BRAIN TUMOUR CLASSIFICATION USING CNNS: FINAL PROJECT REPORT

Marco Mastrangelo
Student# 1009133518
marco.mastrangelo@mail.utoronto.ca

Hussein Ismail
Student# 1008747910
h.ismail@mail.utoronto.ca

Devraj Solanki
Student# 1009065707
devraj.solanki@mail.utoronto.ca

Damilola Aina
Student# 1008932292
dami.aina@mail.utoronto.ca

—-Total Pages: 9

1 INTRODUCTION

Brain tumours are a potentially life-threatening medical condition. The 5-year relative survival rates for certain brain tumours can be as low as 6% (Ostrom et al., 2019). Therefore, the early detection of brain tumours is crucial for improving patient outcomes. Currently, medical professionals use magnetic resonance imaging (MRI) scans to visually identify brain tumours. However, the manual identification of brain tumours can be challenging due to the large number of images produced from one MRI scan (Ullah et al., 2023). Additionally, the large variety of brain tumour sizes, shapes, and locations makes it even more challenging to detect and classify different types of brain tumours (Ullah et al., 2023). Applying deep learning models, specifically convolutional neural networks (CNNs) (which exhibit high proficiency in image classification tasks), can result in faster processing of MRI images and more accurate detection of brain tumours by eliminating human error. This project aims to produce a CNN model that can accurately detect and categorize images into four classes: three types of brain tumours — glioma, meningioma, and pituitary — including images where no tumour is present.

2 ILLUSTRATION OF THE MODEL

The goal of this project is to input an image of an MRI brain scan, preprocess it to improve accuracy and robustness, and then classify what type of tumour is displayed in the MRI, if one exists. The general idea of the model can be seen below in Figure 1.

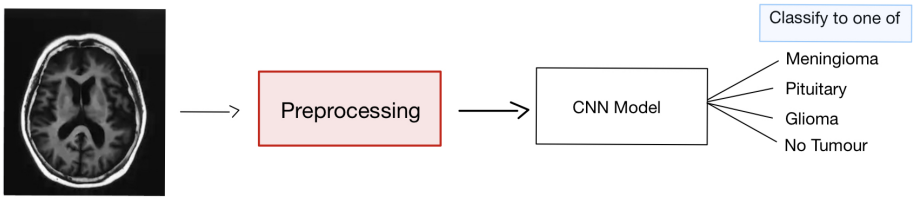

Figure 1: Basic Illustration of Deep Learning Model. Image: Hussein Ismail

3 Background and Related Work

Using deep learning techniques to identify abnormalities in MRI scan images is an ongoing area of research in academia. Classification and/or segmentation of brain tumours in MRI images appears to be a well-documented topic within this area.

A widely known CNN architecture that is used for medical imaging is the *U-Net* architecture, proposed in 2015 by researchers at the University of Freiburg (Ronneberger et al., 2015). U-Net learns the important feature(s) of an image and creates a segmentation mask, highlighting key area(s) of the image where the feature(s) are present. This architecture is widely used in medical imaging segmentation research, with others creating different versions of U-Net.

Researchers today continue to make strides to improve the classification accuracy of these CNN models, with some architectures boasting test accuracies as high as 99.27% (Ullah et al., 2023). This specific network – called *tumourDetNet*, uses 48 convolutional layers with ReLU and Leaky ReLU activation functions.

Recently, researchers have been making use of the concept of transfer learning in their CNN models, as shown in Kumar et al. (2023). This CNN model leverages this idea by using a pre-trained model called *Res-Net 50* to classify tumours as either benign or malignant, and achieves a test accuracy rate of 99.3% and 98.4% respectively for these two types.

Raza et al. (2022) proposes a hybrid model, using all but the last 5 layers of *GoogLeNet* and adding on an extra 15 layers, to achieve a test accuracy of 99.67%.

Even with these impressive advancements in the past years, CNN models still face challenges in becoming applied within the healthcare sector (Xie et al., 2022). For example, there exists a limited amount of publicly available MRI scans to train these networks (due to patient privacy and security concerns). Furthermore, it can be challenging for clinicians and radiologists to understand why a CNN outputs the results it does. For example, a radiologist cannot simply say that a patient has a brain tumour because a CNN says they do. These models can still make errors from time to time and still require human supervision. Xie et al. (2022) explains the current state of using CNNs to detect brain tumours and the challenges concerning putting these models into practice.

