DUAL-TASK VISION TRANSFORMER FOR RAPID AND ACCURATE INTRACEREBRAL HEMORRHAGE CLASSIFICATION ON CT IMAGES *

Jialiang Fan
Franklin College of Arts and Sciences
University of Georgia
Athens
jialiangfan@ieee.org

Guoyu Lu
College of Engineering
University of Georgia
Athens
gl72151@uga.edu

Xinhui Fan
Department of Neurology
The First Hospital of Yulin
Yulin
email@email

ABSTRACT

Intracerebral hemorrhage (ICH) is a severe and sudden medical condition caused by the rupture of blood vessels in the brain, leading to permanent damage to brain tissue and often resulting in functional disabilities or death in patients. Diagnosis and analysis of ICH typically rely on brain CT imaging. Given the urgency of ICH conditions, early treatment is crucial, necessitating rapid analysis of CT images to formulate tailored treatment plans. However, the complexity of ICH CT images and the frequent scarcity of specialist radiologists pose significant challenges. Therefore, we built a dataset for ICH and normal classification and three types of ICH image classification based on the hemorrhage location, i.e., Deep, Subcortical, and Lobar. In addition, we propose a dual-task vision transformer (DTViT) for the automated classification and diagnosis of ICH images. This neural network utilizes the encoder from ViT, employing attention mechanisms for feature extraction from CT images. We incorporated two multilayer perception (MLP)-based decoders within the network to simultaneously identify the presence of ICH and classify three types of hemorrhage locations. Experimental results demonstrate that our proposed multi-classification network performs well on the built real-world test dataset. The code and dataset for this study will be made publicly available upon paper acceptance at: https://github.com/Jialiangfan/ICH-classification.

Keywords Intracerebral hemorrhage classification, vision transformer, dual-task classification.

1 Introduction

Intracerebral hemorrhage (ICH) is type of a severe condition characterized by the formation of a hematoma within the brain parenchyma. Representing 10-15% of all stroke cases, ICH is linked to significant morbidity and mortality rates[1]. Head computerized tomography (CT) is the standard method to diagnose ICH that can obtain accurate images of the head anatomical structure and detect abnormalities. Early ICH classification can provide appropriate early-stage treatment such as blood pressure management, hemostatic treatment, and surgical treatment. However, skilled doctors are often in very short supply in some regions, and they sometimes face a high volume of cases. Additionally, the location of intracerebral hemorrhage (ICH) is crucial for medical professionals when determining treatment strategies. Different types of ICH, such as lobar hemorrhage, subtentorial hemorrhage, and deep hemorrhage, present distinct challenges. Diagnosing the correct type of ICH can be particularly challenging even for experienced radiologists.

In recent years, deep-learning-based computer vision methods have been playing an important role in CT imaging diagnosing. For example, Sahu *et al.* [2] present an automated deep learning-based system for COVID-19 detection using chest CT images, achieving a 95.73% accuracy through advanced preprocessing and the proposed model, highlighting its efficiency over manual diagnostic methods. Huo *et al.* [3] present a general-purpose model for medical image classification that uses hierarchical multi-scale feature fusion to efficiently extract and integrate local and global

---

features, significantly enhancing lesion detection in medical imagery. Huynh *et al.* [4] introduces a loss function named adaptive alended consistency loss for semi-supervised medical image learning that addresses class imbalance by blending class distributions, demonstrating superior performance over existing methods in experiments on imbalanced datasets. For the detection and classification of cerebral hemorrhage, numerous studies have utilized computer vision and deep learning techniques to aid medical professionals and enhance diagnostic efficiency[5, 6, 7, 8]. In literature [5], a deep learning model based on Resnet and EfficientDet that detects bleeding in CT scans is proposed, offering both classification and region-specific decision insights. Literature [6] demonstrates the efficacy of CNN-based deep learning models, particularly CNN-2 and ResNet-50, in classifying strokes from CT images, with future plans to optimize these models for improved diagnostic accuracy and efficiency. Li *et al.*[7] presents a rapid method using microwave signals to detect and classify cerebral hemorrhages, achieving high accuracy through a novel signal processing algorithm combined with neural networks, and demonstrating potential for quicker, life-saving diagnostics in emergency scenarios. However, these studies primarily focus on the task of classifying whether there is cerebral hemorrhage, but neglect further detection and diagnosis, i.e., the location and depth of the hemorrhage, which is more clinically significant.

