Brain Tumour Detection using Deep Learning Techniques

Abstract

This investigation explores how deep learning algorithms may be used to classify MRI scan images as tumorous or non-tumorous. With advancements in CNN models such as ResNet, DenseNet and VGG16, the performance of several approaches were compared, giving a methodical comparison of each model. Major challenges, such as lack of variation in image rotation, size and an imbalance in labelled data, were mitigated with data augmentation to optimise model performance. Hyper-parameters were also tuned to identify the most appropriate setup to leverage each models' capabilities. Finally, metrics including accuracy, confusion matrices, precision and recall were used to evaluate each models' adaptability to new data, with the objective to identify the most appropriate model to diagnose brain tumours from MRI images.

Section 1: Introduction

Brain tumours are one of the most critical neurological disorders. With only 15% of malignant brain tumour patients in the UK surviving their tumours for more than 5 years[1], it is crucial to identify patients as early as possible to ensure it is treated effectively. Careful human diagnosis is a time-consuming task. Moreover, as erroneous identification may directly endanger lives, it is essential to ensure accurate diagnosis to minimise human error. As such, the study of automated brain tumour classification is a field of high interest, especially using Machine Learning, as it resembles a specialised pattern-recognition challenge that can be learned[2].

The aim of this report is to provide methods of automating tumour detection with deep learning techniques using a dataset of 150 positive and 73 negative cases; a total of 223 MRI images.

One of the major advantages of deep learning models is its modularity and variety. Existing models and architectures can easily be customised by adding one's own layers, configurable with parameters and activation functions through transfer learning[3]. They can also vastly vary in complexity. Examples such as ensemble learning techniques[4] used by Müller et al. incorporate multiple CNN architectures in parallel,

---

[1] (1) [cancerresearchuk](https://www.cancerresearchuk.org/about-cancer/brain-tumours/survival)
[2] (2) [State of the art survey on MRI brain tumor segmentation](https://www.sciencedirect.com/science/article/pii/S0730725X13001872?casa_token=FmryUDPkyH4AAAAA:U7siqCnfN3flyNdpbrV67UhC0xe7A2S7c--urP4gfrk7SB74KvvgWmeg7mL25yn167AY8IffPA)
[3] (10) [On Loss Functions for Deep Neural Networks in Classification](https://arxiv.org/pdf/1702.05659)
[4] (15) [An Analysis on Ensemble Learning optimized Medical Image Classification with Deep Convolutional Neural Networks](https://arxiv.org/abs/2201.11440)

aggregating its results to classify medical images, whilst more simpler approaches by Li et al. use Multi-Scale CNNs to identify small and large patterns within the same image, catering to limited datasets.[5]

These examples illustrate the versatility of CNN architectures and its rapid development in medical image classification.

This investigation will explore the suitability of three deep learning techniques, each selected based on a unique property to compare the feasibility of various deep learning techniques.

ResNet50 is a deep CNN with 50 layers. With the pretrained imagenet weights, this model will be used to evaluate the benefits of transfer-learning by extending existing complex models to a medical context.

Next, DenseNet was used. This was for comparison with ResNet50, identifying the differences between two pretrained models.

Finally, a small custom model with two convolution blocks and two fully connected layers will be created to demonstrate the performance of shallow, lightweight networks. Following a sequential stack of Conv and Pooling layers followed by dense layers, it resembles LeNet and VGG architectures in a much smaller form.

With these models in consideration, their accuracy and performance metrics will be compared.

This report will be structured with the following sections - Section 2 will explore the proposed methodology, including preprocessing techniques and model structures, Section 3 will describe the hyper-parameter settings, metrics for evaluation and the observed results. Finally, in Section 4, the findings from this investigation will be discussed, along with proposals for improvements in automation techniques in the field of medical image analysis.

Section 2: Proposed Method

To ensure consistency in the comparison process, all models will follow a standardised framework while allowing variation in architecture of the models themselves. The proposed methodology is comprised of the following steps:

Firstly, to enable evaluation on new unseen data, the holdout method will be used. 20% of the data will be reserved as the testing set, making sure they are never exposed to the models in training and validation.

