DEEP LEARNING FOR BRAIN TUMOR CLASSIFICATION FROM MRI IMAGES

Ehsan Liaqat, Robby Parmar, Jubayer Ahmed, Gopal Sharma

University of Calgary

ABSTRACT

This study assesses the application of deep learning for classifying brain tumors from MRI scans to enhance oncological diagnostics. We tested various deep learning architectures, including DenseNet, VGG16, and ResNet variants and a custom CNN model, with a focus on the impact of transfer learning. The pretrained VGG16 model exhibited exceptional performance with accuracy of 97.83% and F1 score of 0.9759, demonstrating the effectiveness of transfer learning in identifying complex patterns for tumor classification. Our results advocate for integrating advanced deep learning models into clinical settings, which promises significant improvements in early diagnosis and personalized healthcare.

Keywords: Deep Learning, Brain Tumors, MRI, Image Classification, Transfer Learning

1. INTRODUCTION

Incorporating deep learning into medical imaging for brain tumor diagnostics marks a significant advancement in neuro-oncology. MRI scans, despite their complexity and the variability in interpretation, are crucial for tumor detection. Convolutional neural networks (CNNs), particularly DenseNet and VGG16, have shown potential in enhancing diagnostic accuracy. Our research evaluates these models on a specialized dataset, hypothesizing that transfer learning will improve performance. The study seeks to determine the most effective model, thereby contributing to the advancement of automated diagnostics in neuro-oncology and patient care.

2. RELATED WORK

The integration of deep learning into the field of medical imaging, especially for the classification of brain tumors from MRI scans, marks a pivotal advancement in enhancing diagnostic accuracy and treatment planning. Deep learning has recently been used to aid medical diagnosis and significant research has been carried out in the area. The capability of deep learning models to discern intricate patterns within complex medical images has fostered a growing body of research dedicated to exploring these models' effectiveness in various diagnostic applications. This section exposes works that have significantly contributed to the development and application of deep learning techniques in the realm of brain tumor classification, laying the groundwork for the methodologies employed in our study. A couple recent examples of deep learning for medical & tumor diagnosis can be witnessed by A. Çinar et al., investigating tumor detection hybrid convolutional neural network architecture [9]. Another study by S. Kokkalla et al. investigates three class brain tumor classification using deep dense inception residual network [5].

Recent literature has demonstrated the potential of various deep learning architectures in accurately classifying brain tumors from MRI images. These studies have employed a range of techniques and models, each contributing unique insights into the capabilities and challenges associated with deep learning applications in medical imaging.

3. MATERIALS AND METHODS

We analyzed anonymized MRI scans of brain tumors using deep learning architectures, with a focus on VGG16, DenseNet121, ResNet50, ResBet101 and ResNet152 models. A total of 11 different models were investigated, including a custom CNN model.

Two model conditions were tested: untrained (random initial weights) and pretrained (weights from ImageNet, then fine-tuned). Training used mini batch gradient descent (size=32) with momentum(0.9), learning rate(0.0001), and a categorical cross-entropy loss function. We chose categorical cross-entropy as our loss function, prioritizing a balance between precision and recall, which is critical in medical diagnostics. In order to ensure consistency & controlled training, validation & testing across models, the models were split into train, validation & test sets using random seed 42. We adopted a stratified split of the data to maintain the proportion of classes across training (70%), validation (15%), and testing (15%) sets.

Performance metrics, including accuracy, precision, recall, and F1 score, were evaluated, alongside the area under the ROC curve (AUC) to measure classification efficacy. A confusion matrix was generated to assess the true and false positive rates, providing insight into each model's diagnostic precision. The GradCam results on slices of select MRI images were also created to gain further insight.

During training, the best performing model was saved, based on F1 Score. The best model was then used for test results.

We preprocessed the data by resizing images to 224x224 pixels, normalizing pixel values, and augmenting the dataset to increase robustness against overfitting. To ensure reproducibility, we have provided detailed parameter settings and the code in our GitHub repository. Additionally, we compared our models' performance to a baseline to highlight the advancements achieved. The baseline in our study is represented by the simple CNN model. The simple CNN uses 3 convolutional layers and 2 fully connected layers.

4. TABLE SUMMARY OF LITERATURE RESULTS

Table 1. Results Obtained from Literature Works.