Based on the above discussion, we propose a vision transformer (ViT)-based neural network model that achieves both ICH classification and hematoma location classification simultaneously.

2 Materials and methods

In this section, we introduce the dataset, the data preprocessing of ICH images, the construction of the proposed neural network model based on ViT, and the evaluation index. The structure of the proposed neural network is shown in Fig. 1.

2.1 Dataset

The dataset is sourced from the Department of Neurology at The First Hospital of Yulin. It includes 15,936 CT slices from 249 patients with intracerebral hemorrhage (ICH) collected between 2018 and 2021, and 6,445 CT slices from 199 healthy individuals from 2024. The healthy subjects have one set of CT images, while the ICH patients have two sets, captured within 24 hours and within 72 hours of symptom onset. All images were obtained using the LightSpeed VCT scanner from GE, USA, at Yulin First Hospital. All scans are saved as DCM files, featuring a resolution of 512×512 pixels, a slice thickness of 5 mm, and an inter-slice gap of either 5 mm or 1 mm.

In addition, each group of ICH images are manually classified into three different types according to hemorrhage location by physician: Deep ICH, Lobar ICH, and subtentorial ICH. To protect patient's privacy, the dataset only includes patients' gender and CT images. A detailed data description is listed in Table I.

Table 1: Data distribution of the dataset

Attribute	Total	Value	Number	Percentage
Sex	221	Male	117	52.94%
		Female	104	47.06%
ICH	12651	Yes	8244	65.16%
		No	4407	34.84%
Hemorrhage Location	8244	Deep	6093	73.91%
		Lobar	1656	20.09%
		Subtentorial	495	6%

2.2 Data preprocessing

The data preprocessing is composed of three stages: morphological treatment, manual filtering, and data augmentation.

2.2.1 Morphological treatment

Firstly, the pixel array is extracted from the DICOM file. Then, contrast enhancement and normalization are applied to the pixel array, which is finally converted to 8-bit integers. Additionally, when taking CT images, patients are equipped with fixation braces to keep their heads steady, which is also captured by the scanner and may add noise to the image data. Therefore, it is necessary to remove the artifact to purify the data. We binarized the image to highlight key features and applied erosion with a disk-shaped element to reduce noise. Then, Edge columns were zeroed out to eliminate potential artifacts, and flood filling was used to enhance specific areas. Finally, we superimposed the processed binary

| (a) | (b) |

Figure 1: CT image filtering. (a) The original CT image with a fixation device. (b) Processed image with fixation device removed.

image onto one of the windowed images, thereby enhancing the visibility of regions of interest and improving the diagnostic utility of the scans. An instance of the original CT image and the processed result is displayed in Fig. 1.

2.2.2 Manual filtering

The head CT scan usually starts from the base of the brain (near the neck) and covers the entire brain up to the forehead. It means that only part of CT scans can capture the ICH of the patient and present a clear view for diagnoses, and other CT images may be similar between normal people and ICH patients. Therefore, each group of ICH CT images is manually selected by medical specialists to remove meaningless images to refine the dataset and reduce noise.

2.2.3 Data augumentation

As presented in Table 1, The number of images for three types of cerebral hemorrhage locations varies significantly. Therefore, we take data augumentation operation on subtentorial type CT images. Firstly, we augmented the dataset of three types of cerebral hemorrhage CT images proportionally to ensure that the distribution was approximately 1:1:1. Secondly, we also augmented the images without cerebral hemorrhage to achieve a roughly 1:1 ratio between the two categories.

