the need for extensive preprocessing and demonstrates that deep neural architectures can be used to predict molecular biomarkers from routine medical images.

The authors of [21] use multiparametric MRI images of brain gliomas from the TCIA and genomic information from the TCGA and TCIA. The final dataset of 247 subjects included 163 methylated cases and 84 unmethylated cases. Tumor masks for 179 subjects were obtained through previous expert segmentation and for the remaining 68 subjects, they were generated by trained 3D-IDH network and reviewed by two neuroradiologists. The preprocessing steps included: 1) Affine coregistration using Advanced Normalization Tools, 2) skull stripping using Brain Extraction Tool (BET), 3) removal of radiofrequency inhomogeneity using N4 Bias Field Correction, and 4) normalization of intensity to zero-mean and unit variance. The entire preprocessing took about 5 min per dataset.

In this study, transfer learning was applied to predict the MGMT promoter status utilizing a previously trained 3D-IDH network. The decoder part of the network was fine-tuned for voxel-wise dual-class segmentation of the whole tumor, where one represents methylated and two represent unmethylated MGMT promoter types. The authors used a dataset of multiparametric MR images of patients with brain gliomas obtained from the TCIA database, in combination with genomic information obtained from both the TCGA and TCIA databases. The data preprocessing steps applied to the images included: 1) affine coregistration to the SRI24 T2 template using the Advanced Normalization Tools software package, 2) skull stripping using the BET from the Oxford Centre for Functional MRI of the Brain Software Library (FSL), 3) removing radiofrequency inhomogeneity using N4 Bias Field Correction, and 4) normalizing intensity to zero-mean and unit variance. In order to assess the network's performance, the authors implemented a 3-fold cross-validation strategy. The dataset of 247 subjects was randomly shuffled and distributed into three groups, and then the three groups alternated among training, in-training validation, and held-out testing groups. The network's performance was reported only on the hold-out testing group for each fold because it is never seen by the network during the training.

Finally, a 3D-IDH network was trained and fine-tuned for determination of MGMT promoter status using transfer learning method. A three-fold cross-validation approach was implemented on a dataset of 247 subjects with MRI images and known MGMT promoter status. The network achieved a mean testing accuracy of 94.73% across the three folds with a range of 93.98–95.12% (SD 0.66%). Additional evaluation metrics such as sensitivity, specificity, positive predictive value, negative predictive value and AUC were also computed with high performance, ranging from 91.66% to 96.31% (SD 2.06%). The network showed an average Dice score of 0.82

(SD 0.008) for tumor segmentation. The network misclassified 13 cases among the total of 247 subjects.

The authors in [7] aim to build a model for predicting the methylation status of the O6-Methylguanine-DNA-methyltransferase promoter in glioblastoma multiform patients, using a novel radiomics-based machine learning (ML) approach. The following steps were used to build the model:

1. Data source and collection: The pre-processed and segmented multimodal magnetic resonance imaging features from the Cancer Genome Atlas – GBM24 collections were downloaded from the Cancer Imaging Archive public database. Only data entries with MRI modalities such as T1-weighted pre-contrast (T1), T1-weighted post-contrast (T1-Gd), T2, and T2-FLAIR (fluid-attenuated inversion recovery) were selected, resulting in 53 GBM patients included in the study. The 704 radiomics features obtained were classified into seven categories: first-order statistical features, volumetric features, textural features, histogram-based features, morphological features, spatial features, and glioma diffusion properties.

2. A two-stage radiomics feature selection and machine learning classification approach was used to predict the MGMT methylation status in glioblastoma and low-grade glioma (LGG) patients using medical imaging data. Three machine learning models (Random Forest (RF), XGBoost, and Support Vector Machine (SVM)) were incorporated into a genetic algorithm (GA) algorithm for feature selection.

The GA-RF model was found to have the best performance with a sensitivity of 0.894, specificity of 0.966, and accuracy of 0.925 in the GBM dataset. The GA-RF feature set outperformed other feature selection methods with an AUC of 0.93 in identifying MGMT methylation status from radiomics features. In the LGG dataset, the GA-RF model outperformed other models with an accuracy of 0.750, sensitivity of 0.78, and specificity of 0.62. The results indicate the potential of applying the extracted radiomics features for the prediction of MGMT methylation status in both high- and low-grade gliomas.

