Optimizing Intracranial Hemorrhage Detection: A Comparative Study of VGG Deep Learning Architectures with Transfer Learning on CT Brain scans

Yancheng Jin *

School of Data Science, The Chinese University of Hong Kong Shenzhen, Shenzhen, China

* Corresponding author: 120090524@link.cuhk.edu.cn

Abstract. Intracranial Hemorrhage (ICH) is a life-threatening condition that requires rapid and accurate diagnosis. This study applied Visual Geometry Group (VGG) deep learning model to binary classification of cerebral hemorrhage in Computerized Tomography (CT) brain scans, thereby improving the diagnostic process. Using a Kaggle dataset containing 6703 images, this study investigated the validity of VGG11, VGG13, VGG16, and VGG19 architectures, using transfer learning of ImageNet pre-trained weights to improve model performance. This study showed that VGG13 with transfer learning outperformed other models with 99.70% accuracy, indicating the profound impact of pre-trained weights on diagnostic accuracy. VGG models with untrained weights performed differently, with VGG11 achieving 98.61% accuracy, suggesting that less complex models were better suited to specific tasks. Comparative analysis shows that the depth of neural networks and the initialization of pre-training weights are the key factors to optimize performance. The discoveries present encouraging possibilities for integrating Artificial Intelligence (AI) into emergency medical services, with the potential to enhance the detection of cerebral hemorrhage and subsequently improve patient outcomes. While acknowledging limitations associated with dataset diversity and model generalization, the outcomes underscore the importance of tailoring models to specific task requirements. These findings also advocate for additional empirical research to fine-tune and advance the application of AI in healthcare.

Keywords: Intracranial Hemorrhage, VGG Model, CT Scan Classification, Artificial Intelligence in Healthcare.

1. Introduction

Intracranial Hemorrhage (ICH), a sudden severe onset of bleeding within the cranial vault, representing a devastating disease with significant morbidity and mortality [1, 2]. Statistics indicate a global incidence of 24.6 per 100,000 person-years, and it accounts for 40,000 to 67,000 annual cases in the U.S. [3, 4]. The condition's gravity is highlighted by high mortality rates of 35%-52% within the first 30 days, and a stark statistic that only about 20% of survivors regain full functionality six months post-incident, with nearly half of the deaths occurring within the first 24 hours [5, 6]., pinpointing the critical need for immediate and precise medical intervention in emergency care units.

Traditional diagnostic methods, predominantly reliant on Computed Tomography (CT) scans interpreted by medical professionals, face challenges in terms of speed and accuracy, impacted by factors such as human error and variability in expertise. Therefore, the need for enhanced, reliable methodologies is evident due to the variability in scan interpretations. The integration of Artificial Intelligence (AI), specifically deep learning models, presents a promising avenue for improving ICH detection in CT imagery.

The Visual Geometry Group (VGG) model, known for its deep architecture and small convolution filters, has seen some use in various medical imaging areas. Notably, its efficacy has been benchmarked in classifying skin cancer through dermatoscopic images, matching the proficiency of seasoned dermatologists [7]. Its versatility further extends to diagnosing diabetic retinopathy from retinal images [8] and categorizing lung nodules in CT scans, contributing significantly to the precocious detection of lung carcinoma [9].

Parallel to these advancements, deep learning paradigms have carved a niche in ICH detection, attributed to their capacity for accelerated processing and intricate pattern recognition in CT scans, surpassing conventional manual scrutiny. Varied Convolutional Neural Network (CNN) architectures have been the cornerstone of these developments, with research by Liskowski et al. [10] culminating in an automated detection system for hemorrhagic strokes in non-contrast CT images, boasting an accuracy of 94%. However, the deployment of these models is often encumbered by their voracious demand for computational assets and expansive labeled datasets. In spite of the strides in ICH detection utilizing deep learning, there exists a conspicuous lacuna concerning the targeted application of the VGG model within this context. Predominant explorations have harnessed alternative CNN architectures.

This study aims to leverage the VGG models for rapid, accurate binary classification of intracerebral hemorrhage from brain CT scans. Employing VGG variants 11, 13, 16, and 19, the research focuses on training these models on a comprehensive dataset of labeled CT images, followed by a robust evaluation to ascertain diagnostic effectiveness. The comparative analysis seeks to identify the most efficient VGG model in terms of accuracy, speed, and computational practicality, intending to significantly enhance ICH detection in critical medical scenarios.