4 Data Processing

The team collected two datasets with identical labels (no tumor, glioma tumor, meningioma tumor, and pituitary tumor) from Kaggle (Nickparvar, 2021), (Sartaj et al., 2020). These datasets were combined to increase the amount of input data available. Doing so was crucial in ensuring the model was provided with enough high-quality images to increase the model's accuracy. Together, the combined dataset had 2547 glioma tumour images, 2582 meningioma tumour images, 2396 images with no tumour, and 2658 images with a pituitary tumour. This is a balanced dataset which allowed the team to confidently use this dataset without worrying about any significant bias.

The next step was to clean the data. The dimensions of the images were $3 \times 512 \times 512$. Indicating that each image had 3 color channels and had a height and width of 512. However, since the input images were black and white, having 3 color channels was redundant. This was changed to 1 grayscale channel, reducing the number of tuneable parameters which decreases training time. The images were also resized to be 224 in the interest of saving training time. Finally, the values of the grayscale channel were normalized to be between 0 and 1 to improve the stability of the model.

Given the nature of MRI scans being unique depending on each patient and the shape and orientation of their head, it is important that the model is robust and able to deal with variation. This can be addressed by making augmentations to the dataset that introduce additional variety. Random rotations of up to 45 degrees were applied to each of the images so that the model has experience with MRI scans that are oriented imperfectly. Furthermore, a discoloring effect was applied to images with a probability of 0.5. This was added to the images to simulate a faded scan. Finally, Gaussian Blur was applied to simulate a scan where the patient moved slightly. Figure 2 shows an image before augmentation, and Figure 3 shows images after various augmentations were applied.

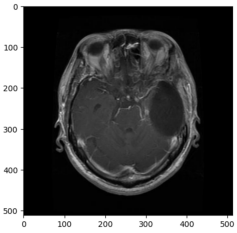

Figure 2: Image before processing. Image: Devraj Solanki

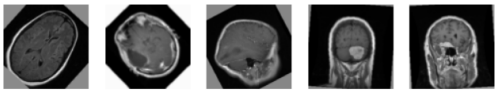

Figure 3: Images after processing. Image: Devraj Solanki

5 Baseline Model

For the baseline model, the team augmented a simple convolutional neural network, originally designed for pneumonia detection from x-ray imagery (Foster, 2019). This model serves as an ideal benchmark for assessing our neural network's performance, as they both deal with medical MRI scans as inputs.

5.1 Model Outline and Implementation

The baseline model consists of 2 convolutional layers, 2 max-pooling layers, 1 flattened layer, 2 fully connected layers, and 1 dropout layer. The original output layer consists of one neuron that labels images a binary classification based on whether they contain pneumonia or not.

For the purposes of this project, the team modified certain aspects of the model. Firstly, the number of input channels was reduced from 3 (RGB) to 1 (Greyscale) due to the fact that the MRI scans are black and white. Next, the input image resolution was increased from 150×150 pixels to 224×224 pixels to match the size of the images after they were processed. Finally, the output layer of the baseline model was augmented to contain 4 neurons so that it could predict between the 4 classes used in the project.

With these changes in place, the team implemented the model in PyTorch. Within the model, the two convolutional layers contained 32 and 64 output channels, respectively. A dropout layer with a dropout rate of 50% was added to the fully connected layers. To train, this model uses the cross-entropy loss function to calculate error, and the Adam optimizer to adjust the weights and biases during backward passes. Also, the hyperparameters for this model were set at a batch size of 32, a learning rate of 0.001, and an epoch size of 10. These hyperparameter choices allowed for efficient train times that allowed the group to obtain baseline results quickly. See model structure below, in Figure 4.

