8-2-2023

Efficient Scopeformer: Toward Scalable and Rich Feature Extraction for Intracranial Hemorrhage Detection

Yassine Barhoumi
Rowan University

Nidhal Carla Bouaynaya
Rowan University

Ghulam Rasool
Rowan University

Recommended Citation

Y. Barhoumi, N. C. Bouaynaya and G. Rasool, "Efficient Scopeformer: Toward Scalable and Rich Feature Extraction for Intracranial Hemorrhage Detection," in IEEE Access, vol. 11, pp. 81656-81671, 2023, doi: 10.1109/ACCESS.2023.3301160.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 July 2023, accepted 30 July 2023, date of publication 2 August 2023, date of current version 9 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3301160

RESEARCH ARTICLE

Efficient Scopeformer: Toward Scalable and Rich Feature Extraction for Intracranial Hemorrhage Detection

YASSINE BARHOUMI[1], NIDHAL CARLA BOUAYNAYA[1], (Member, IEEE), AND GHULAM RASOOL[2], (Member, IEEE)

[1]Electrical and Computer Science Department, Rowan University, Glassboro, NJ 08028, USA
[2]Department of Machine Learning, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

Corresponding author: Yassine Barhoumi (barhou29@students.rowan.edu)

This work was supported in part by the National Science Foundation under Award ECCS-1903466 and Award OAC-2008690.

ABSTRACT The quality and richness of feature maps extracted by convolution neural networks (CNNs) and vision Transformers (ViTs) directly relate to the robust model performance. In medical computer vision, these information-rich features are crucial for detecting rare cases within large datasets. This work presents the ''Scopeformer,'' a novel multi-CNN-ViT model for intracranial hemorrhage classification in computed tomography (CT) images. The Scopeformer architecture is scalable and modular, which allows utilizing various CNN architectures as the backbone with diversified output features and pre-training strategies. We propose effective feature projection methods to reduce redundancies among CNN-generated features and to control the input size of ViTs. Extensive experiments with various Scopeformer models show that the model performance is proportional to the number of convolutional blocks employed in the feature extractor. Using multiple strategies, including diversifying the pre-training paradigms for CNNs, different pre-training datasets, and style transfer techniques, we demonstrate an overall improvement in the model performance at various computational budgets. Later, we propose smaller compute-efficient Scopeformer versions with three different types of input and output ViT configurations. Efficient Scopeformers use four different pre-trained CNN architectures as feature extractors to increase feature richness. Our best Efficient Scopeformer model achieved an accuracy of 96.94% and a weighted logarithmic loss of 0.083 with an eight times reduction in the number of trainable parameters compared to the base Scopeformer. Another version of the Efficient Scopeformer model further reduced the parameter space by almost 17 times with negligible performance reduction. In summary, our work showed that the hybrid architectures consisting of CNNs and ViTs might provide the desired feature richness for developing accurate medical computer vision models.

INDEX TERMS Computed tomography (CT), intracranial hemorrhage, medical imaging, convolutional neural networks, vision transformers, feature maps.

I. INTRODUCTION

Stroke is a general term that describes the disruption of blood flow to the brain, resulting in the loss of one or more brain functions. It is a debilitating and potentially fatal medical condition [1], affecting millions of people worldwide [45], [46]. According to the World Health Organization [44], stroke is the second leading cause of death globally and the third leading cause of disability. In the United States alone, over 795,000 people suffer from strokes each year, with a mortality rate of approximately 15% [44]. There are two main types of stroke: ischemic stroke and hemorrhagic stroke. Ischemic stroke is caused by a blockage or obstruction in a blood vessel supplying the brain, leading to a lack of blood flow and oxygen to the affected area. Hemorrhagic stroke, on the other hand, occurs when there is bleeding in the brain, typically from a ruptured blood vessel or an aneurysm.

Early detection and accurate classification of the type of intracranial hemorrhages using head computed tomography (CT) scans are crucial for patient prognosis and treatment

The associate editor coordinating the review of this manuscript and approving it for publication was Massimo Cafaro.

[47], [48], [49]. Currently, the diagnosis of intracranial hemorrhages relies heavily on the visual interpretation of head CT scans by radiologists and neurologists. This process can be time-consuming and subjective, leading to inconsistencies in the diagnosis and treatment of patients. Diagnosis methods require qualified physicians to manually review and detect any indications of bleeding inside the cranium or the existence of a lesion within the brain tissues. This process may delay critical interventions, which can lead to serious medical complications and extensive brain damage, particularly within the first 24 hours [1], [2]. Moreover, the need for expert medical professionals can make it challenging to diagnose intracranial hemorrhages in rural or remote areas, where access to specialized medical professionals and advanced imaging technologies is limited. This highlights the importance of developing alternative methods to support diagnosis in these underserved areas, which can mitigate the risks associated with delayed or incorrect diagnoses.

In the past, several methods have been employed to diagnose intracranial hemorrhages, with varying degrees of success. Traditional approaches include a manual inspection of CT scans by expert radiologists, using standardized protocols and guidelines to identify hemorrhages [67], [68]. While these methods are effective, they can be labor-intensive and depend significantly on the expertise and availability of medical professionals. Recently, computer-aided diagnosis (CAD) systems have been developed to support radiologists in detecting intracranial hemorrhages. These systems leverage image processing techniques like segmentation, feature extraction, and pattern recognition to automatically identify and classify hemorrhages in head CT scans [69], [70]. While these approaches have yielded encouraging results in enhancing the accuracy and efficiency of intracranial hemorrhage diagnosis, they still necessitate substantial involvement from expert radiologists to validate the findings and make final diagnostic decisions. Despite these advancements, there remains an urgent need for more dependable, robust, and accessible diagnostic tools to address the challenges faced in situations where specialized medical professionals and advanced imaging technologies are scarce. The development of alternative diagnostic methods to support these underserved areas is crucial for minimizing the risks associated with delayed or incorrect diagnoses.

Machine learning algorithms trained to autonomously identify and classify brain hemorrhages can reduce the detection time, allowing quicker and more effective treatment. These emerging computer vision algorithms offer faster and more robust models that can triage patients and help expert physicians and radiologist efficiently use their time [3], [4], [5]. Therefore, developing accurate and efficient machine learning models for intracranial hemorrhage detection is crucial to improving patient outcomes and reducing the burden on the healthcare system.

In recent years, Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have received significant attention in the field of computer vision for their ability to process and analyze large amounts of visual data. CNNs utilize convolution operations and pooling layers to extract discriminative and meaningful features from an image. These features are then used to classify images or recognize objects. On the other hand, ViTs incorporate self-attention mechanisms, borrowed from transformers in Natural Language Processing (NLP), to dynamically weigh the importance of each part of an image when making predictions. This approach has proven effective in image classification and object recognition tasks, demonstrating the potential of ViTs to solve more complex computer vision problems.

Our proposed model combines multiple CNNs and a multi-encoder ViT model to improve the accuracy and efficiency of machine learning-based detection of intracranial hemorrhages. The model's hybrid architecture aims to investigate the dynamics of merging multiple ImageNet-pretrained convolutional neural networks with a vision transformer model. By doing so, we aim to develop a more generalizable model that can reinforce intracranial hemorrhage detection through the inclusion of multiple pretraining techniques and datasets from varying large distributions relevant to the task of hemorrhage detection.

Moreover, the inclusion of self-attention mechanisms borrowed from transformers allows for the incorporation of other medical imaging tasks, which facilitates the integration of these models in larger treatment planning systems [50]. Our work, therefore, aims to develop a more scalable and efficient machine learning model that can encompass large datasets and extract generalizable patterns across multiple applications, enabling the integration of these models into larger clinical settings.

Overall, our research aims to provide an accurate and efficient method for the early detection and diagnosis of intracranial hemorrhages, ultimately improving patient outcomes. By developing a more generalizable machine learning model, our work also has the potential to impact other medical imaging tasks, facilitating the development of more scalable and efficient models for clinical applications.

The model, which we called Scopeformer, accommodates (n) numbers of the CNNs dedicated to the extraction features, and several stacked ViT encoders dedicated to differentially extracting weights from the global feature map. These weights represent inter-feature correlations learned by the model as relevant for the hemorrhage classification problem. The results show that the classification accuracy is proportional to the number of CNNs used to extract the features in the Scopeformer training, leading to higher computational requirements. In this work, we present the Scopeformer model and investigate selective feature engineering methods to generate richer content for classification. We also address the large trainable parameter space issue presented by increasing the number of CNNs for feature extraction. We employ dimensionality reduction techniques on the convolution feature space to control the attention complexity within the ViT encoders. Our experiments resulted in building a scalable and efficient hybrid multi-convolution-based ViT model (n-CNN-ViT) to solve the hemorrhage detection problem. Our contributions can be summarized as:

- Proposed a hybrid architecture called Scopeformer, combining multiple CNNs and a multi-encoder ViT model for intracranial hemorrhage detection in CT images.
- Our proposed model leverages the ImageNet-pretrained off-the-shelf CNNs to generate high-level features for the vision transformer presented as tokens/patches.
- Results showed that the classification accuracy was proportional to the number of CNNs used for feature extraction, leading to higher computational requirements.
- Addressed the issue of large trainable parameter space by employing dimensionality reduction techniques on the convolution feature space to control the attention complexity within the ViT encoders.
- Developed a scalable and efficient hybrid multi-convolution-based ViT model for the hemorrhage detection problem, resulting in improved classification accuracy.

