Machine Learning Report

Abstract

This report presents the findings of using deep-learning image classification model designed to identify brain scans with tumour. 223 images were used to train and evaluate the model: 150 with tumour and 73 without. The DenseNet121 architecture was used for its efficient feature reuse and ability to serve as a powerful pretrained feature extractor. The model was trained using pre-trained ImageNet weights, improving feature extraction and generalisation. The results indicate that such a model is well-suited for classifying medical brain scans, especially with a limited dataset.

Introduction

Brain cancer remains one of the most fatal types of cancer in the world. Malignant brain and central nervous system cancers have contributed to over 250,000 deaths in just 2020 worldwide[1]. However, early detection and treatment have shown to significantly increase survival rates, emphasising the importance of prompt diagnosis.

Traditionally, trained health practitioners identify tumours through analysing MRI and CT scans of the brain's patient. However, the ML algorithms, particularly deep learning models[2], have advanced to a point that they can now help and assist with identifying them. For instance, a recent AI model named CHIEF, developed by Harvard Medical School in 2024, demonstrated up to 94% accuracy in detecting various cancer types, including brain tumours[3].

The aim of this report is to demonstrate and evaluate the suitability of a deep-learning model in classifying medical images with a limited dataset. 223 images were given to build a model to differentiate brain scans based on the presence of tumours. Different architectures like VGG16 and ResNet were considered due to their proven success in image classification tasks. However, DenseNet121 was ultimately chosen for this task due to its powerful pretrained feature extraction capabilities, which are particularly beneficial for small datasets. Combined with fine-tuned, pre-trained weights by ImageNet, the model achieved 93.94% accuracy on unseen test images, suggesting the suitability and success of the chosen approach.

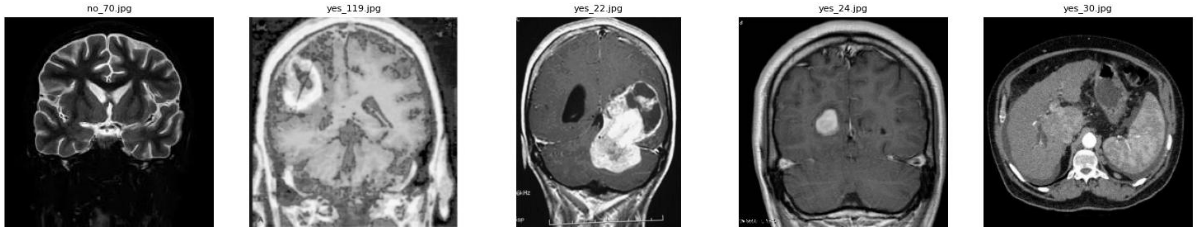
The rest of the report is organised as follows:

- **Proposed Method**: This section provides a description of the methodology, including the pre-processing phase.
- **Experimental Results**: A detailed account of the hyperparameter settings, evaluation process, and the results obtained will be presented.
- **Summary**: This section will highlight and discuss the main findings of the report.
- **References**

Proposed Method

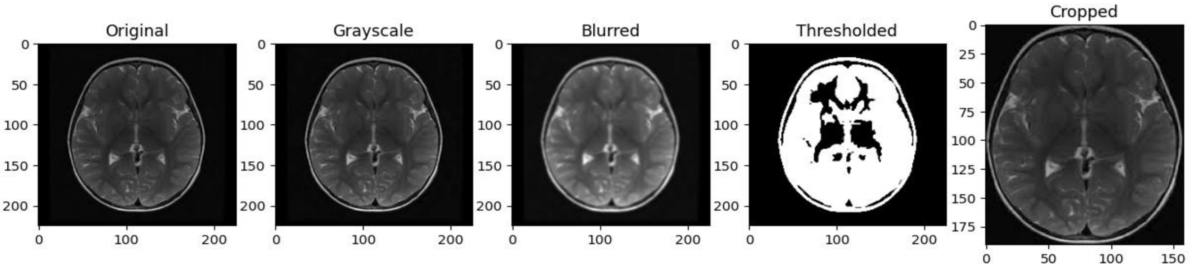
The proposed method started with organising all images by indexing their names and assigning the same image format of .jpg. The dataset, consisted of only 223 images, were manually combed through to remove scans too different from others. These 5 images included a scan of a human torso and 4 brain scans in the frontal plane instead of the horizontal plane.

❖ The outlier images:



The rest of the 218 images varied in size and colour properties, so they were standardised and pre-processed. This process consisted of cropping the brain region and resizing the image to a set resolution of 224x224. The cropping of the brain was done through several steps. First, the given image is converted to greyscale and Gaussian blue is applied to reduce noise. Second, a binary mask is created through a threshold, highlighting the zones of interest. Then the largest contour, the brain region, is identified and cropped accordingly.