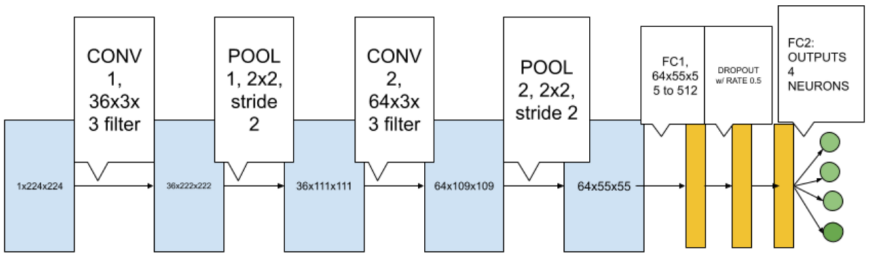

Figure 4: Baseline model visualization. Image: Marco Mastrangelo

6 Model Architecture

The team decided to modify an existing architecture, called *ResNet-18*. This architecture utilizes skip connections to reduce the effects of vanishing gradients. ResNet-18 is the smallest version of the ResNet family, with the largest variation having 152 layers (ResNet-152) (He et al., 2015). This model contains 18 layers – 17 convolutional layers, and a fully-connected output layer. These convolutional layers are split into four main stages, with each stage containing two residual blocks (two convolutional layers per residual block). Each of these residual blocks contains a skip connection to the next residual block in the stage. After all convolutional layers, global average pooling is applied to reduce a $512 \times 7 \times 7$ feature map to a $512 \times 1 \times 1$ vector, which is then fed into one fully connected layer that outputs probabilities of 1000 classes using a softmax function. ResNet-18 takes in RGB images of size $3 \times 224 \times 224$, and the MRI images in the group's dataset are grayscale, of size $1 \times 224 \times 224$. This was modified accordingly. This model was originally made for the ILSVRC competition, where images belong to 1 of 1000 classes. The MRI images each belong to 1 of 4 classes, so the final fully connected layer was adjusted accordingly for this as well. The last modification to the architecture involved freezing the weights first two stages (first 8 convolutional layers) of the model. This was done to speed up training time, which allowed the group more time to tune hyperparameters. Below in Figure 5 is an illustration of the team's chosen model architecture.

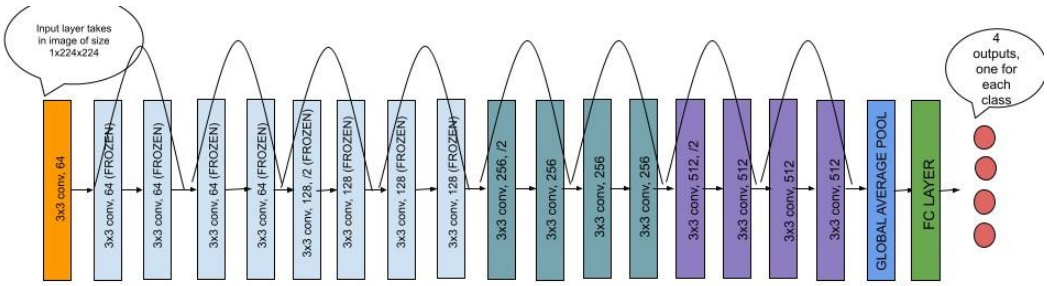

Figure 5: Modified ResNet-18 Architecture. Image: Marco Mastrangelo

Several values were tested for each hyperparameter, as well as different optimizers. The final model that is to be documented in Section 7, was trained using a batch size of 128 and a learning rate of 0.01 for 40 epochs. The model also uses the Adam optimizer and a cross-entropy loss function, as it is a multiclass classification problem. These choices allow for maximizing validation accuracy while minimizing train time.

7 Quantitative and Qualitative Results

The following section describes the quantitative and qualitative results, for both the baseline model and primary model.

7.1 Baseline Model

During evaluation, the baseline model achieved a training accuracy of 88.1% and a validation accuracy of 86.2% while achieving a loss below 0.35 for both the validation and training. With the test set, an 87.3% accuracy was achieved. See Figure 6 below for a training/validation curve of for the baseline model.

