Deep Learning Project

September 29, 2024

Group 4
Rick Jagt, u1232708
Nick Klotz, u2108526
Marleen van Lubeek, u1277963
Lotte Michels, u366836
Mohammed Osman, u974796
Tonny Su, u2106082

Student Contributions:

***Rick Jagt**: Building transfer learning model, Literature review, Implementing hyperparameter tuning, Report: Revision*
***Nick Klotz**: Building improved learning model, literature review for hyperparameter tuning, implementing hyperparameter tuning, Report: baseline and hyperparameter tuning methodology, part of results and discussion for improved model, revision*
***Marleen van Lubeek**: Model performance visualization, Report: Results improved model, Discussion, Conclusion, Revision*
***Lotte Michels**: Implementing code standards (cleaning and commenting), Code debugging, Data and model performance visualization, Report: Layout, Figures, Methodology, Results, Revision*
***Mohammed Osman**: Building baseline model, data visualization, Literature review, Report: Discussion*
***Tonny Su**: Report: Introduction, Transfer learning sections, Results*

1 Introduction

Image classification is one of the most frequently studied topics in the field of computer vision. Using computer vision and deep learning techniques, the task for this assignment is to detect and diagnose various brain tumors. In this report, Section 2 discusses the methodology used for this project. The loading, pre-processing and data visualization steps will be presented. Furthermore, the baseline model, hyperparameter tuning, and the transfer learning model are discussed. Section 3 presents the results of the performance evaluation and the performance comparison of these models mentioned. Section 4 presents a discussion about proposed improvements to further enhance the performance of the models. Section 5 discusses the findings regarding the objective of this assignment. Finally, the references can be found in the added bibliography.

2 Methodology

In this section, the preprocessing, training and optimization steps taken during the project will be presented and justified. For the creation of the neural networks, functionalities from the scikit-learn (Virtanen, 2020) and Keras modules were used (Abadi et al, 2015).

2.1 Loading, pre-processing and data visualization

For this project, the Brain Tumor MRI Dataset from Nickparvar (2024) was used. This dataset comprises 7023 MRI images of the human brain. These images are divided into four categories: glioma, meningioma, no tumor and pituitary. The data was loaded and preprocessed through code provided by the Deep Learning course instructor: images loaded in batches of 32 samples that were rescaled to a 30 by 30 pixel resolution, mapped to grayscale and supplemented with random noise. It was further made sure that the image labels were one-hot encoded.

Next, using the Matplotlib library, two plots were constructed for data visualization and exploration purposes (Hunter, 2007). Firstly, 15 randomly selected samples from the data set were visualized, as shown in Figure 1. This figure indicates the different angles, shapes and sizes of brain representations in the data.

Secondly, a bar chart was created to visualize the distribution of the classes in the data. Figure 2 indicates that the ‘no tumor’ class contains the largest amount of samples (precisely 1996 instances) and the ‘glioma’ class contains the least amount of samples (precisely 1581 instances). All four classes contain between 1500 and 2000 data points. The complete dataset originally comprised 5712 training and 1311 testing samples. A validation set was created out of 1143 random samples from the training data (20%) for hyperparameter tuning.