2. Method

2.1. Dataset Preparation

The dataset utilized for this study is collected from Kaggle [11]. This dataset comprises a total of 6,703 CT images categorized into two classes: Normal and Hemorrhage, encompassing 4,105 and 2,598 images respectively.

The images within this dataset vary in size, ranging from 224×224 pixels to 512×512 pixels, which provides a diverse range of data for the model to learn from. The variation in image size presents a more realistic scenario for the model, as in clinical settings, CT scans may come in varying resolutions and dimensions. Some sample images from both the Normal and Hemorrhage categories are illustrated in Fig. 1.

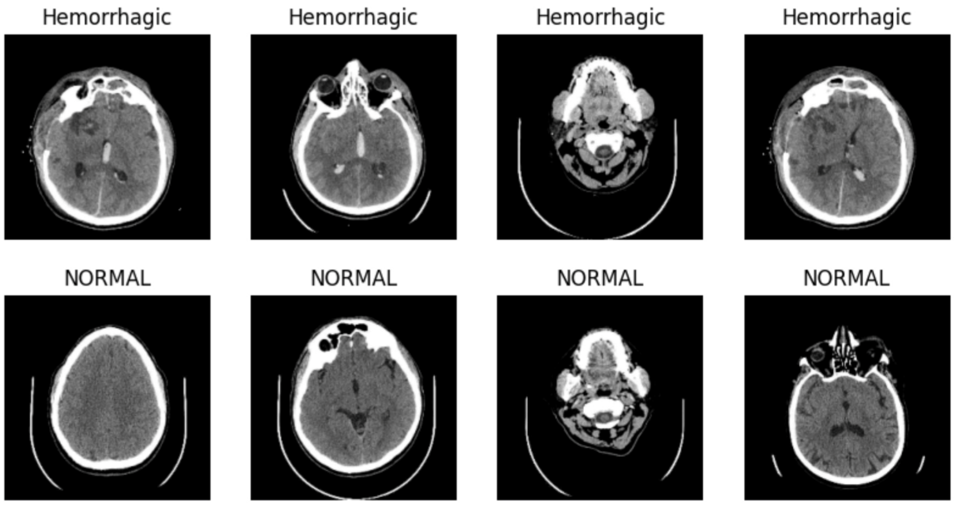

Figure 1. Some sample images of the collected dataset [9].

The data was prepared by loading images from specified directories and resizing them to a standard dimension of 224×224 pixels using the cv2.resize function to ensure uniform input size for the model. The dataset was segregated into training, validation, and testing sets using the train_test_split function from scikit-learn. The labels were converted to a categorical format using np_utils. to_categorical from Keras, facilitating the use of categorical cross-entropy as the loss function during the training phase. This preprocessing phase ensured a structured and consistent dataset, paving the way for effective model training and evaluation [12].

2.2. VGG Introduction

The VGG architecture, emanating from the Visual Geometry Group at Oxford, has significantly impacted the realm of deep learning for image recognition tasks due to its distinctive use of small receptive fields in convolutional layers, which are adept at capturing local patterns in images, and the stacking of these layers facilitates the apprehension of more complex patterns [12]. The architecture is well-structured, comprising a sequence of convolutional layers with 3x3 kernels, trailed by max-pooling layers for spatial down-sampling, and concluding with fully connected layers for class predictions employing a softmax activation function.

In this study, a VGG-based model is harnessed for brain hemorrhage identification from CT images due to its prowess in hierarchical feature learning from image data. The preprocessed image data is inputted into the model; as data traverses through the convolutional layers, features of increasing complexity are captured, which is pivotal for distinguishing intricate anomalies in medical images. The final fully connected layers, adapted to the binary classification task at hand, work on these features to classify the images into 'Normal' or 'Hemorrhage' categories, especially adjusting the neuron count in the last fully connected layer to reflect the binary nature of the classification task [13].

2.3. Implementation Details

The methodology of this research was designed to pinpoint an optimal model framework that excels in the classification of brain hemorrhage images. The process was initiated by leveraging transfer learning, a technique that has been demonstrated to be effective in similar tasks. The subsequent sections provide an outline of the approaches undertaken to realize the objectives of this research.

2.3.1. Utilization of Transfer Learning

Initially, the endeavor to attain the highest accuracy in the model framework led to the adoption of transfer learning. This approach involved utilizing the pre-trained VGG 11, 13,16,19 available in the TensorFlow Keras library (VGG13, 11 is modified from VGG16). The models were initialized with weights pre-trained on the ImageNet dataset, a large-scale image database well-regarded for its utility in training deep learning models. This strategy aimed to harness the already learned feature representations from ImageNet, which could potentially be beneficial for the task of brain hemorrhage image classification.

