Classification Brain Tumor in Hyperparameter-Optimization of VGG-16 Model and Data Augmentation Analysis

Putu Desiana Wulaning Ayu[1*], I Gede Teguh Satya Dharma [2], I Wayan Rizky Wijaya[3], Made Agus Oka Gunawan[4], Ni Putu Eka Apriyanthi[5], Civica Moehaimin Dhewanty[6]

[1,2,3,4,5,6] *Information Technology Department, Politeknik Negeri Bali*

*corr-author: wulaning_ayu@pnb.ac.id

Abstract - This Advancements in computational technology have driven the development of Deep Learning, particularly Convolutional Neural Networks (CNN), in the classification and recognition of digital images. This research focuses on the classification of MRI brain tumor images using the VGG-16 architecture. The primary challenges include gradient vanishing and overfitting due to a small dataset. The objective of the study is to evaluate the performance of the model with various data augmentation techniques and to assess the impact of different dataset compositions (90:10 and 70:30) for training and testing. Two model configurations are used: Model A with 4096 neurons and Model B with 128 and 64 neurons in the first two Dense layers, respectively. The tested augmentation techniques include rotation, flip, Zoom , and their combinations. The results indicate that rotation and Zoom augmentations provide the best performance for both models and dataset compositions. Model A (90:10) achieved an accuracy of 96% with rotation and 92% with Zoom, while Model B (90:10) achieved 94% with rotation and 98% with Zoom. For the 70:30 composition, Model A achieved 94% (rotation) and 90% (Zoom), while Model B achieved 95% (rotation) and 96% (Zoom). This research provides valuable insights into optimizing VGG-16 architecture for brain tumor classification using limited datasets.

Keywords: Deep learning; VGG-16; data augmented; classification.

I. INTRODUCTION

In the medical field, the ability of CNNs to analyse and interpret medical images has paved the way for innovations in diagnostic practices, such as detecting anomalies in radiographs, segmenting tissues in MRI scans, and even predicting patient outcomes based on imaging data [1,2]. These technology advancements have contributed to diagnose for the highest accuracy and precision for medical professionals, enabling more timely and effective patient care[3]. Among the various CNN architectures, VGG-16 has emerged as a leading model due to its depth and simplicity [4,5]. Developed by the Visual Geometry Group at the University of Oxford, VGG-16 comprises 16 weight layers, including 13 convolutional layers and 3 fully connected layers [6, 7]. This model gained prominence for its superior performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014, where it significantly outperformed other models [8]. The appeal of VGG-16 lies in its simple architecture that uses small 3x3 convolutional filters and is trained on large datasets, allowing the model to effectively capture complex features from images. Further studies [9, 10] have shown that using weight initialization techniques like He uniform and dropout in deep learning layers can enhance the performance of VGG-16 by mitigating issues such as vanishing gradients and overfitting. These improvements make VGG-16 a powerful choice for various image classification tasks, including medical image processing, where precision and reliability are crucial. Given its proven efficacy and adaptability, VGG-16 was selected for this research to classify brain tumours using MRI images, aiming to leverage its deep feature extraction capabilities to enhance diagnostic accuracy [11-13].

However, despite the extensive use of CNN-based approaches for brain tumour MRI classification, several limitations remain in existing studies. First, many works focus primarily on reporting overall accuracy, while clinically important metrics such as sensitivity, specificity, and ROC-AUC are often under-reported, limiting interpretability for medical diagnosis [1, 4]. Second, augmentation strategies are commonly adopted as standard practice, but only few studies conduct a structured comparison to quantify which transformations truly improve model generalisation for multi-class MRI variability [16, 14]. Third, hyperparameter and initialization settings are frequently chosen in ad ad-hoc manner or reported without a clear optimization

protocol, making result difficult to reproduce and weakening the methodological contribution. These gaps motivate the need for a systematic investigation of how augmentation design and hyperparameter optimization jointly affect the robustness and clinical relevance of VGG-16 based tumour classification [15, 16].

Therefore, this study adapts VGG-16 specially for brain tumour classification using MRI images, targeting gliomas, meningiomas, pituitary tumours, and non-tumoral cases. Our work is distinguished from prior CNN-based MRI classification studies by two main contributions. First, we perform a structured evaluation of multiple data augmentation strategies to identify transformation that most effectively improve VGG-16 generalisation for brain MRI heterogeneity. Second, we explicitly optimize key hyperparameters-including dropout rates and He uniform weight initialization under a clearly defined experimental protocol to determine the most effective configuration for multi-class tumour recognition. Through this framework, we aim not only to validate VGG-16 effectiveness for brain tumour MRI classification, but also to offer reproducible insights into the role of augmentation and hyperparameter optimization in improving diagnostic robustness and reliability [17,18,19].

II. METHOD

This research aims to achieve the best-performing VGG-16 model for classifying MRI brain tumour images and evaluate its performance. The VGG-16 architecture was chosen due to its effectiveness in image classification tasks. The model is implemented using the Keras framework and tested with various types of augmentation techniques such as rotation, flip, zoom and all combines. Proposed model in this research showed in Fig.1.

A. Dataset

This study utilizes a dataset of MRI images that have been classified and labelled into four different tumour categories glioma, meningioma, pituitary, and no tumour. The dataset is divided into training and testing data. The choice of this dataset is due to its open access and suitability for research, scientific development, and educational purposes. It can be accessed through the official Kaggle website using the following link https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri. This public resource provides valuable tools for researchers and students to advance brain tumour classification studies and develop machine learning models in medical imaging. The sample dataset shown in Fig.1.