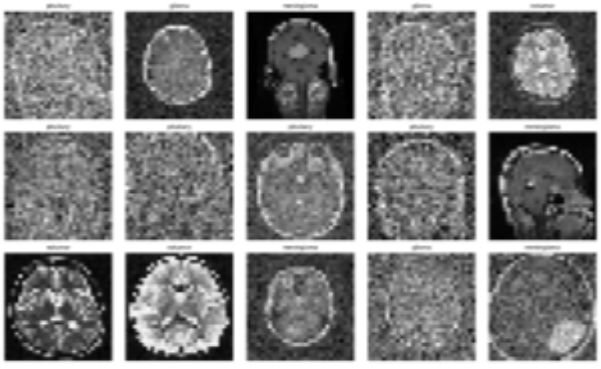

Figure 1: Visualization of 15 random samples

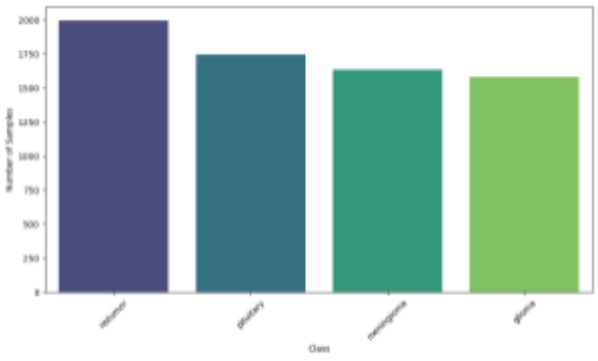

Figure 2: Data class distribution

2.2 Baseline model and hyperparameter tuning

The baseline model was built according to assignment instructions. Hyperparameters were selected based on relevant literature. Several studies made predictions for MRI brain tumor classification and proposed rigorous hyperparameter tuning. A lot of studies focussed on improving transfer learning or building complicated models. Our study uses input of lower dimension and therefore requires less complicated models to prevent overfitting and to reduce computational load. Patil et al. (2021) used the simplest models but did not specifically focus on hyperparameter tuning. However, because they and other studies used different layer configurations with transfer learning and kernel sizes, we added additional convolutional layers and filter sizes as a hyperparameter. Minarno et al. (2021) and Asiri et al. (2024) focused on hyperparameter tuning and compared the number of filters, dense units, dropout rate, and optimizer function. Additionally, we added the activation function and kernel size based on studies with larger models (Asiri et al. 2024). We used ranges on the lower end and close to optimal values found in these and other studies. We considered other hyperparameter optimization methods but due to software constraints, we opted for gridsearch as used in the previously mentioned literature.

2.3 Transfer learning model

For the transfer learning model, the VGG16 architecture was employed for feature extraction. A combination of AlexNet and GoogLeNet simulations have also been considered. First of all, the input shape of the images were reshaped to a 32x32 resolution, as VGG16 does not accept a 30x30 resolution. Next, all the VGG16 model layers were frozen and fully connected layers were added, consisting of a pooling, dense, and output layer. Only the parameters in these added fully connected layers were trainable. Moreover, a gridsearch was run to optimize the model through finding the optimal values for the number of units per dense layer, learning rate and optimizer.

3 Results

As described in the previous section, three tumor classification models were built: a baseline neural network, an improvement of this baseline model, and a transfer learning model. This section will focus on the evaluation and comparison of these models.

3.1 Evaluation of the baseline model

The baseline model showed an accuracy score of 0.810, a precision score of 0.813, a recall score of 0.810, and a f1-score is 0.807 on the validation set. For the test set, the accuracy score is 0.775, the precision score is 0.773, the recall score is 0.775, and the f1-score is 0.766.

In Figure 3, the loss and accuracy plots for the baseline model are presented. Regarding the loss plot, the best result was achieved with a loss of 0.486 after 10 epochs for the validation set. The accuracy plot shows the best result with an accuracy of 0.818 after 10 epochs for the validation set.

In Figure 4, the confusion matrices for the baseline model can be seen. On both the validation and test set, the results indicate that the predictions are more accurate for the notumor and pituitary classes compared to the glioma and meningioma class, with the latter scoring the worst in accuracy.