As introduced in Section 2.1, we conduct experiments on the dataset we made from the real world. The augmented dataset is composed of a total of 30222 images. We divide the dataset as training, validating, and testing according to a ratio of 8:1:1. Furthermore, to enhance the diversity of the dataset, we applied image transformation on the training dataset, including center cropping to the size of 224x224, random rotation of 15 degrees, random sharpness adjustment with a factor of 2, and normalization with the mean value $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$.

2.3 Methods

This section introduces the methodology used in the paper. We discuss the ViT first, present the architecture of the developed neural network model, and introduce the evaluation index.

2.3.1 Vision Transformer

The Transformer, a network architecture based on the self-attention mechanism, has achieved remarkable success in the field of natural language processing (NLP) in recent years. Therefore, Researchers have extended the transformer model to the field of computer vision by segmenting images into patches and feeding them into the Transformer architecture. This approach computes attention weights between different pixel blocks, facilitating effective feature extraction from images. Experimental results demonstrate that Vision Transformers achieve superior and more promising performance compared to traditional CNN-based neural networks.
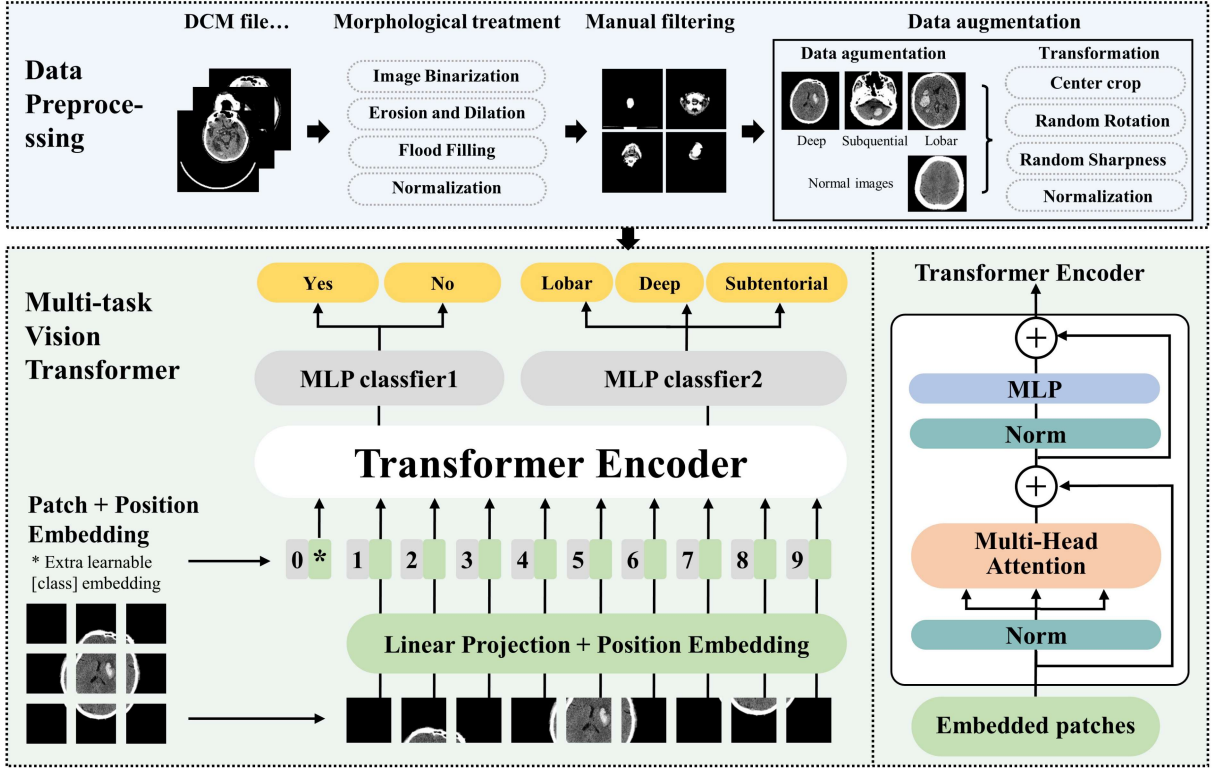
Figure 2: CT image filtering. (a) The original CT image with a fixation device. (b) Processed image with fixation device removed.

