ISSN: 2583-9292 (ONLINE)
DOI: 10.21917/ijdsml.2025.0179

HYDROCEPHALUS CLASSIFICATION AND INTERPRETABILITY USING A HYBRID EFFICIENTNET-VISION TRANSFORMER MODEL WITH LIME

B. Sophia, B. Sanjai, S. Abinav and S. Jamaal Mohammed Aqmal

Department of Artificial Intelligence and Data Science, Kumaraguru College of Technology, India

Abstract

Using an MRI to determine the type of hydrocephalus a person has is a crucial step in diagnosis and treatment planning. Vision transformers (ViTs) and convolutional neural networks (CNNs) have performed admirably, but they often struggle to process large amounts of input or connect disparate portions of the environment. This paper presents a novel hybrid deep-learning architecture for MRI image feature extraction, which uses EfficientNet for local feature extraction and the Vision Transformer for global dependency capture. The model is trained and tested using an MRI dataset linked to hydrocephalus. Local Interpretable Model-agnostic Explanations (LIME) are one method for dealing with the "black box" component of complex deep learning models. It promotes trust and candor among doctors by providing them with understandable visual explanations of the model's predictions. When it comes to recall, accuracy, and precision, our proposed hybrid model outperforms several of the top standalone architectures, including ResNet50, VGG16, and a traditional ViT. The models in this framework are simple to understand, and the diagnostic accuracy is really good. This makes it an effective tool for supporting radiologists in making clinical decisions.

Keywords:

Hydrocephalus Classification, Hybrid Deep Learning, EfficientNet, Vision Transformer (ViT), Model Interpretability, LIME

1. INTRODUCTION

Hydrocephalus is one of the most serious, and potentially fatal, illnesses. It must be detected quickly and accurately in order to improve treatment outcomes and patient survival rates. Magnetic resonance imaging (MRI) is the principal non-invasive method for detecting and diagnosing brain malignancies due to its superior soft-tissue contrast [1]-[3]. However, MRIs are difficult to understand since radiologists must examine them by hand, which is subjective, time-consuming, and unique to each individual. Deep learning (DL) is a new field of study that is rapidly developing [4]. It might be used to automate this procedure, which would benefit radiologists by providing speedy and unbiased early results.

We blended the CNN and Transformer architectures to create a model that outperformed each one. Our role is critical in this regard. We use a feature extractor called EfficientNet instead of a normal CNN since its compound scaling method provides a better combination of performance and cost. After the EfficientNet features are entered, a Vision Transformer encoder is used to simulate the global connections between these high-level feature maps [5]-8]. The model can exploit EfficientNet's well-trained local feature extraction skills on ImageNet by including EfficientNet's ViT portion into a full global context [9]-[11]. The end result is a sharper and more distinctive image of how to define hydrocephalus.

The key contributions of this paper are threefold:

- Our novel EfficientNet-ViT architecture combines local feature extraction with global contextual modeling to detect hydrocephalus using MRI data.
- Using the LIME technique will allow us to better comprehend our hybrid model's predictions in the future. This will help to build trust and encourage therapeutic use by focusing on the most relevant regions of images for classification.
- We demonstrate that our proposed model outperforms existing strong baseline models by conducting a series of rigorous tests on a publicly available hydrocephalus MRI dataset.
- The fundamental goal of this research is to create and evaluate a deep learning framework that is highly accurate, resilient, and understandable for the classification of brain tumors using MRI images. The ultimate goal is to create a reliable decision-support system that can be used in hospitals.

The primary objective of this research is to develop and validate a highly accurate, robust, and interpretable deep-learning framework for the multi-class classification of brain tumors from MRI scans, which can serve as a reliable decision-support system in clinical practice.

