Deep Learning Assignment: Brain Tumor MRI Classification

Angela Barone (2128404), Hsuan Jung Chu (2109692), Xiaojun Liu (2110171),
Weitong Shen (2112815), Yiting Shen (2117507), Yue Wang (2111066)

September 29, 2024

1 Introduction and Overview

Malignant brain tumours are the deadliest cancers, though not all brain tumours are malignant [7]. Even typically benign tumours like meningioma and pituitary tumours can sometimes be malignant. Some gliomas are particularly difficult to detect, with only an 8% survival rate [14], due to their heterogeneity in histological features and MRI appearance [17]. Deep Learning algorithms can assist in classifying brain tumours from MRI scans for faster diagnosis.

This task involves classifying MRI images into four categories: glioma, meningioma, pituitary tumor, or no tumor, making it a multi-class classification problem. The training data consists of preprocessed JPEG images which were already preprocessed in a lower resolution and in grayscale. Further image modification (resolution, color, noise) was restricted by the author of the assignment. The baseline Convolutional Neural Network (CNN) will be enhanced through hyperparameter tuning and SMOTE. Transfer learning with pre-trained models like VGG16, ResNet50, and DenseNet121 will also be explored. Other unexplored steps and methods will be discussed. The project was conducted using Google Colab (A100 GPU) and VSCode.

1.1 Dataset Overview

Both Training and Testing datasets show a similar distribution, with Class 2 ("notumor") having more samples than the other classes. Moreover, other contributors underlined that there are 297 duplicates, mostly belonging to the notumor class [15]. The imbalance and biases will be taken into consideration in our approach. However, as we are not allowed to change the loading and preprocessing performed by the professor, we will not remove the duplicates. Additionally, the publisher of the dataset stated on Kaggle that the SARTAJ dataset contains miss-labeled glioma images. As we lack the domain knowledge, this aspect will be taken into consideration when discussing the results.

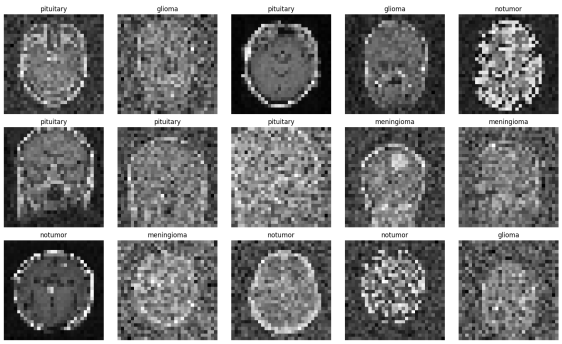

Figure 1: 15 sample images from the dataset along with their corresponding labels.

Figure 2: Class Distribution

1.2 Baseline CNN

Layer (type)	Output Shape	Param #
conv2d_4 (Conv2D)	(None, 30, 30, 32)	320
max_pooling2d (MaxPooling2D)	(None, 15, 15, 32)	0
conv2d_5 (Conv2D)	(None, 13, 13, 32)	9,248
max_pooling2d_5 (MaxPooling2D)	(None, 6, 6, 32)	0
flatten_2 (Flatten)	(None, 1152)	0
dense_4 (Dense)	(None, 32)	36,896
dense_5 (Dense)	(None, 4)	132

Total params: 46,596 (182.02 KB)

Trainable params: 46,596 (182.02 KB)

Non-trainable params: 0 (0.00 B)

