Exploring and Evaluating Modern Techniques in Brain Tumour Classification: A Collaborative Research Study

Ella Vithanage
Artificial Intelligence MSc
Applied AI
De Montfort University, Leicester, UK
p2537413@my365.dmu.ac.uk

Victoria Hektor
Artificial Intelligence MSc
Applied AI
De Montfort University, Leicester, UK
P2629898@my365.dmu.ac.uk

Abdulaziz Alqulayti
Artificial Intelligence MSc
Applied AI
De Montfort University, Leicester, UK
p2811611@my365.dmu.ac.uk

***Abstract*—Brain tumour classification remains a critical challenge in medical imaging, necessitating precise and reliable methods for accurate diagnosis and treatment planning. This paper presents a collaborative study by three authors, each contributing a unique methodology to advance the state-of-the-art in brain tumour classification, aptly name EVA; using our initials for the system. The first approach leverages Convolutional Neural Networks (CNNs) for robust feature extraction and classification, demonstrating significant improvements in accuracy and efficiency, [28]. The second methodology explores Semi-Supervised Transfer Learning and Supervised Reinforcement Learning, harnessing the power of pre-trained models and fine-tuning them with limited labelled data to enhance performance and generalisation, [29]. The third approach utilises Random Forests, a powerful ensemble learning technique, to provide an alternative perspective on tumour classification with an emphasis on interpretability and robustness, [30]. Through a comprehensive literature review and empirical evaluation, this paper not only compares these diverse methodologies but also synthesises current research trends in brain tumour classification. The findings underscore the potential of combining these advanced techniques to achieve superior diagnostic accuracy and reliability, paving the way for future innovations in medical imaging. This article should be read preemptively in conjunction with [28], [29], and [30].**

I.

***Index Terms*—Brain Tumour Classification, Convolutional Neural Networks (CNNs), Semi-Supervised Learning, Transfer Learning, Random Forests, Medical Imaging, MRI, Deep Learning, Ensemble Learning, Diagnostic Accuracy, Feature Extraction, Model Generalisation, Interpretability, Hybrid Models, Patient Outcomes**

II. Introduction

Brain tumours represent a significant and complex challenge in modern medicine due to their aggressive nature and the critical functions of the brain regions they affect. The incidence of brain tumours varies globally, with significant implications for healthcare systems worldwide. According to recent data, brain tumours account for about 3.4% of all cancers, but they have a disproportionate impact due to their high mortality rates and the severe neurological deficits they can cause, [16]. Recent advances in AI and ML have revolutionised the diagnosis and classification of brain tumours. These technologies have significantly improved the accuracy and speed of diagnosis, allowing for better patient outcomes. [16]. The global burden of brain cancer remains substantial, with varying incidence and mortality rates across different regions. A comprehensive study has highlighted that brain cancer incidence is highest in high-income countries, reflecting better diagnostic capabilities, while mortality rates remain a critical concern globally due to the aggressive nature of these tumours and the limited effectiveness of current treatment options, [17].

Statistical data from The Brain Tumour Charity underscores the urgent need for improved diagnostic and treatment strategies. In the UK alone, approximately 11,700 people are diagnosed with a primary brain tumour each year, and brain tumours are the leading cause of cancer-related deaths in children and adults under 40, [18]. The impact on patients and their families is profound, not only due to the physical and cognitive impairments caused by the tumours but also due to the emotional and financial burdens of long-term care. Machine and deep learning approaches have shown promise in enhancing the detection and classification of brain tumours. A systematic review of these technologies indicates that they can effectively distinguish between different types of brain tumours, predict tumour progression, and assist in treatment planning by identifying tumour boundaries with high precision, [18].

A. Abbreviations and Acronyms

CNN: Convolutional Neural Network; MRI: Magnetic Resonance Imaging RL: Reinforcement Learning; AI: Artificial Intelligence; DL: Deep Learning; ML: Machine Learning; DNN: Deep Neural Network; GWO: Grey Wolf Optimization; GSI: Generalised Structural Information; NLP: Natural Language Processing; SVM: Support Vector Machine; RF: Random Forest; ROC: Receiver Operating Characteristic; AUC: Area Under the Curve; TL: Transfer Learning; ReLU: Rectified Linear Unit; RMSProp: Root Mean Square Propagation; SGD: Stochastic Gradient Descent; Adam: Adaptive Moment Estimation; XAI: Explainable Artificial Intelligence; CV:

Cross-Validation; k-NN: k-Nearest Neighbours; PCA: Principal Component Analysis; ICA: Independent Component Analysis; LDA: Linear Discriminant Analysis; QDA: Quadratic Discriminant Analysis; LSTM: Long Short-Term Memory; GRU: Gated Recurrent Unit; GAN: Generative Adversarial Network; AE: Autoencoder; VAE: Variational Autoencoder.

III. Data

The dataset used for this study is the "Brain Tumor MRI Dataset" curated by M. Nickparvar and available on Kaggle. This dataset comprises MRI images of brain tumours, segmented into three categories: glioma, meningioma, and pituitary tumour. Each category contains a substantial number of images, ensuring a diverse and representative sample for training and evaluating classification models. The images are stored in JPEG format and are labelled accordingly, providing a solid foundation for supervised learning tasks. This dataset is particularly valuable for developing and benchmarking brain tumour classification algorithms due to its comprehensive coverage and quality. [12].

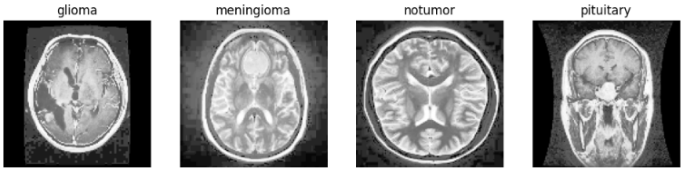

Fig. 1. Depiction of Each Image Category.

IV. Methodologies

A. Convolutional Neural Networks (CNNs):

One author focuses on using CNNs for brain tumour classification, leveraging their capability to automatically extract relevant features from MRI images. This method has shown promise in achieving high accuracy and efficiency in classifying different types of brain tumours. [].

B. Random Forests:

The third methodology employs Random Forests, an ensemble learning technique that combines multiple decision trees to improve classification performance. This method emphasises interpretability and robustness, offering a complementary perspective to the deep learning-based approaches. [].

C. Semi-Supervised Transfer Learning & Supervised Reinforcement Learning:

Another author explores semi-supervised transfer learning techniques, which involve using pre-trained models on large datasets and fine-tuning them with a smaller set of labelled data. This approach aims to maximise the performance and generalisation of the classifier, especially in scenarios with limited annotated data. [].

V. Literature Review

Before moving on to the technical experiment analysis, a thorough literature review was completed, a general overview as well as a more in depth review for each of the employed methodologies is provided in [], [], and []. Brain tumour classification has seen significant advancements in recent years, driven by the development of deep learning and other sophisticated machine learning techniques. This review summarises key research papers in the field, highlighting the various approaches and their contributions.

A. Literature Review on Brain Tumour Classification Techniques

Srinivasan et al. (2024) introduced a hybrid deep CNN model for multi-classification of brain tumours. The model combines different CNN architectures to leverage their strengths in feature extraction and classification, resulting in improved accuracy and robustness in diagnosing brain tumours from MRI images. In another study, Srinivasan et al. (2023) utilised a deep learning technique specifically for grade classification of tumours from brain MRI images. This approach focuses on distinguishing between different grades of tumours, which is crucial for determining the appropriate treatment strategies. [1], [2]. Chatterjee et al. (2022) developed deep spatiospatial models for brain tumour classification. These models consider both spatial and spatio-temporal features within MRI images, enhancing the ability to accurately classify different types of brain tumours. [3].

Rasool et al. (2022) presented a hybrid deep learning model for brain tumour classification. This model integrates traditional machine learning methods with deep learning techniques to improve classification performance, particularly in cases where data is limited or highly imbalanced. [4]. Jain et al. (2024) compared various transfer learning techniques for brain tumour classification using MRI images. Transfer learning involves using pre-trained models on large datasets and fine-tuning them on smaller, task-specific datasets. This approach has been shown to significantly enhance the performance of brain tumour classifiers. [5].

Agarwal et al. (2024) explored deep learning techniques for enhanced brain tumour detection and classification. Their study highlights the effectiveness of deep learning models in improving the sensitivity and specificity of tumour detection, which is critical for early diagnosis and treatment planning. [6]. Zeineldin et al. (2022) investigated the explainability of deep neural networks for MRI analysis of brain tumours. Explainability is essential for gaining clinical trust and ensuring that the models' decisions can be understood and verified by medical professionals. [7]. Haque et al. (2024) introduced NeuroNet19, an explainable deep neural network model for brain tumour classification using MRI data. This model not only achieves high accuracy but also provides interpretable results, making it suitable for clinical applications. [8].

Tandel et al. (2019) provided a comprehensive review of deep learning perspectives in brain cancer classification. This paper summarises various deep learning techniques and their

applications in brain tumour classification, offering insights into current trends and future directions. [9]. Badža and Barjaktarović (2020) utilised convolutional neural networks (CNNs) for the classification of brain tumours from MRI images. Their study demonstrates the efficacy of CNNs in extracting relevant features from MRI data and accurately classifying different tumour types. [10]. ZainEldin et al. (2022) proposed a deep learning approach combined with sine-cosine fitness grey wolf optimisation for brain tumour detection and classification. This hybrid method enhances the optimization process, leading to improved classification accuracy and robustness. [11].