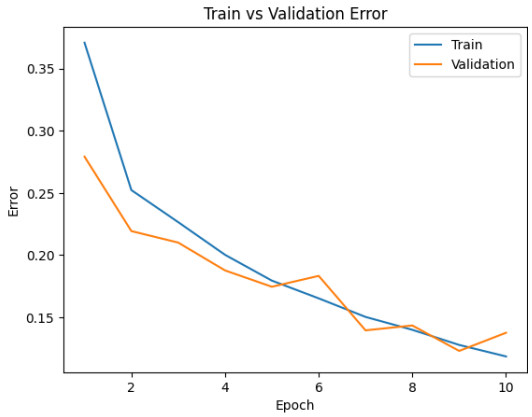

Figure 6: Train and validation error over 10 epochs. Image: Damilola Aina

The high validation and testing accuracy proves that overall, the baseline model is able to adequately classify brain tumour scan imagery. However, the presence of false negatives and positives reveals a significant concern with the efficacy of the baseline model. Seen below is the confusion matrix of the baseline model on the test set.

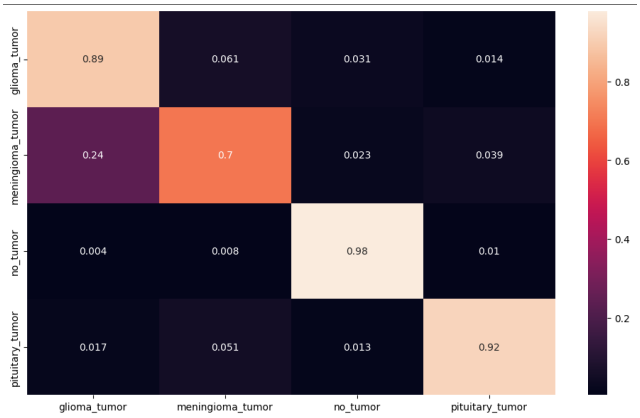

Figure 7: Confusion Matrix obtained from testing baseline model. Image: Damilola Aina

The model is almost perfect in detecting whether a scan has no tumour, achieving a 98% accuracy. The pituitary tumour detection also shows good results, with an accuracy of 92%. However, the model seems to misclassify the glioma and meningioma tumours at a high rate, with about 24% of the meningioma tumours classified as glioma tumours and 6.1% of the glioma tumours classified as meningioma tumours. This indicates that the baseline model may be too simple to differentiate between meningioma and glioma tumours.

To summarize, the baseline model's ability to successfully classify no tumour, and pituitary tumour brain scans contributes to the high validation and training accuracy seen. However, its' challenges with classifying meningioma and glioma tumours may be a significant reason why the model did not perform better on the test and validation set.

7.2 Primary Model

During evaluation, the model achieved a 98.35% train accuracy and a 95.92% validation accuracy. After using a separate test set on the model, the model achieved a test accuracy rate of 96.72% (as seen in Figure 9). Figure 8 shows the train and validation accuracy rate curve for the model over 40 epochs.

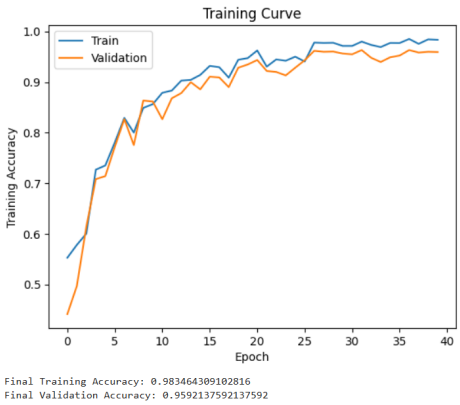

Figure 8: Train and validation accuracy rates over 40 epochs. Image: Devraj Solanki

The high validation and testing accuracy of the model proves that overall, the model performs very well when classifying brain tumour scan imagery. Additionally, the model performed significantly better than the baseline model, further proving that the groups' model has the sufficient complexity and robustness necessary to successfully classify brain tumour imagery.

An important ethical issue to consider however, is the possibility that the model misclassifies a tumour as another type of tumour or not a tumour at all. In the context of a medical setting, this could lead to an improper diagnosis or delayed treatment of the brain tumour; worsening patient outcomes. Therefore, recall was another significant quantitiative measure when evaluating the team's model.

The test set was used when calculating the models' recall rate for each type of tumour (see Figure 9). For glioma tumours, the model achieved a recall rate of 95.10%. For meningioma tumours, the model achieved a recall rate of 94.20%. Lastly, for pituitary tumours, the model achieved a recall rate of 96.81%.