Figure 3: Baseline CNN Algorithm

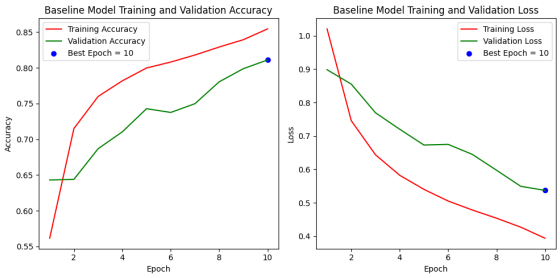

Figure 4: Baseline CNN Algorithm - Accuracy and Loss Plots

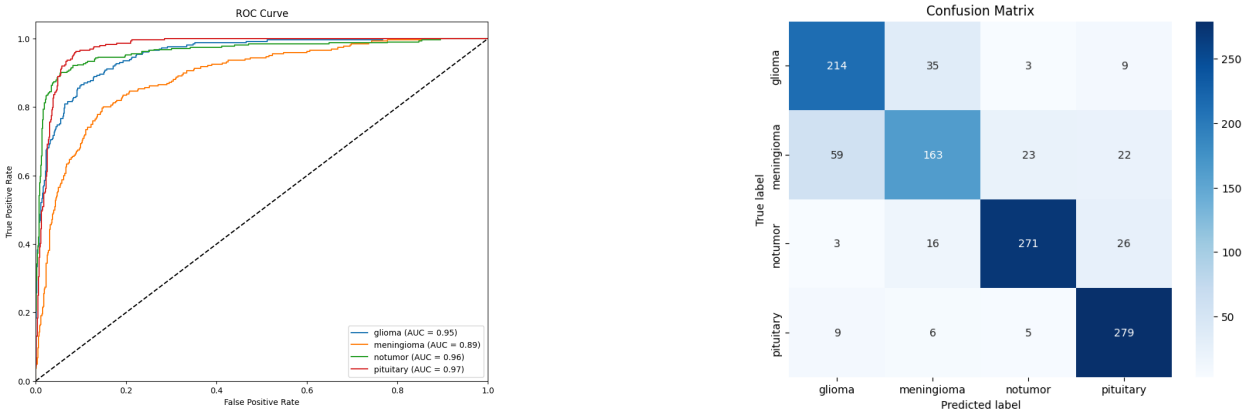

Figure 5: Model Performance on Validation Set

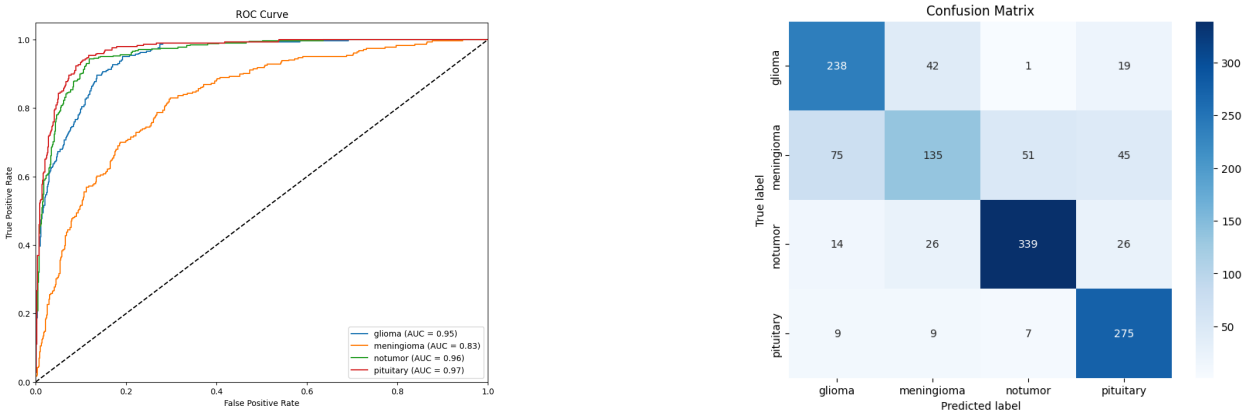

Figure 6: Model Performance on Test Set

Performance measures: Validation Set: accuracy (0.8110), precision (0.8098), recall (0.8110), and F1-score (0.8077); Test Set: accuracy (0.7529), precision (0.7463), recall (0.7529), and F1-score (0.7430).

2 Improved CNN Model

2.1 Model Architecture

The improved Convolutional Neural Network (CNN) is designed to classify grayscale images of size 30×30 into four classes: glioma, meningioma, notumor, and pituitary. The model processes the input images through a series of convolutional and max-pooling layers, employing the ReLU activation function for its efficiency in learning non-linear patterns. 'Same' padding is applied throughout the convolutional layers to retain spatial resolution.