II. BACKGROUND AND LITERATURE REVIEW

A. RSNA INTRACRANIAL HEMORRHAGE DETECTION

In 2019, the Radiological Society of North America (RSNA) provided a large number of brain CT scans of healthy participants and patients with internal cerebral hemorrhage of various types. RSNA held a machine learning challenge to foster the development of autonomous algorithms for multi-class hemorrhage classification [6]. The computerized multi-label classifiers were designed to determine whether there was cerebral bleeding in each 2-dimensional (2D) slice of the input CT image and to give a probability vector with six components relative to classification targets.

B. CONVOLUTION NEURAL NETWORKS (CNNs)

Until recently, in computer vision applications, convolutional neural networks (CNNs) have been the de facto models for extracting high-resolution features for downstream tasks, e.g., classification [7], [8], [9], [10], [11], [12]. The official top-ranking solutions for the RSNA challenge, posted on the Kaggle online community platform, employed multi-stage classification models incorporating convolution-based feature extraction stage [22]. Different stacking and arrangements of convolutional layers yield different features. These features are subject to implementation variations of various configurations, such as the architecture structure, the parameters governing the visual information flow, and the depth of the model [26], [65], [66]. Several off-the-shelf architectures were proposed based on broadening the perceptual field, improving the feature extraction efficacy, and reducing the trainable parameter space for faster and efficient computation [18], [30], [31], [32], [33]. Increasing the depth of the architecture better approximates the target and provides improved feature representations due to the higher non-linearity and the improved receptive field. Enhancements were based on including and optimizing the design of the convolution layers, activation functions, loss functions, regularization methods, and optimization processes [27].

C. VISION TRANSFORMERS (ViTs)

Vision Transformers (ViTs) are increasingly being employed in a wide range of computer vision identification applications [35], [36] and have proven successful in a multitude of vision tasks such as the ImageNet classification challenge [37]. The basic working component of ViTs is the Transformer block [13], originally introduced by Vaswani et al. [14] in the realm of the Natural language process (NLP). The successful implementation of the Transformer model [14] applied to images, known as vision Transformer or ViT, was a milestone in the computer vision field [15]. Various successful implementations of the ViT architecture in the medical field were proposed that outperformed the standard convolution-based models by a large margin [16]. ViT model [15] divides a natural image into equal 3-channel square patches. These patches are flattened and represent uni-dimensional tokens. Each patch represents local semantic information of the raw image, and the model learns to extract patterns from their correlations. Finer patches result in the extraction of higher local correlations and improved semantics due to the large quadratic complexity of the model. However, this complexity results in expensive computations and large data requirements. It was shown that such models only outperform standard CNNs in high data regimes in either pre-training or training [15].

D. CONVOLUTION NEURAL NETWORKS AGAINST VISION TRANSFORMERS

Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) are two models for image recognition in computer vision, with different approaches to feature extraction. CNNs use convolutional layers to learn local patterns, resulting in a pyramid-like structure [23], while ViT uses self-attention mechanisms to process global patterns, resulting in a columnar structure [24]. CNNs handle image scale through pooling layers and are faster and more memory-efficient, while ViT uses a linear projection followed by self-attention to learn a fixed-size representation of the image, making it more effective in learning global patterns and flexible in terms of input size.

While CNNs [30], [39] integrate local information derived from input images by combining multiple convolutional operations, ViTs [15] learn patterns from spatial information and non-local dependencies exploiting the encoder block's multi-head self-attention (MHSA) function [14]. These patterns allow ViT models to gain increasingly rich global context without manually constructing layer-wise local characteristics. Applying attention to all pixels in an image increases the impact of global feature correlations, which allows the model to derive more relevant hidden patterns. As shown in [35], stacking multiple ViT encoders tends to increase the model performance, and with the appropriate training methods, a model constructed of 12 blocks of ViT encoders outperformed a ResNet model consisting of more than 30 bottleneck convolutional blocks on the ImageNet classification task [35]. However, it is shown that increasing

the depth of ViTs via stacking Transformer blocks does not necessarily increase model performance [19]. In fact, ViT model performance plateaus and starts declining beyond certain numbers of stacked encoders [19]. Zhou et al. [19] identified an attention collapse issue and proposed a new mechanism, termed *Re-attention*, that accounts for correlations among the attention multi-heads. The proposed model entitled DeepViT presented delayed plateauing behavior that enables more block stacking to achieve higher performance.

E. ATTENTION-BASED CONVOLUTION NEURAL NETWORKS

In recent years, there has been a growing interest in combining the strengths of convolutional neural networks (CNNs) and self-attention mechanisms to improve various computer vision tasks. The idea of merging these two powerful techniques stems from the desire to leverage the local feature extraction capabilities of CNNs and the long-range contextual understanding provided by self-attention mechanisms. Many studies have explored different ways to integrate self-attention into CNN architectures for tasks such as image classification [54], object detection [55], [56], video processing [57], [58], unsupervised object discovery [59], and unified text-vision tasks [60], [61]. Notably, this idea has also been applied in the field of medical AI. For instance, [62] proposed an attention-gated CNN that effectively integrated attention mechanisms for better medical image segmentation. Similarly, [63] developed a deep learning model combining CNNs and self-attention for more accurate and interpretable Alzheimer's disease diagnosis. In the pursuit of developing robust models with CNNs and the Vision Transformer (ViT), it becomes crucial to recognize the importance of integrating self-attention mechanisms into CNNs. The combination of these techniques not only enhances the potential benefits of the approach but also contributes to the evolving landscape of computer vision and medical AI applications. The existing studies, which demonstrate the successful fusion of CNN features and self-attention mechanisms, provide valuable insights and a strong foundation for future research in this area.

F. SCOPEFORMER: N-CNN-ViT HYBRID ARCHITECTURE

In our previous publications [17], [64], we introduced the Scopeformer, a hybrid architecture combining the strengths of convolutional neural networks (CNNs) and a vision transformer. This architecture was designed to extract high-level features from various inputs. The Scopeformer takes advantage of the ability of CNNs to capture local patterns in data and the ability of the vision transformer to capture global dependencies and relationships between different parts of the input. Combining these two architectures results in an effective and efficient model capable of performing complex feature extraction tasks. The introduction of the Scopeformer represents a significant step forward in the field and opens up new avenues for future research and development.

G. STATE-OF-THE-ART TECHNIQUES FOR RSNA HEMORRHAGE DETECTION

Several state-of-the-art solutions have been proposed for 2D-modeling on the RSNA dataset. Wang et al. [42] achieved an average accuracy of 98.3% using an ensemble approach. Other studies have proposed hybrid convolution models that incorporate ensemble techniques and gradient boosting methods. For example, Asif et al. [43] developed an architecture named Res-Inc-LGBM, which integrates ResNet101-V2, Inception-V4, and LGBM models, resulting in an average accuracy of 97.7%. Burduja et al. [7] reported a 96% accuracy on their 2D evaluation of the ResNeXt-101 32×8d and LSTM-based model. Their method combines a convolutional neural network with a long short-term memory network to capture both spatial features and their sequential correlations from the input data. It is worth noting that these state-of-the-art techniques mainly rely on multi-stage classification and model ensemble techniques. These methods are widely employed in similar competitions to enhance classification performance. However, these approaches do not leverage self-attention mechanisms, which have been shown to be effective in capturing global dependencies and relationships between different parts of the input in Vision Transformers (ViTs). To address this limitation, our proposed hybrid architecture, the Scopeformer, combines the strengths of convolutional neural networks (CNNs) and Vision Transformers (ViTs) to capture both local and global features. The Scopeformer has demonstrated comparable performances compared to existing state-of-the-art techniques on the RSNA dataset, achieving an average accuracy of 98.04%.

III. MATERIALS AND METHODS

This paper presents the Scopeformer, a hybrid multi-CNN vision transformer model, and its improved version, the Efficient Scopeformer. The model processes pseudo-RGB CT scan images through CNNs and a vision transformer to extract high-level features for classification. Our focus is to improve performance and efficiency through modifications to the architecture. The results of these modifications are analyzed and presented in this paper.