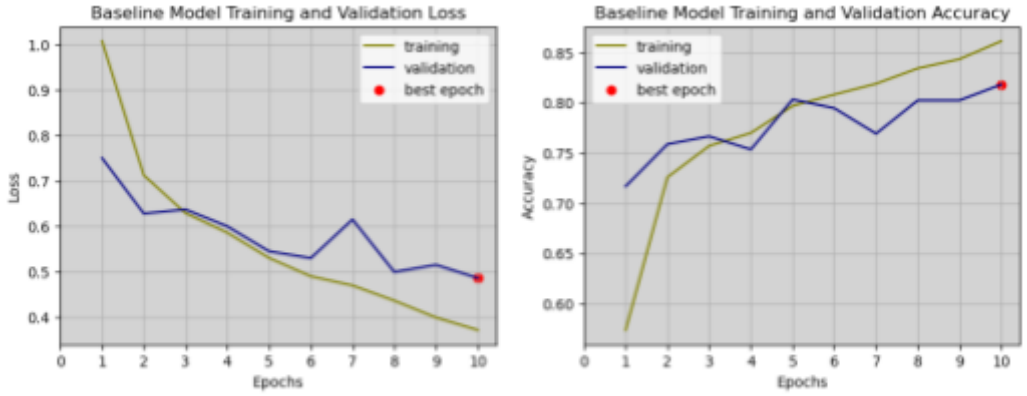

Figure 3: Baseline model loss (left) and accuracy (right) scores

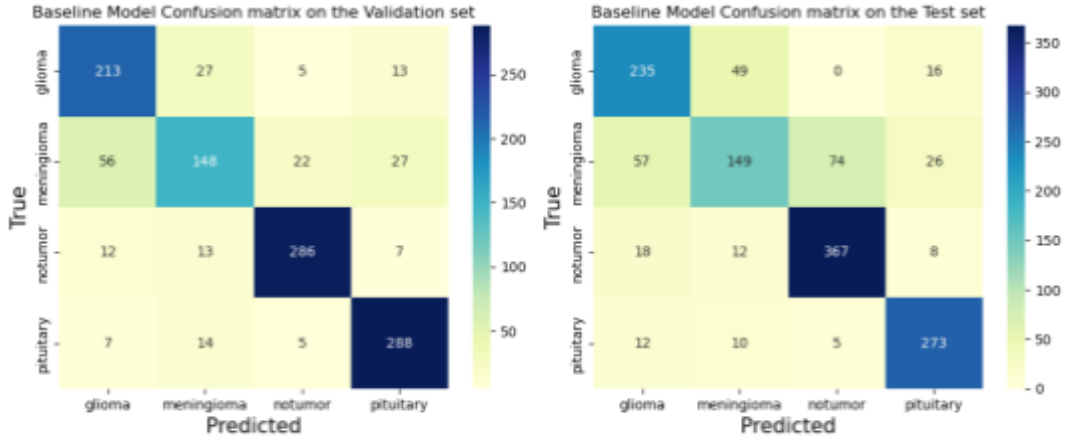

Figure 4: Baseline model confusion matrices for the validation set (left) and test set (right)

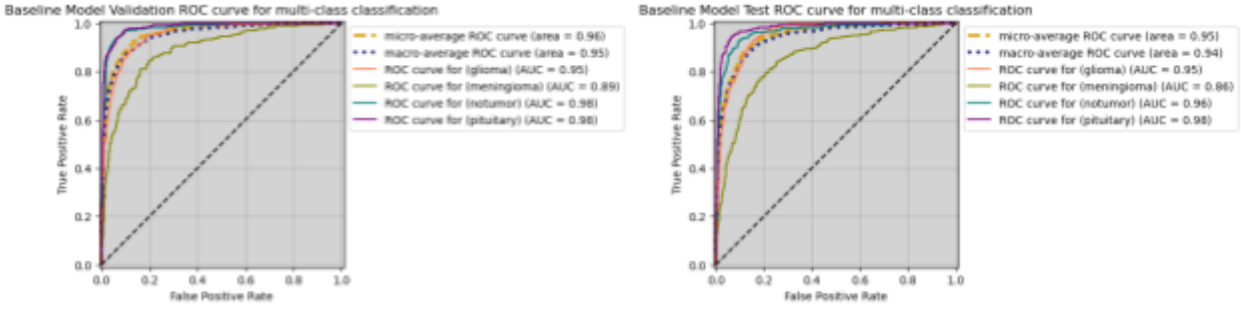

Figure 5: Baseline model ROC curves for validation set (left) and test set (right)

In Figure 5, the ROC curves for multi-class classification for the baseline model are presented. Regarding the test set, the most important results are that the highest AUC achieved is 0.97 for the pituitary class, and an average of around 0.92 AUC regarding all components.