2.2 Handling Imbalances

The overrepresentation of the "notumor" class leads to imbalance, causing the model to favor "notumor" predictions. Additionally, the baseline model confuses Classes 0 and 1 (glioma and meningioma) the most (See Table). This may be due not only to their minority status in the training set, but also to the fact that glioma tumors are highly heterogeneous[14].

To address this issue, we applied several approaches, with our final model incorporating SMOTE (Synthetic Minority Over- sampling Technique) [8] and focal loss.

SMOTE was already successfully employed for image augmentation in the medical field for X-rays [21] to tackle imbalanced datasets. Our strategy was to oversample glioma class only to around 85% of the "notumor" count. Previous attempts at oversampling both meningioma and glioma resulted in model accuracy that was 0.04 lower overall, with a slight decrease in accuracy for each class. Different percentages were attempted.

Another successful strategy was proven to be Focal Loss, which modifies the Cross-entropy Loss by focusing the training on difficult samples (gamma), while also minimizing the influence of easily classified examples (alpha). After applying focal loss, the model's overall accuracy increased from $85.66\% \pm 0.37$ to $87.41\% \pm 0.23$. The most notable improvements are seen in the precision and F1-score for the glioma class, which increased from 0.80 ± 0.1 to 0.84 ± 0.1 and 0.81 ± 0.1 to 0.83 ± 0.1, respectively. Additionally, the recall for meningioma improved significantly from 0.69 ± 0.1 to 0.80 ± 0.1, indicating better sensitivity in identifying this class. The best-performing gamma and alpha were 2 and 0.25 respectively. We explored integrating external datasets [10] to improve the model's performance. Specifically, we attempted to use an open dataset containing MRI images with three classes: glioma, meningioma, and pituitary tumors. Unfortunately, the absence of a "notumor" class in this external dataset presented challenges. As a result, we couldn't fully leverage it for external validation, as a complete validation set would ideally encompass all classes to assess the model's generalization accurately. Validating on a dataset without the "notumor" class might lead to biased performance metrics, overlooking the model's ability to distinguish between tumor and non-tumor cases.

It is worth mentioning that we also experimented with augmenting the glioma class using lower-quality images derived from an external dataset [10]. However, these augmentations did not significantly improve the model's performance. The poorer results might be attributed to inconsistencies between the external dataset's characteristics and our original data. Additionally, augmentation with low-quality external images might have added noise to the training process, reducing the model's ability to learn meaningful features.

2.3 Hyperparameter Tuning

To identify the best hyperparameter configuration, we used Optuna, an open-source framework designed to automate the exploration and optimization of hyperparameter search spaces [9]. We chose Optuna for two reasons. First of all, a previous study compared four Python libraries and found that Optuna outperformed the others in addressing CASH (Combined Algorithm Selection and Hyperparameter Optimization) problems [22]. Finally, Optuna has been successfully employed in previous research related to tumor classification for optimization [20], further validating its effectiveness for our dataset.

Based on the best hyperparameter configuration determined by Optuna in Table 1, we manually adjusted the dropout rate from 0.04 to 0.07, which yielded a better performance. Then, we preferred using focal loss over the suggested cross-entropy as our loss function. As explained in the previous section, unlike the traditional cross-entropy loss, focal loss gives greater emphasis to hard-to-classify and misclassified examples [5].

The training was conducted using Adam optimizer, with a learning rate of 0.000319 (a hyperparameter selected by Optuna), due to its adaptive learning capabilities and effectiveness for complex neural networks. Adam is favored over other optimizers due to its efficient memory usage, ability to manage noisy data, and faster convergence speed.[18] Moreover, Adam was also used in the baseline model and it is a popular choice in works on this dataset [1], further validating our choice. To improve Adam's performance, we initially used two learning rate callbacks in Keras: LearningRateScheduler and ReduceLROnPlateau. The LearningRateScheduler keeps the initial learning rate constant for the first seven epochs, and then decreases it exponentially. ReduceLROnPlateau monitors model performance and reduces the learning rate if no improvement is seen over a set number of epochs. After applying these callbacks in our early hyperparameter tuning stages, our model's accuracy initially improved from 0.63 to 0.68. However, it was not employed in our final model because it did not perform as positively later on.