The composition used in this study is divided into two types: 90% training set and 10% testing set (original dataset composition) and 70% training set and 30% testing set (new dataset composition), as shown in Table I. The purpose of using these different compositions is to evaluate the impact of data distribution on model performance during training and testing, and to determine whether the model recognizes patterns better with more training data or with more testing data for validation. Experiments on the dataset were conducted by combining the training and testing datasets according to their classes, then redistributing them into the new training and testing sets, as shown in Table I. To address class imbalance and improve generalization, we applied data augmentation to the training dataset, and all the distribution showed in Table I. Data augmentation was applied only on the training model, and the techniques include flip, rotate, zoom, and combination (combines the three techniques to produce a broader range of data variations).

B. Model implementation

The base model employed in this study is the VGG-16 architecture, and data augmentation technique were applied. Two variations of the model were developed by adding data augmentation processes and modifying the number of neurons in the first and second dense layers. Model A maintains 4096 neurons in the first two dense layers, as in the original VGG-16 architecture, showed in Table II and III. This large number of neurons allows the model to learn highly complex representations from the data. Model Reduces the number of neurons to 128 in the first dense layer and 64 in the second dense layer. This reduction aims to decrease the model's complexity and test whether a simpler model can achieve comparable or better performance. The A model with fewer neurons may be more efficient in terms of computation time and memory and can help avoid overfitting by forcing the model to learn more general representations. The output layer is adjusted for the classification of four brain tumour classes using the SoftMax activation function. This adjustment allows the model to generate probabilities for each class, which are then used to determine the predicted class. This modification is crucial for the model to correctly process the input and produce relevant output for the multi-class classification task.

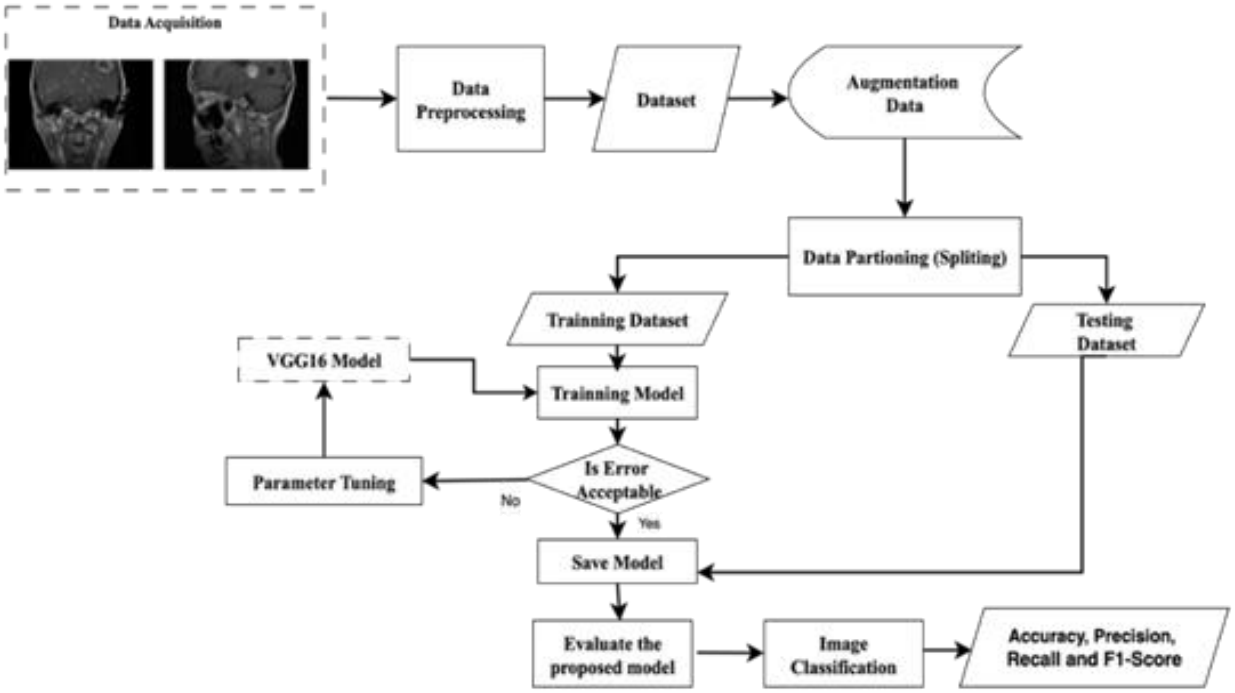


Fig. 1 Flow Diagram proposed model

TABLE I
DATASET DISTRIBUTION

Tumour	Dataset Distribution 90:10		Dataset Distribution 70:30	
	Training (after augmentation)	Testing (original dataset)	Training (after augmentation)	Testing (original dataset)
Glioma	826	101	659	282
Meningioma	822	115	655	268
Pituitary	827	74	638	263
Normal	395	105	333	167
Sub total	2870	395	2285	980
Total	**3265**		**3265**	

C. Training process

The images are resized to 224x224 pixels with 3 colour channels (RGB) to meet the input requirements of the VGG-16 architecture[20,21]. This size is selected as it aligns with standard dimensions commonly used in CNN architectures, enabling the model to effectively capture important detail form the images. To enrich data variation and improve the model's generalization ability, several data augmentation techniques are applied. The training processes for both Model A and Model B (Table III) will be conducted using each augmentation technique as shown in Table IV.