Reference	Model	Dataset	Accuracy (%)
[2]	16-layer VGG-16	Hospitals' dataset from 2010–2015, China	98.00
[3]	CNN-based DL model	REMBRANDT	100.00
[4]	SVM and k-NN classifiers	Figshare, 2017	97.25
[5]	Deep inception residual network	Publicly accessible brain tumor imaging dataset with 3064 pictures	99.69
[6]	CNN model	Publicly released clinical datasets	99.33
[7]	Transfer learning-based classification	Figshare	99.02
[8]	DenseNet	Sartaj Brain Tumor Classification Dataset	92.17
[9]	ResNet50	Sartaj Brain Tumor Classification Dataset	90.72

Deep dense inception residual network study by S. Kokkalla et al. investigates the use of deep inception residual networks, achieving a notable high accuracy [5]. This approach leverages the strengths of both inception modules and residual connections, enhancing the network's ability to learn more complex features without a significant increase in computational complexity.

Transfer Learning-Based Approaches by P. Özlem and C. Güngen explores the application of transfer learning to brain tumor classification [7]. By adapting models pretrained on large, diverse datasets to the specific task of tumor classification, this approach demonstrates the ability to achieve significant accuracy improvements, underscoring the value of leveraging pre-acquired knowledge in enhancing model performance.

CNN Multi-Classification Strategies investigated by E. Irmak details a multi-classification strategy using convolutional neural networks (CNNs) [6]. This technique emphasizes the versatility of CNNs in handling multi-class problems, presenting a methodological advancement in classifying brain tumors into various categories based on their characteristics.

G.S. Tandel et al. employs CNN-based models tailored for brain tumor classification, reporting remarkable accuracy [3]. This study highlights the adaptability of CNN architectures in extracting relevant features from MRI images, even in the presence of significant variability among tumors.

Finally, B. Srikanth and S.V. Suryanarayan utilizes the 16-layer VGG-16 deep neural network, a model renowned for its deep architecture and strong feature extraction capabilities. Despite its relative simplicity, the VGG-16 network demonstrates the profound impact of depth in neural architectures on classification accuracy [2].

These studies collectively underscore the rapid evolution of deep learning in medical imaging analysis, particularly in brain tumor classification.

5. RESULTS AND DISCUSSION

The results from the trained and pretrained models we ran are summarized in table below. Only the best model, with highest F1 score, are selected for determining test results. Class 0 is Healthy and Class 1 is Tumor.

Table 2. Untrained Model Results on test data.

Untrained Models						
Model	Test Loss	Test Accuracy (%)	Accuracy Per Class	Precision	Recall	F1 Score
Densenet	0.6312	65.07	[0.87945205 0.39384615]	0.7442	0.3938	0.5151
Resnet50	0.6257	68.26	[0.82984293 0.5]	0.7032	0.5	0.5844
Resnet101	0.6447	64.2	[0.88188976 0.34627832]	0.7039	0.3463	0.4642
Resnet152	0.6512	63.48	[0.77393617 0.46815287]	0.6336	0.4682	0.5385
VGG16	0.4488	81.3	[0.84575835 0.77076412]	0.7945	0.7708	0.7825
SimpleCNN	0.496	77.1	[0.89790576 0.61363636]	0.8289	0.6136	0.7052

Table 3. Pretrained(Transfer Learning) Model on test data.

Trained Models						
Model	Test Loss	Test Accuracy (%)	Accuracy Per Class	Precision	Recall	F1 Score
Densenet	0.287	89.42	[0.90677966 0.88095238]	0.8997	0.881	0.8902
Resnet50	0.2992	88.26	[0.85983827 0.90909091]	0.848	0.9091	0.8775
Resnet101	0.2594	89.86	[0.91989664 0.87128713]	0.8949	0.8713	0.8829
Resnet152	0.2443	92.17	[0.93005181 0.91118421]	0.9112	0.9112	0.9112
VGG16	0.0586	97.83	[0.98408488 0.97124601]	0.9806	0.9712	0.9759

5.1. Analysis and Discussion of Results

The comparison between untrained and pretrained (using transfer learning) models clearly demonstrates the significant benefits of transfer learning in the context of classifying brain tumors from MRI images. The improvement in performance metrics across the board—test accuracy, precision, recall, and F1 score—highlights the effectiveness of leveraging knowledge from previously trained models.