A. SCOPEFORMER

We present our hybrid n-CNN-ViT model in Figure 1. The model is composed of n number of CNN models stacked to build the feature-extractor backbone. We refer to the n-CNN-ViT model as ''Scopeformer'', derived from the ''Transformer'' (-former) and the word ''Scope-'' for the *selective feature extraction backbone* generated from the convolution blocks with deep receptive fields. Notably, Scopeformer leverages multiple pretrained CNNs to exploit their inherent inductive biases, while significantly reducing the input size for the ViT through concatenation of CNN features.

The Scopeformer model brings significant advancement in ViTs and CNNs. The main difference between a Scopeformer model and ViT resides in employing high-level features

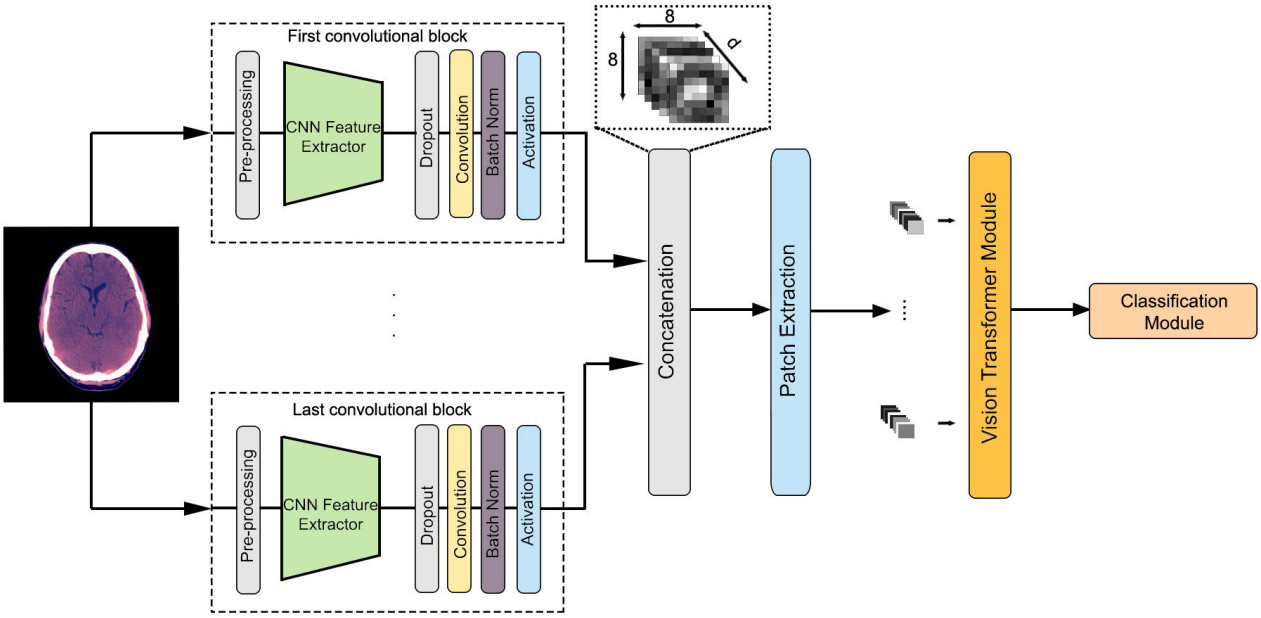
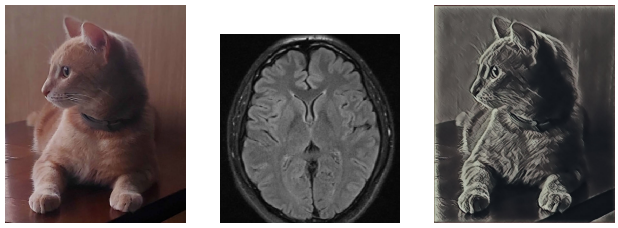


FIGURE 1. A schematic layout of the Scopeformer architecture is presented. The proposed model is composed of four main modules: (1) Scopeformer Backbone, (2) patch extraction, (3) vision Transformer (ViT) encoder, and (4) classification head. A single input image is fed to several CNN models to extract various features and construct feature maps. These feature maps are processed by the patch extraction module and vectorized. The vectors form the input to the Transformer encoder, and the model output is taken from the classification module.

with more semantic information as input to the Transformer encoder, as opposed to the originally proposed ViT model, which inputs raw natural images in the form of small patches. The Scopeformer model takes global convolutional feature maps in the form of smaller but deeper patch sizes. The ViT patch extraction method divides a natural image into patches along the height and width, then flattens every patch and joins all channels into a single 1-D token. Similarly, we pixel-wise divide the feature map along the height and width of the features into $p \times p$ patches, where p is the feature-patch size (with $p = 1$ for all the experiments in this work).

Although we did not specifically test the model without ViT encoders, removing the ViT pipeline reduces the problem to standard settings where a set of CNNs are used to detect hemorrhage [42], [43]. Our research focused on exploring the potential benefits of combining the strengths of both CNN and ViT architectures in a single model, particularly leveraging the self-attention mechanisms of the ViT encoders to capture inter-feature correlations relevant for the hemorrhage classification problem.

The input to the model consists of a tensor with a dimension of $H \times W \times C$, where H represents the height, W represents the width, and C is the number of concatenated channels derived from the RSNA DICOM files. The model executes a concurrent forward pass of the input images through different CNN architectures and stores the output features f. These features are concatenated along the channel axis. The resultant global feature map has a dimension of $h \times w \times c$, where h represents the features height, w represents the features width, and c is the total number of features with $c = n \times f$.



(a) Content image (b) Style image (c) Output image

FIGURE 2. Style transfer method applied on ImageNet dataset. (a) Content image, (b) Style image, and (c) Output image.

The first Scopeformer architecture uses Xception CNNs [18] and several ViT layers. The Xception model comprises several Inception modules composed of depth-wise and point-wise convolutions. In our Scopeformer model, we stack (n) differently pre-trained Xception models [18] in the feature extraction backbone and freeze updates on their weights during training. We use the last inception layers embedded within the Xception models as feature generators. The ImageNet pre-trained Xception CNNs, present high-level features to the ViT block. To this end, we consider that the primary role of the ViT block is to extract *correlations* from depth-wise patches. The global feature map can be generated using one or more Xception blocks stacked in the same Scopeformer as depicted in Fig. 1. Our initial experiments consider stacking raw features from CNN blocks without any further processing.

In our formulation, we tend to diversify the pre-training methods of every Xception CNN. This allows for generating different features specific to each architecture. In the first

phase of model training, we load the ImageNet pre-trained weights in all CNNs using Keras API [40]. In the second phase of training, Xception CNNs are trained to perform different classification tasks, including the RSNA hemorrhage dataset to perform classification. We used hard data augmentation on one of the CNNs and soft data augmentation on the others. We applied style transfer [38] on the ImageNet dataset to induce a grayscale brain-like image shape bias as depicted in Fig. 2. The output dataset was used to pre-train the third CNN. In our experiments, we tested several combinations of the pre-trained CNNs within the Scopeformer architecture.

B. EFFICIENT SCOPEFORMER

After extensively testing the Scopeformer model, we formulated and included several innovations in the feature extractor CNNs and the ViT blocks. We define four modules as presented in Fig. 1. The first module is the Scopeformer Backbone and represents the stack of multiple CNNs contributing to the global feature map. The second module is designed for patch extraction (from the CNN features) to generate ViT tokens. The third module consists of the ViT pipeline. Finally, the fourth module represents the classification head. We discuss these modules in the following sections.

1) MODULE 1: SCOPEFORMER BACKBONE

Efficient Scopeformer uses a variety of CNNs to build the feature extraction block. The backbone CNNs include ImageNet-pre-trained ResNet 152 V2 [30], EfficientNet B5 [31], DenseNet 201 [32], and Xception [18]. The features generated by each CNN are concatenated along the channel axis to form a *global feature map*. However, constructing such a feature map requires that the individual feature maps generated by each CNN have identical height and width. We propose augmenting each CNN with a single trainable 1×1 convolutional layer that projects the features to an appropriate space. The input to the Efficient Scopeformer consists of a tensor with a dimension of $H \times W \times 3$, where H represents the height, W represents the width, and 3 is the number of channels. The image is concurrently fed to four CNNs to generate high-level feature maps. The channel dimension of all four feature maps will be reduced using 1×1 convolution layer to $8 \times 8 \times \frac{d}{4}$, where d is the size of the *global feature map*.

