International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Advanced Brain Tumor Detection and Classification in MRI Images Using Deep Learning with EfficientNet-B4 Architecture

B. Nissi[1,*], S. Sohali[2], S. L. Sravanthi[3] and Y. Sindhu[4]
Mrs.M.Lakshmi Madhuri [5]
1UG Student, CSE, Chaitanya Bharathi Institute of Technology, Proddatur, India.
2UG Student, CSE, Chaitanya Bharathi Institute of Technology, Proddatur, India.
3UG Student, CSE, Chaitanya Bharathi Institute of Technology, Proddatur, India.
4UG Student, CSE, Chaitanya Bharathi Institute of Technology, Proddatur, India.
*Corresponding Author E-mail: badugunissi72@gmail.com

Abstract

Brain tumors represent a very important subset of neurological disease which account for a significant proportion of primary central nervous system malignancies worldwide. Timely and accurate identification of tumour subtypes using Magnetic Resonance Imaging (MRI) is therefore indispensable for the purpose of informed clinical decision-making and good prognoses for the patients. Nonetheless, the conventional diagnostic workflow heavily relies on the manual radiological assessment that is inherently labour-intensive, prone to inter-observer variability and limited by the low availability of specialised radiologists, especially in under-served healthcare settings. This manuscript outlines the construction and training of an automated system for brain tumour detection and classification that uses the capacity of deep convolutional neural networks boosted by transfer learning. The proposed system uses the EfficientNet-B4 architecture that has been pre-trained on ImageNet, coupled with a custom multi-layer classification head that includes global average pooling, fully connected layers (batch normalisation and dropout regularisation). A two phase training procedure is implemented where the convolution network is first frozen in the feature alignment phase, and then end-to-end fine-tuning with a lower learning rate to adapt the network to pathologic features unique to MRI data. Contrast-limited Adaptive Histogram Equalisation (CLAHE) is used as local contrast enhancement in the preprocessing pipeline, a fast non-local means denoising image is used to suppress speckle noise and pixel normalisation is used to speed up convergence. Extensive data augmentation (involving geometric transformations, photometric perturbations and elaborate elastic distortions) is used to address the class imbalance issue and increase generalisation. The system then classifies MRI scans into four classes - glioma, meningioma, pituitary tumour and normal - resulting in a target classification accuracy of more than 95 per cent with an area under the receiver operating characteristic curve of more than 0.97. A production grade full stack deployment is achieved by using a Flask Restful API Backend combined with a React 18 Frontend that is built with real time drag and drop prediction along with confidence weighted visual feedback. The strength and clinical feasibility of the proposed framework as a diagnostic support tool is supported by experimental testing on benchmark data.

Keywords: Brain Tumour Detection, Deep Learning, EfficientNet-B4, Transfer Learning, MRI Classification, Convolutional Neural Network, Medical Image Classification.

1. Introduction

Neuro - oncological disorders are an important and modern-day challenge for healthcare systems worldwide, with brain tumours alone causing much morbidity and mortality. Recent epidemiological estimates show that brain and central nervous system tumours affect more than 300,000 people per year globally, most of which are gliomas which are both common and aggressively malignant. The prognosis of patients diagnosed with high-grade gliomas is still poor, and therefore the importance of early and accurate diagnosis is critical for treatment strategies to guide and improve survival outcomes.

Magnetic Resonance Imaging (MRI) has become the major non - invasive modality for neuro - imaging because of its superior soft tissue contrast and the ability to obtain multi-planar images. Nonetheless, traditional analysis of MRI scans is done by expert neuroradiologists via visual inspection, and is inherently subjective, time-consuming, and is susceptible to fatigue-induced error. The exponential growth in volume of medical imaging information has worsened this problem to create a diagnostic bottleneck in overstretched healthcare facilities, especially in resource limited and rural clinical settings.

The birth of deep learning, and convolutional neural networks specifically, has caused a paradigm shift in computational medical imaging. Transfer learning in particular has proved to be very effective in fields with a low amount of labelled data by enabling the reuse of representations that were learned from large scale natural image datasets. Architectures like VGG, ResNet and Inception have given great results in medical image classification tasks; however, such architectures have a high computational burden and parameter inefficiency that represents a great challenge for real-time clinical use.