Untrained Models: Generally, show lower performance, with test accuracy ranging from 63.48% to 81.3%. The significant variance in test loss and accuracy per class indicates the models struggle to generalize from the training

data without prior knowledge. However, the simple CNN model, which acts as our baselines, despite its low complexity, managed to obtain test accuracy of 77.10%, higher than many other models.

Pretrained Models: Exhibit substantially higher test accuracy, precision, recall, and F1 scores. This improvement is a direct result of initializing the models with weights learned from large, diverse datasets. Such initialization allows the models to start with a higher level of understanding of image features, which is further refined during training on the specific task of brain tumor classification.

Best Performing Model: The **VGG16 Pretrained** model, achieved the highest test accuracy (97.83%), precision (0.9806), recall (0.9712), and F1 score (0.9759). This outstanding performance can be attributed to the VGG16 architecture's depth and capacity to capture complex features in image data, which, when combined with transfer learning, allows it to effectively distinguish between healthy and malignant brain tissue.

Let's look at ROC curves of the untrained and trained models to further analyze the use of transfer learning in our case.

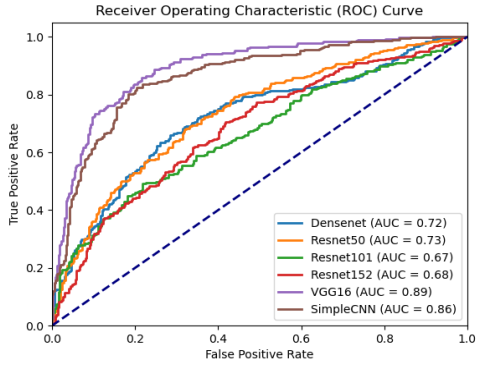

Fig. 1. ROC Curve for Untrained Models.

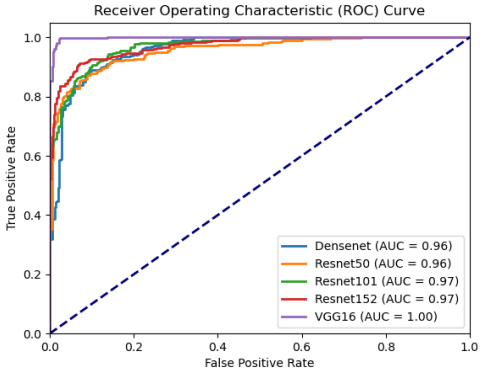

Fig. 2. ROC Curve for Trained Models.

The ROC curves contrast the classification performance of deep learning models on MRI brain tumor scans, showing the true positive rate (TPR) versus the false positive rate (FPR) at various thresholds.

Untrained Models: Display a gradual ROC curve ascent, indicating less effective classification. This slow increase reflects a struggle to differentiate classes without increasing false positives, highlighted by lower AUC values.

Trained Models with Transfer Learning: These models show a sharper initial ROC curve rise, demonstrating their capability to confidently identify brain tumors with minimal false positives.

Insights on Transfer Learning:

- Model Confidence: Transfer learning boosts prediction confidence, likely due to enhanced feature recognition from pretraining on diverse datasets.
- Performance Enhancement: The pronounced ROC curve and higher AUC for trained models suggest transfer learning enhances sensitivity and specificity—key for diagnostics.
- Data Efficiency: It enables better performance even with limited training data, common in medical imaging.

In essence, the ROC curve analysis underscores transfer learning's critical role in improving medical imaging analysis, providing more accurate and reliable diagnostic tools.

Results & Metrics: In the context of medical detection & diagnosis, it is desired to maximize TPR (recall), minimize FPR and maintain a higher F1 score. By having a lower FPR, we can minimize unnecessary medical interventions, treatment and anxiety. With a higher TPR, we can minimize false negative, which may overlook detection of a significant diagnosis. There is also a trade-off between correctly identifying positive cases (recall) and avoiding false positives (precision). A high recall ensures that as many true positive cases as possible are identified, while a high precision ensures that the identified positive cases are indeed true positives. The F1 score balances this trade-off, providing a single metric that reflects both precision and recall, which supports the use of F1 Score to obtain the best performing model.