Early stopping with a patience of 2 epochs is used to prevent overfitting, and halting training if no improvement in validation loss is observed, even though the performance on the training data continues to improve. Higher patience (3, 4) proved to be less effective, leading to worse results in our Accuracy on the test set.

2.4 Results

Although a random seed was used, some variations in the results occurred. It was due to the employment of the Adam optimizer, which can sometimes generalize worse and exhibit more variability in results due to its stochastic nature [8].

The performance of our Best CNN model significantly improved across all metrics compared to the baseline model. Accuracy rose from 0.7529 to 0.8871, and recall improved correspondingly, indicating a reduction in missed tumor cases—essential in medical diagnostics. Precision increased from 0.7463 to 0.8859, and the F1-Score similarly enhanced from 0.7430 to 0.8860.

AUC scores saw notable gains: glioma improved from 0.95 to 0.98, meningioma from 0.83 to 0.96, no tumor from 0.96 to 0.99, and pituitary from 0.97 to 0.99. This highlights better class distinction, especially for meningioma, as evidenced by Fig. 7's confusion matrix. The model achieved a balanced accuracy of 0.8871 ± 0.0053, reducing errors across classes. The ROC Curve in Fig. 7 also shows robust AUC values, particularly for the notumor and pituitary classes.

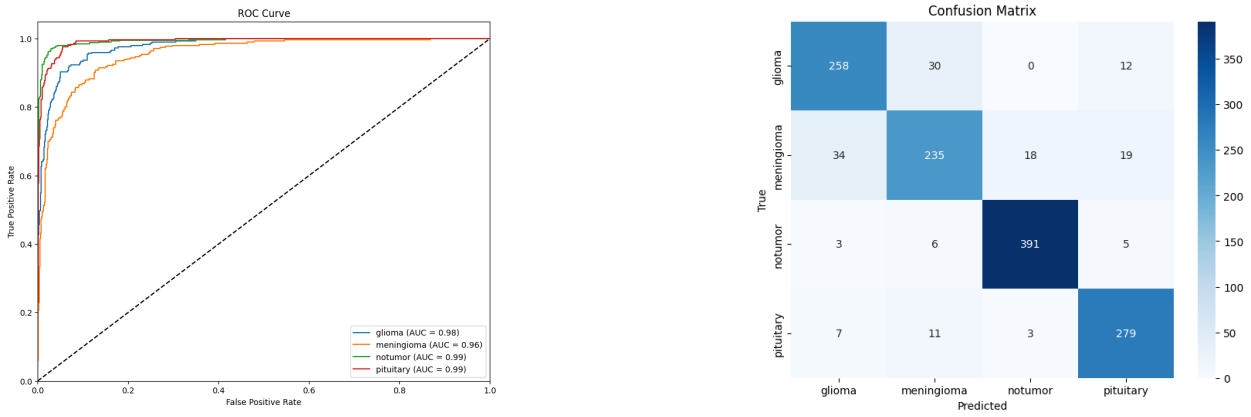

Figure 7: Best Model Performance

2.5 Further Improvements

Our model's performance was heavily impacted by the use of low-resolution images (30 x 30 x 1) in grayscale. Previous studies on the same dataset have shown improved results with higher-resolution images, suggesting that the model's performance could potentially be enhanced by increasing the image resolution. The reliance on low-resolution images also has consequences for data augmentation; since our dataset is relatively small, augmentation techniques were employed to artificially increase its size. However, when applied to low-resolution images, these augmentations can introduce noise and potentially degrade the quality of artificially created data, leading to limited performance.

Therefore, besides these more obvious improvements, we propose the following approaches based on our observations and existing literature.

(1) In terms of preprocessing steps, **automatic image segmentation** techniques could be adopted. By isolating regions of interest (e.g., tumors) from the MRI images, the model could focus on learning relevant features without noise from unrelated structures. U-Net-based architectures have been highly effective in segmentation tasks within medical imaging [16]. However, to our best knowledge there are not conclusive studies on their use on lower-resolution images.

(2) The absence of a "notumor" class in our external dataset limited its usefulness. If we had access to a **complete external dataset** or had been able to incorporate non-tumoral MRI images as a fourth class, we could have significantly expanded the dataset. A larger dataset with diverse real-world examples would likely improve the model's ability to generalize.