2.3.2. Exploring VGG Architectures

Upon achieving a satisfactory level of accuracy with the transfer learning approach, curiosity was piqued regarding the performance of different VGG architectures on the task at hand. This led to an exploration devoid of the initial weights from ImageNet and the pre-trained models. Independent constructions of VGG11, VGG13, VGG16, and VGG19 models were carried out to delve into a comparative analysis. The objective was to discern the performance and behavior of these models concerning the classification of brain hemorrhage images, striving to identify the architecture that most aptly fits the task. Through this comparative exercise, insights were sought on the influence of layer depths on the model’s ability to accurately classify the images, which is pivotal for the overarching goal of this research.

2.3.3. Experimental Setting

In the experiments, this study employed the Categorical Crossentropy loss function due to its prowess in multi-class classification tasks. To minimize this loss, the Adam optimizer was chosen for its renowned efficiency, especially in scenarios with sparse gradients or noisy problems. The model was trained over a span of 30 epochs, a decision made to ensure a balance between significant learning and mitigating the risk of overfitting. For the purpose of evaluation, accuracy stood as the primary metric, complemented by a confusion matrix to offer a granular view of performance. The model's architecture featured a concluding dense layer equipped with a sigmoid activation function, precisely tailored for tasks involving binary classification, such as discerning between "Normal" and "Hemorrhage." During the training process, this study settled on a batch size of 16, aiming to harmonize the pace of training with the model's generalization capabilities. Post the training phase, this study evaluated the model's performance on a separate testing dataset, using metrics like accuracy, loss, and the aforementioned confusion matrix to capture its classification abilities.

3. Results and Discussion

3.1. Transfer Learning with Pre-trained Models on ImageNet

In evaluating the performance of the VGG architectures with pre-trained ImageNet weights, this study achieved remarkable results. For the models VGG11, VGG13, VGG16, and VGG19, the accuracy and loss outcomes were synthesized into a coherent narrative. VGG13, with its rapid convergence, achieved the highest accuracy score of 99.70%, closely followed by VGG19 at 99.60%. The models VGG16 and VGG11 also showed excellent performance, with accuracy scores of 99.30% and 99.10%, respectively (Table 1). The loss values corresponded with the accuracy scores, where VGG13 exhibited a pronounced reduction in loss, the lowest among the models at 435.6731. VGG19 and VGG16 showed consistent and sharp reductions in loss with values of 2.9249 and 2.7973, respectively. Despite its steady convergence, VGG11 recorded the highest loss value of 9536.9268. The accuracy and loss trends over epochs are depicted in the curves presented in Fig. 2, Fig. 3, Fig. 4 and Fig. 5.

Table 1. Accuracy and Loss (With Imagenet)

Model	VGG11	VGG13	VGG16	VGG19
Accuracy	0.9910	0.9970	0.9930	0.9960
Loss	9536.9268	435.6731	2.7973	2.9249