5.2. Analysis of Best Model (VGG16 Pretrained)

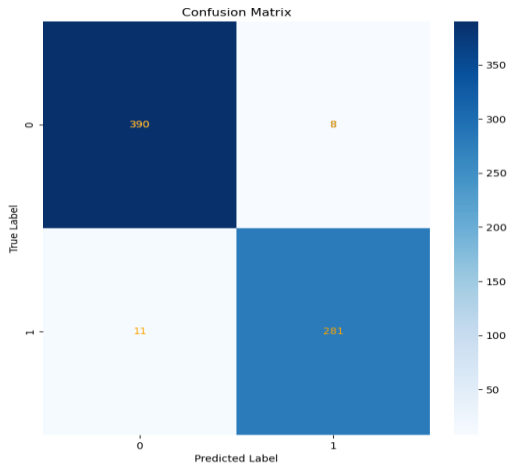

Fig 3. Confusion Matrix for VGG16 Model.

The confusion matrix for brain tumor classification shows the model's high accuracy: '1' for tumor and '0' for healthy. It correctly identified 390 healthy cases and 281 malignant tumors, with only 8 healthy misclassified as malignant (false positives) and 11 tumors missed (false negatives). The low rates of false positives and negatives highlight the model's strong capability to distinguish between healthy and malignant conditions effectively.

The ROC curve analysis demonstrates the VGG16 model's exceptional ability to classify brain tumors from MRI scans, achieving near-perfect accuracy with an AUC of 1.00. This performance aligns with top-tier medical image analysis, suggesting the model's robustness and potential to generalize well. Such outcomes highlight the significant impact of transfer learning in enhancing model effectiveness, particularly evident in the pretrained VGG16's superior performance. This success reinforces the value of transfer learning in medical imaging, where large, diverse datasets are scarce.

5.3. Comparison of Common Models with Literature

Table 4. Comparison of Our Results and Literature Results.

Literature Comparison of Pretrained				
Model	**Test Accuracy (%)**	**Literature Accuracy (%)**	**Difference**	**Reference**
VGG16	97.83	98	-0.17	[2]
Densenet	89.42	94.6	-5.18	[8]
Resnet50	88.26	96.5	-8.24	[9]

Our implementation of the VGG16 model achieved an accuracy slightly lower than that reported in existing literature, which may be attributed to our model optimization strategies and the unique characteristics of our dataset. In contrast, our DenseNet and ResNet50 models did not reach the accuracies found in similar studies, indicating that factors such as dataset diversity, preprocessing, and specific training methodologies could influence outcomes.

5.4. Analysis using Grad-Cam

Grad-CAM, or Gradient-weighted Class Activation Mapping, is an insightful visualization technique for highlighting the regions of an image that are important for predictions made by Convolutional Neural Networks (CNNs). This tool is exceptionally useful in the medical imaging field, particularly for MRI brain scans, as it enhances interpretability and trust in the model's predictions by showing which parts of the image the model is focusing on.

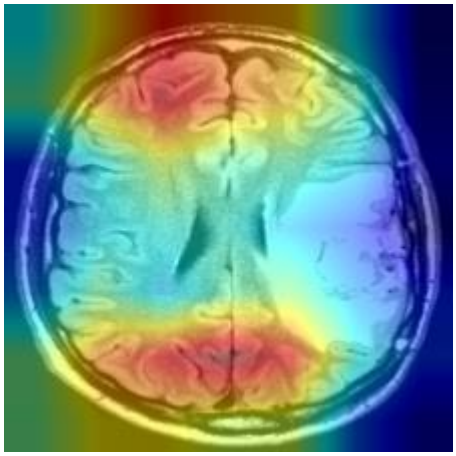

Fig. 5. Grad-Cam for DenseNet Trained Model. Slice Image: Cancer 1609. Actual Class: Tumor. Predicted Class: Tumor.

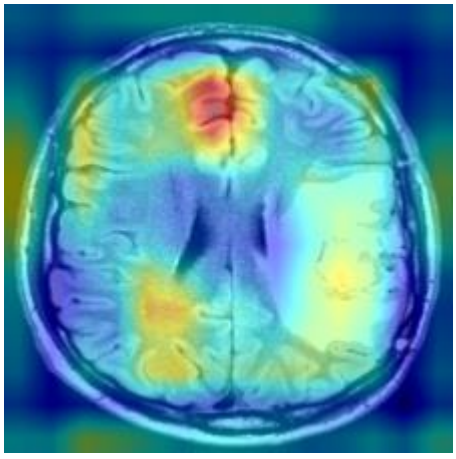

Fig. 6. Grad-Cam for VGG16 Trained Model. Slice Image: Cancer 1609. Actual Class: Tumor. Predicted Class: Tumor.