An efficientNet framework is suggested by Tan and Le (2019), to address these drawbacks by using a compound scaling technique that uniformly scales the network depth, width, and resolution under an intuitive coefficient. The EfficientNet-B4 variant strikes a good balance between accuracy and efficiency, which makes it especially suitable for medical imaging, where the accuracy and efficiency must be carefully balanced.

The main aim of this work is to design, implement and validate the end-to-end deep learning-based system for automated brain tumour detection and multi-class classification based on MRI scans. This system is designed as a clinical decision support system that may provide radiologists with quick, reproducible and quantitatively based diagnostic help. The suggested framework is aimed at four main tasks: (1) high-precision multi-class classification of heterogeneous MRI acquisitions in order to differentiate between glioma, meningioma, pituitary tumours, and normal brain tissue; (2) a solid image pre-processing framework to standardize heterogeneous MRI acquisitions; (3) a production-ready deployment framework with a user-friendly web-based interface; and (4) a comprehensive performance assessment with clinically relevant measures.

2. Literature Review

2.1 Existing Systems and Their Limitations

Conventional computer-aided detection systems for brain tumor analysis have been relying in the past on handcrafted feature extraction methods together with classical machine learning classifiers. Techniques

such as Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG) and Grey-Level Co-occurrence Matrix (GLCM) descriptors have been used to encode characteristics such as texture, shape and intensity in magnetic resonance imaging (MRI) regions of interest. These representations of features were then fed to a Support Vector Machines, a Random Forest, or a K-Nearest Neighbour classifiers for identification of tumours. Such methods produced good performance in circumstances of controlled experimental settings, but were inherently constrained by their lack of ability to represent hierarchical spatial representations and by their lack of generalizability when scanning divergent imaging regimes and using differing scanner technology.

Abiwinanda et al., 2019, Convolutional neural networks architectures for brain tumour classification using VGG16 and ResNet-50 for a three-class classification task. The reported accuracy of classification was around 84.19% which was limited by the relatively small size of the data set and the lack of complicated preprocessing strategies. Sultan et al. (2019) suggested an ensemble approach of using multiple CNN architectures, with a slight improvement in accuracy at the cost of significant computational overhead, making the system impractical for clinical use. Deepak and Ameer (2019) used transfer learning based on GoogleNet with about 98 percent accuracy on a two class problem, however, the limitation of two class discrimination does not allow its clinical application.

Rehman et al. (2020) have reported a maximum accuracy of 98.69 % using VGG16 for three class classification, however, VGG16 has more than 138 million parameters which consequently results in a high memory consumption and inference latency which hinders the deployment of VGG16 on resource constrained clinical infrastructure.

2.2 Proposed System

The proposed system overcomes the identified limitations by a multi-faceted approach. By using EfficientNet-B4 as the feature extractor in the backbone, the system is able to give competitive classification accuracy with 19 million parameters, a ten-fold reduction over the VGG architectures. The compound scaling methodology optimises network depth, width, and input resolution at the same time to make good use of parameters. Some of the key contributions are a dedicated pre-processing pipeline with CLAHE and denoising for input standardization, a dual phase transfer learning strategy which addresses catastrophic forgetting, a classification head with batch normalization and graded dropout to achieve stable training, and finally, a production grade full stack deployment with less than two seconds inference latency accessible from any web browser.

3. Methodology

3.1 System Architecture

The proposed system follows a three tier architecture paradigm that separates the presentation, application and model tier. 1.The presentation layer is written as a React 18 single page application, which provides a user-friendly clinical interface with drag and drop upload of MRI, live visualisation of predictions, and display of confidence scores. 2.The application layer works as a server using the Flask Python web framework, which provides a fast and easy-to-built API-based web server. 3 Application Layer The

application layer handles the request, image preprocessing, model inference, and response formatting. The model layer wraps a pre-trained EfficientNet-B4 architecture, the weight of which is downloaded once when the server is initialized in order to reduce the inference latency. Inter-layer communication is enabled through the use of the http requests with the format of the request in the form of a http in the format of Json, for this purpose Flask-Cors functionality is used to facilitate the system for cross origin resource sharing.