The remaining 80% will be preprocessed, using a combination of random rotations, scaling and contrast. These augmentations prevent overfitting by ensuring data is sufficient for both classes, achieving the main purpose of reducing the disadvantages in the minority dataset. With three augmented copies per image,

[5] (14) [Multi-Instance Multi-Scale CNN for Medical Image Classification](https://arxiv.org/abs/1907.02413)

the tumour-negative class increases from 73 to 228 images. Pretrained models such as ResNet50 also apply preprocessing functions specific to the model as an additional step.

The preprocessed data is further subdivided into 80% for training and 20% for validation. During each epoch, the model will learn from the training set, while the validation set will be used to provide feedback to the model to adjust its weights before proceeding on to the next iteration.

Next, the model's hyperparameters will be selected for tuning. As grid search performs exhaustive combinations of parameter candidates, it is often computationally expensive as dimensions increase. On the other hand, it is effective when the input dataset is small, and hence, will be used in this investigation to systematically determine the best parameter values.

Using the most performant hyperparameters, the model will then be tested on the testing set, which the model has not seen until this stage. With a selection of metrics, the model's performance will be evaluated to determine its adaptability to new data.

Finally, these metrics will be compared across models to determine the most reliable architecture in classifying MRI images.

Details of the hyper-parameter pool, experimental results and the evaluation process will be discussed in the next section.

Section 3: Experimental Results

Section 3.1: Hyperparameters

Section 3.1.1: Grid Search Pool

To determine the most optimal hyper-parameter configurations, a grid search was conducted over the following pool:

- Optimiser: Adam, SGD, RMSprop
- Learning Rate: 0.0001, 0.001, 0.01, 0.1

Each combination was tested, and the best-performing settings were selected based on mean accuracy. Optimisers were chosen based on their unique properties:

- Adam adapts well to complex learning gradients with momentum, allowing it to overcome plateaus.[6]
- Stochastic Gradient Descent often generalises better with a well-tuned learning rate, however, requires careful tuning to perform effectively.[7]
- RMSprop has the potential to converge quickly, employing adaptive learning rates.

The chosen learning rates range from an order of magnitude higher and lower than Keras' default. Learning rates are an order of magnitude higher and lower than Keras' default, resulting in four values of equal spacing on the logarithmic scale, providing ample variation. Too high of a learning rate (0.1), led to unstable training, causing models to not converge. Thus, they tended to rank lower in grid search performance comparisons.

[6] (16) [Comparative Study of First Order Optimizers for Image Classification Using Convolutional Neural Networks on Histopathology Images](https://www.mdpi.com/2313-433X/6/9/92)
[7] (16) [Comparative Study of First Order Optimizers for Image Classification Using Convolutional Neural Networks on Histopathology Images](https://www.mdpi.com/2313-433X/6/9/92)

Here were the grid search results, with the 5 highest ranking parameters for each model:

ResNet50:

rank_test_score	param_model__optimizer	param_model__learning_rate	mean_test_score
1	Adam	0.0010	0.968736
2	Adam	0.0100	0.968728
3	Adam	0.0001	0.953909
4	RMSprop	0.0001	0.952251
5	SGD	0.1000	0.950641

DenseNet121:

rank_test_score	param_model__optimizer	param_model__learning_rate	mean_test_score
1	Adam	0.001	0.943860
2	Adam	0.01	0.931579
3	SGD	0.01	0.926316
4	RMSprop	0.001	0.910526
5	RMSprop	0.0001	0.901754

Custom LeNet:

rank_test_score	param_model__optimizer	param_model__learning_rate	mean_test_score
1	Adam	0.0001	0.931579
2	SGD	0.01	0.922807
3	SGD	0.001	0.915789
4	RMSprop	0.0001	0.912281
5	Adam	0.001	0.896491

Section 3.1.2: Fixed Parameters

The number of epochs was set to 10 after preliminary experiments indicated that loss reduction and accuracy plateaued after this point. As each succeding epoch had diminishing returns, 10 epochs were chosen to reduce the number of hyperparameter combinations to improve grid search runtime.

As the number of batches in preliminary experiments also did not yield significant differences to the model's performance, it was also omitted from the grid search pool discussed in Section 3.1.1.