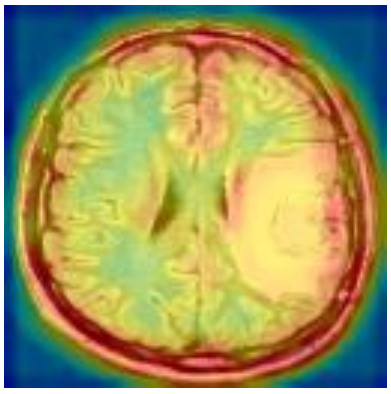

Fig. 7. Grad-Cam for SimpleCNN untrained Model. Slice Image: Cancer 1609. Actual Class: Tumor. Predicted Class: Healthy.

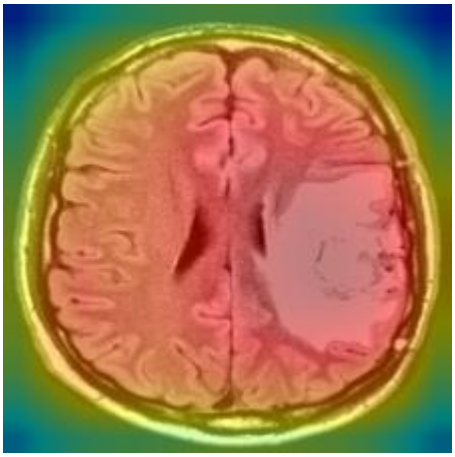

Fig. 8. Grad-Cam for VGG16 untrained Model. Slice Image: Cancer 1609. Actual Class: Tumor. Predicted Class: Healthy.

In our application involving MRI brain scans, Grad-CAM serves multiple purposes:

- It enhances the interpretability of the CNN model, helping us understand the decision-making process of the model.
- It provides insight into the activations that have the greatest influence by looking at the last layer.
- It provides visual evidence to support the model's predictions, which is critical for building trust with healthcare practitioners.
- It allows for model debugging, revealing whether the model is attending to the correct features in the image.
- It supports clinical decision-making by drawing attention to regions of interest that may require further examination.

When applying different CNN architectures like DenseNet and VGG to brain scans, Grad-CAM visualizations have shown that these models can focus on different areas of the image. The architectural nuances of each model influence how they learn and what features they highlight:

- DenseNet is characterized by its dense connections and parameter efficiency, often focusing on more diffuse areas across the image, which may include edges and textures as shown in Figure 5.
- VGG, with its deep layers of convolutions followed by max-pooling, tends to concentrate on more localized, salient features within an image as shown in Figure 6 and Figure 8.

In our case, the visualization indicates that DenseNet is focusing on broader regions, potentially including less relevant areas such as the tumor edges. Conversely, VGG appears to concentrate more on the actual area where a tumor is present, which is precisely the kind of targeted focus desired for accurate tumor detection, witnessed in Figure 8.

The disparity in focus between the two models is telling. If the goal is to identify brain tumors, a model that zeros in on the tumor, as VGG seems to do, would typically be more useful for diagnostic purposes. A model that consistently highlights areas unrelated to the condition of interest may have learned to rely on irrelevant features or correlations that do not generalize well, suggesting a need for further training or refinement of the model.

Ultimately, Grad-CAM visualizations provide critical feedback on the learning patterns of CNNs and are invaluable for refining AI-driven tools in healthcare to ensure they align with clinical objectives and expert knowledge.

6. LIMITATIONS

Within our study, one of the key limitations was that MRI images were not consistent, for example, some MRI images, whether healthy or tumor were captured from the side of the head, instead of top down. That may cause certain areas to be mistaken as cancer, as there are many different components within the human head that may mistakenly show as cancer area to a model. Another major limitation in medical datasets for deep learning is the scarcity of large, labeled datasets, essential for training accurate models as stated by J. Li et al. [10]. Collecting and annotating these datasets is costly, ethically complex, and requires specialized medical expertise, making it challenging to acquire sufficient data.