Rotation: images are randomly rotated within a certain range to add variation in the orientation of the images seen by the model. Flip: images are flipped horizontally or vertically to simulate positional variations that may occur in real data. Zoom: images are randomly zoomed to simulate variations in the distance from which the images were taken. Combination: combines the three techniques to produce a broader range of data variations, which is expected to help the model learn more robust representations. In this study, both Model A and Model B use the same hyperparameter settings to ensure that any differences in the results between the two models can be accurately attributed to architectural changes and data augmentation techniques. The hyperparameters used in the model implementation are as follows in the Table V. The Training process uses the Adam optimizer with a learning rate of 10^{-3}. This optimizer was chosen for its ability to dynamically adjust

the learning rate and its efficiency in achieving convergence. The model is trained for 50 epochs with a batch size of 32. The substantial number of epochs is expected to help the model learn good representations from the training data. The chosen batch size allows for efficient memory usage during training while providing enough updates for stable convergence.

D. Evaluation

Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score [5,22]. These metrics were chosen to provide a comprehensive overview of the model's ability to accurately classify brain tumour types. Experimental data were analysed to identify significant trends and patterns, and to provide insights into the impact of various hyperparameters on model performance.

III. RESULT AND DISCUSSION

The experimental results obtained using the VGG-16 model on the General model showed in Table VI, indicate substantial improvements when different dataset compositions are employed. As observed in Table VI, a detailed analysis of the General model trained with the original dataset (90:10) shows an accuracy of only 28%. However, when using the new dataset composition (70:30), the model's performance significantly improves, achieving an accuracy of 98%. This demonstrates the considerable impact of dataset composition on the overall performance of the VGG-16 model in brain tumour classification tasks.

TABLE II
VGG-16 MODEL CONFIGURATION

Configures Model 16 Weight Layer (VGG-16)		
Base Model	**Model A**	**Model B**
-	Data Augmentation	Data Augmentation
Input (224 X 224 RGB image)		
Conv-64 3x3	Conv-64 3x3 (He uniform, L2)	Conv-64 3x3 (He uniform, L2)
Conv-64 3x3	Conv-64 3x3 (He uniform, L2)	Conv-64 3x3 (He uniform, L2)
MaxPooling		
Conv-128 3x3	Conv-128 3x3 (He uniform, L2)	Conv-128 3x3 (He uniform, L2)
Conv-128 3x3	Conv-128 3x3 (He uniform, L2)	Conv-128 3x3 (He uniform, L2)
MaxPooling		
Conv-256 3x3	Conv-256 3x3 (He uniform, L2)	Conv-256 3x3 (He uniform, L2)
Conv-256 3x3	Conv-256 3x3 (He uniform, L2)	Conv-256 3x3 (He uniform, L2)
Conv-256 3x3	Conv-256 3x3 (He uniform, L2)	Conv-256 3x3 (He uniform, L2)
MaxPooling		
Conv-512 3x3	Conv-512 3x3 (He uniform, L2)	Conv-512 3x3 (He uniform, L2)
Conv-512 3x3	Conv-512 3x3 (He uniform, L2)	Conv-512 3x3 (He uniform, L2)
Conv-512 3x3	Conv-512 3x3 (He uniform, L2)	Conv-512 3x3 (He uniform, L2)
MaxPooling		
Conv-512 3x3	Conv-512 3x3 (He uniform, L2)	Conv-512 3x3 (He uniform, L2)
Conv-512 3x3	Conv-512 3x3 (He uniform, L2)	Conv-512 3x3 (He uniform, L2)
Conv-512 3x3	Conv-512 3x3 (He uniform, L2)	Conv-512 3x3 (He uniform, L2)
MaxPooling		
FC-4096	FC-4096 (He uniform, L2) Dropout (0,2)	FC-128 (He uniform, L2) Dropout (0,2)
FC-4096	FC-4096 (He uniform, L2) Dropout (0,2)	FC-64 (He uniform, L2) Dropout (0,2)
FC-4	FC-4	FC-4

TABLE III
NUMBER OF TRAINING PARAMETERS (IN MILLIONS)

Configuration	General	A	B
Number of training parameters	134	134	17

In the scenario test, where the original dataset is composed of 90% training and 10% for testing data, the model exhibits a higher risk of overfitting and demonstrate limited learning effectiveness from the training data. As shown in Fig. 2, accuracy does not improve beyond the second epoch but rather fluctuates due to the abundance of training data and the scarcity of testing data. This situation can cause the model to perform very well on the training data but poorly on new data. By reducing the training data to 70% (Table III) and increasing the testing data to 30%, the model can mitigate overfitting, though fluctuations in the validation loss still indicate the presence of overfitting in the testing data, while also enhancing generalization as seen in Fig. 3. To provide a visual representation of the performance of the General model during the training process, the following graph illustrates the changes in accuracy and loss for both training and validation data as the number of epochs increases. This graph aims to aid in understanding how the model learns from the training data and how well it generalizes to the validation data.

A. Model A And B

Model A and Model B demonstrated good performance during model training when using specific hyperparameter settings. These models applied a dropout rate of 0.2 on the first two dense layers, which helped reduce overfitting by randomly ignoring 20% of the neurons during training. Additionally, the models implemented weight decay using L2 regularization with a value of 5 x 10^{-4} (0.0005), which served to penalize large weights and prevent the models from becoming overly complex.