In ViT, a raw 2D image is initially transformed into a sequence of 1D patch embeddings. This process mirrors the word embedding technique used in natural language processing. In our research, the 2D CT image patch, denoted as v is reshaped as a sequence of N patches, where P is the patch size set as 16×16. These patches are then linearly projected into vectors of dimension.

In addition, to preserve the spatial context of the CT image patches, a trainable positional embedding is incorporated. This adjustment is essential due to the permutation invariance of self-attention mechanisms.

2.3.2 Model structure

Considering the high-density shadows, variable scales, and diverse locations characteristic of ICH CT images, we adopt the Vision Transformer (ViT) as the encoder of our neural network. Additionally, two decoders have been integrated into the network to achieve multi-task classification. The decoders share the feature extraction results from the ViT encoder applied to the CT images. The extracted image features are then fed into two fully connected layers to classify the images. The structure of the NN model is shown in Figure 2.

The proposed neural network model is compose of an encoder for image feature extraction and two decoders that can simultaneously classify CT images, identifying whether there is a brain hemorrhage and the type of brain hemorrhage.

2.3.3 Evaluation indices

We have taken various experiments and adopt several models to test the performance. The correct classification of images is the criterion of the model performance, which is presented as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

where TP represents true positive instances; TN represents true negative instances; FP represents false positive instances; FN represents false negative instances. The precision metric represents the proportion of correctly predicted

positive instances out of all the instances predicted as positive in a class. Precision is determined using the following equation:

$$Precision = \frac{TP}{TP + FP}. \tag{2}$$

The recall metric quantifies the proportion of positive instances that are correctly identified and the corresponding formula to calculate recall is presented as

$$Recall = \frac{TP}{TP + TN}. \tag{3}$$

The sensitivity metric calculates the proportion of positive instances that are correctly classified, which is defined as

$$Sensitivity = \frac{TP}{TP + TN} \tag{4}$$

In addition, the F1-score metric is the harmonic mean of precision and recall. It is computed as