The models' recall rate for meningioma tumours and glioma tumours are slightly lower than the overall training accuracy, indicating that the model may be struggling to accurately detect these tumour types from brain scan imagery.

However, the models' high recall accuracy shows that the model can successfully detect tumours from the majority of brain scan imagery; proving that the team's model is accurate and possibly implementable from a medical application standpoint.

To start uncovering qualitative findings, a confusion matrix was used to address and uncover trends in the model's prediction patterns, as seen in Figure 9.

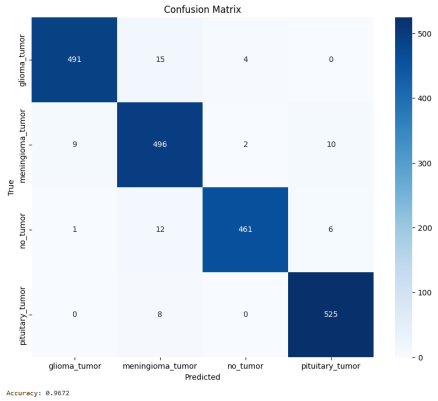

Figure 9: Confusion matrix on the test set for the primary model. Image: Devraj Solanki

As seen above, the model seems to have trouble identifying subtle differences between glioma and meningioma scans. The model incorrectly predicted that 9 meningioma tumours were glioma tumours, and that 15 glioma tumours were meningioma tumours.This can possibly be attributed to the fact that a subtype of a glioma called an oligodendroglioma forms almost exactly where meningiomas form (near the dura mater), which is why the model may misclassify meningioma and glioma tumours (Miami Neuroscience Center).

This nuance in the location of glioma tumours may also explain why the models' recall rate for mengioma and glioma tumours were lower than the overall testing accuracy; since it can be especially difficult to detect and classify between a mengioma and glioma tumour from just brain scans.

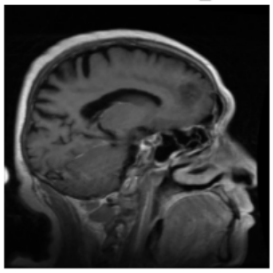

Figure 10: Misclassified glioma tumour. Image: Devraj Solanki

As seen in Figure 10, the model predicts a meningioma tumour, when the true tumour is a glioma (this is most likely an oligodendroglioma). Many more examples of the misclassification of gliomas and meningiomas look similar to this. Errors in classification seen between other classes can be due to image quality and subtle differences in brain structure from person to person, that may cause the model to believe that a specific tumour is present.

8 Evaluating on New Data

The following section describes the end-to-end process that the group carried out to obtain, process, and test the model on brand-new data.

8.1 Obtaining New Data

The group was able to create a new test dataset by combining images from two never-before-seen datasets to achieve the desired 4-class structure. The first new dataset comes from the Southern Medical University, in Guangzhou, China (Cheng, 2015). This set contains 3064 images, with each MRI image labelled as containing a meningioma tumour, glioma tumour, or pituitary tumour. However, images containing no tumour still needed to be acquired. To obtain images for the 'no tumour' class, another Kaggle dataset named *BR35h* was utilized (Hamada, 2020). This dataset contains 1500 MRI scans with no tumour, and 1500 MRI scans that contain some type of tumour. For the group's purposes, only the 'no tumour' class images were taken and added to the previously acquired data as described above.

8.2 Processing New Data

The team decided to test the model on new data using the exact same methods when tested previously on original data. To start, this involved truncating all the counts of images in each class to 500, as the original test dataset had about 500 images per class. Next, the same minor transformations applied to the original test dataset were applied to each of the new images. This includes resizing all images to 224×224, and normalizing pixel values on each image. Because this is a test dataset, no major augmentations such as rotations, flips, or blurring effects were applied, as seen below in Figure 11

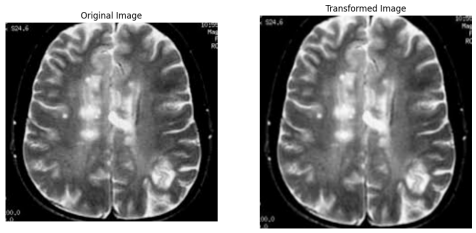

Figure 11: Before and after data processing for image in new dataset. Image: Marco Mastrangelo