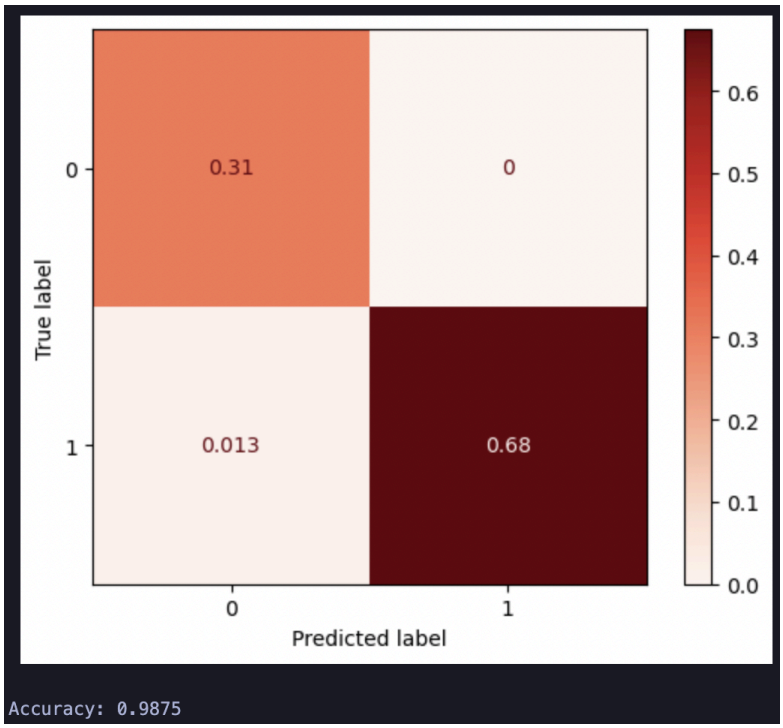
Section 3.2: Model Comparison

To evaluate model generalisation, the best-performing version of each trained model was tested on unseen MRI images. Performance was assessed using multiple metrics, including the following:

Section 3.2.1: Confusion Matrix

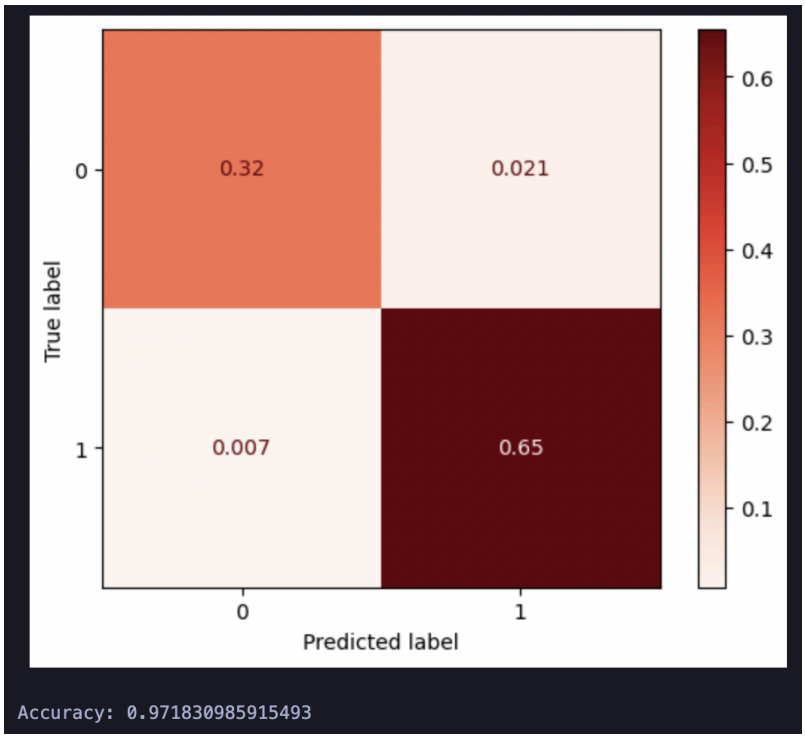
Accuracy only reflects the learning performance of the majority class, made even worse as the distribution skews more[8] [9]. Hence, confusion matrices were used to analyse per-class accuracy.