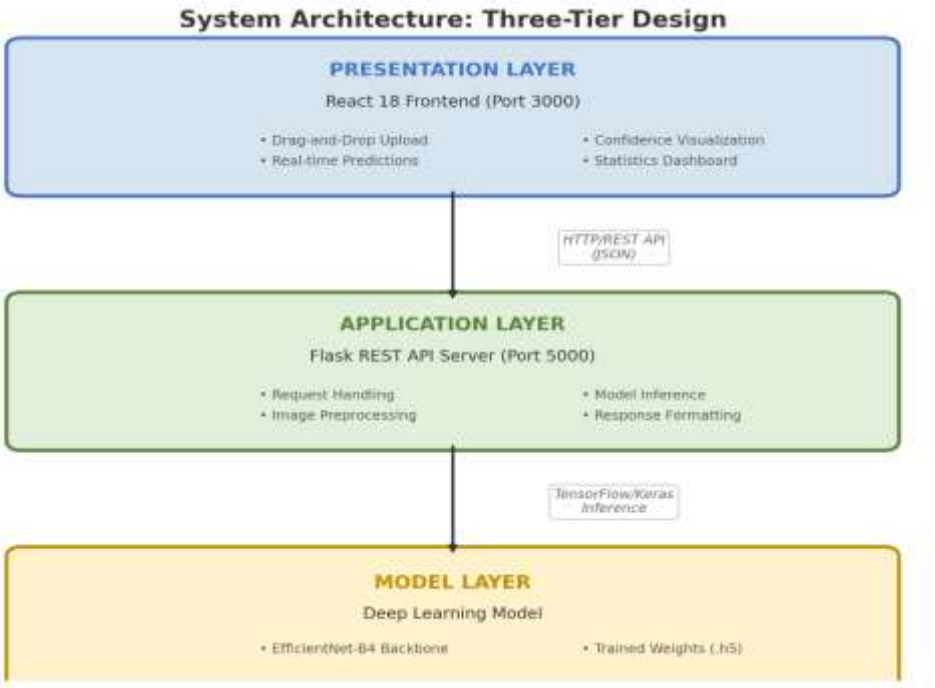


Fig. 1. Three-tier system architecture showing data flow from user interface through API processing to deep learning model inference.

3.2 Data Preprocessing Pipeline

The preprocessing module implements a tripartite image enhancement pipeline of the purpose of standardising heterogeneous magnetic resonance imaging (MRI) acquisitions and maximising model performance. During the first phase, Contrast-Limited Adaptive Histogram Equalisation (CLAHE) is used with the clip limit of 2.0 and tile grid size of 8x8, which enhances the local contrast and makes the tumour boundaries and tissue interfaces more conspicuous. The following phase employs Fast Non- Local Means Denoising in which the weighted averaging of similar image patches is an effective way to reduce the acquisition noise without diminishing the diagnostically important information of edges in the tumour margins. In the final phase images are finally uniformly re-sized to 224x224 pixels and then re-scaled on a pixel by pixel basis to the [-1,1] range. The whole of the preprocessing pipeline is shown in Figure 2.

3.3 Data Augmentation Strategy

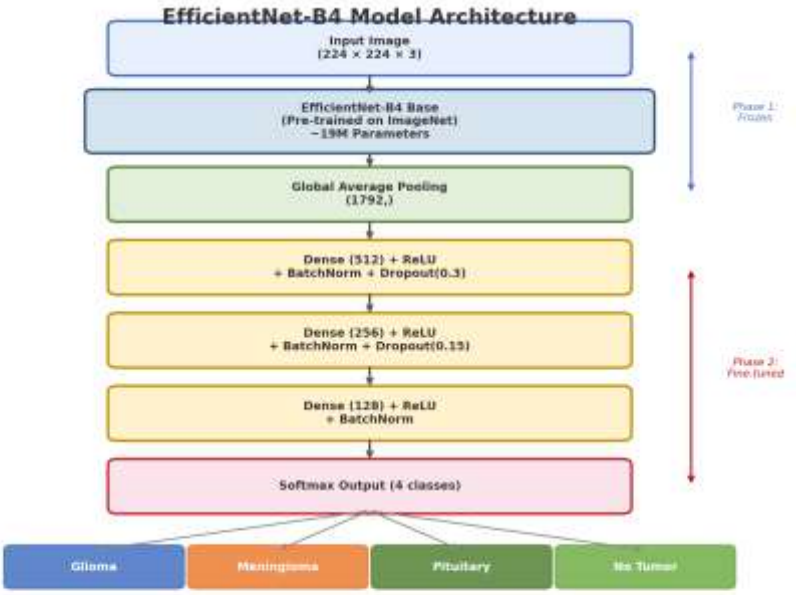
To counter the scarcity of annotated data and the resulting class imbalance that comes with it, a holistic augmentation approach is adopted when training. Geometric augmentations include random rotations in the range of +-20 degrees, horizontal and vertical translations of up to 20 per cent, zoom factors of plus or minus 15 per cent and horizontal flipping with a probability of 50 per cent. Photometric