For weight initialization, these models adopted He uniform initialization on each convolutional layer and the first two dense layers. The optimizer used for training was the Adam optimizer, with a learning rate set to 10^{-3}. During the training process, the learning rate was not reduced because the models already showed good performance on both training and testing data. Training was conducted for 50 epochs, and with each epoch, the models exhibited a trend of continuously improving accuracy and decreasing loss or error. This indicated that the models were able to effectively learn the patterns in the training and testing data during the training process. These settings proved to enhance the models' ability to recognize important patterns and features from the training data, resulting in better classification performance. Our comparison model A dan B for Original Dataset, as shown in Table VII.

Based on the results from Table VII, the experiments and model analysis conducted, it was found that Model B exhibited superior performance compared to Model A in terms of accuracy. Model B achieved its highest accuracy of 98% using the zoom augmentation technique with a 90:10 dataset composition. In contrast, Model A achieved its best accuracy with the rotation augmentation technique, reaching 96% with the same 90:10 dataset composition. Both models showed unsatisfactory performance with the combination augmentation technique, where Model A only achieved 77% accuracy and Model B achieved 78% accuracy, both with a 90:10 dataset composition. Model B was more efficient in the training process and achieved competitive results compared to Model A, which had more parameters. The zoom augmentation technique provided consistent superior results for Model B compared to Model A across both dataset compositions (Fig. 4 and 5).

TABLE II
IMPLEMENTATION OF AUGMENTATION IN THE MODEL

Dataset Composition	Data Augmentation			
	Flip	Rotate	Zoom	Combination
Original	Model A and B	Model A and B	Model A and B	Model A and B
New	Model A and B	Model A and B	Model A and B	Model A and B

TABLE V
HYPERPARAMETER CONFIGURATION

Hyperparameter	Used
Epoch	50
Batch size	32
Optimizer	Adam
Learning rate	10^{-3}
Activation Function	ReLu
Loss function	Categorical Cross-Entropy
Output layer	SoftMax

TABLE VI
GENERAL MODEL TRAINING RESULTS USING THE ORIGINAL DATASET AND THE NEW DATASET

Evaluation	General Model Dataset (90:10)	General Model Dataset (70:30)
Training Accuracy	0.28	**0.98**
Validation Accuracy	0.18	**0.87**
Training Loss	1.34	**0.04**
Validation Loss	1.44	**1.14**
Precision	0.03	**0.87**
Recall	0.18	**0.87**
F1 Score	0.05	**0.87**

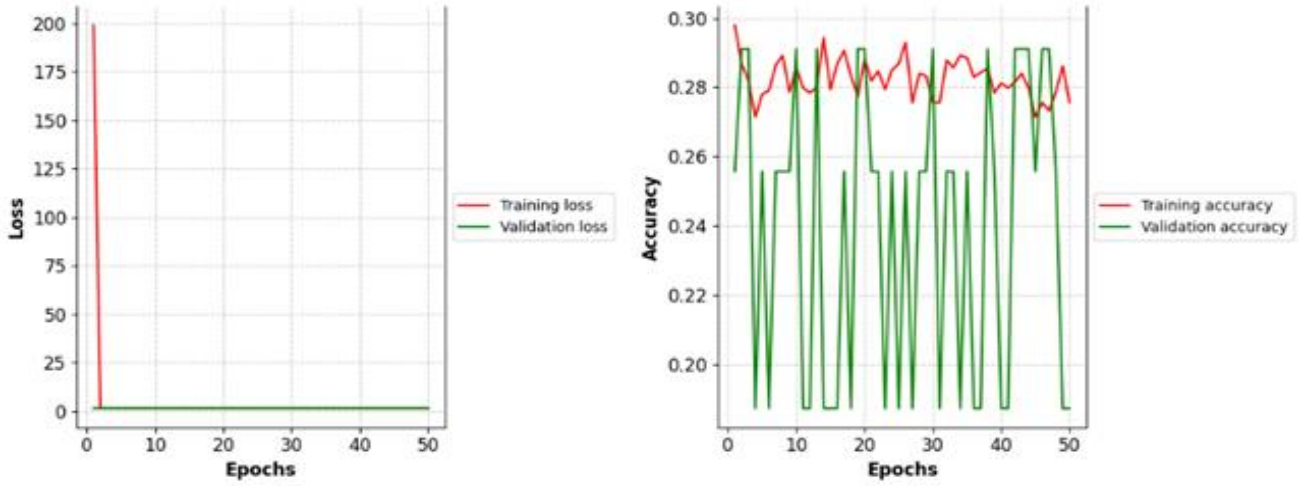


Fig. 2 Accuracy and loss of the general model with the original dataset composition (90:10) during the training process