2.2. Conclusions

In these studies, various preprocessing techniques are used to enhance the predictive capabilities of determining methylation status of the MGMT gene in brain tumors. The preprocessing of magnetic resonance imaging scans is a crucial step in improving the quality and increasing the information content of the imaging data. This is done to obtain a more precise assessment of brain tumors. The commonly employed preprocessing techniques include image registration, normalization, extraction of the entire brain, and segmentation of the tumors. These techniques are executed using software tools such as the FMRIB Linear Image Registration Tool and custom algorithms that utilize convolutional neural networks. Another method

for preprocessing involves the extraction of radiomics features from various MRI sequences, which can be performed using either manual segmentation or open-source platforms like PyRadiomics. These features are then used as inputs for deep and machine learning models, such as ResNet, to predict the methylation status of the O6-methylguanine-DNA methyltransferase gene in brain tumors. The results from these studies indicate a promising level of performance and have potential applications in the clinical setting.

3. Discussion

3.1. Discoveries and analysis

In this section we present a comparison, which summarizes the key characteristics of studies investigating the prediction of MGMT methylation status from brain MRI scans.

The studies mentioned above have used deep or machine learning models to predict various glioma subtypes based on medical imaging data and genetic information. The research in Table 1 provides information about the dataset, preprocessing techniques, model architecture, and evaluation metrics used in the studies, arranged in a clear and organized manner. The studies vary in their specific focus, but all aim to use machine learning and in particular deep learning models to predict different glioma subtypes with high accuracy.

Based on the comparison provided in the table, the authors in [13] reported an accuracy of 94.90% which is the highest among the other studies in the table. Furthermore, they used a dataset of 155 pre-surgery MRI examinations from patients treated at Mayo Clinic between 2007 and 2015, which is a relatively small dataset compared to other studies such as [2] and [10] that used datasets from the Cancer Imaging Archive. The study used only T2 images from the MRI exams, and applied normalization and bias correction as the only preprocessing steps, which were fully automated and computationally efficient. Additionally, the study used ResNet50 architecture, which is a 50-layer deep convolutional neural network, and trained it to classify the images as either methylated or unmethylated tumors. The study found that the ResNet50 architecture performed the best out of all the ResNet architectures tested, achieving an accuracy of 94.90% on the test set. One possible reason for the high performance of the ResNet50 architecture in this study is the small size of the dataset used. With a small dataset, it is possible that the model is able to achieve high performance by overfitting to the training data. Moreover, the study not used any additional data except MRI images, that makes the results only based on the ability of the model to extract the relevant information from the images.

The work of the researchers in [7] is noteworthy for its use of a unique combination of imaging and genomic data which allows for a more comprehensive analysis and prediction of MGMT methylation status. They used a combination of multimodal MRI scans (T1, T1-Gd, T2, and T2-FLAIR) and radiomics features (extracted from

Table 1

Comparative analysis of MRI preprocessing and modeling techniques in radiogenomics

Authors. Chang et al. [2]
Year. 2018
Dataset. MR imaging data and corresponding molecular genetic information retrospectively obtained from the Cancer Imaging Archives and the Cancer Genome Atlas for patients with low- or high-grade gliomas.
Preprocessing. FLIRT for coregistration of the imaging data across different modalities. The FLIRT algorithm is based on linear affine transformation and uses a mutual-information cost function.
Model. 3D Convolutional Neural Network (CNN)
Accuracy. High accuracy in predicting IDH1 mutation (mean 94%), 1p/19q codeletion (mean 92%), and MGMT promoter methylation (mean 83%)
Validation. IDH1 mutation (mean, 94%; range between cross validations, 90– 96%), 1p/19q codeletion (mean, 92%; range, 88–95%), and MGMT promoter methylation (mean, 83%; range, 76– 88%) on 5-fold cross-validation

Authors. Han & Kamdar [10]
Year. 2018
Dataset. MRI scans were obtained from the TCIA and consisted of 5,235 scans from 262 patients; also methylation data from the TCGA for 423 unique patients are used, which were preprocessed to extract methylation sites.
Preprocessing. Scans were preprocessed to remove noise by looking at the distribution of Hounsfield Units in the pixels and only retain the slices that contain the tumor; all images were resized to 128×128 dimensions for consistency.
Model. Convolutional Recurrent Neural Network (CRNN)
Accuracy. Patient-level accuracy of 0.67 on the validation set and 0.62 on the test set
Validation. Examined predictions on the test set