ResNet50:

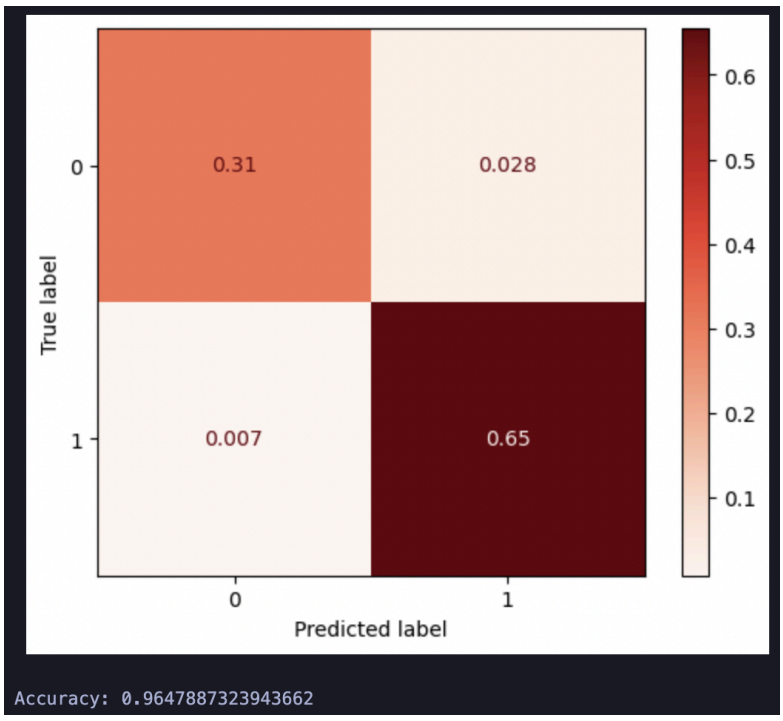


[8] (8) [Review of Classification Methods on Unbalanced Data Sets](https://ieeexplore.ieee.org/abstract/document/9408661)
[9] (17) [A New Evaluation Measure for Imbalanced Datasets](https://crpit.scem.westernsydney.edu.au/confpapers/CRPITV87Weng.pdf)

DenseNet121:



Custom LeNet:



By observing these results, it can be concluded that all models are able to classify new MRI images to an accuracy of above 95%. Looking at the top row of the confusion matrix, all models were able to identify the minority class correctly to a high extent, with high true negative and low false negative results. In particular, ResNet50 was able to score all of the negative cases correctly for new data.

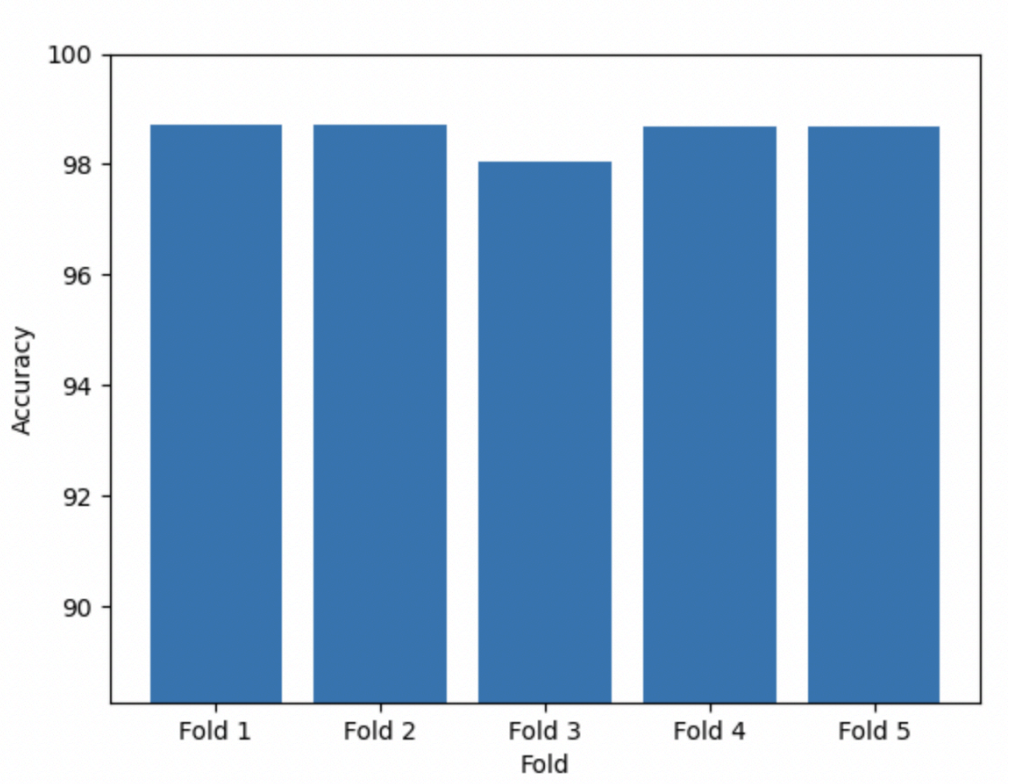
Since the tumour-positive class is the majority, the model may resort to always predicting positive to maintain a high score. A theoretical model that always answers tumour-positive would still achieve an accuracy of 150/223 = 67%. The confusion matrix here can be observed to determine whether this has happened. Looking at the false negatives, it has a significantly lower ratio compared with the true positives. Therefore, recall is very high. This indicates that the model correctly predicted true results when it was actually true, as opposed to guessing true every time.