A notable limitation of deep learning is its struggle with tasks requiring common sense or abstract reasoning beyond its training data stated by B. Zohuri and M. Moghaddam [11]. These models excel in pattern recognition within their learned environment but fail to apply this knowledge to new, unrepresented scenarios.

Additionally, these models struggle with abstract reasoning and applying learned knowledge to new, unseen situations, underscoring a gap between AI capabilities and human cognition. Overcoming these challenges is crucial for leveraging the full potential of deep learning in improving medical diagnostics.

7. CONCLUSION

This research corroborated that deep learning models pre-equipped with transfer learning markedly outshine those trained from scratch, with the pretrained VGG16 model showcasing particularly high diagnostic accuracy for brain tumor classification from MRI scans.
Key limitations to address in future studies are the dataset's limited scope, which may not fully represent the diversity of brain tumors, and the necessity for model transparency to foster clinical trust. To overcome these challenges, enriching the dataset, both in volume and variety, and enhancing interpretability are essential next steps. Furthermore, assimilating data from different imaging modalities like CT and PET could refine the models' diagnostic capabilities.

Moving forward, we should also navigate the regulatory landscape to facilitate the safe integration of AI in healthcare. The collective efforts of data scientists, clinicians, and radiologists will be pivotal in harnessing these advanced models for real-time clinical decision support, contributing to the evolution of precision medicine in oncology.

8. REFERENCES

1. Pkdarabi, "Brain Tumor Detection by CNN PyTorch," Kaggle, 07-Apr-2024. [Online]. Available: https://www.kaggle.com/code/pkdarabi/brain-tumor-detection-by-cnn-pytorch/notebook.
2. B. Srikanth and S.V. Suryanarayana, "Multi-Class classification of brain tumor images using data augmentation with deep neural network," Mater. Today Proc., vol. 2021. doi: https://doi.org/10.1016/j.matpr.2021.01.601.
3. G.S. Tandel et al., "Multiclass magnetic resonance imaging brain tumor classification using artificial intelligence paradigm," Comput. Biol. Med., vol. 122, 103804, 2020, doi: https://doi.org/10.1016/j.compbiomed.2020.103804.
4. C. Öksüz, O. Urhan, and M.K. Güllü, "Brain tumor classification using the fused features extracted from expanded tumor region," Biomed. Signal Process. Control, vol. 72, 103356, 2022, doi: https://doi.org/10.1016/j.bspc.2021.103356.
5. S. Kokkalla, J. Kakarla, I.B. Venkateswarlu, and M. Singh, "Three-class brain tumor classification using deep dense inception residual network," Soft Comput., vol. 25, pp. 8721–8729, 2021, doi: https://doi.org/10.1007/s00500-021-05748-8.
6. E. Irmak, "Multi-Classification of Brain Tumor MRI Images Using Deep Convolutional Neural Network with Fully Optimized Framework," Iran. J. Sci. Technol. Trans. Electr. Eng., vol. 45, pp. 1015–1036, 2021, doi: https://doi.org/10.1007/s40998-021-00426-9.
7. P. Özlem and C. Güngen, "Classification of brain tumors from MR images using deep transfer learning," J. Supercomput., vol. 77, pp. 7236–7252, 2021, doi: https://doi.org/10.1007/s11227-020-03572-9.
8. H. Yahyaoui, F. Ghazouani, I.R. Farah, "Deep learning guided by an ontology for medical images classification using a multimodal fusion," in Proc. Int. Cong. Adv. Technol. Eng. (ICOTEN), Jul. 2021, pp. 1–6, Available: https://ieeexplore.ieee.org/abstract/document/9493469
9. A. Çinar and M. Yildirim, "Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture," Med. Hypotheses, vol. 139, 109684, 2020, doi: https://doi.org/10.1016/j.mehy.2020.109684.
10. J. Li et al., "A Systematic Collection of Medical Image Datasets for Deep Learning," 2021. [eess.IV]. Available: https://arxiv.org/abs/2106.12864.
11. B. Zohuri and M. Moghaddam, "Deep Learning Limitations and Flaws," Modern Approaches on Material Science, 2020, doi: https://dx.doi.org/10.32474/MAMS.2020.02.000138.
12. A. Younis et al., "Brain Tumor Analysis Using Deep Learning and VGG-16 Ensembling Learning Approaches," Applied Sciences, doi: https://doi.org/10.3390/app12147282.