Authors. Korfiatis et al. [13]
Year. 2017
Dataset. 155 presurgery MRI examinations from patients treated at Mayo Clinic between 2007 and 2015; 66 patients had methylated and 89 patients had unmethylated tumors.
Preprocessing. Normalization and bias corrections as the only preprocessing steps, which are fully automated and computationally efficient.
Model. ResNet50 architecture (50 layers) is the best performing model
Accuracy. 94.90% accuracy on the test set
Validation. k-fold cross validation

Authors. Yogananda et al. [21]
Year. 2021
Dataset. MRI images of brain gliomas from the TCIA and genomic information from the TCGA and TCIA
Preprocessing. 1)Affine coregistration using Advanced Normalization Tools, 2) Skull stripping using Brain Extraction Tool, 3) Removal of radiofrequency inhomogeneity using N4 Bias Field Correction, 4) normalizing intensity to zero-mean and unit variance
Model. 3D-IDH network
Accuracy. Mean testing accuracy of 94.73% across the 3 folds with a range of 93.98–95.12% (SD 0.66%), AUC ranging from 91.66% to 96.31% (SD 2.06%)
Validation. 3-fold cross validation

Authors. Chen et al. [4]
Year. 2022
Dataset. 111 patients diagnosed with WHO grade 2–4 glioma who had undergone surgical resection and received plain and enhanced scans from the Affiliated Drum Tower Hospital of Nanjing University Medical School between 2018 and 2020.
Preprocessing. A combination of manual segmentation, and an open-source platform called PyRadiomics to extract features from each patient; then Z-score normalization is applied. The image sequences were registered to the same physical space in order to match the same patient ROIs across each sequence.

Model. ResNet
Accuracy. The final model, which was a combination of T1CE and ADC modalities, with the highest accuracy of 0.91 and AUC of 0.90
Validation. 5-fold cross validation

Authors. Do et al. [7]
Year. 2022
Dataset. Text53 GBM patient data from the Cancer Imaging Archive database with MRI modalities such as T1, T1-Gd, T2, and T2-FLAIR. 704 radiomics features were extracted and classified into 7 categories: statistical, volumetric, textural, histogram-based, morphological, spatial, and diffusion properties.
Preprocessing. Already pre-processed and segmented multimodal magnetic resonance imaging (MRI)
Model. 3 ML models (Random Forest (RF), XGBoost, and Support Vector Machine (SVM)) were incorporated into a GA algorithm for feature selection.
Accuracy. The GA-RF model was found to have the best performance with a sensitivity of 0.894, specificity of 0.966, and accuracy of 0.925 in the GBM.
Validation. 5-fold cross validation

the scans and classified into seven categories). The authors utilized three machine learning models (Random Forest, XGBoost, and Support Vector Machine) and incorporated them into a genetic algorithm for feature selection. The results showed that the GA-RF model had the best performance with a sensitivity of 0.894, specificity of 0.966, and accuracy of 0.925 on the GBM dataset, based on 5-fold cross validation.

The results of this study demonstrate the potential for using machine learning techniques in the analysis of brain MRI scans and radiogenomic data. The combination of imaging and genomic information has the potential to improve the accuracy of patient outcomes prediction, which can inform treatment decisions and improve patient outcomes. Future studies in this direction can focus on further refining the genetic algorithm for feature selection and incorporating additional machine learning models. Additionally, larger datasets can be used to validate the results and explore the potential for using these techniques in a clinical setting. The study provides a significant contribution to the field of radiogenomics and highlights the importance of combining imaging and genomic data to improve patient outcomes.

3.2. Conclusions

The main findings from the comparison tables are closely related to the topic of this paper, which is preprocessing brain MRI scans for predicting MGMT methylation status. The findings suggest that deep and machine learning-based techniques and multi-modal approaches may be more effective for this task, and that larger and more diverse datasets may be more useful for training and evaluating such models. By considering these findings, researchers can make more informed decisions about the most appropriate preprocessing techniques and datasets to use for their specific research questions and goals. Additionally, these findings may also be useful for clinicians and healthcare professionals who are interested in using imaging and other data to predict and manage the treatment of brain tumors in their patients.

4. Incorporating expert knowledge in medical imaging analysis

Incorporating expert knowledge is a crucial aspect for the development of deep and machine learning algorithms for radiogenomics applications using brain MRI scans. The ability to explain and understand the decision-making processes of these algorithms is essential for their clinical implementation and acceptance. Expert knowledge, specifically domain knowledge, can aid in the optimization of imaging data and improve the performance of predictive models.