(3) While **random search** is efficient and performs well for hyperparameter tuning [4], we were unable to use RandomizedSearchCV from sklearn because of compatibility problems with our environment. This limited our ability to properly explore hyperparameters [3]. Access to a more suitable setup, like the powerful computing resources at the university, would have allowed for a full search and likely led to better model optimization.

(4) Optimizing our model with a **cyclical learning rate (CLR)**, essentially changing the learning rate cyclically during training, helping to explore different learning rates, escape local minima, and improve generalization [18]. However, since this approach requires the "TensorFlow-addons" library, we encountered installation issues and could not implement it.

(5) If we were able to merge our dataset with a complementary external dataset, a **Multi-Scale Fusion Convolution Network (MFCN)** could be applied to achieve super-resolution for MRI images [23] [12]. This approach leverages multi-scale feature extraction and fusion techniques, enhancing image quality. By addressing the common issues of lower resolution in MRI scans caused by limited acquisition time or hardware constraints, MFCNs improve the overall detail and clarity of the images, allowing for more accurate diagnostic outcomes. Again, this could be attempted only if the dataset was large enough [13].

3 Transfer learning (VGG16)

This section will be divided into two parts. The first part will explore the implementation of VGG16 starting from the provided low-resolution images (30x30x1). The second part will focus on the application of the model using the original high-resolution images, reduced in size to 128x128x3. This adjustment was necessary as transfer learning models like VGG16 are not suitable for the very low resolution of 30x30x1 images, requiring a larger input size (at least 32x32x3) to use the model at all. Thus, to avoid penalization, we provided both approaches.

3.1 Model 1

Although we also attempted to use ResNet50 and DenseNet121 as well, we decided to focus on VGG16 due to promising results on low-resolution images on cancer images [11]. However, to our best knowledge, there are no studies conducted on transfer learning with VGG16 on low-resolution images in greyscale on tumor classification. We resized our original images to three dimensions: (32x32x3), (64x64x3), and (96x96x3). While this resizing allowed the images to be compatible with our transfer learning model, it did not enhance their quality or add color. This resizing was essentially a form of downsampling, which did not enhance image quality or add color. Instead, the process introduced additional noise due to interpolation and scaling,

altering the images without improving their content for analysis. As we had to freeze the layers until the fully connected layer, no training on greyscale images as Standford's Medical ImageNet [6] or the already succesful CINIC-10 [4] was possible. In fact, some literature found that training transfer learning on similar greyscale images, instead of the general Imagenet, could be beneficial when classifying greyscale images themselves [4]. Initially, we replicated the layer structure from a study on low-resolution medical images [11], but performance was no better than random guessing. Reverting to our original CNN architecture improved accuracy, with the 60x60x3 model performing best (0.72 ± 0.02), followed by 32x32x3 (0.71 ± 1), and 96x96x3 (0.69). This might suggest that larger images improve model accuracy until excessive noise is introduced. The findings are supported by studies on low-quality cancer images, where performance improves with increasing resolution. Unlike our approach, which tested VGG16 by upscaling small images, their study resized larger images to smaller dimensions. Although this method obtained lower results than our best-performing CNN model, our overall accuracy was comparable to the one of our baseline model (0.7628 ± 0.0082).

3.2 Model 2

To further investigate transfer learning with VGG16, we modified certain aspects of the preprocessing. Specifically, we adjusted the input image resolution to 128x128x3 and changed the color mode from 'grayscale' to 'rgb', given that VGG16 is optimized for color images. After implementing these changes, the classification accuracy improved to 0.8154, which is approximately 6% higher than the baseline model but 7% lower than the best CNN model. Inspired by Belaid, O. N., & Loudini, M.[2], we then explored whether adding additional GLCM energy features could enhance performance further. However, contrary to expectations, the accuracy dropped to 0.7796, only slightly higher than the baseline model. This decline in performance may be attributed to the fact that VGG16 has already learned rich and complex feature representations through pretraining on large-scale datasets such as ImageNet. Adding GLCM texture features likely led to redundancy, introducing noise and consequently reducing the overall performance of the model. [19]

With more time, we could have explored the reasons for the lack of improvement, especially focusing on overfitting and class imbalance [8]. Given adequate computational power, we could have also experimented with larger image sizes, such as (224x224x3).