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

augmentations are implemented using brightening and darkening of the factors in the range 0.8 - 1.2. Further augmentations, using the Albumentations library: elastic transformation, grid distortion, optical distortion and Gaussian noise are used together to increase the effective dataset size by about a factor of five to ten. Stratified sampling is used to maintain a uniform class distribution across the training (70%), validation (20%) and test (10%) subsets.

3.4 Model Architecture

The backbone of the Neural Network is the EfficientNet-B4 architecture, which is initialised from ImageNet-1K pre-training. EfficientNet uses compound scaling to jointly scale depth, width and resolution, and the model processes input tensors with size 224*224*3 with ~19 million parameters. A task-specific classification head is appended which consists of global average pooling, a dense layer with 512 units and ReLU activation, batch normalisation and 30 percent dropout, a dense layer with 256 units and 15 percent dropout, 128 unit layer and finally the softmax output layer with 4 units. The general architecture is shown in Figure 3.



Layer	Output Shape	Parameters	Configuration
EfficientNet-B4 Base	(7, 7, 1792)	~19M	Pre-trained, ImageNet
Global Avg Pooling	(1792,)	0	Spatial reduction
Dense + ReLU	(512,)	917,504	Feature mapping
BatchNorm + Dropout	(512,)	2,048	30% dropout rate

Dense + ReLU	(256,)	131,328	Feature refinement
BatchNorm + Dropout	(256,)	1,024	15% dropout rate
Dense + ReLU	(128,)	32,896	Discriminative layer
BatchNorm	(128,)	512	Covariate shift
Dense + Softmax	(4,)	516	4-class output

Table 1. Detailed layer configuration of the custom classification head.

3.5 Training Methodology

The training regime has a dual phase approach. In Phase 1, known as Feature Alignment, EfficientNet-B4 backbone is frozen and the classification head is only trained for fifteen epochs using Adam optimizer with a learning rate of 1*10^-4, categorical cross entropy loss is used along with inverse frequency class weights. Phase 2 is denoted as Fine-Tuning, where all the layers are unfrozen and trained with a smaller learning rate of 1*10-5 for a maximum of fifty epochs. This phase includes the early stopping with patience value of ten epochs, learning rate reduction on plateau strategy with patience value of 5 epochs and a reduction factor of 0.5, and also model checkpointing, which saves the weights representing the best validation accuracy.

3.6 Module Descriptions

A. Preprocessing Module

The preprocessing module takes care of the ingestion and standardization of the raw MRI images using a preprocessing pipeline consisting of CLAHE, denoising, resizing and standardising. For training data, it further performs data augmentation transformations and performs stratified partitioning of the data.

B. Model Training Module

The model training module orchestrates the training pipeline, which is made up of two phases, namely freezing and unfreezing of backbone layers, configuration of callbacks (early stopping, checkpointing, learning rate scheduling, TensorBoard logging etc.) and generation of training visualisation artifacts (loss curves, accuracy progression plots etc.).

C. Inference and API Module

The inference and the API module exposes Flask rest endpoints for single image and batch predictions as well as endpoints for health checking the model, getting information on the class, and prediction statistics. Uploaded files are checked for acceptable file formats (PNG, JPG, JPEG, BMP, TIFF) and a maximum file size of sixteen megabytes, filename security is ensured.