2) MODULE 2: PATCH EXTRACTION

The input dimension of the second module depends on the size of the *global feature map* set by the first module. In our experiments, the resultant *global feature map* is a 3D tensor with a shape of $8 \times 8 \times d$. The patch extraction module splits the features across the height and width channel-wise and extracts $N = \frac{8\times 8}{p^2}$ d-dimensional vectors. We set the patch size to 1×1 and get $N = 64$ tokens representing one local pixel position of features across all the d features. The dimension d is controlled by the projection method used in the previous module and represents a bottleneck of the architecture. Every patch contains semantic information on the local pixel position across all the generated features from the four CNNs. The resultant sequence of flattened patches $X_p \in \mathbb{R}^{64\times d}$ is then used as the input set for the ViT block.

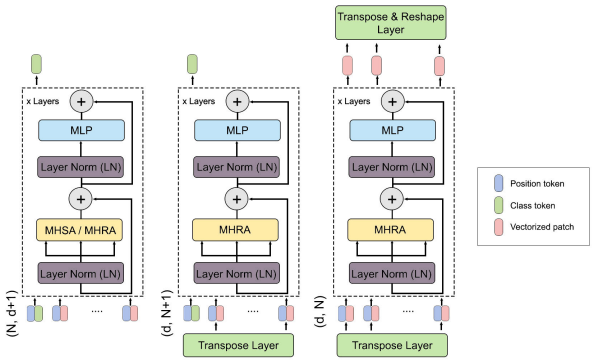


FIGURE 3. ViT Scopeformer configurations. (Left) Baseline Scopeformer Configuration: The first configuration is a ViT block with an input of vectorized patches extracted from the CNNs features. (Center) Deep Scopeformer TR Configuration: The second configuration introduces a transpose layer to transform the channel-wise patches into feature-wise patches. (Right) Efficient Scopeformer Configuration: The third configuration dismisses the token class and uses all the feature tokens as input. The output of the third block will be transposed to retrieve back the dimension of the CNN features, which we feed to the classification module.

3) MODULE 3: SCOPEFORMER ViT

We evaluated three different ViT configurations for the proposed architecture as depicted in Figure 3. These configurations include (1) Deep Scopeformer, (2) deep Scopeformer TR (Transpose), and (3) Efficient Scopeformer.

a: BASELINE SCOPEFORMER CONFIGURATION

In this configuration, we feed a set of vectors generated by the patch extraction layer to ViT encoders. We used trainable position encoding vectors coupled with vectorized patches and a trainable class (CLS) token. The dimension of the input to the ViT encoder block is $Y \in \mathbb{R}^{N\times d+1}$. We used two self-attention variants. The first one is referred to as multi-head self-attention (MHSA) [15] and the second variant as the multi-head re-attention (MHRA) [19]. The key difference resides in the introduction of a trainable transformation matrix. These variants are given by:

$$\text{MHSA}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

$$\text{MHRA}(Q, K, V) = \text{Norm}\left(M^T\left(\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\right)\right)V, \tag{2}$$

where $M \in \mathbb{R}^{h\times h}$ is a learnable transformation matrix, and h is the number of self-attention heads.

b: DEEP SCOPEFORMER TR CONFIGURATION

The second Scopeformer ViT configuration applies a transpose operation to the set of vectors produced by the patch

extraction layer. The output of the transpose layer is summed up with the position-encoded vectors and concatenated with the CLS token. The dimension of the resultant set of vectors is $Y_T \in \mathbb{R}^{d \times N+1}$. We used only the MHRA self-attention variant (Eq. 2) in our experiments.

c: EFFICIENT SCOPEFORMER CONFIGURATION

The third Scopeformer module discards the CLS notion used in previous configurations. In these settings, we use all the features generated by ViT encoders for classification. As such, the dimension of the input and output of ViT encoders remain identical and equal to $Y_T \in \mathbb{R}^{d \times N}$. We use a Transpose and Reshape layer at the ViT output to get the appropriate dimension for the feature map. We use the MHRA self-attention variant to compute self-attention.

4) MODULE 4: CLASSIFICATION MODULE

The classification module in baseline and the deep Scopeformer TR configurations receives a single CLS token. The output of this token is turned into a prediction using a multi-layer perceptron (MLP) with a sigmoid activation function and a single hidden layer. In the efficient Scopeformer configuration, the classification module receives a set of reshaped features $x_t \in \mathbb{R}^{8 \times 8 \times d}$. The classification module applies a 2D average pooling layer, followed by a flattened layer. Finally, the class inference is made via a dense layer with a sigmoid activation function.

C. DATASET

The RSNA dataset was collected by Adam et al. [6] from multiple scanner types used in different institutions worldwide. The dataset is considered the current largest dataset publicly available, aimed to capture complex real-world details of the hemorrhage subtypes. The dataset was publicly released in the 2019 Intracranial Hemorrhage (ICH) detection challenge hosted by the Kaggle platform. The dataset contains 870,301 annotated 16-bit grayscale computer tomography (CT) scans saved in the DICOM format, annotated with five types of hemorrhage. Trained physicians categorized each CT slice with one or more types of a brain hemorrhage. Five different forms of hemorrhages are to be identified in this competition, with an additional class representing the presence of any hemorrhage type in the provided slice. These classes were labeled as Epidural hemorrhage (EDH), Intraparenchymal hemorrhage (IPH), Intraventricular hemorrhage (IVH), subarachnoid hemorrhage (SAH), and Subdural hemorrhage (SDH).

D. PRE-PROCESSING

In this study, we utilized the standardized Hounsfield unit (HU) windowing technique to enhance the contrast of the CT scans and isolate regions of interest. Individual images consist of pixels that have a range of 0 to 2^{16} with a resolution of 256^2, with HU values indicative of the density of the scanned matter [7], [21]. To ensure standardization of the HU ranges across the dataset, we used the Hounsfield ranges provided in the metadata of the RSNA CT scan DICOM files during registration of the scans [25].

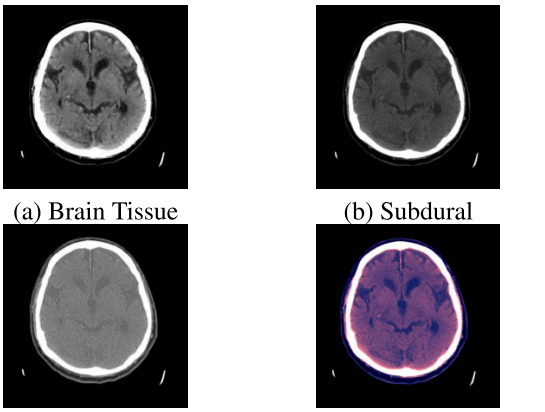


FIGURE 4. Hounsfield unit CT slice conversion and the corresponding stacked 3-channel image.

TABLE 1. Various design configurations of Scopeformer - hyperparameters and learnable parameters.

Model	CNN Blocks	Layers	Feature Size	MLP	Heads	Parameters
Scopeformer (S)	4	8	516	3072	12	34 M
Scopeformer (B)	4	8	512	4096	16	42 M
Scopeformer (M)	4	8	512	5120	16	43 M
Scopeformer (L)/4	4	4	1024	4096	16	51 M
Scopeformer (L)/8	4	8	1024	4096	16	102 M
Scopeformer (L)/16	4	16	1024	4096	16	203 M
Deep Scopeformer (L)/8	4	8	1024	4096	16	102 M
Deep Scopeformer TR (L)/8	3	8	384	4096	16	**6 M**
Efficient Scopeformer	3	8	384	4096	16	**6 M**
Scopeformer [17]	3	12	3072	3072	8	755 M
Scaled Scopeformer [17]	4	8	4096	4096	16	870 M

In our pipeline, we used three HU windows as channels in the input of the Scopeformer model, as depicted in Fig. 4. Specifically, we applied brain windowing with HU values ranging between 40 and 80, subdural windowing with HU values ranging between 80 and 200, and soft tissue windowing with HU values ranging between 40 and 380 [7]. These windowing techniques have been shown to be effective in enhancing the visualization of intracranial hemorrhages in CT scans [51], [52], [53].

To further standardize the input images, we rescaled the HU values between the minimum and maximum values of the respective HU window. This rescaling step ensures that the HU values are normalized and consistent across the input images, facilitating the training of the machine learning model.

E. DESIGN CONFIGURATIONS AND EXPERIMENTS

Details about the various Scopeformer hyperparameter configurations and architectures are presented in Table 1. We present the different proposed Scopeformer variations and details about the number of convolution models used in the feature extraction backbone, the number of ViT layers, the global feature map size, the MLP dimension and the number of heads in each ViT block, and the total number of trainable parameters. We compare our Efficient Scopeformer implementation to our initial model implementation and propose lower trainable parameter space given the configuration hyperparameters. Our experiments comprise four main parts.