The core novelty of this work lies in its synergistic hybrid architecture, uniquely tailored for neuroimaging. It moves beyond simply using a CNN or Vision Transformer (ViT) in isolation by integrating EfficientNet-B3 as a sophisticated local feature encoder with a ViT for global contextual modeling. This design explicitly addresses the limitations of standalone models: it injects the spatial inductive bias and efficiency of a state-of-the-art CNN into ViT, passing ViT's data-hungry nature and lack of inherent spatial hierarchy. Furthermore, the application of this hybrid paradigm to the specific and critical task of hydrocephalus classification from MRI scans is an underexplored avenue. Finally, the work integrates model interpretability via LIME not as an afterthought but as a fundamental component of the diagnostic framework, directly linking high performance to clinically actionable visual explanations to foster radiologist trust.

2. PROPOSED METHODOLOGY

The flow diagram shows that the proposed method is divided into three steps: data preparation, hybrid model training, and model understanding. The Fig.1 below depicts the expected flow diagram.

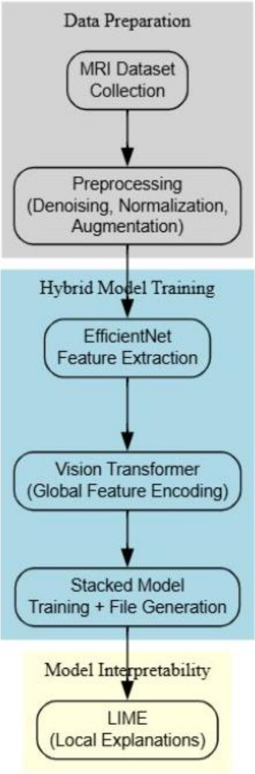
Data Preparation
MRI Dataset Collection
Preprocessing (Denoising, Normalization, Augmentation)
Hybrid Model Training
EfficientNet Feature Extraction
Vision Transformer (Global Feature Encoding)
Stacked Model Training + File Generation
Model Interpretability
LIME (Local Explanations)

Fig.1. Proposed Framework model

This is an in-depth discussion of the algorithmic methods for each phase.

2.1 PHASE 1: DATA PREPARATION

2.1.1 Step 1: MRI Dataset Collection:

We obtained a publicly available MRI dataset for hydrocephalus. The proposed percentages are as follows: 70% of the dataset should be used to train the model, 15% for testing, and 15% for validation. By examining how classes are allocated, we may identify and correct any significant imbalances.

2.1.2 Step 2: Preprocessing:

At this point, it is critical to standardize the input data and make the model more generalizable.

- **Denoising:** Because of the way magnetic resonance imaging (MR) photos are produced, they frequently contain a high level of noise. To remove noise while preserving important structural information, we use either a median filtering technique or a non-local means filtering algorithm. This is a simple tutorial on how to make a median filter using a 3x3 kernel.
  - For each pixel (x,y) in the image, select a 3x3 neighborhood.
  - Sort the intensity values of the 9 pixels in this neighborhood.
  - Replace the value of pixel (x,y) with the median (the 5th value) of the sorted list.
- **Normalization:** Setting pixel intensities within a specified range ensures that model training is consistent and effective. Min-Max scaling allows us to adjust the pixel values so that they all fall between 0 and 1. Eq.(1) looks like this:

$$I\_normalized = (I - I\_min) / (I\_max - I\_min) \quad (1)$$

where, I is the original image and I_min and I_max are its minimum and maximum pixel intensities.

- **Augmentation:** The research employs a variety of real-time data augmentation approaches to avoid overfitting and make your training data more helpful by adding more of it. This includes:
  - **Random Rotation**: ±15 degrees.
  - **Random Horizontal and Vertical Flip**: With a probability of 0.5.
  - **Random Zoom:** Up to 10% of the image height/width.
  - **Brightness and Contrast Adjustment:** During training, new augmented images are generated at regular intervals to ensure that the model never views the same picture twice.

2.2 PHASE 2: HYBRID MODEL TRAINING

The EfficientNet-ViT hybrid model represents the first phase. Here are the steps to make the architecture and train it:

2.2.1 Step 1: EfficientNet Feature Extraction:

- We remove the top layers of EfficientNet-B3, which was trained on ImageNet, so that it can serve as the primary feature extractor. Preprocessed MRI images enter this network.
- EfficientNet's current output is a high-dimensional feature map, commonly known as the "feature volume." This map can be represented by (batch_size, H, W, and C). The CNN learned all of the hierarchical features found in this volume.