8.3 Testing the Model on New Data

The final test accuracy achieved by the model on the new test data was 90.75%, with 94.2% of glioma tumours, 84.8% of meningioma tumours, 88.4% of pituitary tumours, and 95.6% of images with no tumours being accurately detected. A full confusion matrix can be seen below, in Figure 12

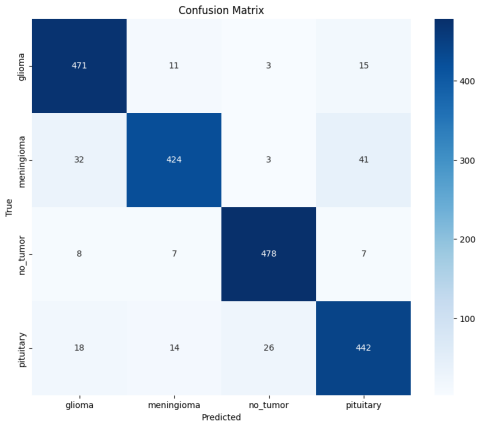

Figure 12: Confusion matrix for new test dataset. Image: Marco Mastrangelo

With the original test accuracy being approximately 96.7% and the new dataset accuracy being 90.7%, it was a slight surprise to the group to see a 6% drop. The model did not show signs of overfitting to the training data, as the final train accuracy and validation accuracy were 98.3% and 95.9%, respectively (not a large deviation between the two). A reason for this accuracy drop may be due to the differing domain conditions between the two sets. The metadata in the new Cheng (2015) dataset shares that the MRI images are 'T1 weighted and contrast-enhanced', meaning that doctors most likely injected a dye into patients arms before the scan to increase prominence/visibility of abnormalities (Ibrahim et al., 2023). The images in the original dataset were not contrast-enhanced, and one can easily tell the difference, as seen below in Figure 13.

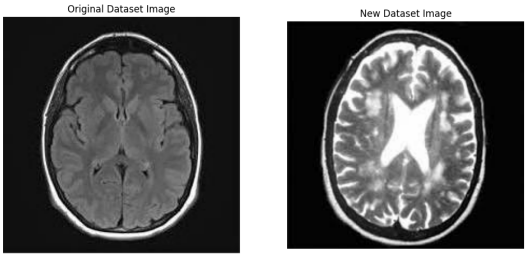

Figure 13: Non-contrast-enhanced image from original dataset vs. contrast-enhanced image from new dataset. Image: Marco Mastrangelo

The model was initially trained, validated, and tested on images that were not contrast-enhanced, so having the model see these new contrast-enhanced images for the first time may have 'tricked' it. Nevertheless, The group still believes that a 90.75% test accuracy on these new images is still a solid performance metric, given the differing domain conditions between the original and new test datasets. In the future, specific augmentations on images in new datasets can be explored to potentially mimic how the images in the training dataset look, to essentially 'reverse' this contrast enhancement effect.

9 Ethical Considerations

The use of deep learning models in the medical industry has been growing at an unprecedented rate. These models are built to mimic human capability, but these advancements are making humans more susceptible to being indolent (AI for Social Good, 2024). The use of this technology, especially a brain tumour classification model, should be used in tandem with human knowledge and influence, as opposed to the model working independently. This is due to the fact that are various ethical considerations that prevents the sole use of deep learning applications for medical purposes.

9.1 Diagnosis Accuracy

With deep learning models, acquiring an accuracy of 100% is generally seen as not feasible. This gives birth to the rise of false negatives and positives. False negatives are terms coined in the event that something is true but labelled as false, while false positives are when something is false but labelled as true (NGSS Engineering Practices).

A false negative diagnosis (meaning that a brain tumour was not detected) would prevent or delay treatment, worsening patient outcomes. On the other hand, a false positive diagnosis (meaning that a brain tumour was falsely detected) would cause unnecessary mental and emotional distress to the patient. This would also mislead patients into undergoing unnecessary medical treatments that pose significant side effects and risks.