$$F1 = \frac{2 * (Precision * Recall)}{Precision + Recall}. \tag{5}$$

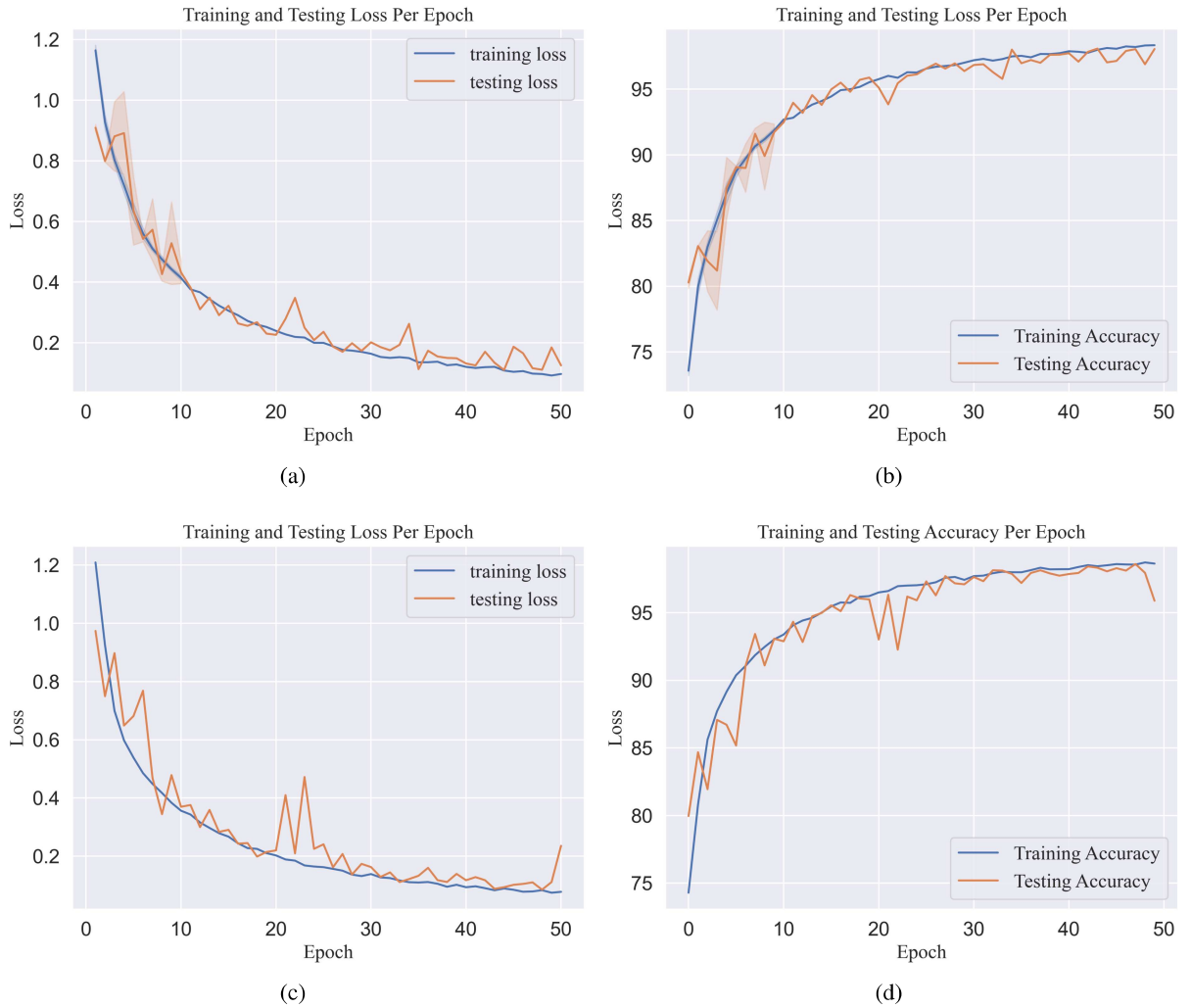We take these indices to evaluate our model's performance.

Figure 3: The accuracy and loss curves in the training and validating process on datasets with and without augmentation. (a) The training and validating losses with augmentation. (b) The training and validating accuracy with augmentation. (c) The training and validating losses without augmentation. (d) The training and validating accuracy without augmentation.

3 Experiments

In this section, we conduct experiments using the built dataset with and without data augmentation to validate the efforts of the augmentation. Performance indexes introduced in Section 2.3.3 are utilized for evaluation. In addition, we also select several classical neural network models as encoders for feature extraction and conduct comparative experiments to assess their performance.

3.1 Environments and parameters

The experimental hardware setup is as follows: The CPU used is an AMD Ryzen 5975WX with 32 cores, and the GPU is an NVIDIA RTX A6000 equipped with 48 GB of memory. The operating system is Ubuntu 20.04. CUDA version 12.2 was utilized for computation. All experiments were conducted using Python version 3.8.18 and PyTorch version 2.2.1.

3.2 DTViT performances

Initially, we trained the model on both the original dataset and the augmented dataset, setting batch sizes of 8 and 32, respectively. We configured the training for 50 epochs, using Stochastic Gradient Descent (SGD) as the optimizer with a learning rate of 0.01. Figure 3 illustrates the evolution of loss and accuracy during the training and validation processes, both with and without data augmentation. As depicted in Figure 3(a), both training and validation losses decrease steadily throughout the training, eventually stabilizing at approximately 0.1. Correspondingly, both training and validation accuracies show steady increases, converging closely by the end of the training period. In contrast, results from the non-augmented dataset display inconsistency: while training loss decreases continuously, validation loss fluctuates and fails to converge, ending above 0.2 and indicating overfitting. Similarly, validation accuracy is unstable compared to training accuracy, with its final value significantly lower, suggesting poor model performance on unseen data.