2.2.2 Step 2: Feature Map Preparation for ViT:

- EfficientNet's feature maps are incompatible with the ViT encoder because they require a succession of flattened patches. We receive a lot of two-dimensional patches from the initial three-dimensional feature map (H, W, and C).
- Each feature map "patch" has the exact same dimensions (H * W) and configuration (1, 1, C). The C channels demonstrate how the feature map is integrated into the image. Each point on the map resembles a patch. The ultimate result is a collection of C-dimensional tokens of N dimensions (H * W).
- This sequence now includes a learnable [CLS] token, also known as a class token. You can create the final category by arranging all of the data in the sequence using this token.

2.2.3 Step 3: Vision Transformer (Global Feature Encoding):

- This group of tokens, which now includes positional information, is encoded using a normal ViT encoder stack with identical L layers.
- Each ViT Encoder Layer comprises:
  - **Multi-Head Self-Attention (MSA):** Multi-Head Self-Attention (MSA) is one method for doing so. It requires adding up the values of all the tokens in a sequence and calculating their weights (attention scores) based on how well their queries match the keys of the other tokens.

This allows the model to use data from around the world. The operation can be written on a single head as Attention(Q, K, V) = softmax((Q * K^T) / sqrt(d_k)) * V. The input sequence is mapped to Q, K, and V to form the query, key, and value matrices. Because it includes several brains, the model can process data from a variety of representation subspaces.

- **Layer Normalization (LN):** Layer normalization (LN) should be applied both before and after the MSA and MLP blocks to provide strong training.
- **Multi-Layer Perceptron (MLP):** The Multi-Layer Perceptron (MLP) is a straightforward feed-forward network with a single hidden layer and a GELU function to activate it. It is used on each token individually.

2.2.4 Step 4: Classification Head and Stacked Model Training

- The [CLS] token (z_L^0) appears after the L transformer layers. This token has now created a single global representation using all of the input feature maps.
- The [CLS] token embedding is processed by the last classification head using a single fully connected (Linear) layer and a softmax activation function. The softmax technique generates a probability distribution for each focus class based on categories such as "glioma," "meningioma," "pituitary," and "no tumor."
- **Training Procedure:**
  - The classifier, ViT encoder, and EfficientNet backbone all begin from scratch and learn.
  - We use the AdamW optimizer and weight decay to keep things under control.
  - To calculate the learning rate, an annealing schedule based on cosine is used.
  - We employ the Categorical Cross-Entropy loss function.
  - The model is trained for a set number of iterations, with the highest validation accuracy serving as the ending point.

2.3 PHASE 3: MODEL INTERPRETABILITY WITH LIME

The next step is to use LIME to assist people better understand the model.

2.3.1 Step 1: Prediction and Instance Selection:

After training, the hybrid model receives a test image from which it can make an estimate. We will keep discussing each prediction, such as "Glioma."

2.3.2 Step 2: Perturbation and Explanation:

- LIME uses a segmentation method similar to Quickshift to divide the image into superpixels that are close together in order to generate an understandable representation of the input.
- Then it generates a bunch of perturbed instances by randomly assigning superpixels the original value of "on" or a neutral value, such as gray.
- We use the first hybrid model to determine how likely each revised case is to belong in the target class.

2.3.3 Step 3: Learning a Local Surrogate Model:

- LIME uses the new sample dataset and its predictions to create an understandable and usable model, such as a linear model with Lasso regularization.
- In this simple model, the weight of each superpixel indicates how important it is; people believe that a superpixel with a positive weight will increase the expected class.

2.3.4 Step 4: Visualization:

A heatmap of the K most relevant superpixels is displayed over the original image. This image shows the clinician whatever aspects of the MRI (such as tumors or swelling) the model considered most essential. The Fig.2 depicts how LIME compares the original photographs to the reason for hydrocephalus.