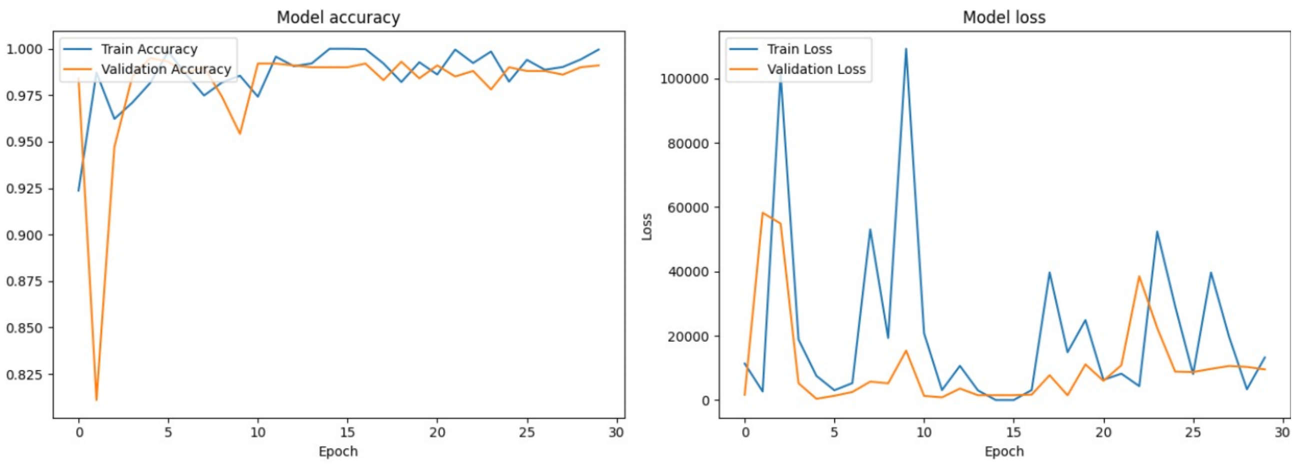

Figure 2. VGG11's accuracy and loss curves (With Imagenet) (Photo credit: Original)

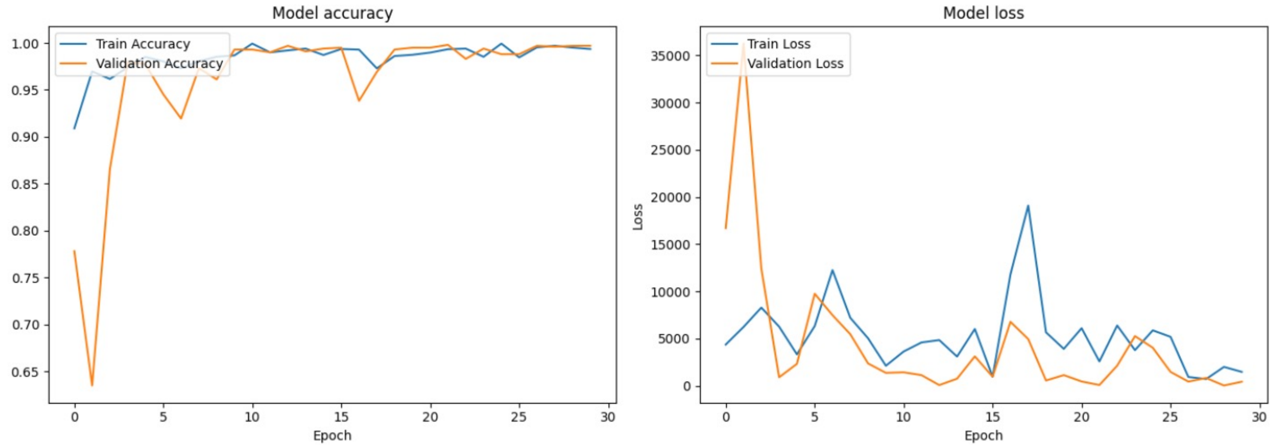

Figure 3. VGG13’s accuracy and loss curves (With Imagenet) (Photo credit: Original)

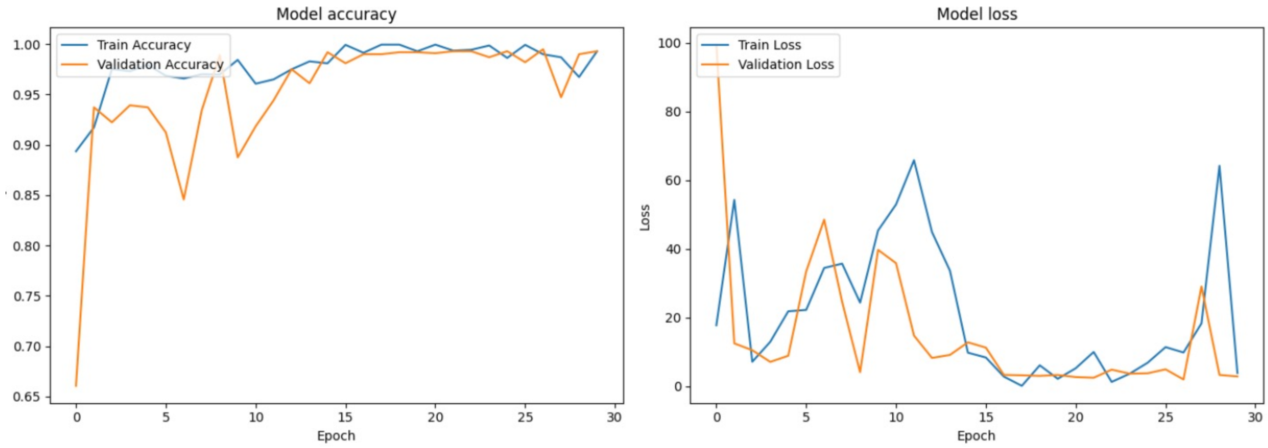

Figure 4. VGG16’s accuracy and loss curves (With Imagenet) (Photo credit: Original)