Whilst compared to Dense and ResNet, which are deep models with over 50 layers, the Custom LeNet performed worse. However, considering it only has less than 10 layers, its difference in accuracy in comparison to these more complex models is less than expected.

Section 3.2.2: K-Fold Cross Validation

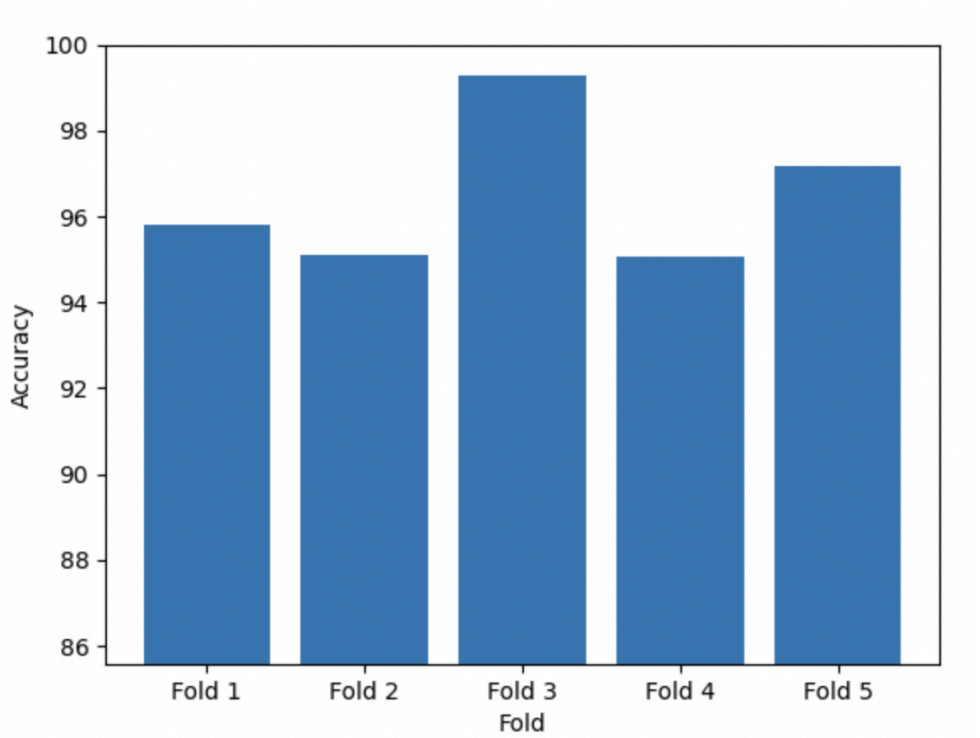
In small datasets, each image has an increased contribution to the model's predictions. Hence, it is important to determine if the model overcompensates for a specific subset of images using K-Fold Cross Validation[10].