3.2 Evaluation of the improved model

The optimal hyperparameters found via gridsearch were 3 convolutional layers, 32 filters per layer with kernel size 5, Relu activation function, Adam optimizer, and 64 units in the dense layer. The resulting model contained 152,772 parameters. On the validation set, the accuracy score is 0.822, the precision score is 0.830, the recall score is 0.822, and the f1-score is 0.825. For the test set, the accuracy score found is 0.822, the precision score is 0.822, the recall score is 0.822, and the f1-score is 0.822.

As can be seen in the loss and accuracy plot for the improved model in Figure 6, the best result has been achieved with a loss of 0.486 and an accuracy of 0.834 after 6 epochs for the validation set. As the training loss keeps improving and the validation does not overfitting occurs with over 6 epochs.

As can be seen in the confusion matrix for the test set in Figure 7, accuracy is highest for the pituitary class, slightly less high for the notumor and meninglioma classes, and lowest for the glioma class. Furthermore, as can be seen in the ROC curves in Figure 8, the highests AUC scores of 0.99 and 0.98 are achieved for the pituitary and notumor classes respectively, and an average of around 0.96 AUC regarding all components. Compared to the baseline model, the improved model achieved an improvement with higher accuracy, similar loss in fewer epochs, and a higher average AUC.

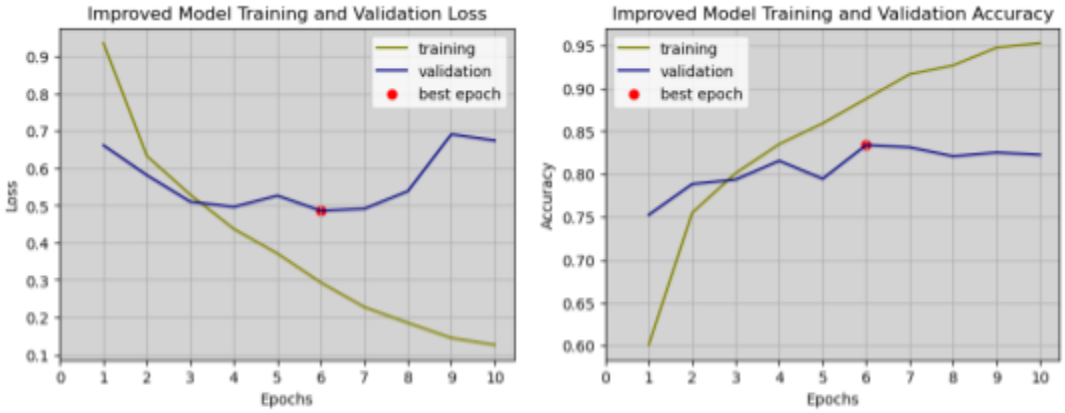

Figure 6: Improved model loss (left) and accuracy (right) scores

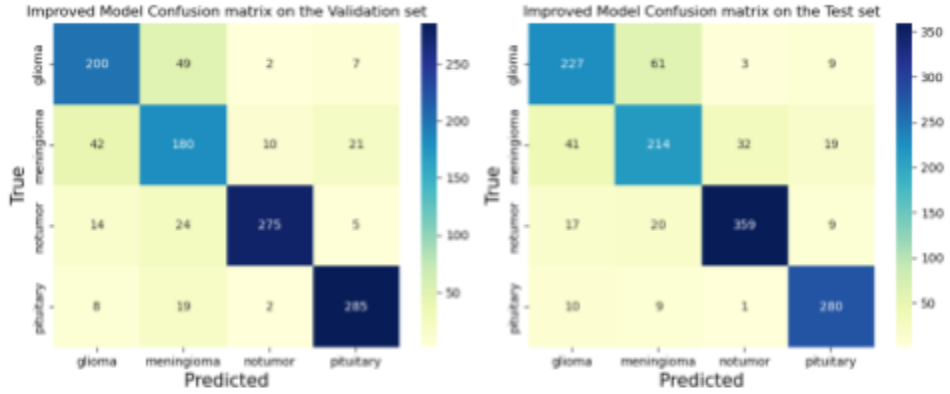

Figure 7: Confusion matrix validation set (left) and test set (right)