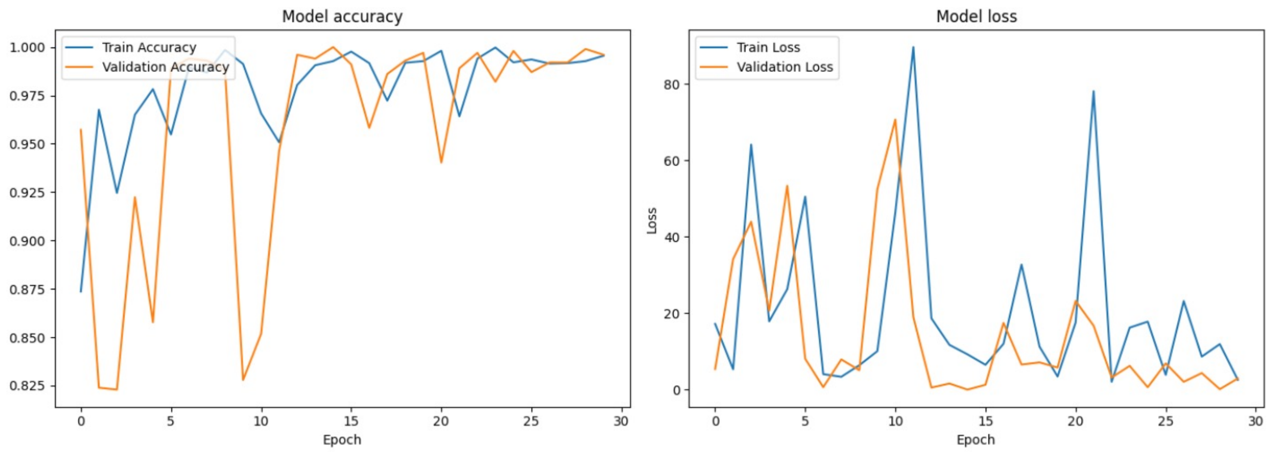

Figure 5. VGG19’s accuracy and loss curves (With Imagenet) (Photo credit: Original)

The effectiveness of the model classification is further substantiated by the confusion matrix data, detailed in Table 2. Notably, VGG19's swift convergence is mirrored in its superior classification accuracy, making only a single misclassification in ICH cases. Remarkably, VGG13 made zero misclassifications for normal cases, underscoring its robustness in distinguishing between classes.

Table 2. Confusion Matrix (With Imagenet)

Number	VGG11	VGG13	VGG16	VGG19
1	386	387	386	389
2	615	616	616	615
3	4	3	4	1
4	1	0	0	1

3.2. Models without Pre-trained Weights

When this study assessed the models without the advantage of pre-trained weights, the results naturally differed. VGG11 emerged resilient, with the highest accuracy of 98.61%, despite minor fluctuations. The model VGG16 showcased relative stability with an accuracy of 88.06%. However, VGG13 and VGG19 encountered considerable setbacks, with accuracies of 82.89% and 75.12%, respectively (Table 3). As for the loss values, VGG11's superior performance continued, exhibiting the lowest loss at 0.0443. VGG19 encountered the highest loss, at 2.5425, reflecting its volatile performance across epochs. These findings are visually represented in the loss and accuracy curves from Fig. 6, Fig. 7, Fig. 8 and Fig. 9.

Table 3. Accuracy and Loss (Without Imagenet)

Model	VGG11	VGG13	VGG16	VGG19
Accuracy	0.9861	0.8289	0.8806	0.7512
Loss	0.0441	0.3493	0.4499	2.5425