Fig. 8 demonstrates the performance of best transfer learning model (image resolution is 128x128x3).'notumor' class was the best-predicted class with 378 correct predictions, followed by 'pituitary' and 'glioma'. meningioma' had the fewest correct predictions, with 192 out of 306.

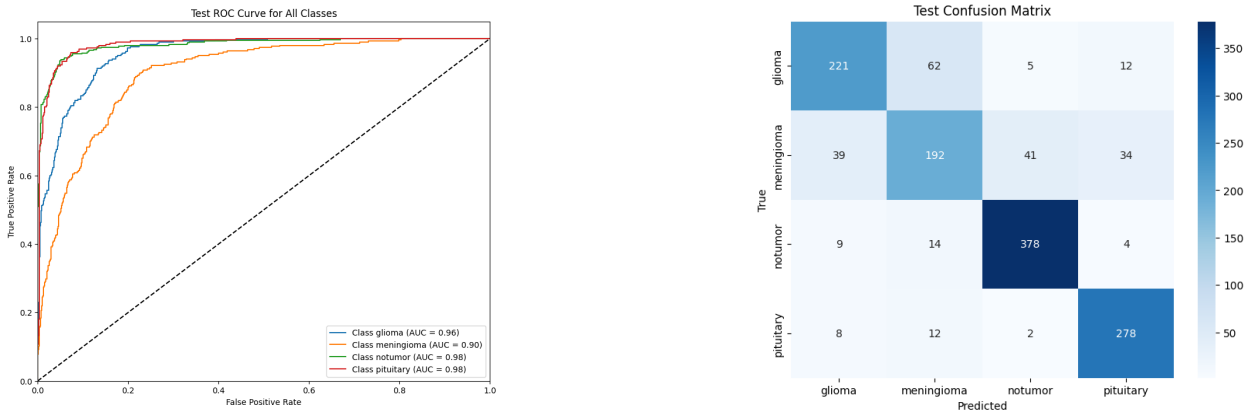

Figure 8: Best Transfer Learning Model Performance

Member	Contributions
Hsuan Jung Chu (2109692)	Baseline model construction, Hyperparameter optimization, Transfer learning (ResNet50), Code organization, Report on Hyperparameter tuning (Optuna)
Weitong Shen (2112815)	Baseline model construction, Hyperparameter optimization, Transfer learning (VGG16), Report organization
Yiting Shen (2117507)	Plotting test results, Hyperparameter tuning, Transfer learning (DenseNet121), improved CNN model and results section
Yue Wang (2111066)	Baseline implementation, Searching for Python packages, Hyperparameter tuning with Optuna, Data processing, Plots in report, Final code collection, Transfer learning (ResNet50)
Angela Barone (2128404)	Plotting class distribution, Hyperparameter tuning, Final CNN with SMOTE, Augmentation with external data, VGG16 Model 1, Report on Introduction, CNN improved model
Xiaojun Liu (2110171)	Plotting test results, Hyperparameter tuning, Transfer learning (VGG16, DenseNet121), Results section

Table 1: Group Members and Contributions

References

[1] Abdullah A. Asiri, Ahmad Shaf, Tariq Ali, Muhammad Aamir, Muhammad Irfan, and Saeed Alqahtani. Enhancing brain tumor diagnosis: an optimized cnn hyperparameter model for improved accuracy and reliability. *PeerJ Computer Science*, 10, 2024.

[2] Ouiza Nait Belaid and Malik Loudini. Classification of brain tumor by combination of pre-trained vgg16 cnn. *Journal of Information Technology Management*, 12:13–25, 6 2020.

[3] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

[4] Antonio Bruno, Davide Moroni, and Massimo Martinelli. Efficient adaptive ensembling for image classification. In *Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR)*, Pisa, Italy, 2024. These authors have contributed equally to this work and share first authorship.

[5] Jianxiang Dong. Focal loss improves the model performance on multi-label image classifications with imbalanced data. In *Proceedings of the 2nd International Conference on Industrial Control Network And System Engineering Research*, pages 18–21, 2020.