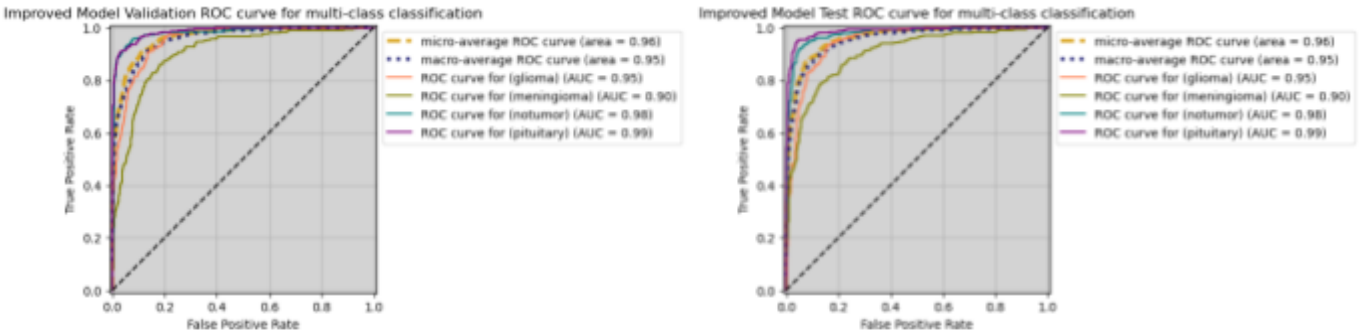

Figure 8: ROC curve validation set (left) and test set (right)

3.3 Evaluation of the transfer learning model

The optimal hyperparameters found via gridsearch were 2048, 1024 and 768 units for the three dense layers respectively, a learning rate of 0.001 and an Adam optimizer. The resulting model contained 44,650,852 parameters, of which 8,133,380 were trainable. On the validation set, the accuracy score is 0.666, the precision score is 0.665, the recall score is 0.666, and the f1-score is 0.660. For the test set, the accuracy score found is 0.673, the precision score is 0.671, the recall score is 0.673, and the f1-score is 0.667. These performance metrics are compared to the baseline and improved model.

In Figure 9, the loss and accuracy plots for the transfer learning model can be seen. Regarding the loss plot, the best result was achieved with a loss of 0.874 after 3 epochs for the validation set. Afterwards overfitting occurs. Compared to the baseline model and improved model, the loss score thus remained highest for the transfer learning model. For the accuracy plot, the best result was achieved with an accuracy of 0.702 after 5 epochs for the validation set. Compared to the baseline model and improved model, the transfer learning model achieved the lowest accuracy score of the three models.

As can be seen in the confusion matrix in Figure 10, the transfer learning model predictions are more accurate for the notumor and pituitary classes, compared to the glioma and meningioma classes. Compared to the baseline model and improved model, the transfer learning model achieved the lowest overall accuracy of the three models.