ResNet50:

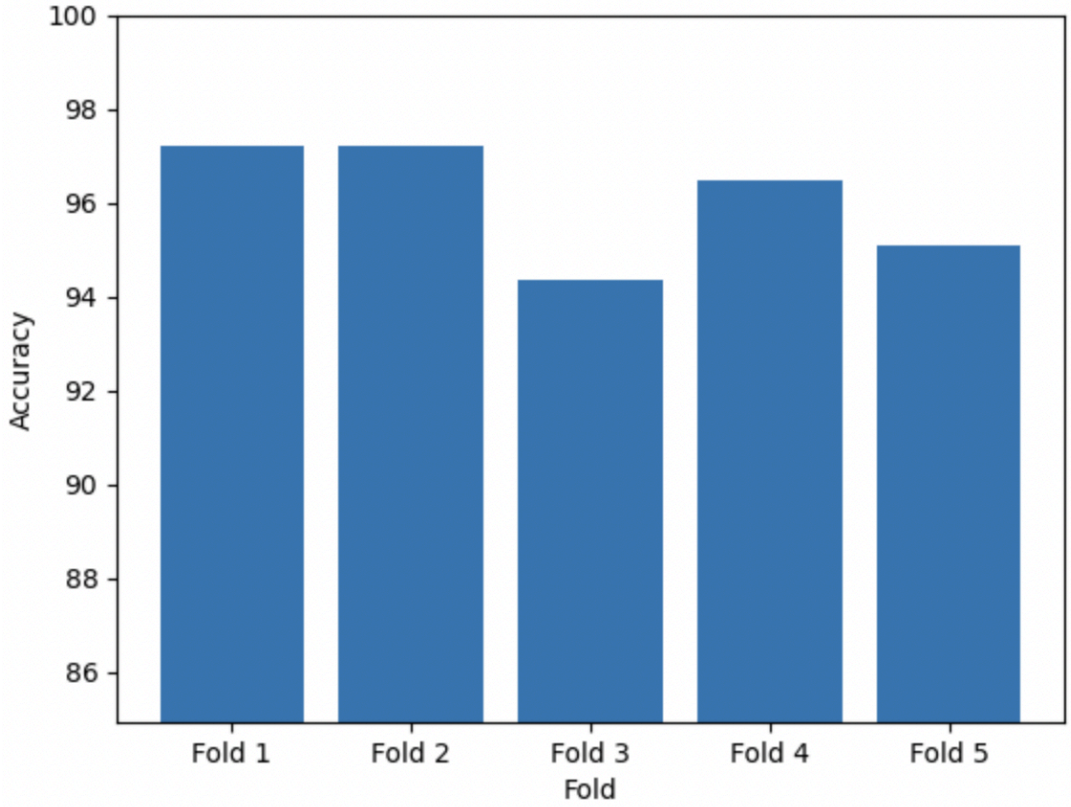


[10] (5) [Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification](https://ieeexplore.ieee.org/document/7544814)

DenseNet:



Custom LeNet:

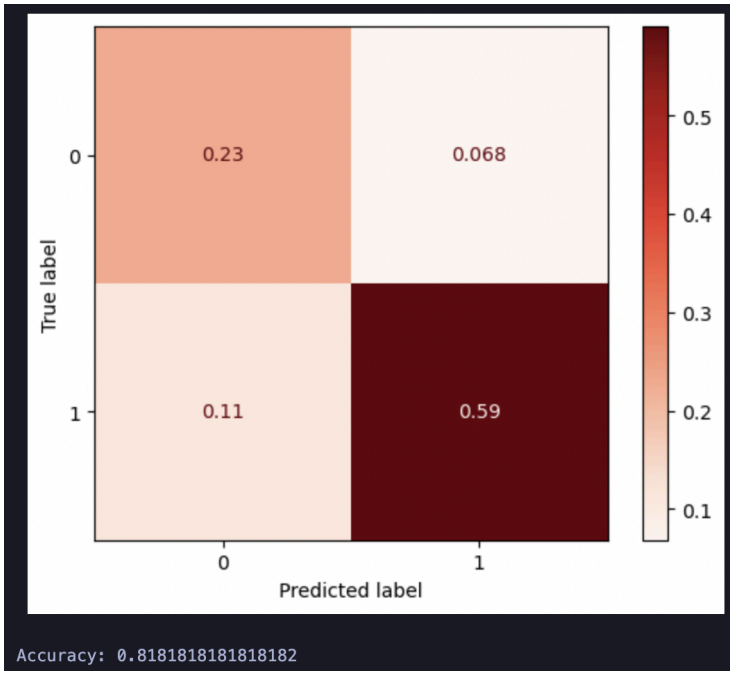


Observing these accuracies-per-fold, we can see if the model has inconsistent performance across folds to identify biases toward a subset of data. Whereas in ResNet, with a near-even accuracy across all folds, in Densenet, fold 3 has a significantly higher accuracy than the rest, suggesting that it has overfitted into a subset of data.