B. Method 1 Literature Review

Convolutional Neural Networks have been developed in recent years as a branch of Deep Learning within Artificial Intelligence. Bringing various applications to several domains such as within radiology and medical fields, CNNs are versatile by using convolution, pooling and fully connected layers [25]. They work by using a mathematical design whereby grid patterns which exist in images can be processed. They have been credited in their use for image processing and in an article titled 'Image processing for medical diagnosis using CNN' [25], it is explained that an advantage of using CNNs is the ability to be able to identify and extract features, however they can be computationally complex resulting in real-time processing being a substantial problem for outputting results. In another article [26], brain tumour detection is attempted by using a CNN by using a binary and multiclass classification system, achieving results of 97.8% and 100% classification accuracy. This shows that CNNs are a great resource for precision in classifying brain tumours, being the main rationale for using this method. It's suggested that using GPU is optimal for the execution time when using Convolutional Neural Networks [3], this makes use of high performance computation which is necessary for the execution and deployment of CNNs especially in real-time image processing. [24], [27].

C. Method 2 Literature Review

In developing the semi-supervised reinforcement learning algorithm, drawing on the extensive survey by Cheplygina et al. (2018) which highlights the utility of semi-supervised learning (SSL) in medical imaging. This approach leverages both labelled and unlabelled data to enhance classifier performance, a strategy we adapt to improve the efficiency and accuracy of our reinforcement learning model. [13]. Additionally, the insights provided by Inés et al. (2021), which demonstrate the effectiveness of combining transfer learning with semi-supervised learning techniques for biomedical image classification. Their AutoML method, implemented in the ATLASS tool, enhances model performance by leveraging both labelled and unlabelled data, which is a strategy we adapt to improve the efficiency and accuracy of our reinforcement learning model. [14]. Drawing on the innovative approach by Ge et al. (2020), which successfully employs deep semi-supervised learning for brain tumour classification, helps to inform this method. Their method effectively utilises both labelled and unlabelled data through a novel graph-based framework, augmented by GAN-generated synthetic MRIs to improve classification accuracy. [15].

D. Method 3 Literature Review

Random Forest method is a commonly employed ensemble learning technique. It is remarkably adequate For medical image examination. Random Forest Presented by Breiman (2001), [20]. It creates numerous decision trees with the utilisation of various parts from the feature subsets and dataset leading to improving the precision and also helping to minimise the overfitting in comparison with single decision trees. RF in medical imaging has shown substantial promise. Ghongade and Wakde (2017), [21], displayed its efficacy in interpreting and classifying breast cancer using a computer-aided diagnosis solution, underlining its capability to manage overfitting well and high-dimensional data. Also, Menze et al. (2015), [22]. used RF to classify brain tumour varieties using MRI images, which showed trustworthiness and increased precision. All of These papers highlight the RF's ability to handle complicated patterns found in medical imaging datasets, which drives us to the conclusion that it's a robust and trustworthy instrument for medical diagnostics tasks.

VI. CNN Implementation, Results & Analysis

A. Data Pre-Processing

To process the image data to be able to use CNNs, the images must be accessible and resized for the use in Google Colab. The data was accessed via mounting the Google drive, meaning the images wouldn't have to be uploaded each time. The data was then split into three sets: training, testing and validation. This enables for the Convolutional Neural Network to be trained using the training set where the model can learn patterns such as the features that each individual tumour presents, validation is apparent when parameters are changed through trial and error and finally the test set enables generalisability to unseen data.

B. Model Architecture

The architecture design that is employed for the CNN method uses:

- Input layer: To receive the images as inputs and be able to use them within the CNN.
- Convolutional layers: Three convolutional layers are utilised so it fits within the area of deep learning, they detect patterns and features in the images.
- Pooling layers: Three layers are deployed in this system's architecture, these layers aim to reduce the computational complexity and prevent overfitting.
- Dropout layer: To further prevent overfitting, a dropout layer is used, it randomly selects half of the input units to be set to 0.
- Flatten layer: This aids the model to be able to use dense, connected layers for classification.

- Fully Connected Layers: These connect every neuron in one layer to the next using the weighted sum and bias. By using these layers, classification is achievable within image processing for this task.
- Output Layer: This provides the final output for the model which is a classification accuracy based on which items are correctly classified.

C. Training and Optimisation

To train the CNN, the training set is utilised, this is where the system is able to recognise patterns and learn recursively through iterations (epochs). To optimise this procedure, GPU is used so the processing time is reduced, however this caused problems on Google Colab which has limited GPU usage. Intuitively, the more epochs that are ran, the greater performance this will have on results. As the system is computationally complex, only a finite number of instances are supplied, giving the basis of a model that could be developed with more time to execute and test.

- Benchmark 1 - Using one epoch to test the system seemed the most optimal way to exhibit results under time constraints: The training classification accuracy results in the loss being 1.0550 and accuracy being 50% (2sf) with the validation set, this results in the loss being much higher, at 1.3812 and accuracy being 45% (2sf).
- Benchmark 2 - Using two epochs: The best results of the second run through the system situated in the training loss being 0.7970 and the accuracy being 67% (2sf). The validation set results were the loss being 0.9035 and accuracy being 66% (2sf).
- Benchmark 3 - Using three epochs: The best results with the third epoch results in the training loss being 0.6325 and accuracy being 75% (2sf) and validation loss being 0.9053 and accuracy being 71% (2sf).

D. Visualisation and Results

In the analysis of benchmark results for tumour categorisation using supervised learning on labelled data, three benchmarks were evaluated based on different epoch counts. Benchmark 1, with one epoch, achieved a generalisability of 62% (to 2 significant figures) and a loss of 0.9982, which is lower than the test or validity set results. Benchmark 2, with two epochs, showed an overall accuracy of 49% (2sf) and a loss of 1.6382. Benchmark 3, with three epochs, resulted in an overall accuracy of 63% (2sf) and a loss of 1.1953. The analysis indicates that the highest training and validation accuracy of 75% was observed in the third benchmark, while the best test accuracy of 63% was also achieved in the third benchmark. This suggests that increasing the number of epochs is likely to improve performance accuracy in categorising tumours. Therefore, it is recommended to conduct further testing with more epochs, under less rigid time constraints, and to explore different layers and compiling features to enhance the system's performance.

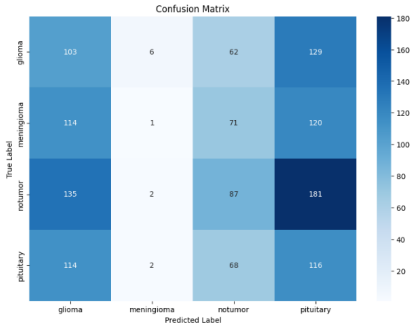

Fig. 2. Method One Best Benchmark confusion matrix.

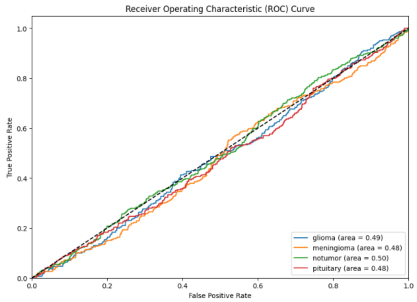

Fig. 3. Method One Best Benchmark Curve Visualisation.

VII. Semi-Supervised Transfer Learning & Supervised Reinforcement Learning Implementation, Results & Analysis

A. Data Pre-Processing

The dataset employed, sourced from Kaggle's Brain Tumour MRI Dataset, included MRI scans categorised into glioma, meningioma, no tumour, and pituitary tumour. Pre-processing involved resizing images to 128x128 pixels, normalisation, and applying data augmentation techniques such as rotation, zoom, and horizontal flipping to enhance model robustness and prevent overfitting.

B. Model Architecture

The first model leveraged the VGG16 convolutional neural network for feature extraction, augmented with custom fully connected layers tailored for the classification task. This transfer learning approach utilised the pre-trained weights of VGG16, reducing the training data requirements. Additionally, an EfficientNetB0 architecture was explored, fine-tuned to improve classification accuracy.

C. Training and Optimisation

Training was conducted on Google Colab, utilising its GPU capabilities. The Adam optimiser and categorical cross-entropy loss function were used for multi-class classification. To enhance the model's performance, hyperparameter tuning was performed using Keras Tuner, adjusting parameters such as learning rate and dropout rate. The training process incorporated callbacks like ReduceLROnPlateau and EarlyStopping to prevent overfitting and ensure optimal learning.

D. Visualisation and Results

Visualisation tools such as confusion matrices and learning curves were employed to monitor model performance. The best results achieved were as follows; Benchmark One training accuracy reaching around 82% and validation accuracy around 83% accuracy achieved was 31%, with the confusion matrix highlighting significant misclassifications, particularly among tumour classes. Classification reports provided detailed metrics, including precision, recall, and F1-score, revealing areas for improvement. The results underscore the complexity of brain tumour classification and the potential benefits of refined model architectures and advanced training techniques.

Fig. 4. Method Two Best Benchmark.

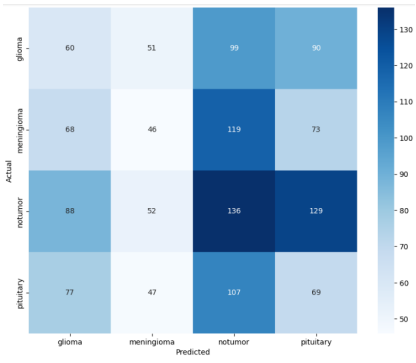

Fig. 5. Method Two Best Benchmark Confusion Matrix.

VIII. Random Forests Implementation, Results & Analysis

A. Data Pre-Processing

The images from the dataset had different sizes. Because of the model requirement of united size, images were resized to 128x128 pixels. Then, features were extracted from the images using ORB, SIFT, and HOG. After that, images were transformed to grayscale, normalised, and flattened. Finally, each feature was integrated with the flattened images, and the labels encoded for the training.

B. Model Architecture

The model uses RandomForestClassifier as the primary model with other Classifiers. It is a method that creates numerous decision trees. By integrating the predictions from these trees, this model improves its accuracy. RandomForest is considered adequate for complicated tasks such as brain tumour classification, which makes it a sensible option.

C. Training and Optimisation

Training the model is accomplished by splitting the dataset into training and validation sets. The model was fine-tuned using hyperparameters such as tree depth, minimum samples per split, and the number of trees. Validation accuracy is employed to guarantee the most promising model accuracy.