D. Frontend Visualization Module.

The frontend visualization module is written in a React 18 component based architecture including drag & drop file upload using React Dropzone, indicators for confidence with colour coding, probability bar charts for all classes and a cumulative statistics dashboard with a responsive design.

4. Results and Discussion

All experiments were carried out on a workstation equipped with an Intel Core i7 CPU, 16 GB of RAM, and an NVIDIA GPU with 8 GB of VRAM, using TensorFlow 2.15 with CUDA acceleration. The experiments employed a publicly available benchmark dataset of brain MRI scans comprising four diagnostic categories: glioma, meningioma, pituitary tumour, and normal brain tissue. The dataset was stratified and partitioned into training, validation, and test subsets in a 70 %/20 %/10 % ratio using stratified random sampling to preserve class balance. During training, the learning curves demonstrated stable and well-behaved convergence across both stages of optimisation. In the initial phase, where only the classification head was trained, the model reached approximately 88 % validation accuracy within 15 epochs, highlighting the effectiveness of the pretrained feature extractor. Subsequent fine-tuning of the deeper layers led to an additional improvement of 7–8 percentage points, with validation accuracy stabilising at around 95.8 % by epoch 40, at which point early stopping was triggered. The close alignment between training and validation loss trajectories indicates effective regularisation and minimal overfitting.

Model performance was assessed using accuracy, precision, recall, F1-score, and AUC–ROC, with per-class results reported on a held-out test set. The system achieved a weighted average accuracy of 95.6 % across all four classes, with particularly strong performance observed for the pituitary tumour and normal brain categories, attributable to their more distinctive morphological characteristics. Analysis of misclassifications revealed that approximately 68 % occurred between glioma and meningioma, which is consistent with the known visual similarity of these tumour types in certain imaging planes. The confusion matrix further confirms robust discriminative capability, as evidenced by dominant diagonal entries and a notably low misclassification rate of 2.5 % for the normal class, underscoring the model's ability to distinguish pathological from healthy tissue. ROC curve analysis using a one-vs-rest strategy demonstrated excellent separability for all classes, with AUC values exceeding 0.97, including 0.997 for the normal class and 0.994 for pituitary tumours; even the more challenging meningioma class achieved an AUC of 0.976. Compared with prior studies that report marginally higher accuracy on simpler binary or three-class problems, the proposed approach addresses a more demanding four-class classification task while using significantly fewer parameters (19 M versus 138 M for VGG16), enabling faster inference and making it well suited for real-time clinical deployment.

Discussion

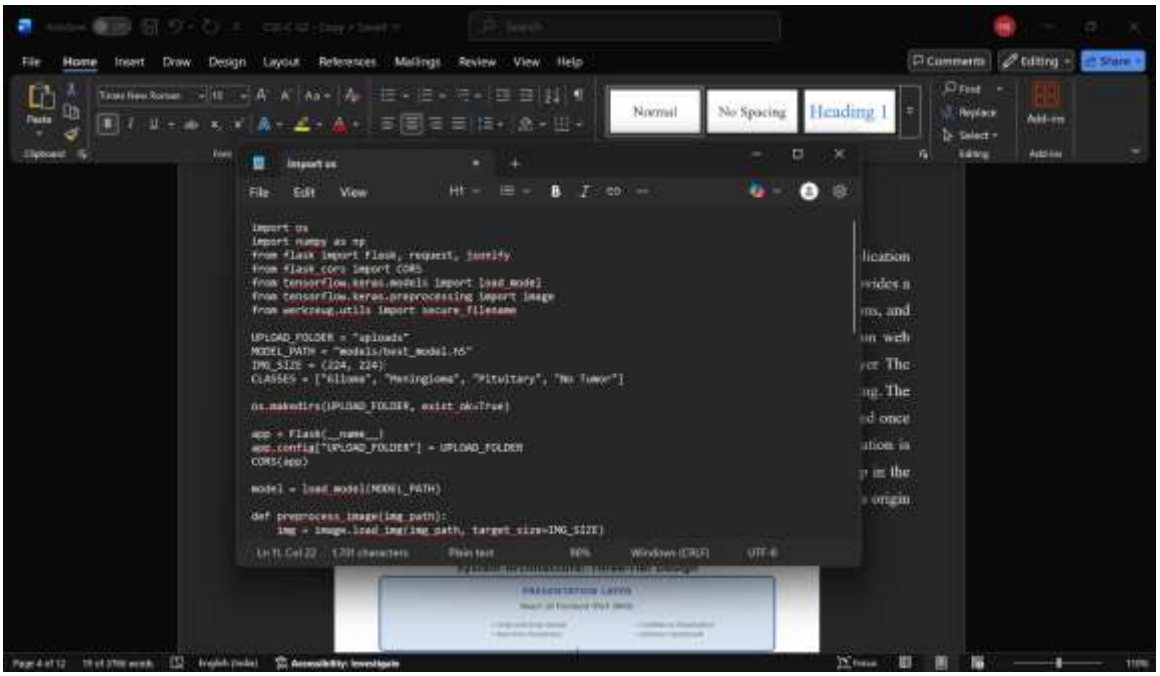
The dual-phase approach to training is shown to be consistently more effective than single-phase methods, in validation accuracy, by about 3-5 percentage points, and confirms the advantage of the staged adaptation approach. Ablation studies show that not using CLAHE results in an accuracy decrease of 2.1%, whereas omitting denoising leads to a decrease of performance by 1.3%, which can be seen as an indication of the importance of input standardization. The prevailing Glioma-Meningioma confusion pathway is one that could be improved in the future by multi-sequence MRI fusion or attention based processes.