Further, we apply the trained model on the testing dataset to get the performance, which is composed of 3023 CT images, including 1342 for normal images and 1681 ICH images with Deep 606, Lobar 479, and Subtentorial 596. The confusion matrix is shown in Fig. 4, where Fig. 4(a) shows the result of Classifier 1 for normal and ICH patients classification, and Fig. 4(b) shows the results of Classifier 2 for three types of ICH classification. It can be seen that the true negative cases take 1331 over 1332, and the true positive cases take 1691 over 1691 cases, achieving 99.23% of the accuracy. Similarly, the classification of ICH types also achieved desirable results with a testing accuracy of 97.37%. In addition, the detailed data, including precision, recall, F1 score, and specificity, is displayed in Table 2.

Table 2: Performance comparison of DTViT with and without data augmentation on the testing data

Classifier	DA	Accuracy	Precision	Recall	F1 Score	Specificity
Classifier 1	Yes	0.992	0.996	0.987	0.991	0.997
Classifier 1	No	0.974	0.953	0.991	0.972	0.960
Classifier 2	Yes	0.974	0.962	0.969	0.966	0.995
Classifier 2	No	0.877	0.756	0.987	0.856	0.986

3.3 Comparative experiments

We conducted comparative experiments using classical computer vision models, including VGG-16 [9] and ResNet18 [10], on our augmented dataset. The results, presented in Table 3, indicate that both models perform exceptionally well on the testing dataset. Specifically, ResNet18 achieves an accuracy of 0.999 on Classifier 1, while VGG16 reaches perfect accuracy, correctly classifying all test images on Classifier 1. The performance on Classifier 2, though slightly lower, remains impressive: ResNet18 and VGG16 achieve accuracies of 0.997 and 0.995, respectively. Overall, both ResNet18 and VGG16 slightly outperform the Vision Transformer (ViT) model.

4 Discussion

In this paper, we first collect CT images from intracerebral haemorrhage (ICH) patients and normal people, which are sourced from real-world patient data from Yulin First Hospital, China. Furthermore, medical specialists put effort into data filtering and labelling of three types of ICH haemorrhage.
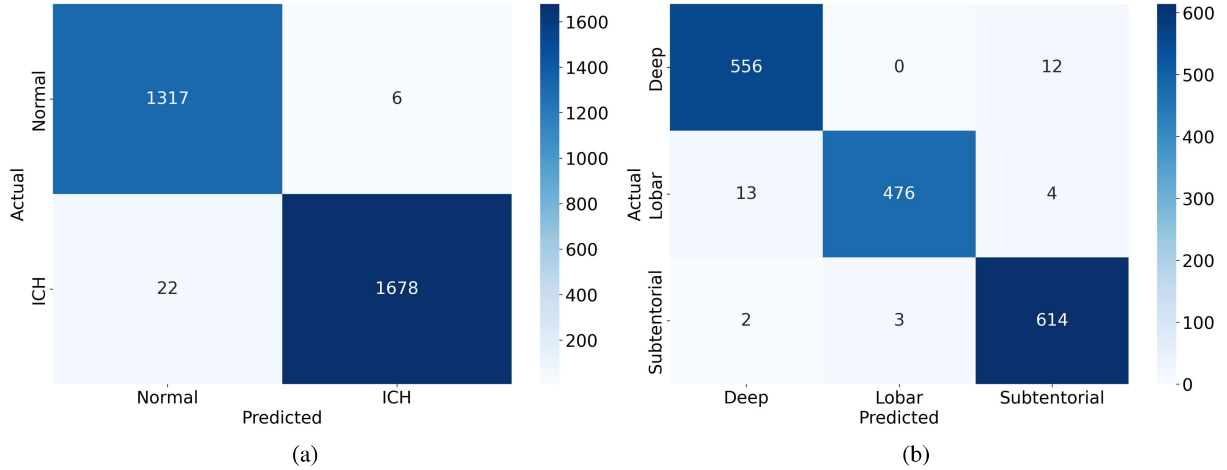
Figure 4: The confusion matrixes of two classifiers on the testing dataset. (a) The confusion matrix of classifier 1 for normal and ICH classification. (b) The confusion matrix of classifier 2 for three types of ICH classification.