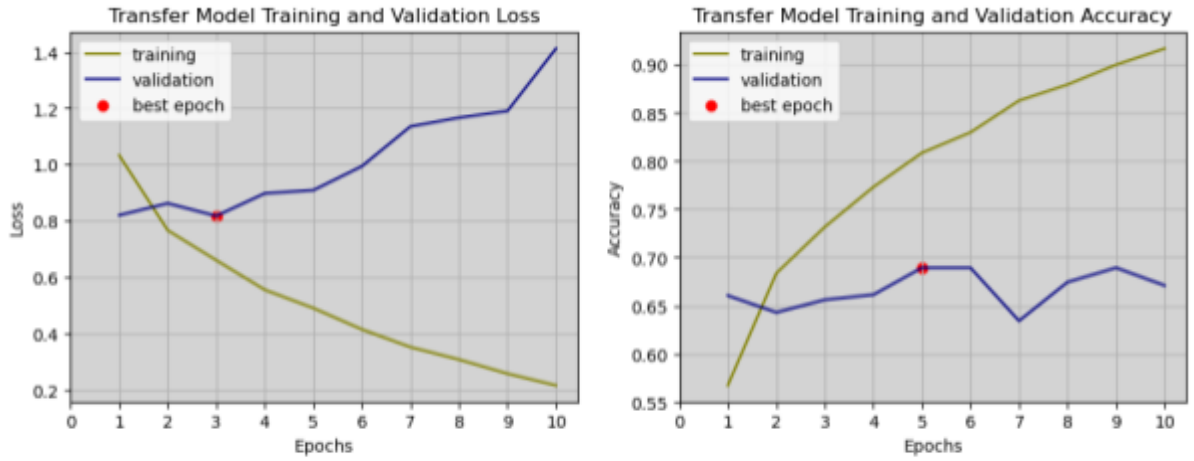

Figure 9: Transfer model loss (left) and accuracy (right) plots

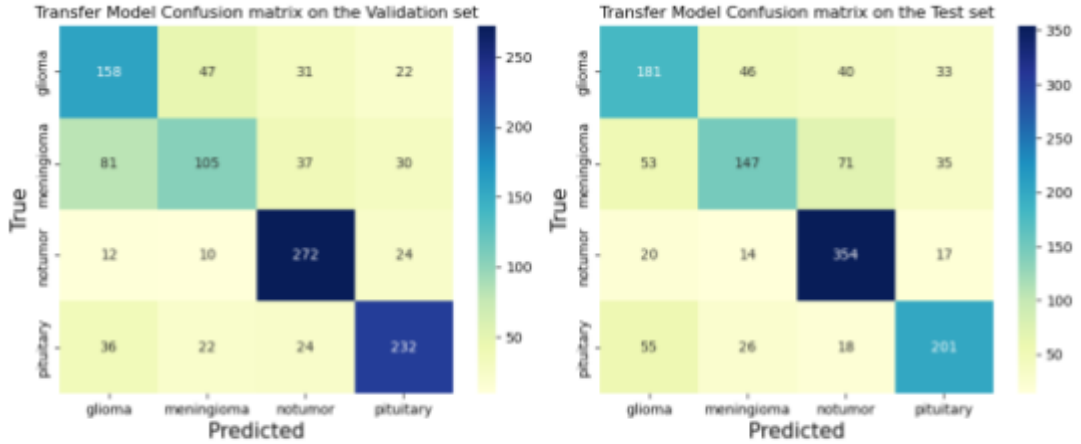

Figure 10: Transfer model confusion matrices on the validation set (left) and test set (right)

Regarding the ROC curves for multi-class classification for the transfer learning in Figure 11, the most important results are that the highest AUC achieved is 0.93 for the notumor and pituitary classes, and an average of around 0.87 AUC regarding all components. Compared to the baseline model and improved model, the transfer learning model achieved the lowest average AUC of the three models.