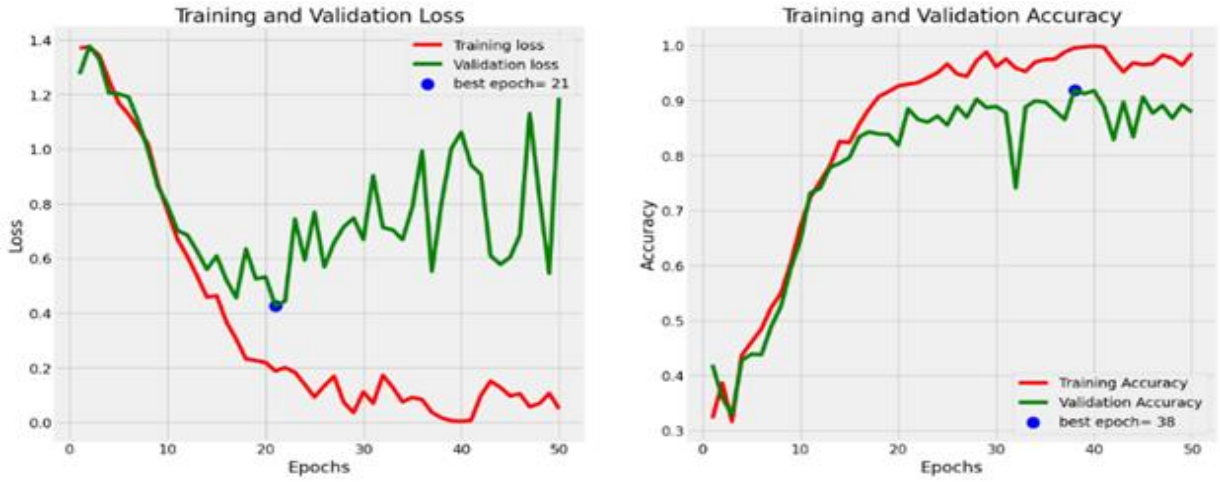


Fig. 3 Accuracy and loss of the general model with the new dataset composition (70:30) during the training process

TABLE II
TRAINING AND EVALUATION RESULTS OF MODEL A AND B USING THE ORIGINAL DATASET

Evaluation	Model A (Original Dataset) 90:10				Model B (Original Dataset) 90:10			
	Rotate	Flip	Zoom	Combination	Rotate	Flip	Zoom	Combination
Training Accuracy	**0.96**	0.91	0.92	0.77	0.94	0.92	**0.98**	0.78
Validation Accuracy	0.68	0.62	**0.69**	0.41	0.69	0.62	**0.71**	0.42
Training Loss	3.59	3.81	**3.02**	4.00	1.30	1.42	**0.86**	1.16
Validation Loss	5.72	5.96	**4.87**	6.13	3.84	3.38	4.46	**2.95**
Precision	0.71	**0.72**	0.63	0.55	0.72	0.66	**0.76**	0.48
Recall	**0.68**	0.63	0.66	0.44	0.69	0.60	**0.71**	0.41
F1 Score	**0.83**	0.59	0.61	0.40	0.65	0.57	**0.67**	0.38

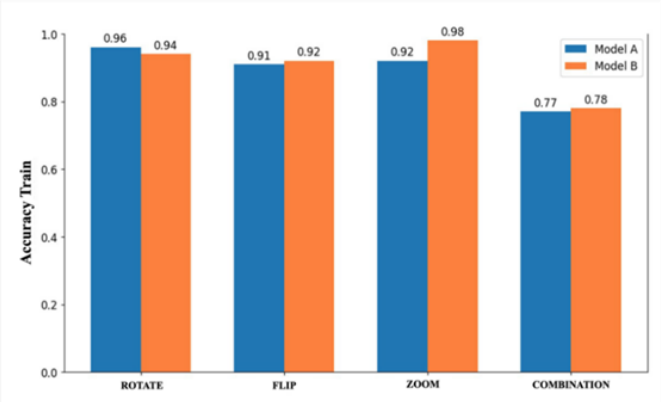


Fig. 4 Accuracy results of models a and b with the original dataset composition of 90:10

The use of the new dataset with a 70:30 composition provided overall better results compared to the original 90:10 dataset. This is evident from the decrease in loss values and the increase in accuracy in both models. The new dataset (augmentation) (70:30) helped reduce overfitting, as the models were able to learn better with a more balanced proportion of training and validation data. This is reflected in the training accuracy and loss graphs for models A and B, where the validation loss and accuracy were closer to the training loss and accuracy, as shown in Table VIII. Additionally, the classification evaluation results for both models A and B showed higher and more consistent precision, recall, and F1 scores across all augmentation techniques compared to the original dataset composition (90:10). As a limitation, this study employs a stratified hold-out split rather than K-fold cross validation. This choice was made to maintain a fixed data distribution for a controlled comparison of augmentation and hyperparameter configurations. Future work will incorporate K-fold cross validation or repeated stratified experiment to provide more robust generalization estimates across diverse data partitions.

TABLE VIII
TRAINING AND EVALUATION RESULTS OF MODEL A AND B USING THE NEW DATASET

Evaluation	Model A New Dataset 70:30				Model B New Dataset 70:30			
	Rotation	Flip	Zoom	Combination	Rotation	Flip	Zoom	Combination
Training Accuracy	**0.94**	0.88	0.90	0.61	0.95	0.88	**0.96**	0.76
Validation Accuracy	**0.86**	0.83	0.82	0.60	0.86	0.81	**0.86**	0.73
Training Loss	3.80	4.10	**3.62**	5.45	1.69	1.58	**1.50**	1.81
Validation Loss	4.06	4.25	**3.94**	5.49	2.02	**1.78**	1.86	1.87
Precision	**0.86**	0.82	0.82	0.59	0.87	0.82	**0.88**	0.75
Recall	**0.87**	0.82	0.82	0.58	**0.87**	0.80	**0.87**	0.74
F1 Score	**0.86**	0.82	0.81	0.58	**0.87**	0.80	**0.87**	0.74