Embedding expert knowledge into the preprocessing and analysis of brain MRI scans can address the unique challenges present in medical images, such as high inter-class similarity and limited labeled data. For example, incorporating anatomical information can improve the registration of multi-modal imaging data, while incorporating radiomic features can enhance the representativeness of the imaging data. Additionally, incorporating text reports accompanying images can provide additional clinical information for the decision-making process [4].

Furthermore, incorporating expert knowledge into the training and validation process can also improve the interpretability of the models. This can be achieved through the use of methods such as feature importance analysis and decision tree visualization. These methods allow for the identification of the most important features used by the model in its decision-making process and can provide insight into how the model is using the expert knowledge. This can help researchers and clinicians understand how the model is making its predictions, which can ultimately lead to more trust in the model's predictions. Additionally, the use of expert knowledge can also lead to the development of more robust models that are able to generalize better to unseen data.

It is also worth mentioning that expert knowledge does not need to be only from a radiologist, other experts from different fields such as computer vision, medical physics, or medical informatics can also bring valuable insights to optimize the algorithms. For example, computer vision experts can help in preprocessing the images, medical physics experts can help in understanding the underlying physics of the images and medical informatics experts can help in understanding the clinical context of the images. Collaboration between different experts can lead to a more comprehensive approach to radiogenomics.

In conclusion, incorporating expert knowledge in medical imaging analysis is crucial for the development of deep and machine learning algorithms for radiogenomics applications using brain MRI scans. Integrating domain expertise into the preprocessing, analysis, and validation of the data can lead to more accurate and interpretable models that are better suited for clinical implementation. Collaboration between experts from different fields can also bring valuable insights that can optimize the algorithms further.

5. Conclusions

In conclusion, the use of preprocessing techniques for brain MRI scans has been shown to be a useful tool for radiogenomics applications, particularly in the prediction of MGMT methylation status. A comparative analysis of the studies listed in the table, such as [2, 3, 7, 10, 13, 21], reveal that each study used different preprocessing techniques, models and achieved different levels of accuracy. The results discussed in [7] is particularly noteworthy for its use of a unique combination of imaging and genomic data, which allows for a more comprehensive analysis and prediction of MGMT methylation status, and the high accuracy achieved by the GA-RF model with a sensitivity of 0.894, specificity of 0.966, and accuracy of 0.925 based on 5-fold cross validation of the GBM dataset. This study demonstrates the potential for combining multiple data sources to improve predictions in medical imaging.

It has been demonstrated that the integration of various types of data can provide a more comprehensive understanding of the underlying biology of brain tumors and potentially enhance the diagnostic and therapeutic decision-making process. Therefore, it is important for future research to focus on investigating the potential of incorporating multiple data modalities in radiogenomics applications, in addition to the development of advanced preprocessing techniques that can optimize the quality and information content of imaging data for improved prediction accuracy. Furthermore, with the advancement of deep learning techniques, there is a growing potential to integrate these models into preprocessing techniques, which can lead to more accurate and efficient predictions, and also enable to explain the decision-making process. It is important to consider the use of deep and machine learning models in future research in order to fully exploit the potential of radiogenomics applications in the diagnosis and treatment of brain tumors.

References

[1] D. Abler, V. Andrearczyk, V. Oreiller et al., Comparison of MR preprocessing strategies and sequences for radiomics-based MGMT prediction, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th Int. Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part II, Springer-Verlag, Berlin, Heidelberg, 367–380. https://doi.org/10.1007/978-3-031-09002-8_33.

[2] P. Chang, J. Grinband, B. Weinberg et al., Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas, Am. J. Neuroradiol. 39(7) (2018) 1201–1207.

[3] S. Chen, Y. Xu, M. Ye et al., Predicting MGMT promoter methylation in diffuse gliomas using deep learning with radiomics, J. Clin. Med. 11 (2022) 3445, https://doi.org/10.3390/jcm11123445.

[4] X. Chen, X. Wang, K. Zhang et al., Recent advances and clinical applications of deep learning in medical image analysis, Med. Image Anal. 79 (2022) 102444, https://doi.org/10.1016/j.media.2022.102444.

[5] X. Chen, M. Zeng, Y. Tong et al., Automatic prediction of MGMT status in glioblastoma via deep learning-based MR image analysis, Biomed. Res. Int. 2020 (2020) 9258649, https://doi.org/10.1155/2020/9258649; PMID: 33029531; PMCID: PMC7530505.