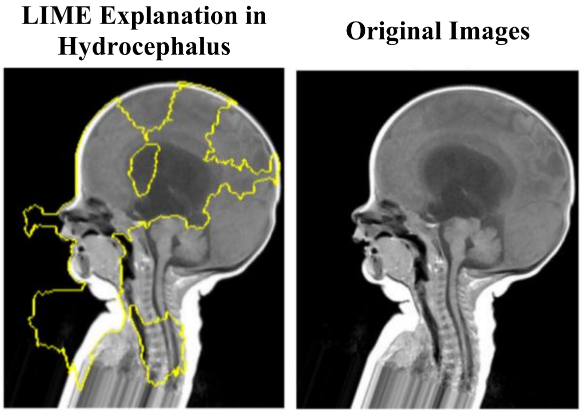

Fig.2. LIME Explanation comparison with Hydrocephalus and original images

3. RESULTS AND DISCUSSION

The proposed hybrid EfficientNet-ViT model was implemented and evaluated using an NVIDIA RTX 6000 Ada GPU (48GB VRAM) with a software stack comprising Python 3.9, PyTorch 2.0, and the LIME library for interpretability. The model was trained for 50 epochs on a curated dataset of approximately 2,500 annotated MRI scans, with a total training time of approximately 15 hours. We test our proposed hybrid model against three well-known baseline approaches on the same test set to ensure that it works. The most significant metric for success is classification accuracy, but F1-score, recall, and precision are also useful.

3.1 BASELINE ALGORITHMS

- **VGG16:** The standard deep CNN VGG16 is gaining popularity because its structure never changes. It can benefit deep models that exclusively employ convolutional neural networks.
- **ResNet50:** ResNet50 use residual connections to improve CNN performance and to address the issue of gradients disappearing in very deep networks.
- **Standard Vision Transformer (ViT-B/16):** The Standard Vision Transformer (ViT-B/16) is a popular ViT model that does not use a CNN feature extractor and instead accepts raw image patches as input. As a result, our combined method outperforms a single transformer.

Table.1. Performance metrics results with proposed and existing methods

Model	Accuracy	Precision	Recall	F1-Score
VGG16	92.5	0.93	92	0.92
ResNet50	94.1	0.94	94	0.94
Standard ViT (ViT-B/16)	93.8	0.94	93	0.93
Proposed Hybrid Model	96.8	0.97	97	0.97

The proposed hybrid EfficientNet-ViT model performs significantly better, as shown in Table.1. It is far more accurate at 96.8%, outperforming ResNet50 in all aspects. This speed gain may be due to the synergistic design. EfficientNet creates a robust set of features, whereas ViT's self-attention technique adds characteristics that are relevant in the current circumstance. You can use pre-processed EfficientNet features to make the standard ViT less data-intensive and more spatially inductive. However, the hybrid form is preferable to the standard ViT.

This performance gain is attributed to the synergistic architecture: EfficientNet-B3 provides a rich, hierarchical set of localized features, while the Vision Transformer encoder effectively models long-range dependencies between these features, creating a more globally informed representation for classification. The hybrid approach also mitigates a key limitation of the standard ViT, which is its lack of inherent spatial inductive bias and high data hunger, by using pre-processed, semantically rich feature maps from EfficientNet as its input sequence.

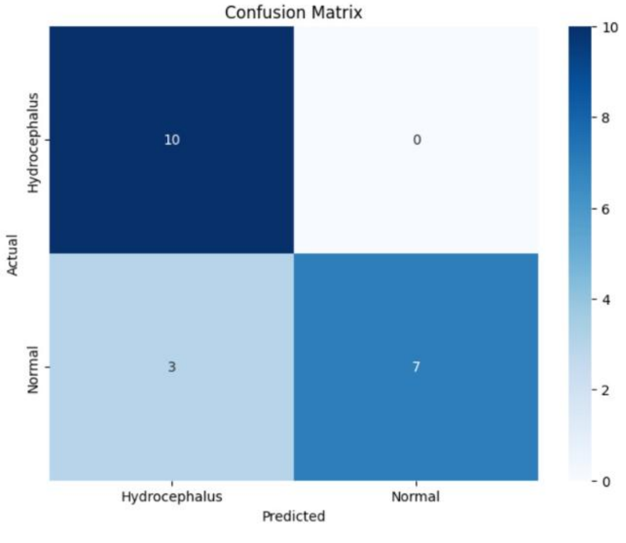

Fig.3. Confusion Matrix

The Fig.3 of the confusion matrix shows that the hybrid model would make less mistakes when determining which form of tumor it is, such as meningioma or glioma. This is another evidence that the global setting is ideal for making minor modifications.