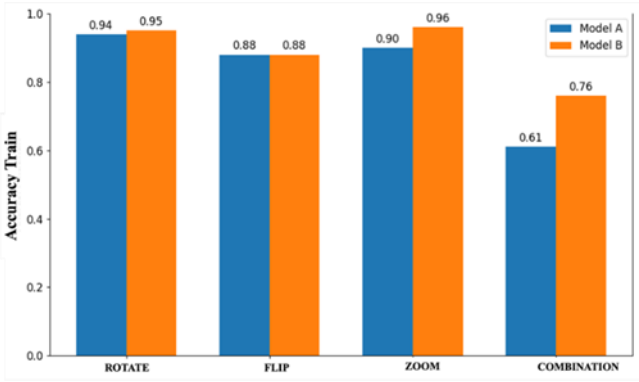
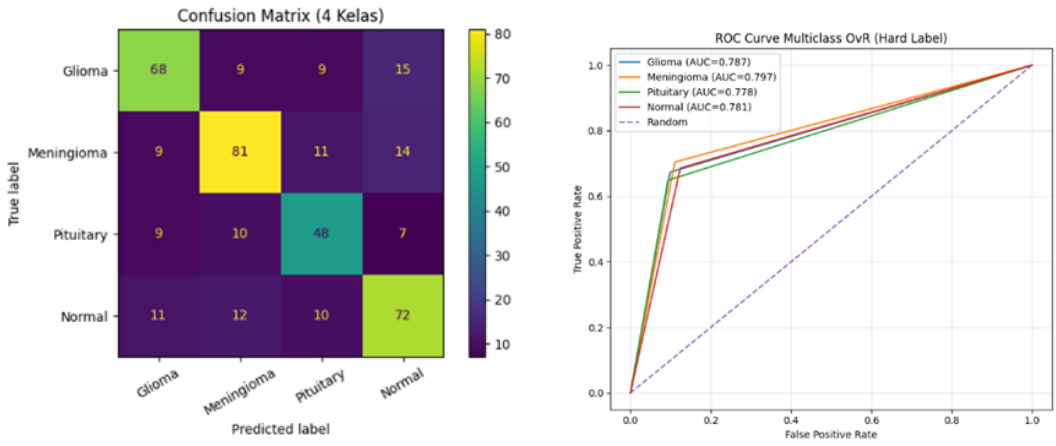


Fig. 5 Accuracy results of models a and b with the new dataset composition of 70:30

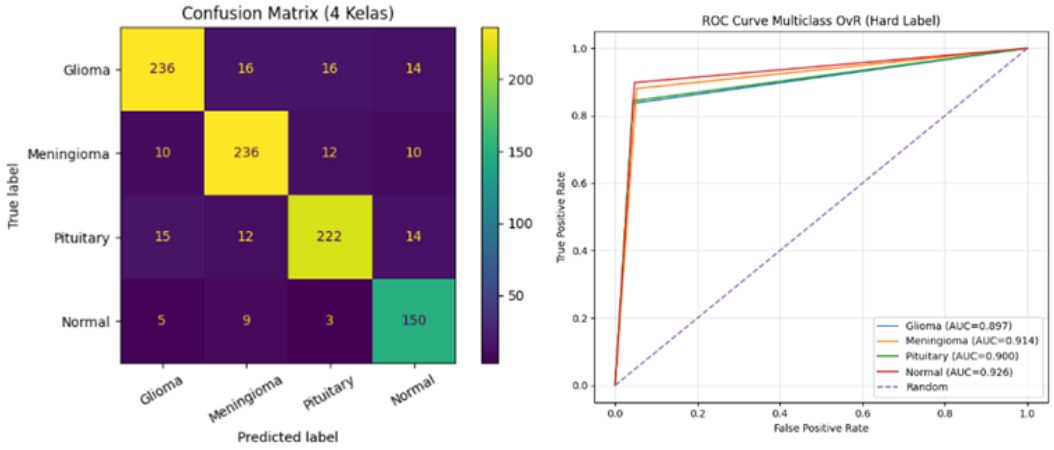
Model A demonstrates the performance of a four-class classifier (glioma, meningioma, pituitary tumour, and normal) evaluated on a 90:10 test split comprising 395 samples. The model correctly classifies 269 cases, resulting in an overall accuracy of approximately 68.1%. Class-wise recall is relatively uniform but moderate: glioma achieves 68/101 (0.67), meningioma 81/115 (0.70; the highest), pituitary 48/74 (0.65; the lowest), and normal 72/105 (0.69), indicating the absence of severe class bias but also suggesting limited discriminative learning across categories. Precision exhibits a comparable trend: glioma attains 68/97 (0.70), meningioma 81/112 (0.72; the highest), pituitary 48/78 (0.62; the lowest), and normal 72/108 (0.67), implying that pituitary predictions are the least reliable and are frequently contaminated by other classes. The most prominent error for glioma is misclassification as normal (15 cases), which may reflect subtle or low-contrast tumour features resembling healthy tissue. Meningioma is primarily confused with normal (14 cases) and pituitary (11 cases), suggesting overlap in appearance or indistinct lesion boundaries. Pituitary tumours are often misidentified as meningioma (10 cases) and glioma (9 cases), highlighting shared structural or textural characteristics among tumour types. Normal images are misclassified across all tumour categories (11 glioma, 12 meningioma, and 10 pituitary), indicating that normal anatomical variations may be erroneously interpreted as pathological patterns. Overall, the misclassifications are distributed across classes rather than concentrated in a single category, suggesting that the main limitation arises from inter-class similarity, potentially compounded by class imbalance (normal samples remain fewer than each tumour class even after augmentation) and intrinsic visual overlap in MRI tumour morphology. The confusion matrix and ROC curves model A are presented in Fig. 6.