D. Visualisation and Results

The best model, which is Random Forest, was assessed on the test data and has reached high accuracy. The classification report and confusion matrix delivered precise performance metrics. For visualisation, learning curves and a confusion matrix heatmap were used to present the model's efficacy in classifying brain tumours. The outcomes underlined that the RandomForest model is conceivable for clinical application with supporting proof of its classification accuracy among different tumour types. It achieved a test accuracy of 0.92

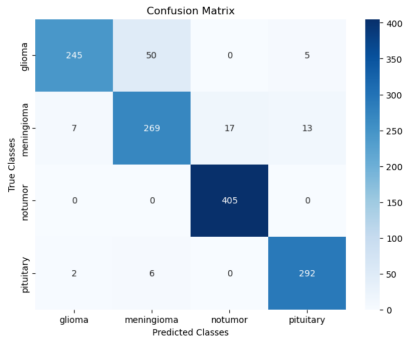

Fig. 6. Random Forest Confusion Matrix

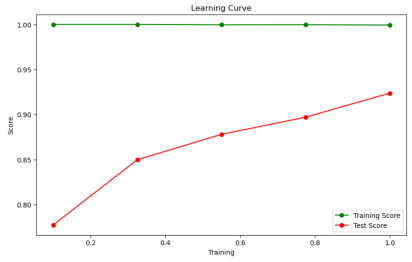

Fig. 7. Random Forest Learning Curve

IX. Comparison, Analysis & Conclusions

A. Methodology Comparison

B. Analysis

C. Conclusion

Acknowledgements

We wish to thank Nathanael Baisa for his support, patience, teaching, and guidance throughout the Applied AI unit. Without his help, we surely would not have produced such a strong piece of research. We also wish to thank all the teaching staff within the Artificial Intelligence MSc programme, who have taught us in ways we could understand, providing us with a solid foundation to succeed in this and all other modules.

References

[1] Saravanan Srinivasan, Divya Francis, Sandeep Kumar Mathivanan, Hariharan Rajadurai, Basu Dev Shivahare, and Mohd Asif Shah. A hybrid deep cnn model for brain tumor image multi-classification. *BMC medical imaging (Online)*, 24, 01 2024.

[2] Saravanan Srinivasan, Mercy Bai, Sandeep Kumar Mathivanan, V. Muthukumaran, J. Chinna Babu, and Lucia Vilcekova. Grade classification of tumors from brain magnetic resonance images using a deep learning technique. *National Library of Medicine - Diagnostics (Basel)*, 13:1153–1153, 03 2023.

[3] Soumick Chatterjee, Faraz Ahmed Nizamani, Andreas Nürnberger, and Oliver Speck. Classification of brain tumours in mr images using deep spatiospatial models. *Scientific Reports*, 12:1505, 01 2022.

[4] Mohammed Rasool, Nor Azman Ismail, Wadii Boulila, Adel Ammar, Hussein Samma, Wael M. S. Yafooz, and Abdel-Hamid M. Emara. A hybrid deep learning model for brain tumour classification. *Entropy*, 24:799, 06 2022.

[5] Jawahar Jain, Mihika Kubadia, Monika Mangla, and Prachi Tawde. Comparison of transfer learning techniques to classify brain tumours using mri images. *Engineering Proceedings*, 59, 01 2024.

[6] Monika Agarwal, Geeta Rani, Ambeshwar Kumar, Pradeep Kumar K, R. Manikandan, and Amir H Gandomi. Deep learning for enhanced brain tumor detection and classification. *Results in Engineering*, 22:102117–102117, 04 2024.

[7] Ramy A. Zeineldin, Mohamed E. Karar, Ziad Elshaer, ·Jan Coburger, Christian R. Wirtz, Oliver Burgert, and Franziska Mathis-Ullrich. Explainability of deep neural networks for mri analysis of brain tumors. *International Journal of Computer Assisted Radiology and Surgery*, 17:1673–1683, 04 2022.

[8] Rezuana Haque, Md. Mehedi Hassan, Anupam Kumar Bairagi, and Mohammed Shariful. Neuronet19: an explainable deep neural network model for the classification of brain tumors using magnetic resonance imaging data. *Scientific Reports*, 14, 01 2024.

[9] Gopal S. Tandel, Mainak Biswas, Omprakash G. Kakde, Ashish Tiwari, Harman S. Suri, Monica Turk, John Laird, Christopher Asare, Annabel A. Ankrah, N. N. Khanna, B. K. Madhusudhan, Luca Saba, and Jasjit S. Suri. A review on a deep learning perspective in brain cancer classification. *Cancers*, 11:111, 01 2019.

[10] Milica M. Badža and Marko Č. Barjaktarović. Classification of brain tumors from mri images using a convolutional neural network. *Applied Sciences*, 10:1999, 03 2020.

[11] Hanaa ZainEldin, Samah A Gamel, El-Sayed M El-kenawy, Amal Alharbi, Doaa Khafga, Abdelhameed Ibrahim, and Fatma M Talaat. Brain tumor detection and classification using deep learning and sine-cosine fitness grey wolf optimization. *Bioengineering (Basel)*, 10:18–18, 12 2022.

[12] Msoud Nickparvar. Brain tumor mri dataset, 2021.

[13] Veronika Cheplygina, Marleen de Bruijne, and Josien P.W. Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 05 2019.

[14] A. Inés, C. Domínguez, J. Heras, E. Mata, and V. Pascual. Biomedical image classification made easier thanks to transfer and semi-supervised learning. *Computer Methods and Programs in Biomedicine*, 198:105782, 01 2021.

[15] Chenjie Ge, Irene Yu-Hua Gu, Asgeir Store Jakola, and Jie Yang. Deep semi-supervised learning for brain tumor classification. *BMC Medical Imaging*, 20, 07 2020.

[16] Reham Kaifi. A review of recent advances in brain tumor diagnosis based on ai-based classification. *Diagnostics*, 13:3007, 01 2023.

[17] Irena Ilić and Milena Ilić. International patterns and trends in the brain cancer incidence and mortality: An observational study based on the global burden of disease. *Heliyon*, 9:e18222–e18222, 07 2023.

[18] Brain Tumour Charity. Source k: Brain tumour statistics, 06 2023.

[19] Novsheena Rasool and Javaid Iqbal Bhat. Brain tumour detection using machine and deep learning: a systematic review. *Multimedia tools and applications*, 67, 05 2024.

[20] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[21] R D Ghongade and D G Wakde. Computer-aided diagnosis system for breast cancer using rf classifier. *International Conference on Wireless Communications and Signal Processing*, 1, 03 2017.

[22] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-Andre Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Herve Delingette, Cagatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, Jose Antonio Mariz, Raphael Meier, Sergio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34:1993–2024, 10 2015.

[23] W3Schools. W3.css templates, 2020.

[24] Sunil Kumar, R Dhir, and N Chaurasia. Brain tumor detection analysis using cnn: A review — ieee conference publication — ieee xplore, 2021.

[25] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9:611–629, 06 2018.

[26] Paolo Arena, Adriano Basile, Maide Bucolo, and Luigi Fortuna. Image processing for medical diagnosis using cnn. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 497:174–178, 01 2003.

[27] Sasanka Potluri, Alireza Fasih, Laxminand Kishore Vutukuru, Fadi Al Machot, and Kyandoghere Kyamakya. Cnn based high performance computing for real time image processing on gpu, 07 2011.

Brain Tumour Classification using a Convolutional Neural Network

Ella Vithanage
MSc Artificial Intelligence
De Montfort University, Leicester, UK
Applied AI
P2537413@my365.dmu.ac.uk

***Abstract*—This document underpins the usage of Convolutional Neural Networks (CNNs) to classify Brain Tumours using Artificial Intelligence techniques. MRI scans are used for the basis of the application of CNNs and are classified using Supervised Learning in Python on Google Colab. The system provided is a simple CNN model which manages to classify the images to a 75% accuracy in training and 63% accuracy during testing. Limitations are outlined within computational complexity and time to produce results which, if amended, would help to create a greater model for more generalisability.**

***Index Terms*—Convolutional Neural Networks, Deep Learning, Data Preprocessing, Supervised Learning, Python, Google Colab, Computer Vision**

I. Introduction

The recent advancements of CNNs have revolutionised many fields of AI, with various convolution types such as 1-D, 2-D and multidimensional, as well as classic and advanced models [1]. The real-life applications of this method within the branch of deep learning include image classification [2], object detection [3] and robotics [4]. As well as its applications to image processing, it can also be used for speech processing which is useful in broader subject areas such as natural language processing [5]. Many articles outline that these systems may have limitations in terms of computer complexity and managing to use large data sets with a low processing time.

Research that has been undertaken on categorising brain tumours using MRI image scans have been tackled in different ways. For instance, one article uses a modified U-Net model aiming to improve accuracy and efficiency [6]. More specifically, when using deep learning techniques, Probabilistic Neural Networks can be used alongside Particle Swarm Optimisation (PSO) and Ant Colony Optimisation [7]. Other methods include Residual Networks (ResNet), Sub-pixel Convolution, Histogram Equalisation and Gray Level Co-occurrence Matrix [8]. Collaboratively, this report contributes to the working of EVA (Enhanced Visionary Analytics) which provides several different methods to tackle the same problem of brain tumour classification which is defined in more depth in the 'problem definition' section of this report.

A. Abbreviation's and Acronyms

AI: Artifical Intelligence, CNN: Convolutional Neural Networks, DL: Deep Learning, NLP: Natural Language Processing

II. System Proposal

The model defined is a basic CNN which works to classify brain tumours from four labeled categories: 'glioma', 'meningioma', 'pituitary' and 'no-tumour' [9]. As the model is using labeled data, it works by using a supervised learning method where the data is split into categories and aims to situate the data in the correct places. Confusion matrices will be utilised to show the split of what data is correctly classified, and can be used to underpin any data that is being overfitted.