❖ Stages of preprocessing:



The images were then randomly distributed to 3 separate sets: training, validation and testing. Training set consisted of 152 images (70%) and was used to train the model by adjusting the parameters to reduce the loss. Validation set was made up of 33 images (15%) and was used to tune the parameters and prevent overfitting. The final 33 images (15%) made up the testing set which was used to verify the model's ability to generalise. Such a split (70%-15%-15%) was chosen to have sufficiently enough images for accurate validation and testing scores, while the small training set was augmented to artificially create more dataset. This was done dynamically during the training of the model where images were randomly rescaled, rotated, dimmed or brightened, shifted, and flipped. It is important to note that the other 2 image sets were only rescaled during validation and testing to normalise pixels while preserving the features of original real-world brain scans.

The actual model is based on DenseNet121 architecture which was chosen for its ability to learn robust and compact feature representations during pre-training on ImageNet. These learned features provide a strong foundation that generalizes well to small datasets, helping to mitigate overfitting. Additionally, DenseNet121's efficient design leads to a lower parameter count and reduced computational overhead compared to other architectures like VGG16 or ResNet, making it well-suited for applications with limited data

Experimental Results

Model Hyperparameters

- **Feature Extraction Backbone:**

The model uses a pre-trained DenseNet121 as its base. By freezing these layers, the model leverages the robust feature representations learned from a large dataset (ImageNet). This approach is especially useful when the available dataset is limited because it prevents the model from overfitting by reducing the number of trainable parameters from 7,300,161 parameters to 262,657 parameters.

- **Custom Model Head:**

A custom classification head was added to the frozen base model. First, a Global Average Pooling (GAP) layer was implemented, reducing spatial dimensions while preserving key features. This is followed by a fully connected (dense) layer with 256 neurons and a ReLU activation, introducing non-linearity to capture complex patterns. This setup ensures efficient feature extraction and effective learning for the classification task.

- **Batch Size**

Batch size of 32 was chosen as a compromise between very small batch sizes that might introduce excessive noise in gradient estimates, and very large batch sizes that tend to not generalise well. This balance is supported by research, including Keskar et al. (2017)[4]. It is shown that such moderate batch sizes tend to converge to a flatter minima which is associated with better performance in generalisation.

- **Weight Adjustment**

The class weights were adjusted to compensate for the class imbalance in training data. The brain scans without tumour only made up 33% of the dataset. This led the model to have bias towards guessing the more abundant class to minimise loss. By giving more weight to smaller class, such a bias was slightly compensated. However, the adjustment of “no-tumour” class was limited to a weight of 1.2 compared to 1.0 of “yes-tumour" class. This was done to prevent frequent false negatives that are especially harmful in medical field.

- **Dropout Regularisation:**

Dropout was applied to help prevent overfitting. By randomly dropping a portion of the neurons during training, dropout forces the model to learn redundant representations. On one hand, if the dropout is too high, it might remove too much information which hinders learning. On the other hand, too little dropout may not provide enough regularization. Therefore, value of 0.4 was chosen to strike a balance between these two extremes.

- **Output Layer and Activation:**

The output layer is designed for binary classification, utilizing a sigmoid activation function. This choice ensures that the model's output can be interpreted as a probability, which is well-suited for binary decision tasks.

- **Loss Function and Optimizer:**

Binary cross-entropy is used as the loss function since it directly measures the performance of a binary classifier. The Adam optimizer was chosen for its adaptive learning rate properties, which help in efficiently navigating the loss landscape during training.

- **Learning Rate Scheduling:**

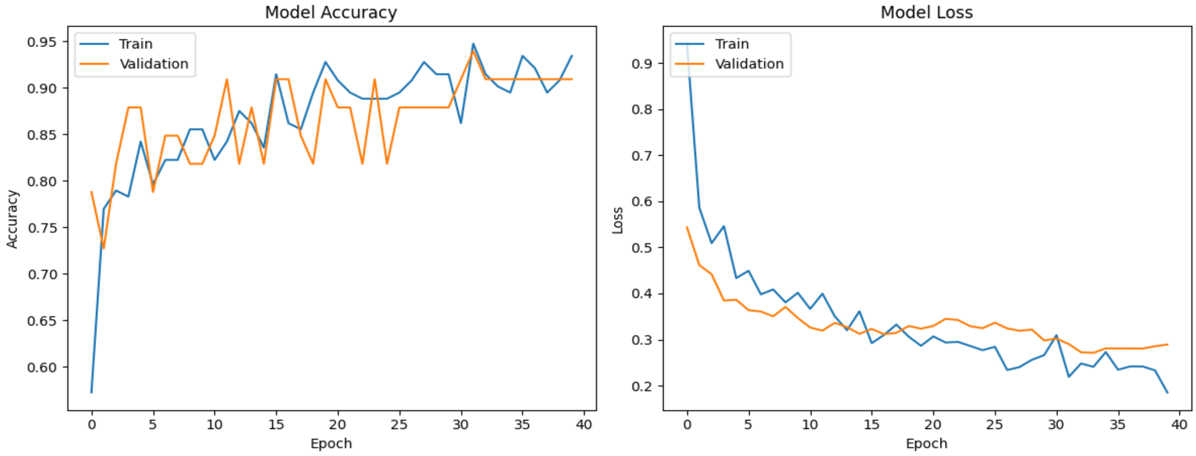
A learning rate scheduler monitors the validation loss and reduces the learning rate by 0.6 if progress stalls for 7 epochs. This strategy helps maintain training stability and avoids overshooting minima during optimisation. Lowering the learning rate in response to plateauing performance allows the model to fine-tune the weights more delicately in later stages of training. This technique essentially allows the model to start off bold and finish precise.