[6] Stanford Center for Artificial Intelligence in Medicine Imaging. Medical imagenet (medimagenet). https://aimi.stanford.edu/medical-imagenet. Accessed on [your access date].

[7] Sarah Ali Abdelaziz Ismael, Ammar Mohammed, and Hesham Hefny. An enhanced deep learning approach for brain cancer mri images classification using residual networks. *Artificial Intelligence in Medicine*, 102:101779, January 2020.

[8] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain. Effective class-imbalance learning based on smote and convolutional neural networks. *Applied Sciences*, 13(6):4006, 2023.

[9] Johnsymol Joy and Mercy Paul Selvan. A comprehensive study on the performance of different multiclass classification algorithms and hyperparameter tuning techniques using optuna. In *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, pages 1–5, 2022.

[10] Deniz Kavi. Brain tumor dataset. https://www.kaggle.com/datasets/denizkavi1/brain-tumor, 2023. Accessed: [Insert Access Date].

[11] MD Reyad Hossain Khan, Abdul Hasib Uddin, Abdullah-Al Nahid, and Anupam Kumar Bairagi. Skin cancer detection from low-resolution images using transfer learning. *Khulna University, Khulna-9208, Bangladesh*, 2021.

[12] Zhengchun Lin, Siyuan Li, Yunzhi Jiang, Jing Wang, and Qingxing Luo. Feedback multi-scale residual dense network for image super-resolution. *Signal Processing: Image Communication*, 107:116760, 2022.

[13] Chang Liu, Xi Wu, Xi Yu, YuanYan Tang, Jian Zhang, and JiLiu Zhou. Fusing multi-scale information in convolution network for mr image super-resolution reconstruction. *BioMedical Engineering OnLine*, 17:114, 2018.

[14] National Brain Tumor Society. Brain tumor facts. Accessed: 2024-09-26.

[15] Masoud Nickparvar. Brain tumor mri dataset. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset/discussion/482896, 2022.

[16] Rafael Núñez-Martín, Ricardo Cubedo Cervera, and Mariano Provencio Pulla. El tumor del estroma gastrointestinal y la aparición de segundos tumores: revisión de la bibliografía. *Medicina Clínica (English Edition)*, 149(8):345–350, 2017.

[17] Quinn T. Ostrom, Hayley Gittleman, Lindsay Stetson, Smita Virk, and Jill S. Barnholtz-Sloan. Epidemiology of gliomas. In *Gliomas*, volume 163 of *Cancer Treatment and Research*, pages 1–14. Springer, 2016.

[18] Mohaimenul Azam Khan Raiaan, Sadman Sakib, Nur Mohammad Fahad, Abdullah Al Mamun, Md Anisur Rahman, Swakkhar Shatabda, and Md Saddam Hossain Mukta. A systematic review of hyperparameter optimization techniques in convolutional neural networks. *Decision Analytics Journal*, page 100470, 2024.

[19] Ahmed Saihood, Hossein Karshenas, and Ahmad Reza Naghsh Nilchi. Deep fusion of gray level co-occurrence matrices for lung nodule classification. *PLOS ONE*, 17(9):1–26, 09 2022.

[20] Christy Atika Sari, Eko Hari Rachmawanto, Erna Daniati, Fachruddin Ari Setiawan, Agoes Santika Hyperastuty, and Ery Mintorini. Breast tumor classification using adam and optuna model optimization based on cnn architecture. *Journal of Soft Computing Exploration*, 5(2):153–165.

[21] D Schaudt, R von Schwerin, A Hafner, P Riedel, M Reichert, M von Schwerin, M Beer, and C Kloth. Augmentation strategies for an imbalanced learning problem on a novel covid-19 severity dataset. *Scientific Reports*, 13(1):18299, 2023.

[22] Shashank Shekhar, Adesh Bansode, and Asif Salim. A comparative study of hyper-parameter optimization tools. *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6, 2021.

[23] Qiling Tang, Yangyang Liu, and Haihua Liu. Medical image classification via multiscale representation learning. *Artificial Intelligence in Medicine*, 79:71–78, 2017.