In the first set of experiments, we evaluate the size effect of various variants of Scopeformer on the classification accuracy, where four variants are evaluated; small (S), base (B), medium (M), and large (L). We keep the number of ViT layers fixed (equals 8) and increase the complexity of the model by configuring the MLP size residing in the ViT blocks for S, B, and M variants and increasing the feature size for the L variant. The number of trainable parameters drastically increases from the smallest (S) to the largest (L) variants.

In the second set of experiments, we investigate the effect of the number of ViT encoder blocks on the model performance. Based on preliminary results conducted in the first set of experiments, we conduct our ablation study on the large Scopeformer variant (L) with a feature size of 1024 and an MLP dimension of 4096. We consider three experiments where we gradually stack in an end-to-end fashion 4, 8, and 16 ViT encoders, forming three models named Scopeformer (L)/4, Scopeformer (L)/8, and Scopeformer (L)/16, respectively. Given the largest model parameters reside within the ViT architecture, the total number of trainable parameters is linearly scaled to the number of ViT blocks we use.

The third set of experiments examines the transition from the originally proposed ViT model [15] to a different version called DeepViT [19]. We test this configuration on the highest performing model from the previous two sets of experiments; Scopeformer (L)/8 with a global feature map size of 1024, 8 layers of ViT encoders, and an MLP dimension of 4096. The model version, entitled Deep Scopeformer (L)/8, has a slightly higher number of trainable parameters.

The final experiment introduces three different ViT configurations to our Scopeformer architecture as depicted in Figure 6. We add these configurations to the highest-performing model from the previous three parts of the study; Deep Scopeformer (L)/8 with a global feature map size of 1024, 8 layers of ViT encoders, and an MLP dimension of 4096. We introduce and compare a set of three Scopeformer configurations, as presented in *section 2.2.3*; Baseline Scopeformer configuration, Deep Scopeformer-TR configuration, and Efficient Scopeformer configuration.

F. PRE-TRAINING EFFICIENT SCOPEFORMER

In all the experiments, we initially pre-trained the Scopeformer model using the ImageNet-1k dataset [20]. Later, we train all models using the RSNA dataset [6]. In the first module (convolutional backbone), we freeze $\approx$ 70% of the layer weights in each CNN and keep top $\approx$ 30% trainable along with the newly introduced 1×1 convolution layer. In our last experiment using the Efficient Scopeformer model, we pre-trained the backbone neural network on the RSNA dataset for hemorrhage classification for 150 epochs on top of the defaulted pre-training on ImageNet-1k. In this experiment, denoted as Efficient Scopeformer (p), we freeze weights of the feature extraction block during training.

G. THE LOSS FUNCTION

Following guidelines from the RSNA Intracranial Haemorrhage Challenge (ICH), we adopted a weighted version of the *multi-label logarithmic loss* function for our model training. The weighting was introduced to amplify the importance of classifying the first class representing all types of hemorrhages, with a coefficient of 2, at the expense of the rest of the classes, which have coefficients of 1. The evaluation of the loss value with respect to a single instance represents the weighted average over all the binary losses computed on each class individually. The ICH represents a multi-label classification problem, i.e., the input image can be classified into multiple classes, using binary labeling for each class to indicate its presence or absence. In our formulation, we applied multi-label hot encoding on the dataset to assign a binary value on each class for every CT slice. The *multi-label logarithmic loss* function is defined as follows:

$$L_{\text{multi-BCE}}(y, \tilde{y}) = -\sum_{n=1}^{6} \alpha_{\text{n}} (y_n \log \tilde{y_n} + (1 - y_n) \log (1 - \tilde{y_n})), \quad (3)$$

where α_{n} represents the coefficient of the target classes, y_n represents the ground-truth of each class n, and $\tilde{y_n}$ is the corresponding predicted probabilities.

H. EVALUATION METRICS

The official model evaluation metric in the RSNA IHC was the *weighted accuracy*. We evaluate the overall performance of the models based on three metrics, (1) the classification accuracy on the RSNA dataset, (2) the visual evaluation of the global feature richness of the embedding layer generated by the convolution backbone, and (3) the ratio of the model size function to the total number of trainable parameters.

IV. RESULTS AND DISCUSSION

A. THE EFFECT OF BACKBONE NEURAL NETWORK MODEL SIZE AND PRETRAINING TECHNIQUES

We gradually stack *n* various pre-trained Xception models in the feature extraction backbone. We freeze all these architectures in the backbone to prevent updates on their weights during training. We pre-trained the CNN models on diversified pre-training schemes, including ImageNet-1k natural image dataset (I) and the generated style transfer-base dataset (S). Table 2 compares different models and the corresponding performances on the hemorrhage classification

TABLE 2. Classification performance of ViT-based models on the RSNA validation dataset. (I) represents pre-training on ImageNet. (S) represents pre-training on ImageNet with style transfer images.

Model	ViT input dimension	Validation accuracy	Loss
ViT	256×256×3	94.33%	0.1822
1-CNN-ViT (S)	7×7×1024	96.95%	0.08272
2-CNN-ViT (I-I)	7×7×2048	97.22%	0.07984
2-CNN-ViT (S-S)	7×7×2048	97.26%	0.07934
2-CNN-ViT (I-S)	7×7×2048	97.46%	0.07754
3-CNN-ViT (I-I-S)	7×7×3072	98.04%	0.07050

task. While the ViT component in the n-CNN-ViT models were trained on the convolution features generated by the convolution backbone, the ViT model was trained on the raw images dataset. The input dimension of the ViT block represents the full-resolution image or the set of features before splitting into patches. Results show that extracting features using convolution models to train the ViT model is a better alternative to the raw dataset. The Scopeformer model exploits the pre-training for generating high-level features useful for the ViT architecture. The use of CNNs leverages the need for high data regimes since the ViT model is used to fit these high-level features and extract semantic correlations instead of learning the spatial features in training. Furthermore, results show that the classification accuracy is proportional to the number of CNN models used in the Scopeformer training, i.e., as we stack feature extraction architectures in the backbone of the model, we get higher performances on hemorrhage classification. We further boost these performances by selectively varying the pre-training methods for each CNN architecture. We hypothesize that increasing the feature map size in the ViT input allows for increased semantic correlation extractions by the ViT block. Furthermore, diversifying the inductive biases derived from differently pre-trained CNN architectures may lead to a different set of feature maps, which contributes to a richer feature map and leads to observed improved performances.

B. THE EFFECT OF SCOPEFORMER SIZE

Tables 3 and 4 show the results of experiments performed with different variants of the Scopeformer model. Table 4 depicts different results obtained on individual classes of the S, B, and M models. We propose four sizes of the Scopeformer model, S, B, M, and L, with a reduced number of trainable parameters compared to our initial implementation of the Scopeformer model involving several Xception-based CNNs. The key component to the parameter reductions is linked to the trainable 1×1 *convolutional layer* placed after each convolution architecture in the feature extraction backbone before concatenation. In this experiment, we gradually increase the model complexity of S, B, and M variants by varying the MLP dimension and the number of self-attention heads within the ViT module as depicted in table 1.

TABLE 3. Performance of the different Scopeformer variants.

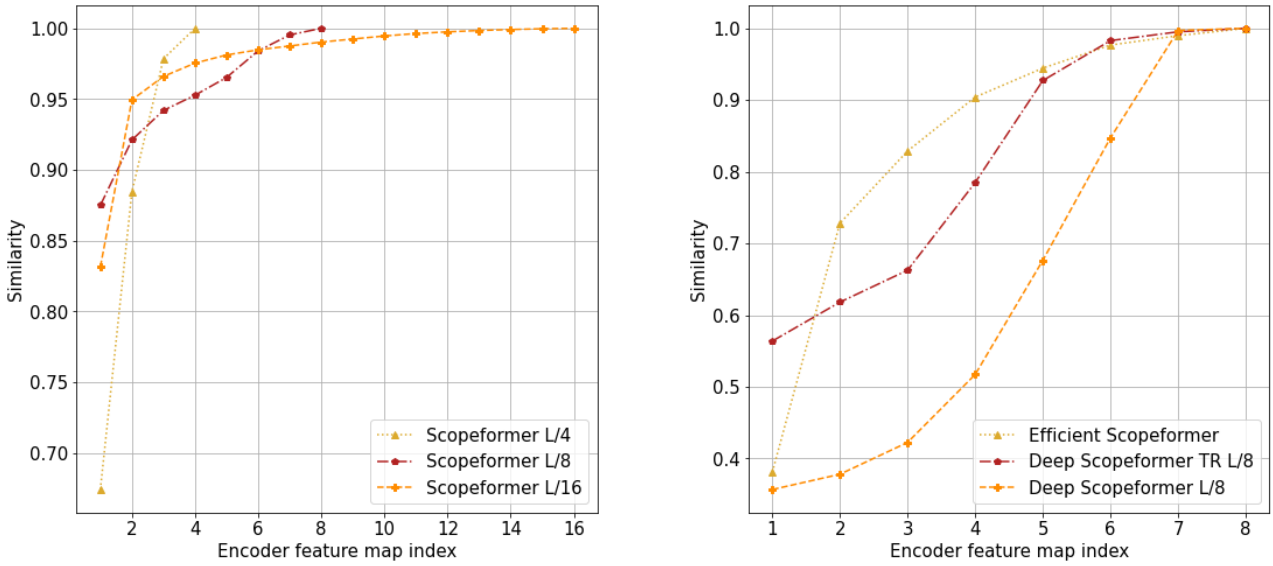
Model	Accuracy	Loss	Recall	Trainable Parameters
Small (S)	93.00%	0.1703	84.95%	34M
Base (B)	93.92%	0.1461	89.29%	42M
Medium (M)	93.88%	0.2285	88.44%	43M
Large (L)/4	93.12%	0.1378	87.81%	51M
Large (L)/8	**94.69%**	**0.1197**	**89.33%**	102M
Large (L)/16	92.57%	0.1395	87.34%	203M