[6] K. Dang, T. Vo, L. Ngo and H. Ha, A deep learning framework integrating MRI image preprocessing methods for brain tumor segmentation and classification, IBRO Neurosci. Rep. 13 (2022) 523–532, https://doi.org/10.1016/j.ibneur.2022.10.014; PMID: 36590099; PMCID: PMC9795279.F.

[7] D. Do, M. Yang, L. Lam et al., Improving MGMT methylation status prediction of glioblastoma through optimizing radiomics features using genetic algorithm-based machine learning approach, Sci. Rep. 12 (2022) 13412, https://doi.org/10.1038/s41598-022-17707-w.

[8] K. Doi, H. MacMahon, S. Katsuragawa et al., Computer-aided diagnosis in radiology: potential and pitfalls, Eur. J. Radiol. 31 (1999) 97–109.

[9] N. Goel, A. Yadav and B. M. Singh, Medical image processing: A review, in: 2016 2nd Int. Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity (CIPECH), Ghaziabad, India, 2016, 57–62, https://doi.org/10.1109/CIPECH.2016.7918737.

[10] L. Han and M. Kamdar, MRI to MGMT: predicting methylation status in glioblastoma patients using convolutional recurrent neural networks, in: Pacific Symp. Biocomputing 2018: Proc Pacific Symp., World Scientific, 331–342, 2018.

[11] M. Hegi, A. Diserens, T. Gorlia et al., MGMT gene silencing and benefit from temozolomide in glioblastoma, N. Engl. J. Med. 352(10) (2005) 997–1003, https://doi.org/10.1056/NEJMoa043331; PMID: 15758010.

[12] K. Jae, L. Pyo, K. Jae and K. Gi, Comparison of pre-processed brain tumor MR images using deep learning detection algorithms, J. Multimed. Inf. Syst. 8(2) (2021) 79–84, https://doi.org/10.33851/JMIS.2021.8.2.79.

[13] P. Korfiatis, T. Kline, D. Lachance et al., Residual deep convolutional neural network predicts MGMT methylation status, J. Digit. Imaging 30(5)(2017) 622–628.

[14] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[15] S. Masoudi, S. Harmon, S. Mehralivand et al., Quick guide on radiology image preprocessing for deep learning applications in prostate cancer research, J. Med. Imaging (Bellingham) 8(1) (2021) 010901, https://doi.org/10.1117/1.JMI.8.1.010901; PMID: 33426151; PMCID: PMC7790158.

[16] B. Menze, A. Jakab, S. Bauer et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34(10) (2015) 1993–2024, https://doi.org/10.1109/TMI.2014.2377694; PMID: 25494501; PMCID: PMC4833122.

[17] M. Salvi, U. Acharya, F. Molinari and K. Meiburger, The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis, Comput. Biol. Med. 128 (2021) 104129, https://doi.org/10.1016/j.compbiomed.2020.104129.

[18] L. Shui, H. Ren, X. Yang et al., The era of radiogenomics in precision medicine: An emerging approach to support diagnosis, treatment decisions, and prognostication in oncology, Front. Oncol. 10 (2021) 570465, https://doi.org/10.3389/fonc.2020.570465; PMID: 33575207; PMCID: PMC7870863.

[19] L. Wang, Z. Li, C. Liu et al., Comparative assessment of three methods to analyze MGMT methylation status in a series of 350 gliomas and gangliogliomas, Pathol. Res. Pract. 213(12) (2017) 1489–1493, https://doi.org/10.1016/j.prp.2017.10.007; PMID: 29103769.

[20] M. Weller, R. Stupp, G. Reifenberger et al., MGMT promoter methylation in malignant gliomas: ready for personalized medicine?, Nat. Rev. Neurol. 6(1) (2010) 39–51.

[21] C. Yogananda, B. Shah, S. Nalawade et al., MRI based deep-learning method for determining glioma MGMT promoter methylation status, Am. J. Neuroradiol., 42(5) (2021) 845–852, https://doi.org/10.3174/ajnr.A7029, http://www.ajnr.org/content/early/2021/03/04/ajnr.A7029.

Received on March 17, 2023
Accepted on May 7, 2023

Mariya Miteva

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: mmiteva@fmi.uni-sofia.bg

Maria Nisheva–Pavlova

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: marian@fmi.uni-sofia.bg

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia
BULGARIA