- **Epoch Count**

The model ran for 40 epochs. This number of epochs resulted in the best results as any fewer epochs would lead to failure to learn and underfitting while more epochs lead to memorisation and overfitting. Such a choice was guided by monitoring validation loss and accuracy trends, ensuring optimal performance.

Performance Metrics

The model trained on the training dataset for 40 epochs. The Adam optimizer was set to an initial learning rate of 0.001. The model's performance was evaluated using accuracy and loss.



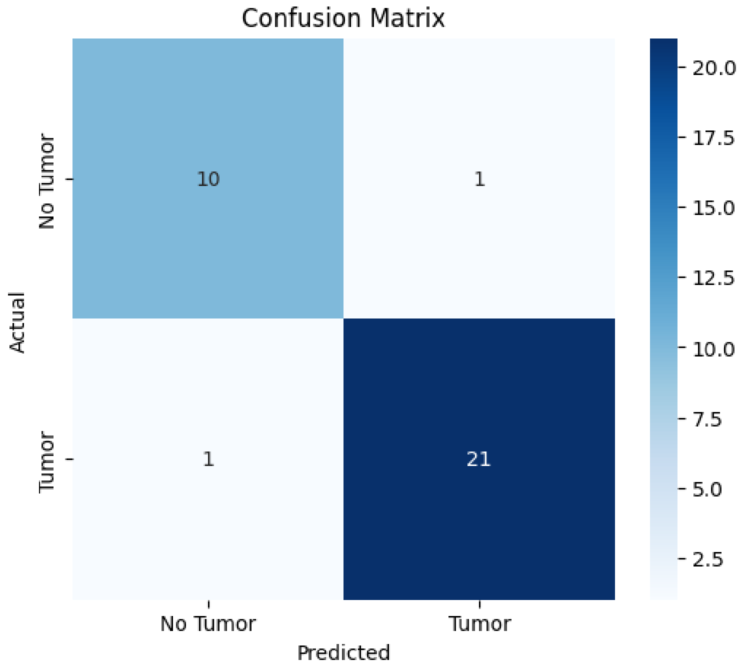
The accuracy plot shows that the training accuracy reached approximately 0.95, while the validation accuracy plateaued at around 0.90. The loss plot indicates that both training and validation loss steadily decreased, with the validation loss stabilizing around 0.30. The learning rate was reduced to 0.0006 on plateau at epoch 24, which helped stabilize the validation performance. The final training accuracy was 0.9274, and the final validation accuracy was 0.9091. The final training loss was 0.1972, and the final validation loss was 0.2891.

The model achieved good performance on the given task, with a validation accuracy of approximately 90%. There is some evidence of overfitting, but the gap of around 3% in accuracy is not substantial given a limited dataset.

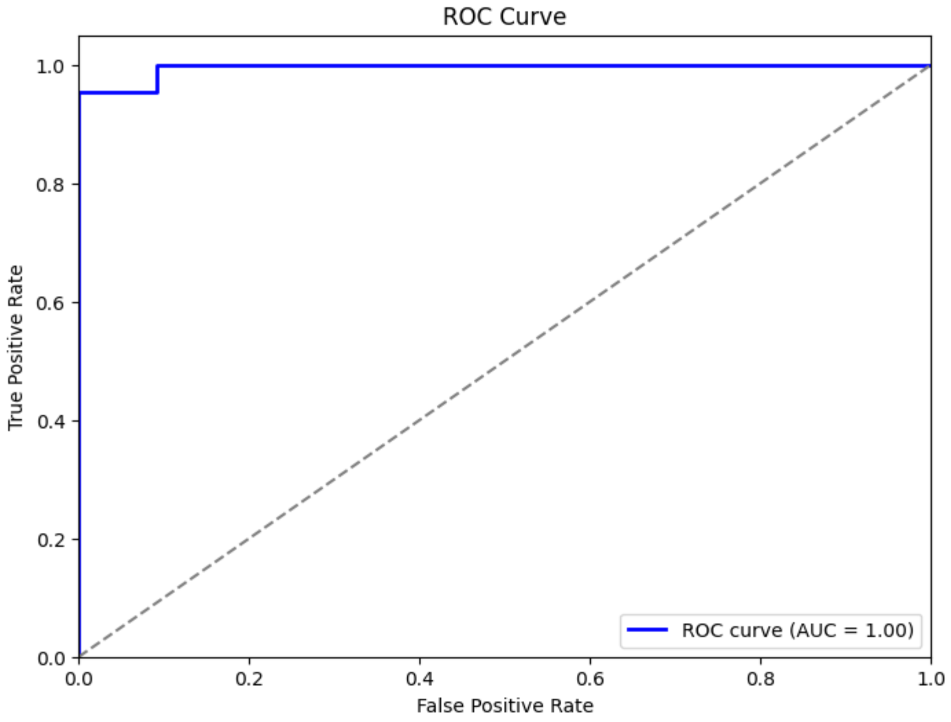
The confusion matrix shows the results of the model's predictions on the testing set of 33 previously unseen images:

- Accuracy: 93.94%
- Precision: 95.45%
- Recall: 95.45%
- F1-score: 95.45%
- Specificity: 90.91%

The model demonstrates high accuracy in and classifying tumours suggesting great ability in generalising. It exhibits excellent precision and recall for detecting tumour cases, indicating that it is effective in minimizing both false positives and false negatives. The high F1-score suggests that the model is reliable in identifying tumor cases, while the specificity of 90.91% indicates good performance in recognizing non-tumor cases.



The Receiver Operating Characteristic (ROC) curve for the model is shown below:



ROC curve is used to distinguish how well a binary classification model can distinguish between the two classes, presence or absence of tumours in this case. Tihe curve shoots up to a very high score early on indicating an exceptional performance. Additionally, the area under the curve (AUC) is nearly 1.00, meaning that the model is excellent at separating the two classes. However, such a score suggests that the model is overfitting. Despite the high accuracy score on test images, the near-perfect AUC score, combined with a gap in validation and training scores, indicates that the model may have learned patterns too specific to the training data, reducing its ability to generalise to unseen samples.

Points for Improvement

The problem of overfitting can be addressed in multiples ways. By implementing an early stopping, the model could be stopped at earlier epoch before it started to stagnate and overfit. A larger dataset could be obtained to train the model on a more diverse variety of brain scans and prevent rapid overfitting. Alternatively, a more aggressive augmentation of training set could also be implemented to artificially boost the diversity of images.

Summary

The report presents a deep learning model designed to identify brain tumours in medical scans. Using a limited dataset of 223 images (150 with tumours and 73 without), the method implemented a DenseNet121 architecture with pre-trained ImageNet weights. The model achieved excellent scores in accuracy, precision, recall, F1, and specificity. The model generalises well and classifies unseen brain scans with an accuracy score of 93.94%. Additionally, the ROC curve and the AUC showed exceptional separation. However, the experimental results also suggest that the model is overfitting. The model, therefore, can be improved by obtaining a larger dataset, implementing early stopping, and more aggressively augmenting the training data.

References:

1. Mousavi, S., Seyedmirzaei, H., Shahrokhi Nejad, S. *et al.* Epidemiology and socioeconomic correlates of brain and central nervous system cancers in Asia in 2020 and their projection to 2040. *Sci Rep* **14**, 21936 (2024). https://doi.org/10.1038/s41598-024-73277-z

2. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60-88.[1] https://doi.org/10.1016/j.media.2017.07.005

3. Harvard Medical School. (2025). *New Artificial Intelligence Tool for Cancer*. Retrieved from https://hms.harvard.edu/news/new-artificial-intelligence-tool-cancer.

4. Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2017). *On large-batch training for deep learning: Generalization gap and sharp minima.* arXiv preprint arXiv:1609.04836. Retrieved from https://arxiv.org/abs/1609.04836.

Declaration

AI-supported/AI-integrated use is permitted in this assessment. I acknowledge the following uses of GenAI tools in this assessment:

[ ] I have used GenAI tools for developing ideas.
[ ] I have used GenAI tools to assist with research or gathering information.
[ ] I have used GenAI tools to help me understand key theories and concepts.
[ ] I have used GenAI tools to identify trends and themes as part of my data analysis.
[ ] I have used GenAI tools to suggest a plan or structure for my assessment.
[✓] I have used GenAI tools to give me feedback on a draft.
[ ] I have used GenAI tool to generate images, figures or diagrams.
[✓] I have used GenAI tools to proofread and correct grammar or spelling errors.
[ ] I have used GenAI tools to generate citations or references.
[ ] Other [please specify].
[ ] I have not used any GenAI tools in preparing this assessment.

I declare that I have referenced use of GenAI outputs within my assessment in line with the University referencing guidelines.