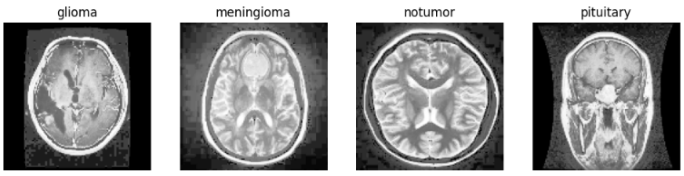

Fig. 1. Examples of the labelled MRI brain scan dataset.

III. Methodology

A. System Methodology Outline

The outline of the proposed CNN system is as followed, the methodology exists in the creation of the system, which is defined as:

- Problem Definition:
  The first step in creating a CNN system is to define the problem which is wished to be answered by the system.
- Data Preprocessing:
  Data preprocessing is the second step in building a CNN, the data must be accessible to the environment it is working in. It also must be split into test, validation and training sets.
- Architecture Design:
  The CNN must be designed in terms of the parameters. This includes: input layer, convolutional layers, activation function, pooling layers, connected layers and the output layer.
- System Compiling:
  To check if the model works, it must be ran and provide indication that it situates in the way that it is meant to. Compiling features are utilised here.
- Model Training:

To use a CNN the model must be trained, this is so the system is able to learn from trial and error, using deep learning.

- System Evaluation:
  By evaluating the system using the accuracy of classification, this can show any areas that require further improvement.
- Hyperparameter Tuning:
  To enhance the accuracy of results for the CNN model, the parameters must be fine-tuned so it has a greater chance of providing better classification levels.

IV. Problem Definition

The problem at hand is within classifying brain tumours to their correct labels. This would be useful as a tool in aiding medical professionals to help diagnose tumours without as much human input. As stated in the introduction, CNNs have great applications to image classification and this is the rationale behind using this advanced AI technique. The main aim is to have a system that can provide classifications of four labelled datasets. Three of the datasets provide images of tumours and the final, provides images of brains without any tumour. By using labelled data, the problem can be approached by using supervised learning where the system aims to classify pre-existing labelled data. To achieve the aim, the model must be able to distinguish between the different types of tumours that are presented.

V. Data Preprocessing

To use the visual, image data in Google Colab, this data must be easily accessible. Through trial and error this was accessible by using the Google Drive and mounting to avoid loading the data each time and avoid RAM usage conditions on Colab. This process, although tedious, is necessary for being able to gain access to the data set. The preprocessing that is also required, is rescaling the images so it is able to be read in Python. Splitting the data into testing, training and validation sets is needed within building a CNN model. The purpose of this split within training is to use this set to train the CNN, this is where the model learns to recognise patterns and features within the data. The way this works in the system is by iteratively adjusting parameters by attempting to minimise the loss function on the training data. In the validation set, the parameters are tuned and used to evaluate the model's performance during training, this can be used to help to reduce the likelihood of overfitting, stopping the training early and selecting different models/settings to be evaluated based on performance. Finally, the test set is used to evaluate the final performance of the model, providing an unbiased estimate of the generalisability of the system to unseen data.

VI. Architecture Design

The design of the system is a main element in the creation and deployment of a CNN model. As CNNs mainly operate via trial and error through corrections iteratively, the parameters that are used are a crutial point of interest for how the system will perform down the line. The initial design entailed the following:

- Input layer: This recieves the MRI scans as inputs and uses the dimensions and colour channels to be able to interpret the images.
- Convolutional layers: The model uses three convolutional layer which each have filters.
- Pooling layers: Three pooling layers are used in this systems architecture.
- Dropout layer: This prevents overfitting by randomly selecting half of the input units to be nullified, which enforces the model to learn more robust features.
- Flatten layer: This is important for the processing of dense layers and reducing computational complexity.
- Fully Connected Layer: These layers connect the neurons from one layer to the other, which allows for the images to be processed.
- Output layer: The output layer gives the accuracy levels of how well the system performs based on the precision of correctly classified images.

See Fig. 11 for the full list of parameters used within the CNN system.

VII. System Compiling

To evaluate the performance of the model in general, it is important that it is able to be compiled which is an indication of whether the system is working or not. If error messages are thrown out, this can show that there are problems in the way the data is being processed. The compilation of this system includes:

- Optimiser: The optimiser chosen to perform in this system is 'Adam', which is an adaptive learning rate algorithm.
- Loss Function: The loss function is Categorical Cross-entropy, which is suitable to multi-class classification problems such as the one being tested upon in this paper.
- Metrics: Accuracy is used as the metric to provide the proportion of correctly classified tumours.

These elements are important for the way that the results are analysed and executed, as accuracy is used for the metrics, this is the means of evaluation that the system will be compared upon.

VIII. Model Training

To train the data, the system learns through the number of epochs that the system is able to run through. As CNNs require great computational engines, this provides a problem for the time constraints that need to be adhered to. Google Colab also provides problems in the RAM allowance and disconnecting when all the allowance is used. Using the 'model.fit' command in Python allows the data to be fitted in the most practical way and when the complexity is reduced, enables a conventional way of training and fitting data. With the first run of the system resulting in an hour per epoch, it was clear that adjustments must be made for any results to be presented. From the

architecture shown, the model must be simplified so it's able to generate results in a faster processing time. GPU is also used to speed up time to output the accuracy of the model, however provides limitations as there is a usage limit of this element on Colab.

The most sufficient means of evaluation was decided to be through benchmark tests of one (Benchmark 1), two (Benchmark 2) and three (Benchmark 3) epochs due to the heavy computational requirements of executing a CNN. This method of evaluation is only able to consider very finite runs which will not allow for generalisability, however will provide a basis for a simple CNN working model that could be used to categorise medical images by extracting features through training.

IX. System Evaluation

To evaluate the system, the generalisability of the precision of accuracy using metrics must be considered. By using a test set, the system can provide indications of whether it is performing well or if there are any further improvements that can be made.The results executed by the test sets relate to how well the system is able to be used on unseen data and therefore how reliable it is in categorising on a greater scale than the training set alone.

X. Hyperparameter Tuning

To improve the system, hyperparameters must be ammended. In this case, as the system is being deployed under time constraints, the majority of fine tuning techniques are under the basis of reducing computational complexity and time to produce results. To improve the model, epochs can be changed, as well as some of the compiling features. As seen, with the time of computation being great, the only evaluation suggested is on a very finite number of epochs and with more time for development, could situate in much more strenuous results.

XI. Results and Analysis

As the CNN model provides computational complexity, the most suitable method for evaluation is using benchmarks where the epochs are increased iteratively. With each epoch taking around an hour, this situated in pressures to produce results given the time constraints that must be adhered to.

A. Benchmark 1: One Epoch

Training and Validation:

The training classification set accuracy resulted in the loss of 1.0550 and accuracy being 50% (to 2 significant figures), within the validation set, the results concluded that the loss is much higher, at 1.3812 and accuracy is slightly lower giving 45% (to 2 significant figures).

Testing:

The system results in an overall accuracy level of 62% and loss of 0.9982, lower than the test or validity set results.

Summary:

- Training Loss = 1.0550, Training Accuracy = 50%
- Validation Loss = 1.3812, Validation Accuracy = 45%
- Test Accuracy = 62%
- Test Loss = 0.9982

See Fig. 8 for the classification report for Benchmark 1.

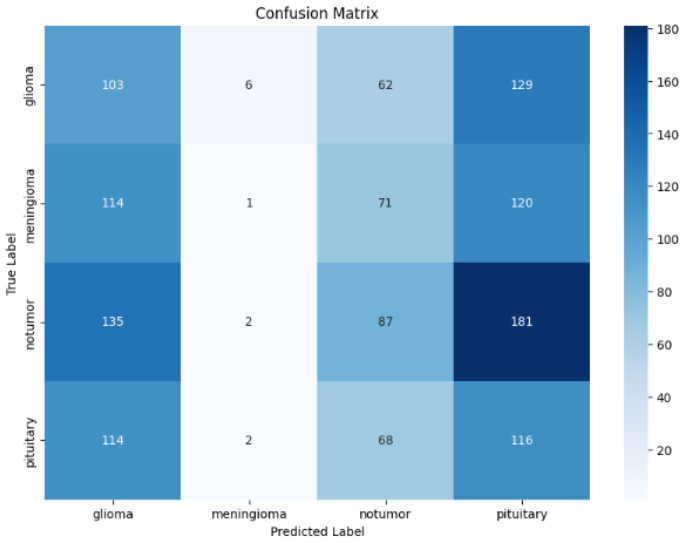

Fig. 2. Confusion Matrix to show the spread of data for Benchmark 1.

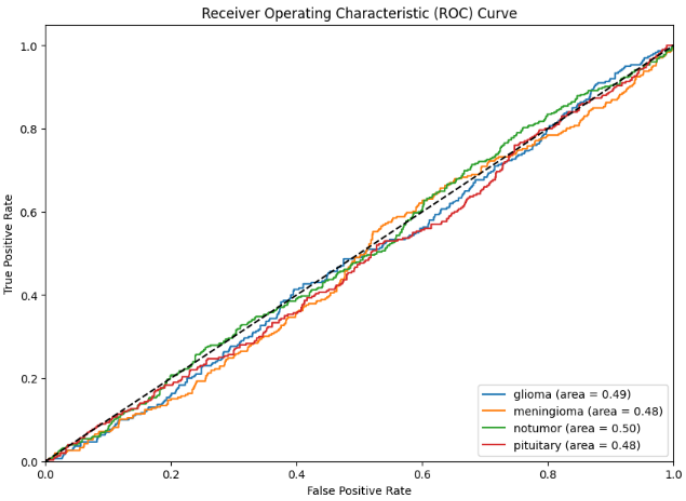

Fig. 3. ROS to show the spread of data for Benchmark 1.

B. Benchmark 2: Two Epochs

Training and Validation

The best results from two epochs being ran through the system situated in the training loss being 0.7970 and accuracy being 66% (to 2 significant figures). The validation results suggest that the loss was 0.9035 and the accuracy is 66% (to 2 significant figures).