In table 3, we note that the base model outperforms the small and medium variants. However, in Table 4, we observe that the Base model performs better on IPH, IVH, and SAH classes, whereas the Small model shows higher accuracy for epidural, SDH, and all classes. Based on these observations, we hypothesize that the improved performance observed on higher MLP dimensions indicates the ability of the model to encompass a larger amount of information and extract useful semantics for classification. However, the model shows signs of overfitting when the MLP dimension reaches 5120. Based on these results, we build our large *Scopeformer (L)/8* model by adopting the configuration of the base variant with a global feature dimension $d = 1024$. The feature size increment resulted in a proportional increment of the model trainable parameters. The large model (L)/8 performed the best among the proposed variants. The improved performance observed on larger ViT sizes while increasing the input feature embedding space indicates richer information brought by these added features, where the model extracted useful semantics for classification. Increasing the feature space improved some of the classes at the expense of others, as evident from Table 4.

C. THE EFFECT OF NUMBER OF ViT ENCODERS

We evaluate the effect of the number of ViT encoders on the Scopeformer (L)/8 model using 4, 8, and 16 encoders. As presented in Table 1 and Table 3, the number of parameters scales linearly with the number of encoders. We note that using 8 ViT encoders yields better results than a shallower model with 4 ViT encoders. However, a deeper model with 16 ViT encoders drastically reduces the model performance. We conclude that increasing the depth of the ViT model does not scale linearly and that there is a critical number of ViT encoders where the model performs optimally.

In Figure 5, we plot the cosine similarity between the features generated by each ViT encoder and the last layer of the model. We observe that similarities across features generated by each ViT encoder rapidly increase for all proposed models. These similarities further increase in models with higher numbers of ViT encoders. We believe that the increased similarities among the features of the Scopeformer(L)/16



(a) Scopeformer variants

(b) Deep Scopeformer variants

FIGURE 5. Cosine similarity of the ViT encoder feature maps with respect to the last encoder feature map. We observe the increased similarities across ViT encoder features function to the depth of Scopeformer models.

model may have contributed to the performance decline observed in Table 3. Similarly, reduced similarities among ViT features observed on Scopeformer(L)/4 may explain the observed sub-optimal performance. From these results, we conclude that the cosine similarity can be a good metric for model performance, as reduced or increased similarities may indicate sub-optimal performances of the Scopeformer model. Shallow models presenting reduced similarities may hint at higher performances by stacking more ViT layers, whereas deeper models may require additional data to reduce similarities across ViT features to perform optimally. The results also suggest that there is an optimum number of ViT encoders for the Scopeformer model based on the complexity of the dataset and the effectiveness of the convolution backbone networks.

D. THE EFFECT OF TWO DIFFERENT SELF-ATTENTION VARIANTS

The Deep Scopeformer (L)/8 builds on the Scopeformer (L)/8 model by replacing the MHSA layer with an MHRA layer. The additional trainable matrices M add an insignificant number of parameters to the Scopeformer (L)/8 model. In Figure 5 (b), we note substantial dissimilarities among ViT encoders' features for the *Deep Scopeformer (L)/8* model. The result may imply an increased feature richness acquired by the model from the additional inter-correlations of the MHRA heads. This configuration resulted in an accuracy improvement by $+1.11\%$ as shown in Table 5.

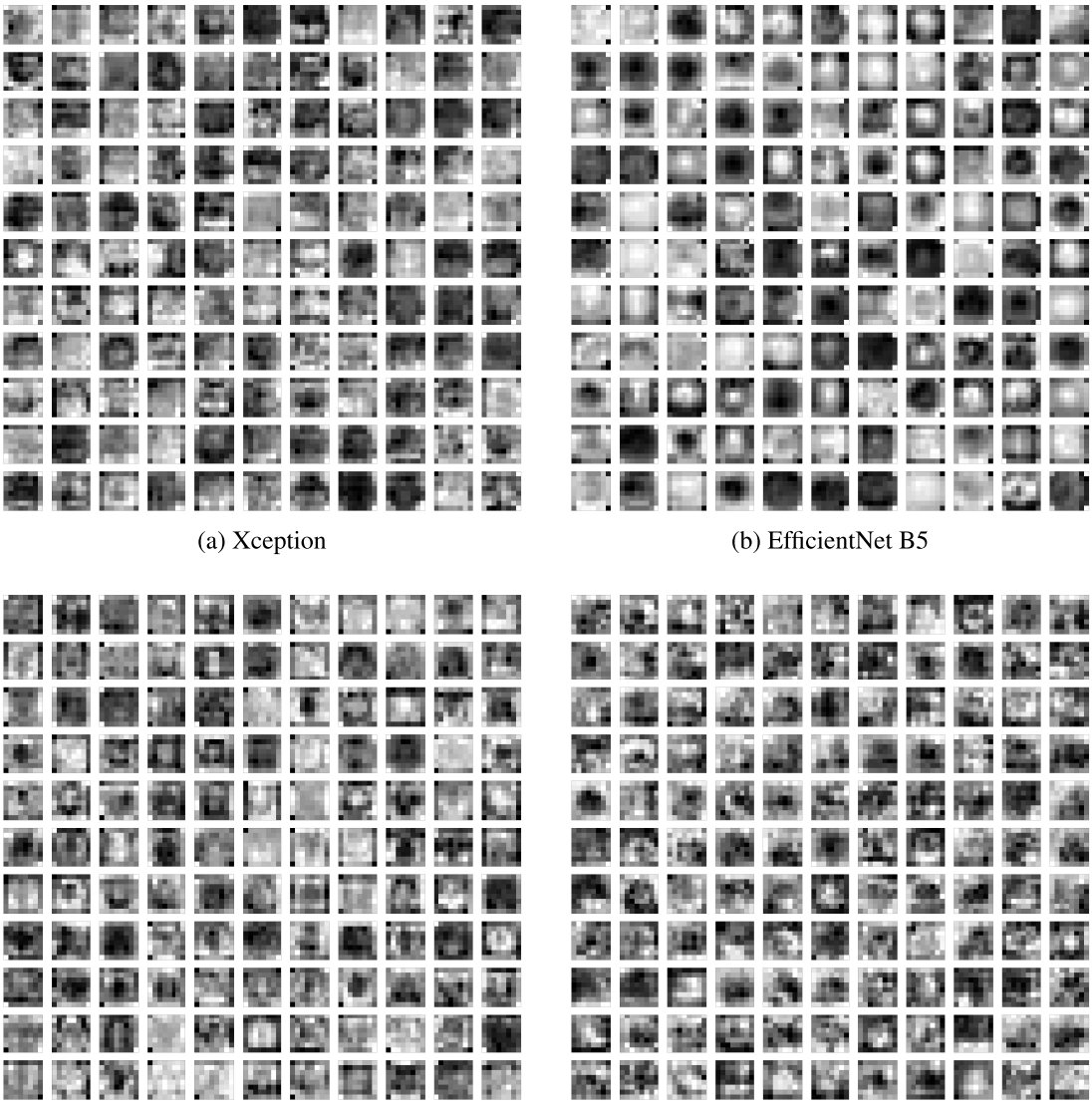
E. ViT SCOPEFORMER CONFIGURATIONS

We address the self-attention computational complexity problem by introducing a transpose layer before the ViT module. The attention weights matrix in *Deep Scopeformer (L)/8* has a dimension of 1024^2. In the second and third ViT configurations, the attention weights matrices have dimensions of 65^2 and 64^2, respectively. The use of the transpose layer has substantially contributed to the reduction of the number of trainable parameters as indicated in Table 1. This is due to the MHRA quadratic reduction in computation complexity. Additionally, transposing the input sequence effectively preserved the feature content retrieved by the feature extractor module and conserved the classification performance. Table 5 shows the performance of the three proposed configurations. The proposed *Efficient Scopeformer* variant performed relatively better than the *Deep Scopeformer (L)/8* for a lower trainable parameter space. We speculate that the role of the ViT module in this configuration is to improve the global feature map that was previously optimized by the convolution backbone. The global feature map improvement resides in using attention computations to generate new features characterized by inter-correlations among all features generated by the convolution networks.