Model B demonstrates strong discriminatory performance with an overall accuracy of 86.1 percent, corresponding to 844 correct predictions out of 980 test samples. The diagonal elements dominate, indicating that most instances are correctly assigned across all tumour categories and the normal class. Class wise sensitivity or recall is high and well balanced. Glioma achieves 83.7 percent with 236 correctly identified out of 282 samples. Meningioma reaches 88.1 percent with 236 out of 268. Pituitary attains 84.4 percent with 222 out of 263. Normal is the highest at 89.8 percent with 150 out of 167. These results indicate effective detection of both pathological and non-pathological cases, with particularly robust recognition of meningioma and normal images. Precision values confirm reliable positive predictions for tumour classes. Glioma precision is 88.7 percent based on 236 correct glioma predictions among 266 predicted glioma cases. Meningioma precision is 86.4 percent from 236 correct among 273 predicted meningioma cases. Pituitary precision is 87.7 percent from 222 correct among 253 predicted pituitary cases. Normal precision is lower at 79.8 percent, reflecting 150 correct normal predictions among 188 predicted normal cases. This reduction in normal precision suggests that the model tends to over assign the normal label relative to its true frequency. Despite this, the F1 scores remain consistently high, namely 86.1 percent for glioma, 87.2 percent for meningioma, 86.0 percent for pituitary, and 84.5 percent for normal, indicating a stable trade-off between precision and recall. The distribution of errors reveals clinically plausible inter class overlap among tumour subtypes. Glioma is most often confused with meningioma and pituitary, with 16 cases each, suggesting shared imaging characteristics. Meningioma errors are fewer and mainly shift to pituitary with 12 cases, and to glioma or normal with 10 cases each. Pituitary misclassifications are spread across glioma with 15 cases, meningioma with 12 cases, and normal with 14 cases. Importantly, confusion between tumour and normal categories is limited. Only 38 tumour cases are predicted as normal in total, while only 17 normal cases are predicted as tumours. This indicates that the model maintains strong separation between abnormal and normal brain MRI, and that remaining errors predominantly reflect similarity among tumour types rather than ambiguity between tumours and normal tissue. The confusion matrix and ROC curves model B are presented in Fig. 7.



(a) **(b)**

Fig. 6 (a) Confusion matrix and (b) ROC curve of model A



(a) **(b)**

Fig. 6 (a) Confusion matrix and (b) ROC curve of model B

IV. CONCLUSION

The training and evaluation outcomes of Model A and B, across both 90:10 and 70:30 dataset splits, indicates that the zoom augmentation technique consistently yields the highest performance for both models. Model A, with zoom augmentation and a 90:10 dataset composition, achieved a training accuracy of 92% and a test accuracy of 69%, while with a 70:30 dataset composition, it reached a training accuracy of 90% and a test accuracy of 82%. Model B also showed good performance with zoom augmentation, achieving a test accuracy of 98% with the 90:10 dataset and 96% with the 70:30 dataset. Zoom augmentation is more consistent with the real clinical variability present in brain MRI, especially changes in tumour size, local field-of-view, and slight difference in positioning between scans. By slightly zooming in and out, the model learns to be more robust to variations in lesion scale and coverage of the brain region, which enhances its ability to localise and classify tumour of different size. In contrast, the combination augmentation technique yielded the worst results for both models and dataset compositions, with test accuracies as low as 78% for Model B with the 90:10 dataset and 61% for Model A with the 70:30 dataset, and it demonstrated highly unstable loss values. In conclusion, zoom augmentation is the best choice to enhance model performance in brain tumour classification, particularly with the 70:30 dataset composition, with hyperparameter configuration in Table V, which shows better generalization parameter

evaluation. On the other hand, the combination augmentation technique should be avoided as it tends to degrade model accuracy and stability. However, this study has a several limitations, the experiment was conducted using a single stratified hold-out split of the augmented dataset, without K-fold cross-validation or external validation on independent cohorts, which may limit the generalisability of the reported performance. Future work will address these limitations by incorporating K-fold cross-validation and repeated experimental runs, as well as external validation on multi-centre datasets to obtain more robust generalisation estimates. In addition, extending the comparison to other modern architectures such as ResNet, DenseNet, Efficient Net, and ensemble methods may further improve predictive performance.

REFERENCES

[1] S. Agarwal, D. Jain, S. Gupta, S. Sawhney, and Y. Mittal, "Brain Tumor Detection and Classification Using Deep Learning," *Proceedings - IEEE 2023 5th International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2023*, pp. 635–640, 2023, doi: 10.1109/ICAC3N60023.2023.10541584.

[2] S. Abirami and G. K. D. P. Venkatesan, "Biomedical Signal Processing and Control Deep learning and spark architecture based intelligent brain tumor MRI image severity classification," *Biomed Signal Process Control*, vol. 76, no. February, p. 103644, 2022, doi: 10.1016/j.bspc.2022.103644.

[3] R. Vankdothu and M. Abdul, "Measurement : Sensors Brain tumor MRI images identification and classification based on the recurrent convolutional neural network," *Measurement: Sensors*, vol. 24, no. August, p. 100412, 2022, doi: 10.1016/j.measen.2022.100412.