Testing

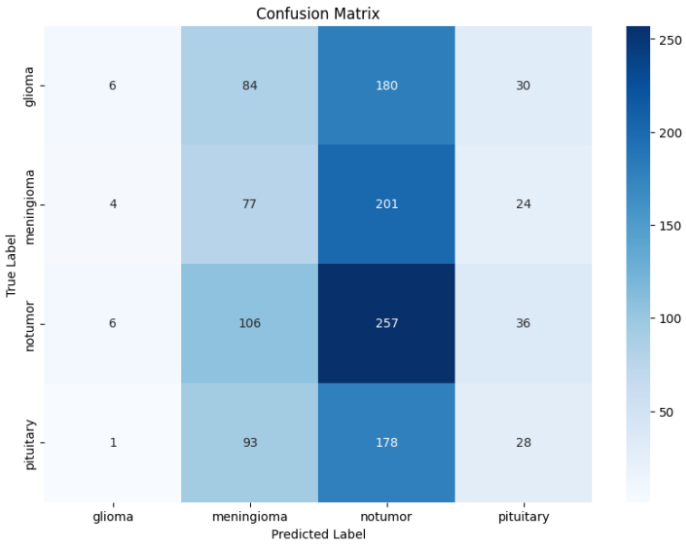

Fig. 4. Confusion Matrix to show the spread of data for Benchmark 2.

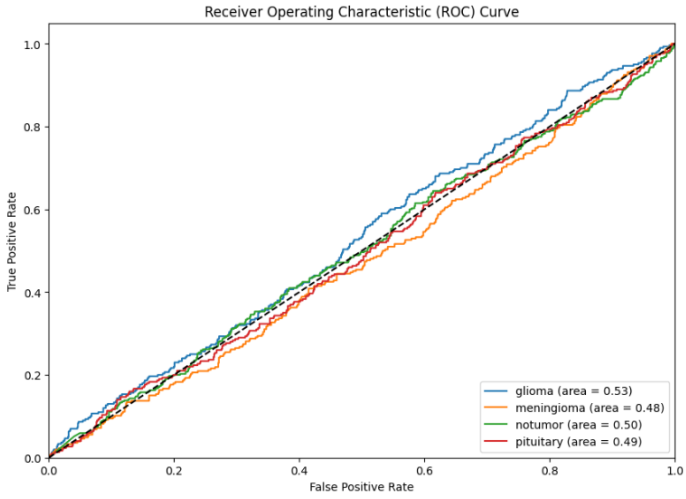

Fig. 5. ROS to show the spread of data for Benchmark 2.

This run, the system concluded in an overall accuracy of 49% (to 2 significant figures) and loss of 1.6382. This indicates that the model is overfitting, which is common in the training phase.

Summary:

- Training Loss = 0.7970, Training Accuracy = 67%
- Validation Loss = 0.9035, Validation Accuracy = 66%
- Test Accuracy = 49%
- Test Loss = 1.6382

See Fig. 8 for the classification report for Benchmark 2.

C. Benchmark 3: Three Epochs

Training and Validation

For the third benchmark, the results show the training loss being 0.6325 and the classification accuracy is 75% (to 2 significant figures). The validation loss is 0.9053 and accuracy results with 71% (to 2 significant figures).

Testing

The overall results for this benchmark subsided in the accuracy level of classification being 63% (to 2 significant figures) and the loss being 1.1953.

Summary:

- Training Loss = 0.6325, Training Accuracy = 75%
- Validation Loss = 0.9053, Validation Accuracy = 71%
- Test Accuracy = 64%
- Test Loss = 1.1953

See Fig. 10 for the classification report for Benchmark 3.

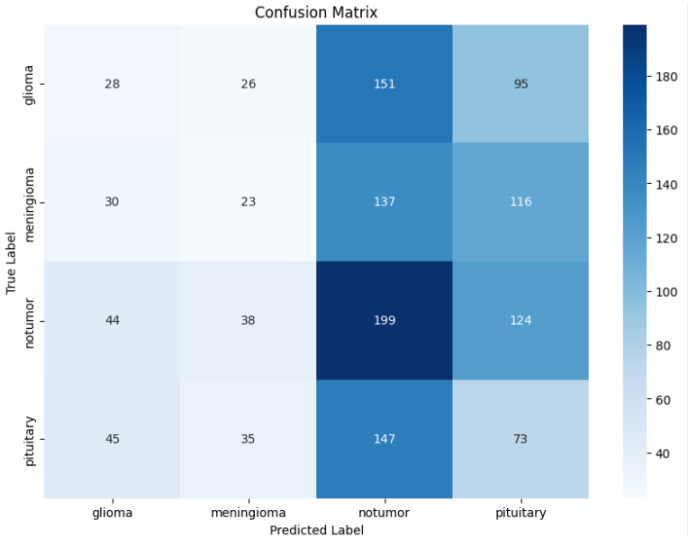

Fig. 6. Confusion Matrix to show the spread of data for Benchmark 3.

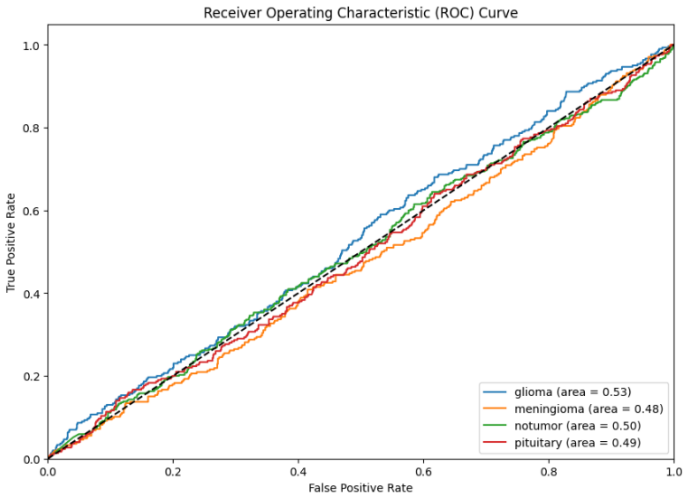

Fig. 7. ROS to show the spread of data for Benchmark 3.

XII. Conclusions and Recommendations

In conclusion, presented is a valid CNN model which works to categorise brain tumours. The system, although simple,

works to be able to utilise data after bring preprocessed, and use computer vision techniques to be able to correctly classify the tumours using supervised learning.

A. Training and Validation

After one individual epoch, the system has little opportunity to learn from the training data. The loss is relatively high and accuracy is fairly low (50%), this indicates that the model has began to learn but is far from optimal solutions being generated. In the validation phase, the loss is even higher and accuracy lower, this suggests that the model is not generalising to unseen data well, as it has only ran through one epoch. In the second benchmark, there is more time to learn from the training data and the loss decreases to 0.7970 whist the accuracy improves to 67%, indicating the model is learning. In the validation set, the loss also was decreased and accuracy improved, which indicates better generalisability. Finally, with three epochs, the loss further decreases and the accuracy increased to 75% being the best results seen within all the benchmarks. The validation loss is slightly higher than the training loss but still an improvement on the previous benchmark, the accuracy in this phase also increases to 71% which shows that incrementally, it is improving.

B. Testing

The test results indicate how the CNN is able to perform when given unseen data. For the first Benchmark, the test accuracy is 62% and suggests that the model is fairly good at generalising new data, as it is higher than the validation accuracy, this could indicate that there is variability due to the size of the test set. The loss in this instance is lower than the validation loss, but higher than training loss. This suggests that the model may have learnt some valuable patterns, but is not optimal after executing one epoch. For two epochs, the test accuracy surprisingly drops (to 49%), this inconsitency may suggest that the model has overfitted the training data in the second epoch, resulting in poorer generalisation of the test set. The loss is also significantly higher than both the training and validation loss which further supports the notion of overfitting. When using three epochs, the accuracy improves to be 63% accurate which aligns with the validation accuracy closer, this suggests that the more epochs that are ran through on the system, the better chance of a higher accuracy level being achieved. The loss again is higher than both the training and validation loss and this shows there is still room for improvement for the system.

C. Overall Analysis

- Generalisation: The validation loss and accuracy improves with more epochs, indicating better generalisation. The validation loss does not decrease as consistently as the training loss, which could indicate that the model has overfitted data in the training phase.
- Learning Progression: As the epochs increase, the training loss decreases. This trend shows that the system is learning each time the epoch is increased.
- Convergence: To monitor the overall convergence of the system, additional epochs could be used and the metrics could be monitored.
- Overfitting: The discrepencies in the second epoch test results suggest that the model has overfitted the data, this occurs when the model learns patterns specific only to the training data and then cannot generalise this well to new data.
- Improvements: The system seems to adapt in the third epoch, to fix the issue of overfitting within the second epoch. This suggests that the model is able to adapt itself by using the CNN.

D. Recommendations and Further Work

To improve this system, the model should be tested with far more epochs which seem to improve the chances of convergence to the optimal solution for the system. The system could have early stopping implemented based on validation loss to prevent overfitting when the validation loss stops improving. The learning rate, batch size and other hyperparameters could be further fine tuned to see the implications of these features on execution time and performance. Within testing, the training and validation metrics should be monitored to see the impact of this on the testing set and hence, the generalisability of the model. The epochs could be increased with more step ups such as 10's or 100's to see the impact, if hyperparameter changes reduce computational time. More regularisation techniques such as dropout and weight decay could be employed on the mode to prevent overfitting and improve generalisation.

Further work could be approached by using the system with given improvements and trialing it on different data sets. This will verify whether the system is generalisable or not.

XIII. Critical Analysis

This documentation provides a comprehensive overview of the applications of Convolutional Neural Networks (CNNs) in classifying brain tumours using AI techniques. The paper discusses appropriate methodology, system proposal, architecture design and results obtained from training and testing of the CNN model.

Overall the report provides a valid and sufficient evaluation of methodology used and results acquired. However there are aspects that could be further developed and improved on:

- Technical Details

Whist listing many well-known terminologies, this report lacks depth for technical details. Providing more information about the specific CNN architecture used may enhance the reader's understanding of the approach being tackled.