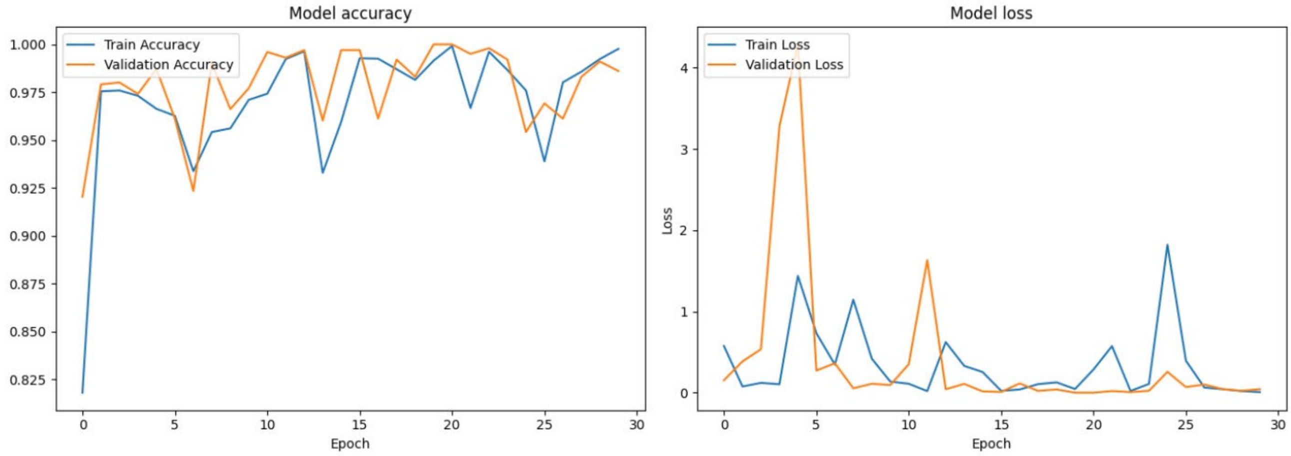

Figure 6. VGG11's accuracy and loss curves (Without Imagenet) (Photo credit: Original)

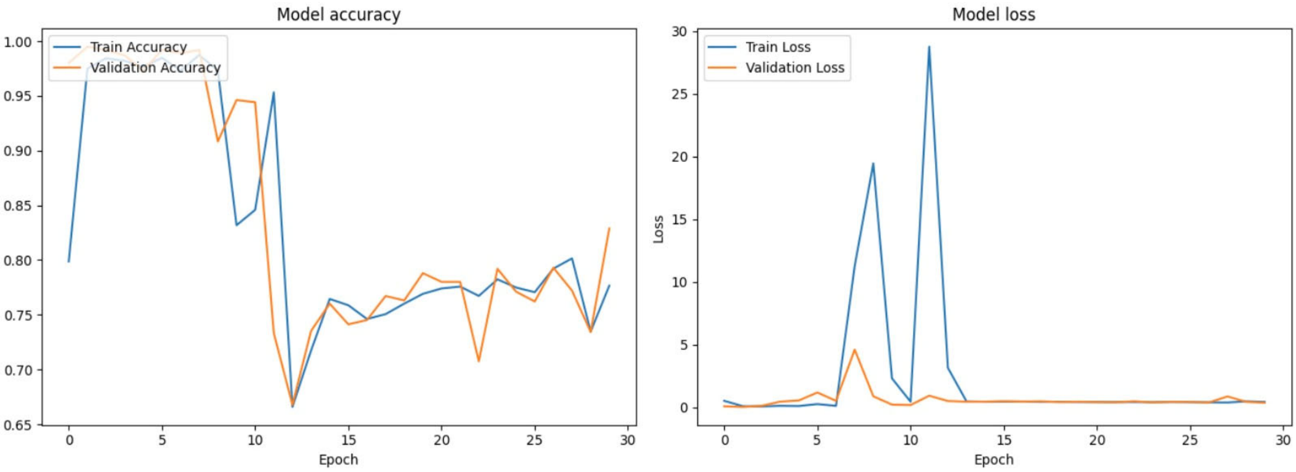

Figure 7. VGG13's accuracy and loss curves (Without Imagenet) (Photo credit: Original)