Table 3: Performance comparison of two classifiers with and without data augmentation

Classifier	Encoder	Accuracy	Precision	Recall	F1 Score	Specificity
Classifier 1	ViT	0.989	0.997	0.979	0.988	0.998
Classifier 1	ResNet18	0.999	1	0.998	0.999	1
Classifier 1	VGG16	1	1	1	1	1
Classifier 2	ViT	0.971	0.97	0.967	0.968	0.988
Classifier 2	ResNet18	0.997	1	0.993	0.996	1
Classifier 2	VGG16	0.995	1	0.988	0.994	1

Based on the built dataset, we propose the dual-task vision transformer (DTViT) model based on the vision transformer for dual-task classification. The proposed model is composed of an encoder to extract information from CT images and two decoders for different classification tasks, i.e., classification of normal and ICH images and classification of three types of ICH based on the location of the hematoma. The proposed DTViT achieves 99.7% of the training accuracy and 99.5% of the testing accuracy on the augmented dataset. We also compared the proposed model with several state-of-the-art models, i.e., Resnet18, Resnet34, Vgg16, and Alexnet, and the results show that the proposed model achieves a desirable result.

There is some previous research focusing on ICH classification or detection[7, 5, 6, 8]. For example, literature[7] proposes an ICH classification and localization method using a neural network model, achieving the accuracy of 97.4%, while the input signals are microwave signals and the hardware requirements are relatively high. Also, a CT-image-based deep learning method is proposed in [5], which is based on the EfficientDet and achieves an accuracy of 92.7%. Similar work using computer vision methods for ICH detection and classification includes [11, 12], where literature [11] achieves ICH classification using a CNN-based architecture called EfficientNet; literature [12] uses ResNet-18 for ICH classification with the accuracy 95.93%.

However, to the best of our knowledge, there is neither such a dataset describing the location of hematoma of ICH CT images nor such classification model simultaneous determining whether a CT image is of a cerebral hemorrhage or normal, and classifies the three types of cerebral hemorrhage.

The primary aim of creating the Intracerebral Hemorrhage (ICH) CT image dataset and developing the classification model is to harness state-of-the-art computer vision technology to assist physicians in diagnosing and treating patients with cerebral hemorrhage. Cerebral hemorrhage is a severe, acute medical condition affecting approximately two million individuals annually, often associated with higher mortality and morbidity and limited treatment options. Early diagnosis and tailored treatment strategies based on the hemorrhage location are crucial for patient outcomes, as different hemorrhage sites require varied treatment approaches. Therefore, the classifier developed in this study, which categorizes hemorrhage based on its location, is clinically significant as it aids healthcare professionals in quickly and accurately determining treatment plans.

Limitations: This study faces several limitations. First, the dataset, derived from real clinical data, is challenging to obtain and inherently imbalanced, which may limit the model's accuracy and generalizability. Future efforts will focus on expanding and balancing the dataset through continued data collection. Second, the current data augmentation technique involves mere replication of the dataset. Future improvements will explore the use of generative models, such as diffusion models, for data enhancement.

Future Directions: In subsequent work, we aim to develop a multimodal diagnostic dataset for cerebral hemorrhage, integrating clinical data such as blood pressure, lipid profiles, and bodily element levels to enhance diagnostic accuracy. Furthermore, we plan to create a cerebral hemorrhage classification and diagnosis model based on Artificial Intelligence-Generated Content (AIGC), which will improve diagnostic efficiency for physicians.