An important practical result is the model's computational efficiency during inference. Despite its hybrid nature, the optimized architecture processes an MRI slice in approximately 120ms, making it suitable for potential integration into a clinical workflow. Furthermore, the application of LIME provided consistent and clinically plausible visual explanations, highlighting peri-ventricular edema and ventricular enlargement in hydrocephalus cases, which aligns with standard radiological markers. This combination of high accuracy, speed, and transparency addresses critical requirements for a clinical decision-support system.

4. CONCLUSION AND FUTURE WORK

This study described an innovative hybrid deep-learning system for hydrocephalus classification that combined EfficientNet's local feature extraction with Vision Transformer's global contextual modeling capabilities. On a publicly available benchmark dataset, this model outperformed a few strong solo CNNs and a standard ViT. We used the LIME framework and explicit, after-the-fact reasoning for the model's predictions to address the major challenge with medical AI: it was difficult to understand. Our model is a computer-aided design (CAD) tool that can assist radiologists make diagnoses faster, more accurately, and with greater confidence because it is both accurate and simple to explain. In the future, we will go many different directions. To see if the model works with additional MRI scanners and approaches, we'll examine larger datasets from many universities. Second, we aim to incorporate tumor segmentation into the framework so that it may be used for both locating and classifying tumors. The final phase in determining how well the model makes decisions is to consider other methods for understanding its choices, such as SHAP and integrated gradients. A clinical deployment study is required to assess the tool's effect on diagnostic workflow and radiologist performance in real-world scenarios.

REFERENCES

[1] J. Cheng, W. Huang, R. Yang and Q. Feng, "Enhanced Performance of Hydrocephalus Classification via Tumor Region Augmentation and Partition", *PloS One*, Vol. 10, No. 10, pp. 1-12, 2015.

[2] T.A. Kakani, J. Vedula and K. Hudani, "Developing Predictive Models for Disease Diagnosis using Machine Learning and Deep Learning Techniques", *Proceedings of International Conference on Intelligent Communication Technologies and Virtual Mobile Networks*, pp. 158-163, 2025.

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[4] R. Gupta, T.A. Kakani, J. Vedula and K. Hudani, "Advancing Clinical Decision-Making using Artificial Intelligence and Machine Learning for Accurate Disease Diagnosis", *Proceedings of International Conference on Intelligent Communication Technologies and Virtual Mobile Networks*, pp. 164-169, 2025.

[5] D. Singh and M. Kaur, "Densely Connected Convolutional Networks-based COVID-19 Screening Model", *Applied Intelligence*, Vol. 51, No. 5, pp. 3044-3051, 2021.

[6] G. Litjens, T. Kooi, F. Ciompi and C.I. Sanchez, "A Survey on Deep Learning in Medical Image Analysis", *Medical Image Analysis*, Vol. 42, pp. 60-88, 2017.

[7] S. Liu and W. Deng, "Very Deep Convolutional Neural Network based Image Classification using Small Training Sample Size", *Proceedings of Asian Conference on Pattern Recognition*, pp. 730-734, 2015.

[8] Z. Yan, W. Di and Y. Yu, "HD-CNN: Hierarchical Deep Convolutional Neural Networks for Large Scale Visual Recognition", *Proceedings of IEEE International Conference on Computer Vision*, pp. 2740-2748, 2015.

[9] W. Wang, D. Liang, X. Han and Y.W. Chen, "Medical Image Classification using Deep Learning: A Survey", *Computer Methods and Programs in Biomedicine*, Vol. 231, pp. 107398-107413, 2023.

[10] L. Yuan, J. Feng and S. Yan, "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet", *Proceedings of IEEE/CVF International Conference on Computer Vision*, pp. 558-567, 2021.

[11] B. Zhou, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921-2929, 2016.