[4] T. Loganayagi, M. Sravani, B. Maram, T. Venkata, and M. Rao, "Hybrid Deep Maxout-VGG-16 model for brain tumour detection and classification using MRI images," *J Biotechnol*, vol. 405, no. August 2024, pp. 124–138, 2025, doi: 10.1016/j.jbiotec.2025.05.009.

[5] P. D. W. Ayu, G. A. Pradipta, R. R. Huizen, E. S. W. Kadek, and I. G. E. Artana, "Combining CNN Feature Extractors and Oversampling Safe Level SMOTE to Enhance Amniotic Fluid Ultrasound Image Classification," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 1, pp. 251–262, 2024, doi: 10.22266/ijies2024.0229.24.

[6] S. Sharma and K. Guleria, "A Deep Learning based model for the Detection of Pneumonia from Chest X-Ray Images using VGG-16 and Neural Networks," *Procedia Comput Sci*, vol. 218, pp. 357–366, 2022, doi: 10.1016/j.procs.2023.01.018.

[7] A. Victor Ikechukwu, S. Murali, R. Deepu, and R. C. Shivamurthy, "ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 375–381, 2021, doi: 10.1016/j.gltp.2021.08.027.

[8] M. Tounsi, E. Aram, A. T. Azar, A. Al-Khayyat, and I. K. Ibraheem, "A Comprehensive Review on Biomedical Image Classification using Deep Learning Models," *Engineering, Technology and Applied Science Research*, vol. 15, no. 1, pp. 19538–19545, 2025, doi: 10.48084/etasr.8728.

[9] L. Sánchez-moreno, A. Perez-peña, L. Duran-lopez, and J. P. Dominguez-morales, "Ensemble-based Convolutional Neural Networks for brain tumor classification in MRI : Enhancing accuracy and interpretability using explainable AI," *Comput Biol Med*, vol. 195, no. November 2024, p. 110555, 2025, doi: 10.1016/j.compbiomed.2025.110555.

[10] M. Afroj, M. R. H. Mondal, M. R. Hassan, and S. Akter, "MobDenseNet: A hybrid deep learning model for brain tumor classification using MRI," *Array*, vol. 26, no. May, p. 100413, 2025, doi: 10.1016/j.array.2025.100413.

[11] M. A. Ali, F. Dornaika, I. Arganda-Carreras, R. Chmouri, and H. Shayeh, "Enhancing MRI brain tumor classification: A comprehensive approach integrating real-life scenario simulation and augmentation techniques," *Physica Medica*, vol. 127, no. October, p. 104841, 2024, doi: 10.1016/j.ejmp.2024.104841.

[12] K. Chiranjeevi and D. Victosudhageorge, "Brain tumor detection and classification using deep learning algorithms: A survey," *AIP Conf Proc*, vol. 3162, no. 1, pp. 1–19, 2025, doi: 10.1063/5.0241733.

[13] H. Anh and K. C. Santosh, "ScienceDirect An expert voting system for brain tumor classification using MRI images," vol. 260, pp. 316–324, 2025, doi: 10.1016/j.procs.2025.03.207.

[14] I. K. Seneng, P. D. W. Ayu, and R. R. Huizen, "Comparative Analysis of Augmentation and Filtering Methods in VGG19 and DenseNet121 for Breast Cancer Classification," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 3, pp. 1131–1146, 2025, doi: 10.52436/1.jutif.2025.6.3.4397.

[15] T.-W. Wang, Y.-C. Shiao, J.-S. Hong, W.-K. Lee, M.-S. Hsu, H.-M. Cheng, H.-C. Yang, C.-C. Lee, H.-C. Pan, W. C. You, J.-F. Lirng, W.-Y. Guo, and Y.-T. Wu, "Artificial Intelligence Detection and Segmentation Models: A Systematic Review and Meta-Analysis of Brain Tumors in Magnetic Resonance Imaging," *Mayo Clinic Proceedings: Digital Health*, vol. 2, no. 1, pp. 75–91, 2024, doi: 10.1016/j.mcpdig.2024.01.002.

[16] R. Adhitama Putra, G. Angga Pradipta, and P. Desiana Wulaning Ayu, "Gallbladder Disease Classification from Ultrasound Images Using CNN Feature Extraction and Machine Learning Optimization," *Journal of Electronics, Electromedical Engineering, and Medical*

Informatics, vol. 7, no. 4, pp. 1089–1111, 2025, doi: 10.35882/jeeemi.v7i4.1030.

[17] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J Big Data*, vol. 6, no. 1, p. 60, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[18] P. Desiana Wulaning Ayu, G. Angga Pradipta, R. Rudolf Huizen, K. W. Eka Sapta, and I. Gede Edy Artana, "Modified of Single Deepest Vertical Detection (SDVD) Algorithm for Amniotic Fluid Volume Classification," 2023. doi: 10.30595/juita.v11i2.18435.

[19] P. D. W. Ayu, G. A. Pradipta, I. M. D. Susila, D. P. Hostiadi, and M. Liandana, "Deep Learning Based Detection and Classification of Amniotic Fluid Echogenicity Type for Enhanced Prenatal Diagnosis," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 1, pp. 246–267, 2025, doi: 10.22266/ijies2025.0229.18.

[20] A. A. Aouragh and M. Bahaj, "Comparison Results of Hybrid CNN-Machine Learning Algorithms Architectures for Monkeypox Images Classification," *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology, IRASET 2023*, pp. 1–6, 2023, doi: 10.1109/IRASET57153.2023.10153062.