Our Efficient Transformer module improved the global features map correlations and contributed to better performance. We note that for the model *Efficient Scopeformer (P)* pre-training the convolution block on the target dataset and freezing the entire block during training produces better performance than end-to-end training with around 30% trainable parameters of the Efficient Scopeformer's convolution block. We argue that backbone CNNs and ViTs present different dynamics that require different model training settings.

1) GLOBAL AND ViT FEATURE MAPS

Figures 6, 7 and 8 present convolution features generated by three Scopeformer architectures for an epidural example; Scopeformer (L)/8, Deep Scopeformer (L)/8, and



(a) Xception (b) EfficientNet B5

(c) DenseNet201 (d) ResNet152V2

FIGURE 6. Feature maps visualization of an epidural type hemorrhage example for the model Scopeformer (L)/8. These features represent intermediate internal representations that the model may find useful and use in its own learning and decision-making. These features may not be directly interpretable to the human visual system.

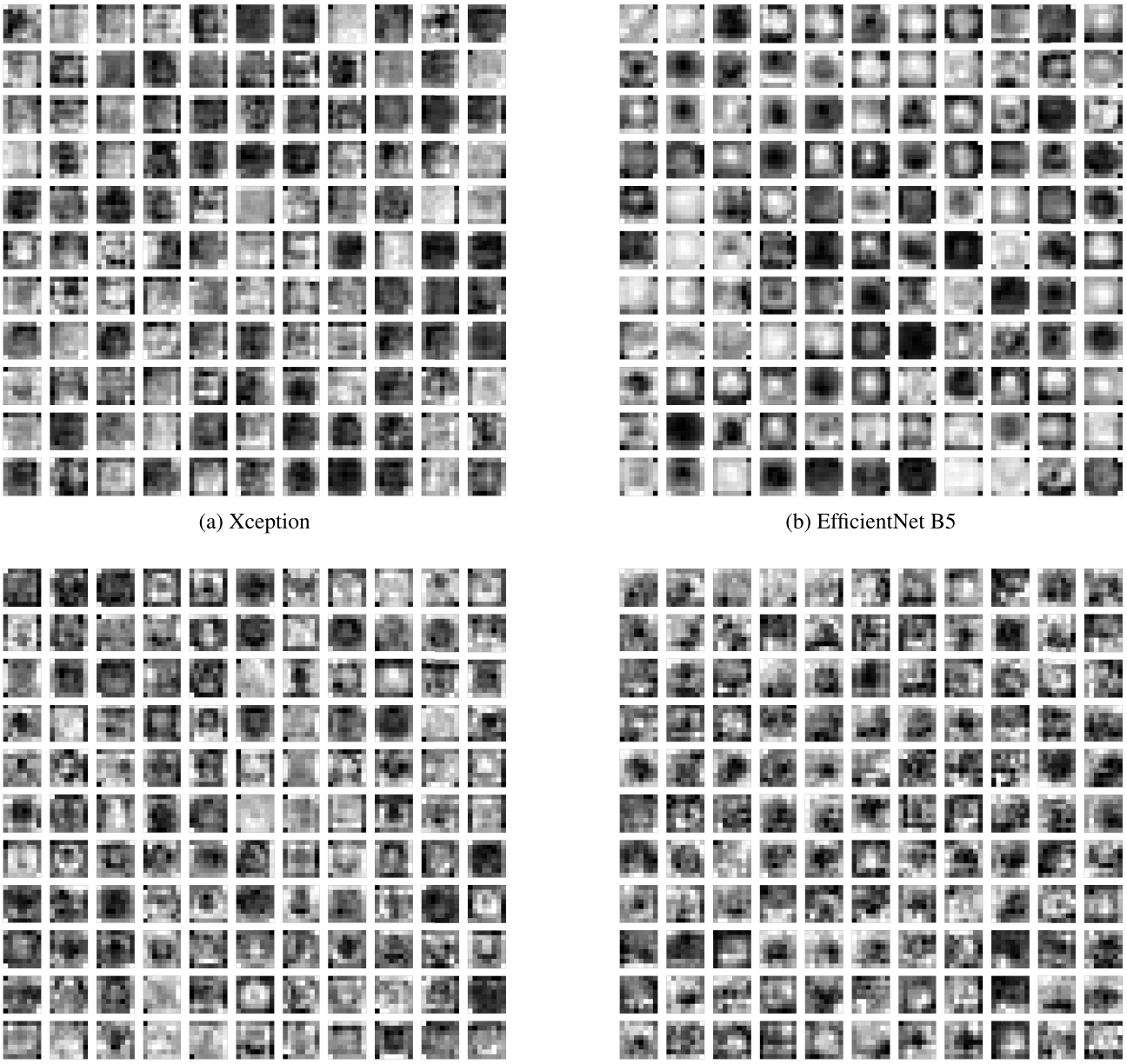
TABLE 4. Model performance on individual target classes.

	Accuracy			
	Large	Medium	Base	Small
All	**71.34%**	60.26%	70.5%	70.83%
Epidural	96.98%	90.18%	95.73%	**98.08%**
IPH	85.94%	71.10%	**87.28%**	85.95%
IVH	90.5%	70.73%	**91.72%**	90.13%
SAH	**78.69%**	65.49%	78.57%	77.04%
SDH	**77.08%**	60.78%	74.35%	74.54%

TABLE 5. Model performance for different Scopeformer modalities.

	Accuracy	Loss	Trainable Parameters
Scopeformer (L)/8	94.69%	0.1197	102M
Deep Scopeformer (L)/8	96.03%	0.1088	102M
Deep Scopeformer TR (L)/8	95.40%	0.1176	**6M**
Efficient Scopeformer	95.77%	0.1160	**6M**
Efficient Scopeformer (P)	**96.94%**	**0.0833**	**5M**

Efficient Scopeformer. We observed high variability of the features generated by each CNN architecture. Furthermore, we observe that there is no apparent similarity among the features generated by different CNNs for all Scopeformer variants. Subsequently, the resultant global feature map has low redundancy and higher feature richness. However, among these models, we note that the DenseNet model showed the highest feature redundancy across the observed



(a) Xception (b) EfficientNet B5

(c) DenseNet201 (d) ResNet152V2

FIGURE 7. Feature maps visualization of an epidural type hemorrhage example for Deep Scopeformer (L)/8. These features represent intermediate internal representations that the model may find useful and use in its own learning and decision-making. These features may not be directly interpretable to the human visual system.