- Benchmarking Approach

Due to time constrictions, the only method of evaluation constituted in comparing different, finite number of epochs. To improve on this approach, experiments could be made with different hyperparameter changes which would provide a more comprehensive analysis of the system's behaviour.

- Discussion of Results

A brief discussion of the results of using different epochs are discussed, this could be improved by providing more detailed and insightful feedback. The graphs could have further analysis when portrayed.

XIV. Acknowledgements

Acknowledgements must be made to Victoria Hecktor and Abdulaziz Alqulayti who have been invaluable to the creation of this system through a magnitude of support. It has been an honour to work aside such compelling people, who have brought great insight into the formalisation of the group project and aiding the journey to creating this proposed model. Nathanael Baisal also deserves commendation for bringing such an interesting project to light, and for teaching valuable and applicable skills which will bring great opportunities for the future.

References

[1] viii. Li, Z., Liu, F., Yang, W., Peng, S. and Zhou, J., 2021. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 33(12), pp.6999-7019.

[2] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[3] Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[4] Levine, S., Pastor, P., Krizhevsky, A. and Quillen, D., 2016. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. The International Journal of Robotics Research, 37(4-5), pp.421-436.

[5] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J. and Chen, T., 2018. Recent advances in convolutional neural networks. Pattern recognition, 77, pp.354-377.

[6] Pedada, K.R., Rao, B., Patro, K.K., Allam, J.P., Jamjoom, M.M. and Samee, N.A., 2023. A novel approach for brain tumour detection using deep learning based technique. Biomedical Signal Processing and Control, 82, p.104549.

[7] Selvy, P.T., Dharani, V.P. and Indhuja, A., 2019. Brain tumour detection using deep learning techniques. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol, 169, p.175.

[8] Nazir, M., Shakil, S. and Khurshid, K., 2021. Role of deep learning in brain tumor detection and classification (2015 to 2020): A review. Computerized medical imaging and graphics, 91, p.101940.

[9] M. Nickparvar, "Brain Tumor MRI Dataset," www.kaggle.com, 2021. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset (accessed May 14, 2024).

XV. Appendices: Figures

Classification Report	precision	recall	f1-score	support
glioma	0.22	0.34	0.27	300
meningioma	0.09	0.00	0.01	306
notumor	0.30	0.21	0.25	405
pituitary	0.21	0.39	0.27	300
accuracy			0.23	1311
macro avg	0.21	0.24	0.20	1311
weighted avg	0.21	0.23	0.20	1311

Fig. 8. Classification Report for Benchmark 1.

Classification Report	precision	recall	f1-score	support
glioma	0.35	0.02	0.04	300
meningioma	0.21	0.25	0.23	306
notumor	0.31	0.63	0.42	405
pituitary	0.24	0.09	0.13	300
accuracy			0.28	1311
macro avg	0.28	0.25	0.21	1311
weighted avg	0.28	0.28	0.22	1311

Fig. 9. Classification Report for Benchmark 2.

Classification Report	precision	recall	f1-score	support
glioma	0.19	0.09	0.13	300
meningioma	0.19	0.08	0.11	306
notumor	0.31	0.49	0.38	405
pituitary	0.18	0.24	0.21	300
accuracy			0.25	1311
macro avg	0.22	0.23	0.21	1311
weighted avg	0.23	0.25	0.22	1311

Fig. 10. Classification Report for Benchmark 3.

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 254, 254, 32)	896
max_pooling2d (MaxPooling2D)	(None, 127, 127, 32)	0
conv2d_1 (Conv2D)	(None, 125, 125, 64)	18496
max_pooling2d_1 (MaxPooling2D)	(None, 62, 62, 64)	0
conv2d_2 (Conv2D)	(None, 60, 60, 128)	73856
max_pooling2d_2 (MaxPooling2D)	(None, 30, 30, 128)	0
dropout (Dropout)	(None, 30, 30, 128)	0
flatten (Flatten)	(None, 115200)	0
dense (Dense)	(None, 512)	58982912
dense_1 (Dense)	(None, 4)	2052

Fig. 11. Parameters for System Architecture.

Enhancing Brain Tumour Classification with Semi-Supervised Deep Learning and Supervised Reinforcement Learning

Victoria Hektor
Artificial Intelligence MSc
De Montfort University
Leicester, United Kingdom
p2629898@my365.dmu.ac.uk

***Abstract*—This report explores the application of advanced artificial intelligence methods for brain tumour detection using MRI scans and Deep Learning (DL) techniques. The project integrates semi-supervised reinforcement learning with transfer learning, specifically employing the VGG16 convolutional neural network, [11]. Three benchmarks were established to evaluate the effectiveness of various techniques, including standard supervised learning, enhanced data augmentation, class weighting, and a custom hybrid approach combining supervised and reinforcement learning. The primary development environment was Google Colab, [13], with image data stored and accessed via Google Drive. Despite some computational limitations inherent to Colab, the findings provide valuable insights into the performance of different machine learning strategies for brain tumour classification. Benchmark One demonstrated the highest accuracy at 28%, while subsequent models incorporating more sophisticated techniques yielded marginal improvements, highlighting the need for more robust computational resources and refined methodologies. [12]. The results underscore the complexities of medical image classification and the potential of reinforcement learning when meticulously implemented.**

***Index Terms*—Brain tumour detection, MRI classification, Machine learning, Deep learning, Reinforcement learning, Transfer learning, VGG16, EfficientNetB0, Image augmentation, Google Colab, TensorFlow Keras, Hyperparameter tuning, Confusion matrix, Classification report, Medical imaging, Supervised learning, Data preprocessing, Class imbalance, Model evaluation, GPU computing**

I. Introduction

Brain tumours pose significant challenges in medical imaging and diagnosis due to their intricate nature and potential for rapid progression. Precise and early detection is paramount for effective treatment and improved patient outcomes. Recent advancements in machine learning and deep learning have revolutionised the field, offering unprecedented accuracy in brain tumour classification. This project, EVA (Enhanced Visionary Analytics), seeks to harness these advancements by integrating semi-supervised reinforcement learning with transfer learning, specifically utilising the VGG16 convolutional neural network. By combining these sophisticated methodologies, EVA aims to improve the accuracy and reliability of brain tumour diagnosis using MRI scans, thereby contributing to more effective and timely medical interventions.

II. Literature Review

The classification of brain tumours using MRI data has been extensively studied, with various methodologies being explored to enhance accuracy and efficiency. Nickparvar's dataset [1], a widely used resource in the research community, provides a robust foundation for training and evaluating classification models. Papers With Code [2] has benchmarked several models on this dataset, highlighting the ongoing advancements and challenges in this field. Recent studies have introduced sophisticated models, such as the Vision Transformers Ensembling technique, which has shown significant promise in improving classification performance [3]. Another innovative approach is the hybrid deep CNN model proposed by Srinivasan et al. [4], which integrates multiple deep learning architectures to enhance multi-class classification of brain tumour images. Furthermore, the framework developed by Ullah et al. [5], which combines deep learning with Bayesian optimisation and quantum theory-based algorithms, represents a cutting-edge advancement in the application of machine learning to medical imaging.

The reviewed articles collectively contribute to the advancement of brain tumour classification using deep learning techniques. Ge et al. (2020) and Cheng et al. (2016) both emphasise the significance of addressing dataset imbalances and utilising data augmentation to enhance model robustness. Ge et al.'s novel semi-supervised learning framework effectively combines labelled and unlabelled data, showcasing the potential of leveraging unannotated datasets in medical imaging. Cheng et al.'s implementation of cost-sensitive learning further highlights the importance of tailored approaches to handle class imbalances. [7], [8].

He et al. (2015) provide a foundational innovation with their introduction of residual networks, which have since been widely adopted in various image recognition tasks, including medical imaging. The ability to train deeper networks without degradation significantly improves the performance and reliability of classification models. [9]. Yu et al. (2022) demonstrate the efficacy of supervised learning models and highlight the importance of thorough cross-validation to achieve high

classification performance. [10]. The integration of semi-supervised learning, data augmentation, and deep residual networks presents a comprehensive approach to enhancing brain tumour classification. Future research should focus on refining these techniques and exploring their synergistic potential to further improve diagnostic accuracy and clinical outcomes in medical imaging.

III. System Proposal

A. Data

The dataset used in this study is sourced from Kaggle, specifically the "Brain Tumour MRI Dataset" curated by Nickparvar [1]. This dataset comprises MRI scans categorised into four classes: glioma, meningioma, no tumour, and pituitary tumour. The images are preprocessed to ensure consistency in size and quality, which is crucial for training robust machine learning models. The dataset is divided into training, validation, and test sets to facilitate the evaluation of model performance.

B. Computational Resources

This project utilised Google Colab as the primary development environment. Google Colab provides a convenient platform for running Python code, particularly for machine learning and deep learning applications. The images used for brain tumour classification were stored in Google Drive. By mounting Google Drive within the Colab environment, the images were accessed seamlessly, enabling efficient data loading and processing. This setup leveraged Colab's GPU capabilities, which are essential for handling the computational demands of deep learning models. [13].

C. Implementation

The implementation of the EVA model involves several key steps that leverage advanced machine learning techniques to accurately classify MRI images into four categories: glioma, pituitary tumour, meningioma, and no tumour. Initially, the dataset is pre-processed using data augmentation methods such as rotation, zoom, and horizontal flipping to enhance model robustness and prevent overfitting. Images are resized to 128x128 pixels and normalised to ensure consistent input to the neural network. The model utilises a pre-trained VGG16 convolutional neural network, renowned for its efficacy in image classification tasks, which is fine-tuned by adding custom fully connected layers tailored to the specific classification needs. The transfer learning approach harnesses the pre-existing knowledge of VGG16, reducing the amount of data and computational resources required for training. The training process is optimised using the Adam optimiser and the categorical cross-entropy loss function, suitable for multi-class classification.