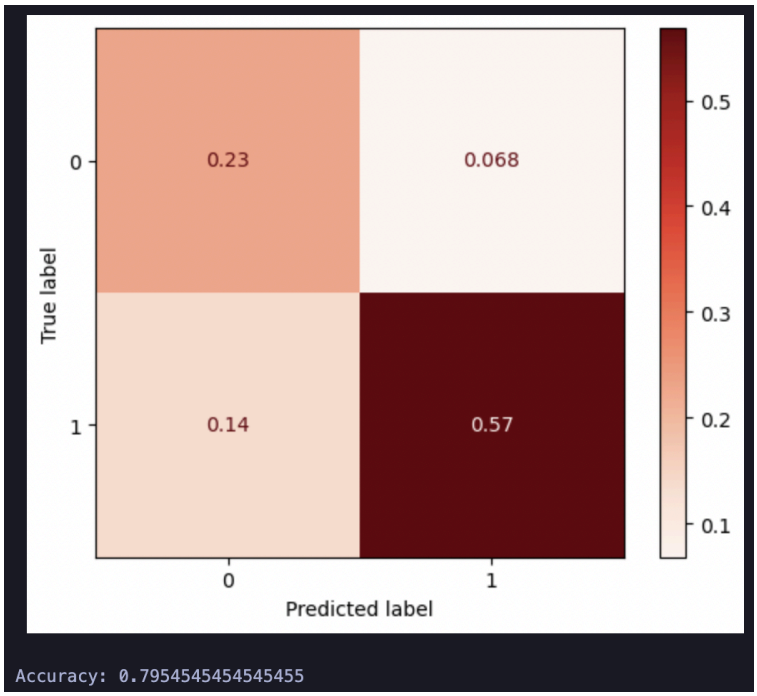
Section 3.3: Augmented vs Non-Augmented

Models were also tested for performance comparisons when using augmented and non-augmented counterparts to evaluate whether there was any significant increases in accuracy performance as a result of data augmentation.

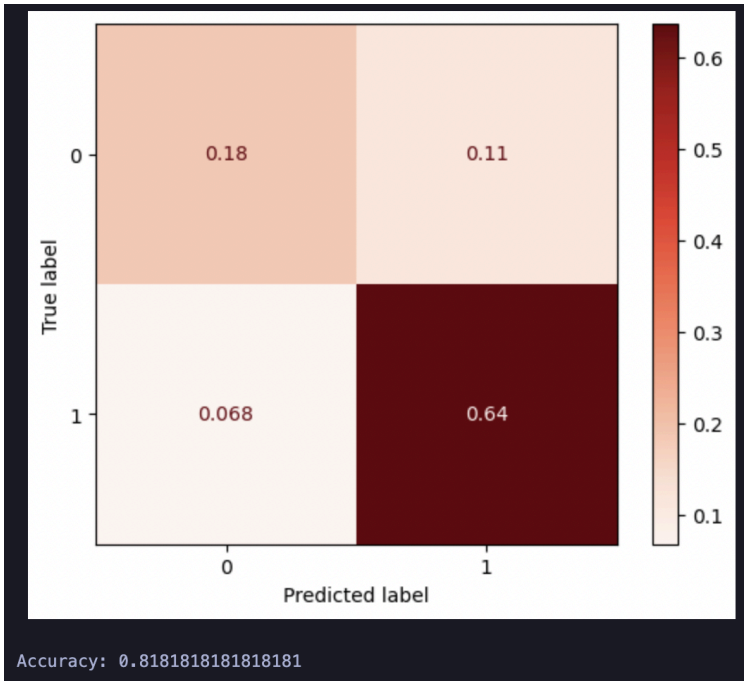
ResNet (without augmentation):



DenseNet (without augmentation):



Custom LeNet (No Augmentation)



Overall, it appears that without augmentation, the models perform significantly worse, indicating that augmentation has successfully improved these models' performance in generalising predictions.