9.2 Limitations

With every model, having a wide variety of data to test is crucial, especially in the medical field, where every individual may vary based on a number of different factors. Due to this issue, more brain scans would be required to make a more accurate and applicable model. However, accessing more brain scans may be an issue, as most of these scans are not publicly accessible. Acquiring consent for the usage of this data could be seen as a tedious process, while using this data without proper consent could result in confidentiality issues within the organization and patients in question.

10 Discussion of Results

Overall, the final model architecture proved to be a success in regards to classifying brain tumour types. On the original dataset, the training, validation, and testing accuracies of 98.3%, 95.9% and 96.7% were achieved, respectively. These accuracies are in the same range as other research papers that solve the same classification problem, such as in Kumar et al. (2023), described in Section 3. The performance of the final model on the new data also proved to be respectable, with a test accuracy of 90.75% achieved. Although this is less than the original test accuracy, it is likely that the difference is a consequence of differing domain conditions, rather than a case of overfitting, outlined in Section 8.

During the course of this project, the group initially intended to enhance the model's complexity by integrating a segmentation model (UNet) on top of the classification model. With this integration, the group aimed to not only classify the tumour but also to provide a predicted location of the identified tumour. However, the implementation of this model presented significant challenges, particularly with respect to the extensive time required for training to achieve a satisfactory level of segmentation accuracy. Due to these constraints, the group suspended the planned integration and will explore this idea further in the future.

REFERENCES

Team AI for Social Good. Artificial intelligence – the double-edged sword that can make humans vulnerable and lazy. *AI Blog*, 2024.

Jun Cheng. Southern medical university dataset, 2015. URL https://figshare.com/articles/dataset/brain_tumor_dataset/1512427?file=7953679.

Lyron Foster. Building a medical image classifier with deep learning and python. *Medium*, 2019.

Ahmed Hamada. Br35h :: Brain tumor detection 2020, 2020. URL https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection?select=no.

Kaiming. He, Xiangyu. Zhang, Shaoqing. Ren, and Jian Sun. Deep residual learning for image recognition. *Computer Vision and Pattern Recognition*, 2015.

Michael. Ibrahim, Bita. Hazhirkarzar, and Arthur Dublin. Gadolinium magnetic resonance imaging. *StatPearl*, 2023.

Sandeep. Kumar, Shilpa. Choudhary, Arpit. Jain, Karan. Singh, Ali. Ahmadian, and Mohd Bajuri. Brain tumor classification using deep neural network and transfer learning. *Brain Topography*, 2023.

Miami Neuroscience Center. Brain tumor types.

Team NGSS Engineering Practices. Practices of science: False positives and false negatives.

Msoud Nickparvar. Brain tumor mri dataset, 2021. URL https://www.kaggle.com/dsv/2645886.

Quinn. Ostrom, Giono. Cioffi, Haley. Gittleman, Nirav. Patil, Kristin. Waite, Carol. Kruchko, and Jill. Barnholtz-Sloan. Cbtrus statistical report: Primary brain and other central nervous system tumors diagnosed in the united states in 2012–2016. *Neuro-Oncology*, 1 -100, 2019.

Asaf. Raza, Javed. Khan, Ijaz. Ahmad, Salama. Ahmed, Yousef. Daradkeh, Danish. Javeed, Ateeq. Rehman, and Habib Hamam. A hybrid deep learning-based approach for brain tumor classification. *MDPI Electronics*, 2022.

Olaf. Ronneberger, Philipp. Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 2015.

Bhuvaji. Sartaj, Kadam. Ankita, Bhumkar. Prajakta, Dedge. Sameer, and Kanchan Swati. Brain tumor classification (mri), 2020. URL https://www.kaggle.com/dsv/1183165.

Naeem. Ullah, Ali. Javed, Ali. Alhazmi, Syed. Hasnain, Ali. Tahir, and Rehan. Ashraf. A unified deep learning model for brain tumor detection and classification. *PLOS ONE.*, 2023.

Yuting. Xie, Fulvio. Zaccagna, Leonardo. Rundo, Claudia. Testa, Raffaele. Agati, Raffaele. Lodi, David. Manners, and Caterina Tonon. Convolutional neural network techniques for brain tumor classification (from 2015 to 2022): Review, challenges, and future perspectives. *Diagnostics (Basel)*, 2022.