- Data Preparation and Augmentation: The dataset is augmented using the ImageDataGenerator class from Keras, which applies transformations like rotation, zoom, horizontal flip, width and height shifts, and rescaling. This helps in improving the model's ability to generalise by providing a diverse set of training images.
- Model Architecture: The system leverages the VGG16 pre-trained model from Keras. The top layers of VGG16 are excluded to allow the addition of custom dense layers. This transfer learning approach uses the pre-trained weights of VGG16 for feature extraction, providing a strong foundation for the model.
- Training with Callbacks: To optimise the learning process, the model is trained with two crucial callbacks: ReduceLROnPlateau and EarlyStopping. ReduceLROnPlateau reduces the learning rate when a metric has stopped improving, which helps in fine-tuning the model. EarlyStopping stops the training process if there is no improvement in the validation loss for a specified number of epochs, preventing overfitting.

To further refine the model, reinforcement learning techniques are integrated following the supervised learning phase. A custom environment is defined where the model's actions, such as classifying images, are rewarded or penalised based on correctness. This hybrid approach allows the model to interact with the environment, receiving rewards for correct classifications and penalties for incorrect ones, thus enhancing its decision-making process through reinforcement learning.

IV. Analysis Techniques

The visualisations chosen for this implementation are integral to monitoring and fine-tuning the EVA model. Training and validation accuracy and loss curves provide real-time insights into the model's learning process, helping to detect overfitting and underfitting by comparing performance on training data versus unseen validation data. A consistent divergence between these curves signals potential issues with model generalisation, prompting adjustments in the training regimen or data augmentation strategies. The confusion matrix is another visual tool, offering a detailed breakdown of the model's classification performance across each category. By displaying the true versus predicted labels, it highlights specific areas where the model struggles, such as misclassifying glioma as meningioma. This detailed error analysis informs targeted refinements in the model architecture or preprocessing steps. Additionally, classification reports provide comprehensive metrics like precision, recall, and F1-score for each class, enabling a nuanced understanding of the model's strengths and weaknesses. Sample prediction visualisations, juxtaposing predicted and actual labels, offer qualitative insights that complement the quantitative metrics, ensuring a holistic evaluation of model performance. These visual tools collectively guide iterative improvements, enhancing the EVE model's accuracy and reliability for brain tumour classification. To ensure robust evaluation and analysis of the model's performance, several techniques and metrics are employed:

- Data Augmentation Analysis: By visualising augmented images and ensuring diversity, the system verifies that the augmentation process is effectively increasing the vari-

ance within the training set. This is crucial for preventing overfitting and improving generalisation.

- Model Performance Metrics: The system employs standard classification metrics such as accuracy, precision, recall, and F1-score to evaluate the model's performance. These metrics provide a comprehensive view of how well the model is performing across different classes.
- Confusion Matrix: A confusion matrix is generated to understand the model's performance in terms of true positives, true negatives, false positives, and false negatives. This helps identify specific classes where the model might be struggling.
- Learning Curves: Plotting learning curves for both training and validation loss/accuracy over epochs helps in diagnosing underfitting or overfitting. These curves indicate how the model's performance evolves with each epoch, providing insights into whether the model needs more epochs, data, or a different architecture.
- Hyperparameter Tuning: The system systematically varies key hyperparameters such as learning rate, batch size, and dropout rates to find the optimal configuration. This tuning process helps in identifying the best parameters that enhance model performance.

V. Testing & Analysis

Benchmark One

A. Confusion Matrix:

The confusion matrix for Benchmark One reveals the following key observations:

- Glioma: The model correctly classifies 64 instances but misclassifies a significant number of instances into other categories, particularly 'notumor' and 'pituitary'.
- Meningioma: The model correctly identifies 70 instances but has a high rate of misclassification, especially into 'notumor'.
- Notumour: This category has the highest correct classification (142) but still suffers from substantial misclassifications.
- Pituitary: The correct classification count is 95, with notable confusion with 'notumor'.

B. Learning Curves:

The training and validation accuracy curves indicate a consistent improvement over epochs, with training accuracy reaching around 84% and validation accuracy around 82%. However, there is a slight gap between the training and validation accuracies, suggesting some degree of overfitting. The loss curves show a decreasing trend for both training and validation loss, with the training loss reaching below 0.4 and validation loss around 0.5. This suggests that the model is learning effectively but still has room for improvement in generalisation.

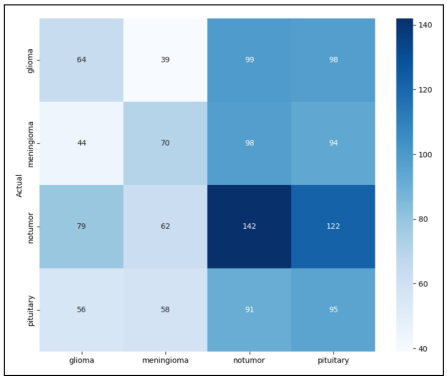

Fig. 1. Benchmark One Confusion Matrix

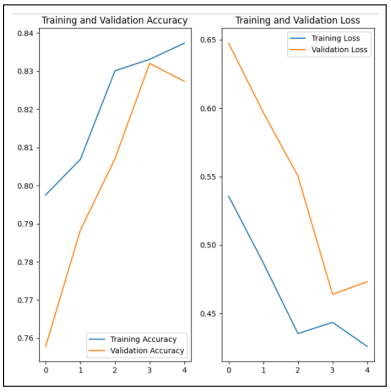

Fig. 2. Benchmark One Training Accuracy & Validation

C. Classification Report:

- Glioma: Precision (0.26), Recall (0.21), F1-score (0.24)
- Meningioma: Precision (0.31), Recall (0.23), F1-score (0.26)
- Notumor: Precision (0.33), Recall (0.35), F1-score (0.34)
- Pituitary: Precision (0.23), Recall (0.32), F1-score (0.27)
- Overall: Accuracy is 28%, with macro and weighted averages around 28-29%.

TABLE I
Classification Report - Benchmark One

Class	Precision	Recall	F1-Score	Support
Glioma	0.26	0.21	0.24	300
Meningioma	0.31	0.23	0.26	306
Notumor	0.33	0.35	0.34	405
Pituitary	0.23	0.32	0.27	300
Accuracy	0.28	0.28	0.28	1311
Macro avg	0.28	0.28	0.28	1311
Weighted avg	0.29	0.28	0.28	1311

Benchmark Two

Enhanced data augmentation techniques were implemented, including increased rotation and shift parameters, as well as brightness adjustments. These changes helped create a more diverse training dataset, thereby reducing overfitting. To handle class imbalances, class weights were applied, giving more

weight to underrepresented classes. This adjustment aimed to balance the training process and improve performance across all classes. Additionally, the prediction function was improved for better error handling, optional visualisation, and flexibility in parameters.

D. Confusion Matrix:

The confusion matrix for Benchmark Two reveals the following observations:

- Glioma: The model correctly classifies 60 instances, showing a slight improvement from Benchmark One but still misclassifies many instances into 'notumor' and 'pituitary'.
- Meningioma: The model correctly identifies 46 instances, which is a decrease compared to Benchmark One, indicating that the changes did not improve classification for this category.
- Notumor: This category remains the highest for correct classification with 136 instances, showing a small decrease from Benchmark One but still the best performing class..
- Pituitary: Correct classification count is 69, which is slightly worse than Benchmark One but shows a different pattern of misclassification.

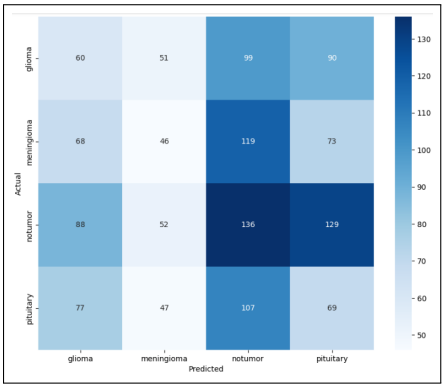

Fig. 3. Benchmark Two Confusion Matrix

E. Learning Curves:

The training and validation accuracy curves show an improvement over epochs, with training accuracy reaching around 82% and validation accuracy around 83%. There is a slight improvement compared to Benchmark One, with a smaller gap between the training and validation accuracies, suggesting reduced overfitting. The loss curves also show a decreasing trend for both training and validation loss, with the training loss reaching below 0.5 and validation loss around 0.6. The validation loss has improved, indicating better generalisation compared to Benchmark One.

Fig. 4. Benchmark Two Training Accuracy & Validation

F. Classification Report:

- Overall: Accuracy is 24%, with macro and weighted averages around 23-24%

TABLE II
CLASSIFICATION REPORT - MODEL 2

Class	Precision	Recall	F1-Score	Support
Glioma	0.00	0.00	0.00	300
Meningioma	0.00	0.00	0.00	306
Notumor	0.31	1.00	0.47	405
Pituitary	0.00	0.00	0.00	300
Accuracy	0.31	0.31	0.31	1311
Macro avg	0.08	0.25	0.12	1311
Weighted avg	0.10	0.31	0.15	1311

These metrics show that while there has been some improvement in certain classes, overall performance is still suboptimal. The most notable improvement is the slight increase in validation accuracy and the correct identification of the glioma tumour in the test example. For the third benchmark, an attempt at employing a custom reinforcement learning technique to see if this improves the outcome.

BENCHMARK THREE

Utilising a custom hybrid approach combining supervised learning and reinforcement learning techniques. Initially, data augmentation and preprocessing were performed using the ImageDataGenerator class from TensorFlow Keras to enhance the training dataset's diversity and robustness. Then constructed a deep learning model based on the EfficientNetB0 architecture, fine-tuning it to classify brain tumour images accurately. Hyperparameter tuning was facilitated by Keras Tuner, specifically using the Random Search strategy to optimise model parameters such as learning rate and dropout rate. To incorporate reinforcement learning elements, we introduced a custom callback that interacted with a simulated environment to provide additional rewards, thus integrating a reinforcement learning component into the training process. Finally, we evaluated the model's performance using a confusion matrix and a classification report, providing detailed insights into the classification accuracy across different tumour types.