5 Conclusion

In this paper, we have constructed a dataset for Intracerebral Hemorrhage (ICH) based on real-world clinical data. The dataset comprises CT images that have been initially processed and manually classified by medical experts into categories of normal and ICH images. Moreover, the ICH images have been categorized into three types-Deep, Subcortical, and Lobar-according to the hemorrhage's location.

Additionally, we have introduced a dual-task vision Transformer (DTViT) aimed at the classification of Intracerebral hemorrhage. This innovative neural network model incorporates an encoder, which utilizes the cutting-edge vision Transformer architecture, and two decoders that are designed to classify images as ICH or normal and to determine the hemorrhage type. Our experiments have demonstrated that the DTViT has achieved a remarkable accuracy rate of 99.7% on the test data. We have also compared the DTViT with traditional models such as ResNet, VGG, and AlexNet.

To our knowledge, this research is the first to have developed both a specialized dataset and a neural network model tailored for hemorrhage location classification in a clinical context. This contribution holds substantial potential for enhancing clinical diagnosis and treatment planning.

Acknowledgments

References

[1] D Rajashekar and J.W. Liang. Intracerebral hemorrhage, Feb 2023. Updated 2023 Feb 6. In: StatPearls [Internet]. 2024 Jan-.

[2] Hemlata P. Sahu and Ramgopal Kashyap. Fine_denseiganet: Automatic medical image classification in chest ct scan using hybrid deep learning framework. *International Journal of Image and Graphics*, 0(0):2550004, 0.

[3] Xiangzuo Huo, Gang Sun, Shengwei Tian, Yan Wang, Long Yu, Jun Long, Wendong Zhang, and Aolun Li. Hifuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomedical Signal Processing and Control*, 87:105534, 2024.

[4] Tri Huynh, Aiden Nibali, and Zhen He. Semi-supervised learning for medical image classification using imbalanced training data. *Computer methods and programs in biomedicine*, 216:106628, 2022.

[5] Luis Cortés-Ferre, Miguel Angel Gutiérrez-Naranjo, Juan José Egea-Guerrero, Soledad Pérez-Sánchez, and Marcin Balcerzyk. Deep learning applied to intracranial hemorrhage detection. *Journal of Imaging*, 9(2):37, 2023.

[6] Yung-Ting Chen, Yao-Liang Chen, Yi-Yun Chen, Yu-Ting Huang, Ho-Fai Wong, Jiun-Lin Yan, and Jiun-Jie Wang. Deep learning–based brain computed tomography image classification with hyperparameter optimization through transfer learning for stroke. *Diagnostics*, 12(4):807, 2022.

[7] Qinwei Li, Lunxiao Wang, Xiaoguang Lu, Dequan Ding, Yang Zhao, Jianwei Wang, Xinze Li, Hang Wu, Guang Zhang, Ming Yu, et al. Classification and location of cerebral hemorrhage points based on sem and ssa-ga-bp neural network. *IEEE Transactions on Instrumentation and Measurement*, 2024.

[8] Nicolas Raposo, Maria Clara Zanon Zotin, David J Seiffge, Qi Li, Martina B Goeldlin, Andreas Charidimou, Ashkan Shoamanesh, Hans Rolf Jäger, Charlotte Cordonnier, Catharina JM Klijn, et al. A causal classification system for intracerebral hemorrhage subtypes. *Annals of neurology*, 93(1):16–28, 2023.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Aniwat Phaphuangwittayakul, Yi Guo, Fangli Ying, Ahmad Yahya Dawod, Salita Angkurawaranon, and Chaisiri Angkurawaranon. An optimal deep learning framework for multi-type hemorrhagic lesions detection and quantification in head ct images for traumatic brain injury. *Applied Intelligence*, pages 1–19, 2022.

[12] Miguel Altuve and Ana Pérez. Intracerebral hemorrhage detection on computed tomography images using a residual neural network. *Physica Medica*, 99:113–119, 2022.