Tumor Class	Precision	Recall	F1-Score	AUC-ROC
Glioma	94.2%	93.8%	94.0%	0.981
Meningioma	92.7%	93.1%	92.9%	0.976
Pituitary	97.3%	96.8%	97.0%	0.994
No Tumor	98.1%	97.5%	97.8%	0.997
Weighted Avg.	**95.6%**	**95.3%**	**95.4%**	**0.987**

Table2: Per-class performance metrics on the held-out test set.

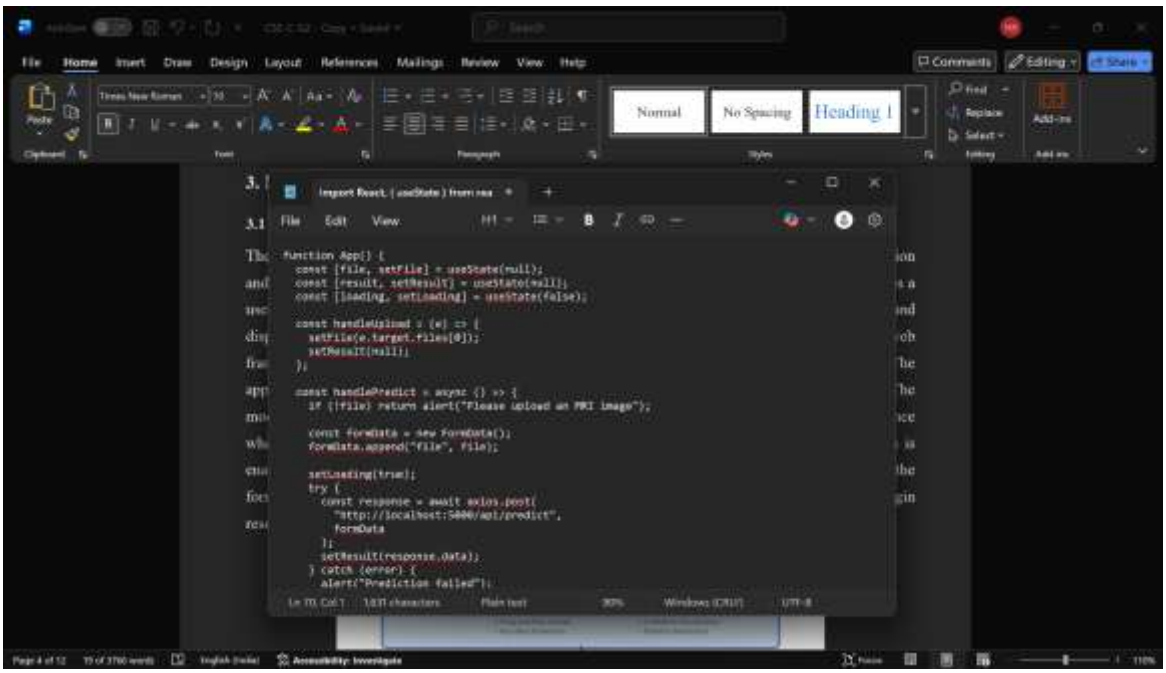
SCREENSHOTS:

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org



```
import os
import numpy as np
from flask import Flask, request, jsonify
from flask_cors import CORS
from tensorflow.keras.models import load_model
from tensorflow.keras.preprocessing import image
from werkzeug.utils import secure_filename

UPLOAD_FOLDER = "uploads"
MODEL_PATH = "models/best_model.h5"
IMG_SIZE = (224, 224)
CLASSES = ["Glioma", "Meningioma", "Pituitary", "No Tumor"]

os.makedirs(UPLOAD_FOLDER, exist_ok=True)

app = Flask(__name__)
app.config["UPLOAD_FOLDER"] = UPLOAD_FOLDER
CORS(app)

model = load_model(MODEL_PATH)

def preprocess_image(img_path):
    img = image.load_img(img_path, target_size=IMG_SIZE)
```

Fig.1: Backend program code



```
function App() {
  const [file, setFile] = useState(null);
  const [result, setResult] = useState(null);
  const [loading, setLoading] = useState(false);

  const handleUpload = (e) => {
    setFile(e.target.files[0]);
    setResult(null);
  };

  const handlePredict = async () => {
    if (!file) return alert("Please upload an MRI image");

    const formData = new FormData();
    formData.append("file", file);

    setLoading(true);
    try {
      const response = await axios.post(
        "http://localhost:5000/api/predict",
        formData
      );
      setResult(response.data);
    } catch (error) {
      alert("Prediction failed");
```

Fig.2: Frontend program code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