G. Classification Report:

The classification report reveals the model's performance across four classes: glioma, meningioma, no tumour, and pituitary. Precision, recall, and F1-score are notably low for glioma, meningioma, and pituitary classes, with values of 0.00 for all metrics, indicating that the model failed to correctly classify any instances of these classes. In contrast, the "no tumour" class achieved a recall of 1.00 and a precision of 0.31, resulting in an F1-score of 0.47. The overall accuracy of the model stands at 0.31, suggesting a significant imbalance and misclassification issue. The macro average and weighted average for precision, recall, and F1-score also highlight the model's inadequacies in handling the diverse tumour types effectively.

H. Confusion Matrix:

The confusion matrix corroborates the findings of the classification report. It demonstrates that the model predominantly classified all test samples as "no tumour," ignoring the actual presence of glioma, meningioma, and pituitary tumours. This matrix provides a visual representation of the misclassification pattern, highlighting a critical flaw in the model's ability to differentiate between the tumour types.

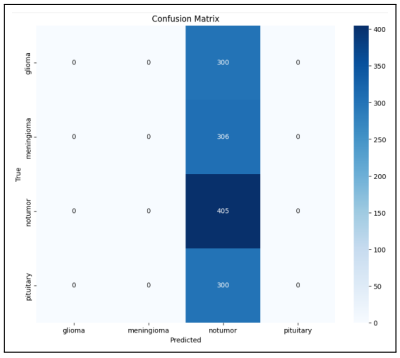

Fig. 5. Benchmark Three Confusion Matrix

I. Training Epochs and Rewards:

The training log, showcasing the reward mechanism introduced through reinforcement learning elements, indicates variability in the rewards across epochs. Despite the rewards system, the model's accuracy plateaued around 0.27 to 0.28, with minimal improvements across epochs. The consistency in validation loss and accuracy further underscores the model's inability to learn and adapt effectively to the classification task.

J. Model Accuracy and Loss:

The accuracy and loss plots for the model reflect minimal changes across the training epochs. The training accuracy slightly fluctuates between 0.2665 and 0.2792, while the validation accuracy remains nearly constant at 0.2796. The loss plot indicates a significant drop in the initial epoch, stabilising afterwards, suggesting that the model converges quickly but fails to generalise well, as indicated by the stagnant validation metrics.

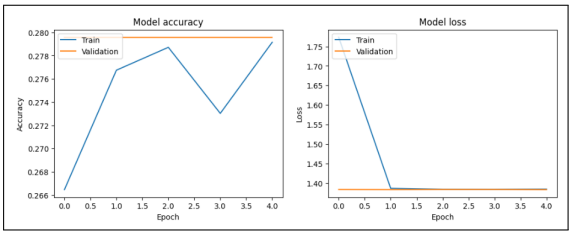

Fig. 6. Benchmark Three Model Accuracy, Validation & loss

K. Model Evaluation:

The final evaluation on the test data resulted in a test loss of 1.3798 and an accuracy of 0.3089. This outcome aligns with the training and validation performance, reiterating the model's overall inefficacy in distinguishing between the different tumour types accurately.

L. Hyperparameter Tuning Results:

The hyperparameter tuning process identified an optimal learning rate of 0.01 and a dropout rate of 0.3, with the best validation accuracy recorded at 0.2714. Despite the tuning efforts, the marginal improvement in accuracy underscores the need for a more sophisticated approach, possibly involving a different architecture or enhanced preprocessing techniques to address the model's deficiencies.

TABLE III
TRAINING RESULTS

Epoch	Time	Loss	Accuracy	Val_Loss	Val_Accuracy
1/10	20s 37ms	0.7073	0.7507	4.6563	0.2334
3/10	5s 25ms	0.3170	0.8888	4.4243	0.3089
5/10	5s 25ms	0.1556	0.9492	7.2835	0.2586
7/10	5s 25ms	0.0947	0.9788	17.3369	0.2288
9/10	5s 26ms	0.0530	0.9851	23.6810	0.2334

VI. Overall Analysis, Comparison & Conclusions

The three benchmarks provided distinct insights into the performance of brain tumour classification models, highlighting their respective strengths and weaknesses.

A. Considerations

While Google Colab offers substantial benefits, such as free access to GPUs and ease of integration with Google Drive, the project encountered limitations related to computational resources. Colab's restricted RAM and processing power posed significant challenges for this computer vision task, which requires extensive computational capabilities. The constrained environment limited the ability to train more complex models or perform extensive hyperparameter tuning. These limitations highlighted that, despite Colab's advantages, it may not be the most suitable platform for large-scale computer vision

projects. Future endeavours should consider more robust computational resources to overcome these constraints and achieve more efficient and effective model training and evaluation.

B. Overall Analysis

Benchmark One serves as the primary reference, with an accuracy of 28%. The confusion matrix indicates reasonable performance in the 'notumor' class but significant misclassifications in the other tumour classes. The learning curves suggest a slight overfitting, with a training accuracy reaching 84% and a validation accuracy around 82%. The classification report confirms that the model struggles with 'glioma', 'meningioma', and 'pituitary' classes, evidenced by low precision, recall, and F1-scores. Benchmark Two implemented enhanced data augmentation and class weighting to address class imbalances. The overall accuracy decreased to 24%, indicating a deterioration in performance. Despite this, the learning curves show a slight improvement in generalisation, as evidenced by the smaller gap between training and validation accuracies. However, the confusion matrix and classification report reveal persistent issues, particularly with the 'glioma' and 'meningioma' classes, which experienced further declines in performance metrics. Benchmark Three utilised a custom hybrid approach combining supervised learning with reinforcement learning elements. Despite an innovative methodology, the overall accuracy of 31% shows only marginal improvement over Benchmark One. The confusion matrix highlights a critical flaw, with the model predominantly misclassifying samples into the 'notumor' category. The training and validation metrics remained largely stagnant, reflecting the model's inability to adapt effectively. The hyperparameter tuning process identified optimal values, yet the improvements were insufficient to significantly enhance overall performance.

C. Comparison

Comparing the three benchmarks, it is evident that Benchmark One, despite its limitations, provides a baseline performance that subsequent models struggled to surpass significantly. Benchmark Two's enhancements in data augmentation and class weighting did not translate into improved accuracy, instead highlighting the challenges in addressing class imbalances effectively. Benchmark Three, while innovative in incorporating reinforcement learning, failed to leverage this approach to achieve a notable performance boost, demonstrating that the integration of reinforcement learning requires more sophisticated implementation and tuning.

D. Conclusion

In conclusion, Benchmark One remains the best-performing model in terms of accuracy, serving as a reference point for future improvements. The comparative analysis underscores the complexity of brain tumour classification, where even advanced techniques such as reinforcement learning require meticulous implementation and tuning to be effective. Future efforts should focus on refining model architectures, enhancing data preprocessing, and exploring more robust reinforcement learning strategies to achieve better generalisation and classification accuracy across all tumour types.

VII. Acknowledgment

I would like to express my sincere gratitude to Ella Vithanage and Abdulaziz Alqulayti for their invaluable collaboration and support throughout this project. Their insights and contributions have been instrumental in its success. I am profoundly grateful to Nathaneal Baisa for his inspiring and interactive teaching style, which has been a constant source of motivation and inspiration for this work. Additionally, I extend my heartfelt thanks to all the teachers at De Montfort University who teach on the Artificial Intelligence MSc. Their dedication and thought-provoking assignments have significantly enriched my academic journey and contributed to my growth as a researcher.

References

[1] Msoud Nickparvar. Brain tumor mri dataset, 2021.
[2] Papers With Code. Papers with code - brain tumor mri dataset benchmark (classification), 2022.
[3] Sudhakar Tummala, Seifedine Kadry, Syed Ahmad Chan Bukhari, and Hafiz Tayyab Rauf. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Current Oncology*, 29:7498–7511, 10 2022.
[4] Saravanan Srinivasan, Divya Francis, Sandeep Kumar Mathivanan, Hariharan Rajadurai, Basu Dev Shivahare, and Mohd Asif Shah. A hybrid deep cnn model for brain tumor image multi-classification. *BMC Medical Imaging (Online)*, 24, 01 2024.
[5] Muhammad Sami Ullah, Muhammad Attique Khan, Anum Masood, Olfa Mzoughi, Oumaima Saidani, and Nazik Alturki. Brain tumor classification from mri scans: a framework of hybrid deep learning model with bayesian optimization and quantum theory-based marine predator algorithm. *Frontiers in oncology*, 14, 02 2024.
[6] Kameswara Rao Pedada, Bhujanga Rao A., Kiran Kumar Patro, Jaya Prakash Allam, Mona M. Jamjoom, and Nagwan Abdel Samee. A novel approach for brain tumour detection using deep learning based technique. *Biomedical Signal Processing and Control*, 82:104549, 04 2023.
[7] Chenjie Ge, Irene Yu-Hua Gu, Asgeir Store Jakola, and Jie Yang. Deep semi-supervised learning for brain tumor classification. *BMC Medical Imaging*, 20, 07 2020.
[8] A. Inés, C. Domínguez, J. Heras, E. Mata, and V. Pascual. Biomedical image classification made easier thanks to transfer and semi-supervised learning. *Computer Methods and Programs in Biomedicine*, 198:105782, 01 2021.
[9] Veronika Cheplygina, Marleen de Bruijne, and Josien P.W. Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 05 2019.
[10] Zhengyu Yu, Qinghu He, Jichang Yang, and Min Luo. A supervised ml applied classification model for brain tumors mri. *Frontiers in Pharmacology*, 13, 04 2022.
[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 04 2015.
[12] Novsheena Rasool and Javaid IQBAL Bhat. Unveiling the complexity of medical imaging through deep learning approaches. *Chaos Theory and Applications*, 5:267–280, 2023.
[13] Google. Google colaboratory, 2019.