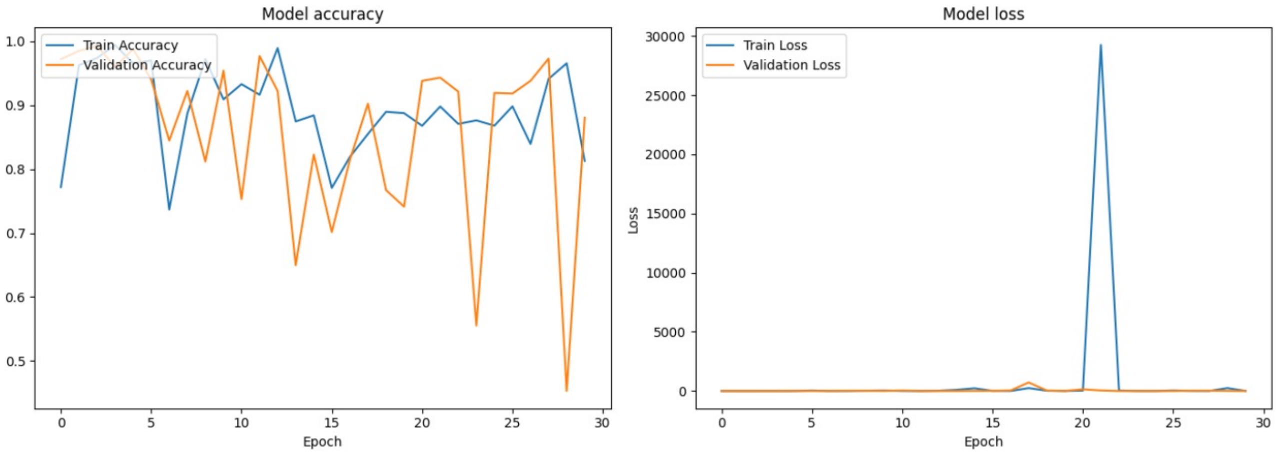

Figure 8. VGG16's accuracy and loss curves (Without Imagenet) (Photo credit: Original)

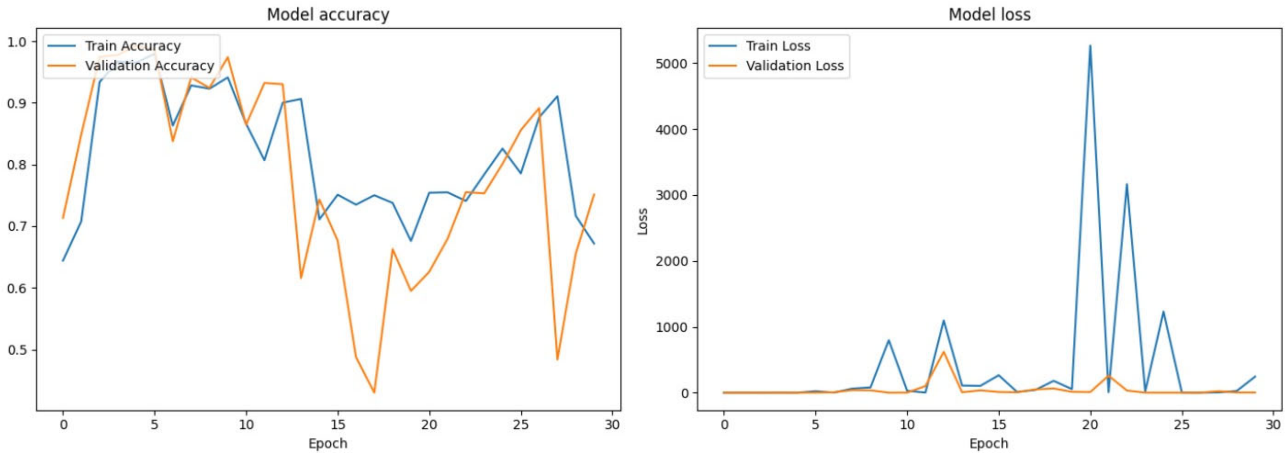

Figure 9. VGG19's accuracy and loss curves (Without Imagenet) (Photo credit: Original)

The challenges faced by the models in accurately classifying images without pre-training are quantified in the confusion matrix presented in Table 4. VGG19's difficulties are highlighted by its substantial number of misclassifications: 101 actual ICH cases and 129 normal cases. In stark contrast, VGG11's robustness is evident as it misclassified only 19 ICH cases, further proving the model's efficacy even without the aid of pre-trained weights.

Table 4. Confusion Matrix (Without Imagenet)

Number	VGG11	VGG13	VGG16	VGG19
1	371	367	296	289
2	615	461	598	487
3	19	23	94	101
4	1	155	10	129

3.3. Discussion and Analysis

The comparative analysis underscores the transformative effect of transfer learning on model performance. VGG13, when equipped with pre-trained ImageNet weights, emerged as the most robust, leading in both accuracy and loss metrics. The stark performance decline observed in VGG19 without pre-training underscores the essential role of ImageNet-derived features.