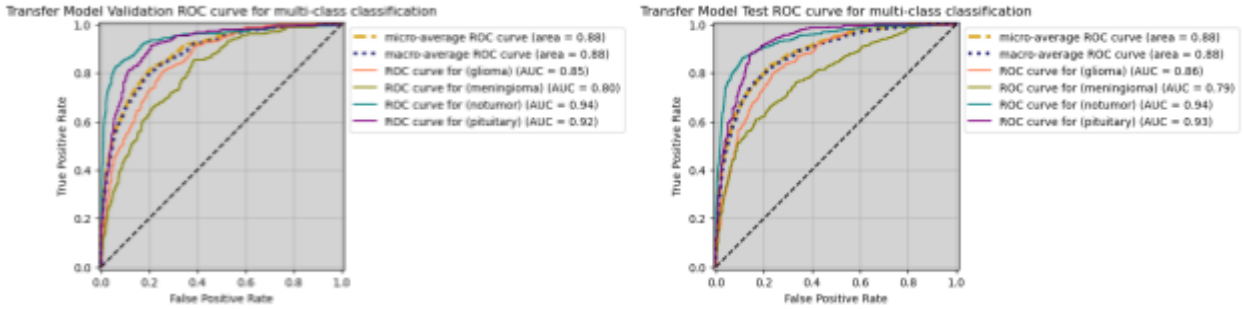

Figure 11: Transfer learning model ROC curves for validation set (left) and test set (right)

4 Discussion

We can conclude that the improved model performed best. It had highest scores on accuracy, precision, recall and f1-score. It also had the highest AUC values, which substantiates this observation. This was to be expected as the improved model used hyperparameter tuning to improve the baseline model and could not perform worse. It performs reasonably well at the classification task. However, a recall score of 0.822 is likely too low to reliably use in a clinical setting as this would result in an unacceptable amount of missed diagnoses. Of the 3 models, the transfer learning model performed the worst, which was somewhat unexpected. The reason for this could be that the transfer learning model might be trained on sets that do not match our target, therefore not making good predictions. Possibly, because of lower data quality of our dataset compared to higher ones likely used to build the external models.

In order to improve future models, higher quality data should be used as this would likely increase the ability to learn of our models. Furthermore, combining CNNs with LSTM units could capture spatial and temporal features in sequential data. Additionally, using more powerful transfer learning models like ResNet, DenseNet or EfficientNet could significantly improve classification accuracy by leveraging deeper architectures optimized for medical imaging tasks (He et al., 2016; Huang et al., 2017; Tan & Le, 2019). Advanced data augmentation techniques, such as elastic deformations, or generating synthetic data with GANs could help address class imbalance and improve model robustness (Frid-Adar et al., 2018). Furthermore, applying optimization methods like Bayesian optimization and enhancing regularization techniques, such as batch normalization or adaptive loss functions, could refine model performance and mitigate issues like class imbalance or overfitting (Lin et al., 2017).

5 Conclusion

In conclusion we found that the improved model performed best in comparison to the baseline and the transfer learning model. Better detection of brain tumors is important, hospitals and doctors will be able to perform better, because of early and more accurate detection. For future research there are still multiple techniques that can be explored to improve these models for recognising brain tumors.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems* SoftwareSoftwareSoftware. https://www.tensorflow.org/

Ait Amou, M., Xia, K., Kamhi, S., & Mouhafid, M. (2022, March). A novel MRI diagnosis method for brain tumor classification based on CNN and Bayesian Optimization. In *Healthcare* (Vol. 10, No. 3, p. 494). MDPI.

Asiri, A. A., Shaf, A., Ali, T., Aamir, M., Irfan, M., & Alqahtani, S. (2024). Enhancing brain tumor diagnosis: an optimized CNN hyperparameter model for improved accuracy and reliability. *PeerJ Computer Science*, *10*, e1878.

G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 2018, pp. 289-293, doi: 10.1109/ISBI.2018.8363576.

Minarno, A. E., Mandiri, M. H. C., Munarko, Y., & Hariyady, H. (2021). Convolutional neural network with hyperparameter tuning for brain tumor classification. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*.

Multiclass Receiver Operating Characteristic (ROC). (z.d.). Scikit-learn. https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

Nickparvar, M. (2024). *Brain Tumor MRI Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

Patil, S., Kirange, D., & Nemade, V. (2020). Predictive modelling of brain tumor detection using deep learning. *Journal of Critical Reviews*, *7*(04), 1805-1813.

T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2999-3007, doi: 10.1109/ICCV.2017.324.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (ICML)*, 6105-6114.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). *SciPy 1.0: Fundamental algorithms for scientific computing in Python*. *Nature Methods, 17*(3), 261-272. https://doi.org/10.1038/s41592-019-0686-2