Section 4: Summary

An overarching issue present during this investigation was the imbalance of labelled data across classes in the given dataset. This is reflective of how large amounts of detailed and well-annotated data can often be scarce in the medical field due to cost, privacy and ethical concerns.[11] With deep CNNs requiring large datasets to generalise well, shallow networks have been preferred in the analysis of medical images[12].

This experiment demonstrates that shallow datasets do not necessarily hinder a model's ability to classify images accurately. In fact, observations in this investigation suggest that given appropriate hyperparameters, they can perform just as well as deep networks.

On the other hand, medical image analysis, especially involving decisions of whether a patient needs to be treated urgently, requires higher-level analysis based on domain-specific expertise. Considering these factors, it is very difficult to achieve this using automation. In reality, the accuracies approaching 99%, achieved by the ResNet50 model during this investigation, still does not guarantee assurance that patients will be identified with absolute certainty. Therefore, through this investigation, it can be concluded that image-classification is feasible to a highly accurate degree as a tool to speed up patient diagnosis. However, there may still be challenges to implement it practically as a fully automated way to diagnose real patients with tumours.

---

[11] (4) [Overcome medical image data scarcity by data augmentation techniques: A review](https://ieeexplore.ieee.org/abstract/document/10005544?casa_token=h-qaYs71mTEAAAAA:DyFISUxcHoZP52SyTU9twXsl-6dHva5mt4mayvuzPkVh_My4WEDfZdO1AyvbYWC9PWxALKewVw)

[12] (3) [Medical Image Analysis using Convolutional Neural Networks: A Review](https://link.springer.com/article/10.1007/s10916-018-1088-1)

References

(1) [cancerresearchuk](https://www.cancerresearchuk.org/about-cancer/brain-tumours/survival)
(2) [State of the art survey on MRI brain tumor segmentation](https://www.sciencedirect.com/science/article/pii/S0730725X13001872?casa_token=FmryUDPkyH4AAAAA:U7siqCnfN3flyNdpbrV67UhC0xe7A2S7c--urP4gfrk7SB74KvvgWmeg7mL25yn167AY8IffPA)
(3) [Medical Image Analysis using Convolutional Neural Networks: A Review](https://link.springer.com/article/10.1007/s10916-018-1088-1)
(4) [Overcome medical image data scarcity by data augmentation techniques: A review](https://ieeexplore.ieee.org/abstract/document/10005544?casa_token=h-qaYs71mTEAAAAA:DyFISUxcHoZP52SyTU9twXsl-6dHva5mt4mayvuzPkVh_My4WEDfZdO1AyvbYWC9PWxALKewVw)
(5) [Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification](https://ieeexplore.ieee.org/document/7544814)
(6) [Improving classification accuracy using data augmentation on small data sets](https://www.sciencedirect.com/science/article/pii/S0957417420305200?casa_token=FJqo_oNv4XQAAAAA:xt4Cr8IeJ9v7GZirLN7Hr-b7ZQiwpYmR6FvzJPS-5IXn929lLkFAQ7OpdqVfCCGeQjHspngQNQ#b0125)
(7) [Deep learning](https://www.nature.com/articles/nature14539)
(8) [Review of Classification Methods on Unbalanced Data Sets](https://ieeexplore.ieee.org/abstract/document/9408661)
(10) [On Loss Functions for Deep Neural Networks in Classification](https://arxiv.org/pdf/1702.05659)
(14) [Multi-Instance Multi-Scale CNN for Medical Image Classification](https://arxiv.org/abs/1907.02413)
(15) [An Analysis on Ensemble Learning optimized Medical Image Classification with Deep Convolutional Neural Networks](https://arxiv.org/abs/2201.11440)
(16) [Comparative Study of First Order Optimizers for Image Classification Using Convolutional Neural Networks on Histopathology Images](https://www.mdpi.com/2313-433X/6/9/92)
(17) [A New Evaluation Measure for Imbalanced Datasets](https://crpit.scem.westernsydney.edu.au/confpapers/CRPITV87Weng.pdf)