Conversely, the absence of pre-trained weights revealed a significant divergence in model performance. Notably, the relative success of VGG11 without transfer learning intimates that its architecture might be inherently more attuned to the specific tasks at hand, achieving a balance between complexity and generalizability.

However, the substantial performance variability exhibited by models such as VGG13 and VGG19 when trained from scratch hints at overfitting and a heightened sensitivity to the learning rate. Their deeper architectures and extensive parameters may predispose them to capture extraneous noise or overly specific dataset details, resulting in erratic training behavior.

This analysis reinforces the notion that initializing models with weights from a dataset like ImageNet can provide a substantial head start in terms of accuracy and convergence, especially for complex models designed to capture a wide array of features. In contrast, starting without the benefit of pre-trained weights often results in more pronounced performance differentials, highlighting the importance of model architecture and initialization in the broader context of machine learning applications.

3.4. Limitation

This study, while insightful, has inherent limitations. First, Dataset Diversity is a concern. The datasets used might not encompass the full range of brain hemorrhage presentations. Limited by dataset size, there's a risk of overestimating model performance due to potential overfitting on a less diverse dataset. Second, Model Generalization is a challenge. Although certain models, like VGG11, showcased commendable performance, their efficacy in real-world, diverse clinical scenarios remains to be validated. The controlled nature of the dataset might mask challenges that arise in practical applications. Lastly, Hyperparameter Tuning plays a pivotal role in deep learning outcomes. The study's results are based on a specific set of hyperparameters, and variations in these could lead to different performance metrics. An exhaustive exploration of hyperparameters might offer alternative insights into model behavior and performance.

4. Conclusion

This research embarked on an exploratory journey into the intricacies of VGG architectures, evaluating their prowess in classifying brain hemorrhage images with and without the leverage of transfer learning. The results highlighted the power of transfer learning, with VGG13's near-perfect accuracy standing testament. However, the study also illuminated VGG11's surprising resilience without pre-trained weights, challenging the notion of deeper architectures always yielding superior outcomes. VGG19's struggles without prior knowledge underscored the potential pitfalls of overfitting in deeper networks. These findings emphasize the importance of customizing deep learning approaches based on dataset characteristics and the specific task at hand. In the ongoing effort to harness AI for healthcare, grounded empirical research continues to be pivotal in ensuring that technological advancements result in tangible clinical benefits.

References

[1] Parizel, P., Makkat, S., Van Miert, E., Van Goethem, J., Van den Hauwe, L., & De Schepper, A. Intracranial hemorrhage: principles of CT and MRI interpretation. European radiology, 11, 1770 - 1783, 2001.

[2] Qiu, Y., Chang, C. S., Yan, J. L., Ko, L., & Chang, T. S. Semantic segmentation of intracranial hemorrhages in head CT scans. In 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS) (pp. 112-115). IEEE, 2019.

[3] van Asch C. J. J., Luitse M. J. A., Rinkel G. J. E., et al. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. The Lancet Neurology, 2010, 9 (2): 167 - 176.

[4] Aguilar M. I., Freeman W. D. Spontaneous intracerebral hemorrhage. Seminars in Neurology, 2010, 30 (5): 555 - 564.

[5] Broderick J., Brott T., Tomsick T., et al. Guidelines for the Management of Spontaneous Intracerebral Hemorrhage in Adults. Stroke, 2007, 38 (6): 2001 - 2023.

[6] Elliott J., Smith M. The Acute Management of Intracerebral Hemorrhage: A Clinical Review. Anesthesia & Analgesia, 2010, 110 (5): 1419 - 1427.

[7] Esteva A., Kuprel B., Novoa R. A., et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 2017, 542 (7639): 115 - 118.

[8] Gulshan V., Peng L., Coram M., et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 2016, 316 (22): 2402 - 2410.

[9] Ardila D., Kiraly A. P., Bharadwaj S., et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 2019, 25 (6): 954 - 961.

[10] Liskowski P., Krawiec K. Segmenting retinal blood vessels with deep neural networks. IEEE Transactions on Medical Imaging, 2016, 35 (11): 2369 - 2380.

[11] Mahjoubi M. A. Brain CT Hemorrhage AMINE Dataset. Kaggle. Available at: https://www.kaggle.com/datasets/mahjoubimohamedamine/braincthemorrhageaminedataset. Accessed on: October 15, 2023.

[12] Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556, 2014.

[13] Litjens G., Kooi T., Bejnordi B. E., et al. A survey on deep learning in medical image analysis. Medical image analysis, 2017, 42: 60 - 88.