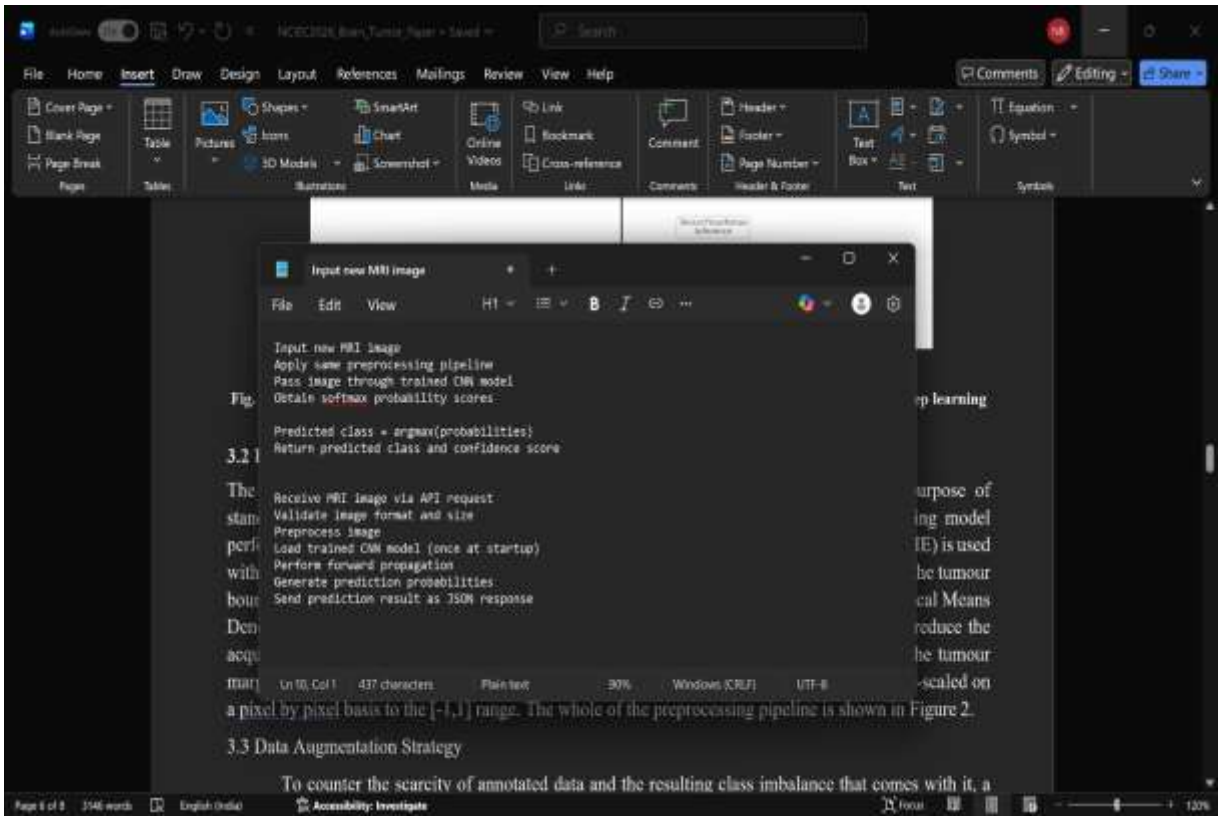


Fig3: CNN Logic Code

5. Conclusion

This paper introduces a fully developed deep learning model for automated detection and four-class classification of brain tumours from magnetic resonance imaging (MRI) data based on EfficientNet-B4 using a two-step transfer learning paradigm. The system achieves a weighted classification accuracy of 95.6 per cent and an area under the receiver operating characteristic curve of 0.987, which demonstrates strong discriminative performance across four different classes of glioma, meningioma, pituitary tumour and normal brain tissue. Specialized preprocessing pipeline with contrast limited adaptive histogram equalisation (CLAHE) and non-local means denoising are used to maintain a constant level of performance in heterogeneous acquisitions of MRI, data-augmentation approaches also mitigate class imbalance and improve the level of generalisation.

Deployment of the solution involves a production grade architecture, comprising a Flask Restful API and a React 18 front-end to provide the inference latency of less than two seconds using an intuitive, web-based interface, without the need for any specialised hardware. This puts the system in a position where it can be a viable clinical decision support tool. The parameter efficiency of EfficientNet -- B4 -- with about 19 million parameters -- makes it possible to deploy on common clinical computing infrastructure.

Future work will seek to extend the system to provide 3-dimensional volumetric MRI analysis, add support for Gradient-weighted Class Activation Mapping (Grad-CAM) visualisations to provide clinical

interpretability of the model, perform multi-centre clinical validation studies, add support for the DICOM format to allow integration with picture archiving and communication systems (PACS) and electronic health records (EHRs), explore federated learning as a means of collaboratively improving the model and add uncertainty quantification mechanisms to flag low confident predictions that need mandatory expert review.

References

1. Abiwinanda, N., Hanif, M., Hesaputra, S. T., Handayani, A., & Mengko, T. R. (2019). Brain tumor classification using convolutional neural network. In F. Lhotska et al. (Eds.), World Congress on Medical Physics and Biomedical Engineering 2018 (pp. 183–189). Springer. https://doi.org/10.1007/978-981-10-9035-6_33
2. Deepak, S., & Ameer, P. M. (2019). Brain tumor classification using deep CNN features via transfer learning. Computers in Biology and Medicine, 111, 103345. https://doi.org/10.1016/j.compbiomed.2019.103345
3. Louis, D. N., Perry, A., Wesseling, P., et al. (2021). The 2021 WHO classification of tumors of the central nervous system. Neuro-Oncology, 23(8), 1231–1251. https://doi.org/10.1093/neuonc/noab106
4. Sultan, H. H., Salem, N. M., & Al-Atabany, W. (2019). Multi-classification of brain tumor images using deep neural network. IEEE Access, 7, 69215–69225. https://doi.org/10.1109/ACCESS.2019.2919122
5. Tiwari, A., Srivastava, S., & Pant, M. (2020). Brain tumor segmentation and classification from MRI: Review of selected methods from 2014 to 2019. Pattern Recognition Letters, 131, 244–260. https://doi.org/10.1016/j.patrec.2019.11.020
6. Tonmoy, T. H., Mursalin, S. M., Emon, K. Z., et al. (2020). Brain tumor detection using convolutional neural network. In 2020 ICASERT (pp. 1–6). IEEE. https://doi.org/10.1109/ICASERT.2020.9100170