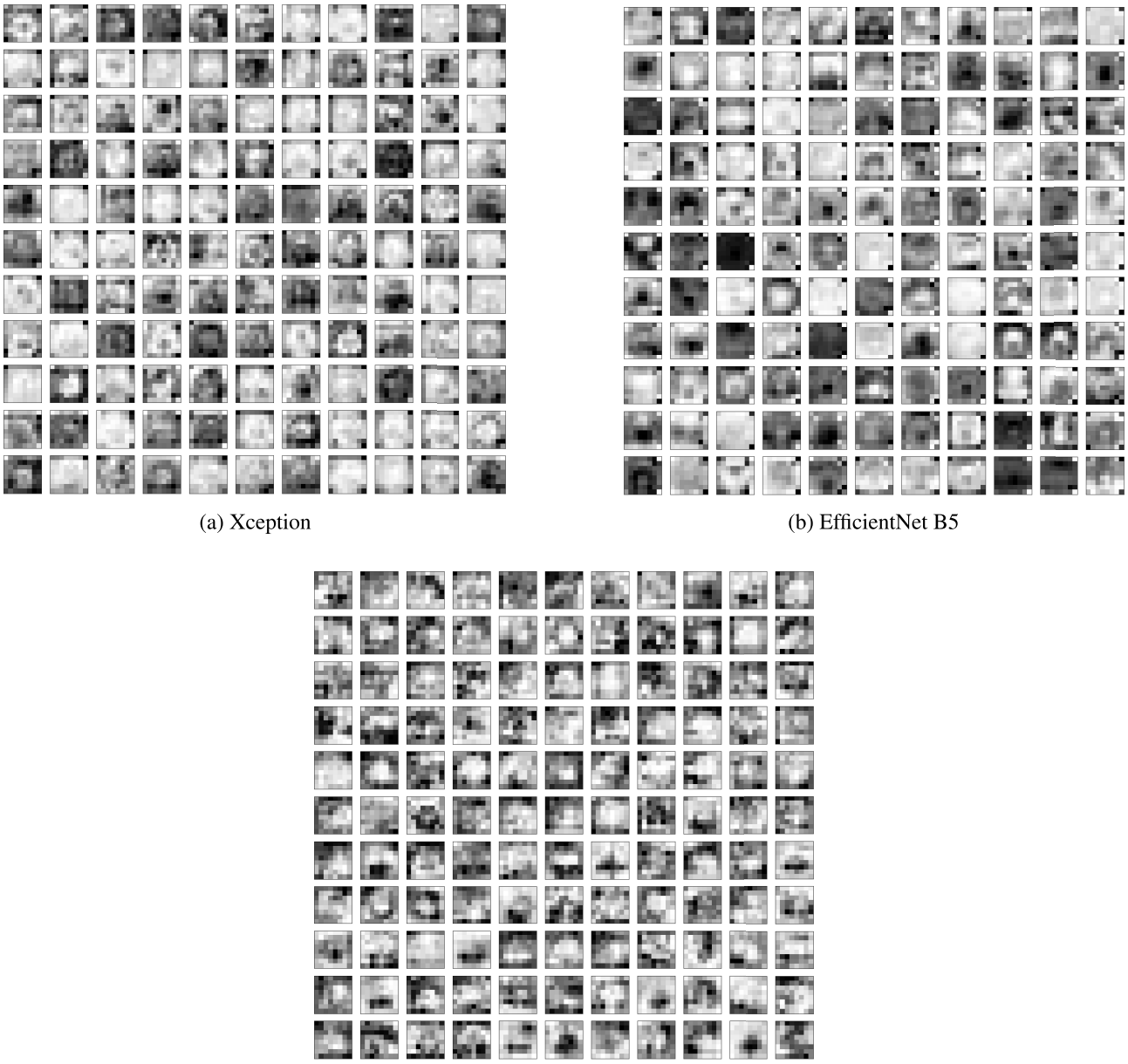
features. Therefore, we conducted an ablation study on the *Deep Scopeformer TR*, which resulted in removing the DenseNet201 model from the *Efficient Scopeformer* model backbone.

2) ATTENTION PATTERNS VISUALIZATIONS

Figure 9 shows the attention patterns visualizations of the 16 MHRA heads concerning the first and last ViT encoders. In the first ViT layer, we observe that the model extracts high correlations among features derived from every CNN architecture. This observation suggests the high similarities among the input features of every CNN model. Each head learns different correlation patterns among the set of features. However, deeper into the model, we observe that the model learns to extract global correlation patterns across all the CNN features. The generated set of features adds information about the relevance of every feature to the rest of the features, which contributes towards the observed higher performance.

F. OVERVIEW OF SCOPEFORMER POTENTIAL AND EFFECTIVENESS

Our original investigation aimed to demonstrate the feasibility and effectiveness of the CNN + ViT architecture in



(a) Xception

(b) EfficientNet B5

(c) ResNet152V2

FIGURE 8. Feature maps visualization of an epidural type hemorrhage example. Efficient Scopeformer.

various configurations, focusing on the potential benefits in accuracy, generalization, and robustness. The results of our experiments showed that the Scopeformer model was competitive and had the potential to achieve comparable or even better performance than state-of-the-art methods. In this work, we concentrated on improving the model and exploring different configurations in an experimental study rather than strictly comparing it with existing methods. As our proposed Scopeformer model presented a novel approach that combined multiple CNNs and a ViT model in a unique architecture, a direct comparison with existing methods might not have fully captured the essence of our contribution. Our primary goal was to pave the way for higher data regimes and multi-disciplinary medical AI applications within a single model, such as hemorrhage detection, organ-at-risk identification, and tumor detection, all residing in the brain and utilizing different types of imagery like CT, PET, and MRI.

In selecting the specific CNN models used in our study, we were guided by several factors. We conducted ablation studies to analyze the impact of different CNN architectures on the overall performance of our model. By including a variety of CNN models, we aimed to investigate how each model contributed to the results. We also ensured that the output dimensions of the selected CNN models were uniform, specifically 8×8, for building the Efficient Scopeformer. This uniformity allowed us to concatenate the features easily and effectively. Furthermore, the Xception model was chosen to initiate our study with preliminary results using multiple off-the-shelf CNN architectures. We initially used the Xception

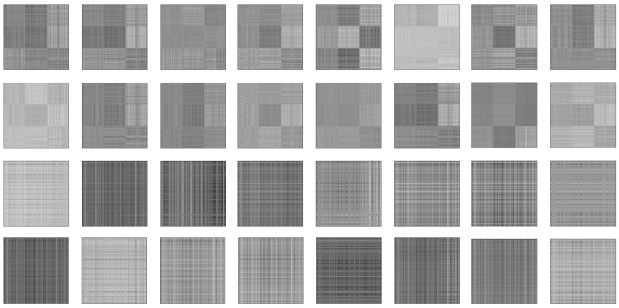


FIGURE 9. Attention pattern visualization of the Efficient Scopeformer model. The first and second rows represent the 16 attention heads of the first encoder layer. The third and fourth rows represent the 16 attention heads of the last encoder layer. Each attention map has a dimension of 384 × 384.

model by itself to highlight the effect of pretraining different CNNs without changing the architecture. It also served as a reference point for comparison with the other models used in our study. This approach allowed us to investigate the effectiveness of different CNN architectures and their impact on the overall performance of the hybrid model.

We also considered the time complexity of our proposed methodology, particularly when comparing the hybrid CNN-ViT model with a standalone ViT model applied directly on raw images. The forward pass of one or multiple CNNs, with frozen weights, has a time complexity in the order of milliseconds. If the feature space of the CNNs is similar to the image patches input to the ViT model, the computational complexity will be roughly the same in both cases. In our preliminary experiments, we verified this by testing our model with roughly identical input sizes to the ViT in both cases (with and without the CNNs) and found no significant issues with computation. Regarding the convergence time, it may vary between the models, but it is essential to note that training ViTs on raw images can be time-consuming. In our study, we focused on training the model on features and maintaining the training time within the acceptable range, comparable to that of training ViTs on raw images. Additionally, our model was designed to be efficient in terms of the number of trainable parameters, which significantly reduced the time complexity. Our main goal was to improve the model's ability to learn from the features and their intra and inter correlations, as opposed to learning from raw images, which enabled enhancing generalizability and potential use in multi-domain medical AI fields. Moving forward, we will continue to refine the Scopeformer architecture and investigate additional applications within the medical AI field.

V. CONCLUSION

We proposed a set of convolutional-based ViT models called Scopeformer to address the challenging problem of classifying types of hemorrhage in brain CT scans. We defined a range of model architectures for both CNNs and ViTs. We explored the effect of using multiple off-the-shelf CNN models on the global feature richness of the architecture and investigated a feature projection method to reduce the large redundant feature space into a lower and more efficient one. We conducted a parametric optimization study to evaluate the size effects on model performance and efficiency. We implemented three ViT configurations to evaluate the re-attention module within the Scopeformer model and the channel-wise versus feature-wise patch extraction of the global feature map. Results show increased richness of the features due to different CNN architectures. The re-attention module increased dissimilarities of ViT features resulting in improved performances and allowing deeper models. With our proposed feature-wise patch extraction method, the model size was reduced 17 times with comparable performance. Our Efficient Transformer module improved the global features map correlations and contributed to better performance. Furthermore, we observed that pre-training the convolution block on the target dataset and freezing the whole block during training produces better results than end-to-end training with 30% trainable parameters of the Efficient Scopeformer's convolution block.

REFERENCES

[1] J. Elliott and M. Smith, "The acute management of intracerebral hemorrhage: A clinical review," *Anesthesia Analgesia*, vol. 110, no. 5, pp. 1419–1427, May 2010.

[2] J. M. Wardlaw, "Overview of cochrane thrombolysis meta-analysis," *Neurology*, vol. 57, pp. S69–S76, Sep. 2001.

[3] T. Gong, R. Liu, C. L. Tan, N. Farzad, C. K. Lee, B. C. Pang, Q. Tian, S. Tang, and Z. Zhang, "Classification of CT brain images of head trauma," in *Proc. 2nd IAPR Int. Workshop Pattern Recognit. Bioinf. (PRIB)*, Singapore, Oct. 2007. [Online]. Available: https://doi.org/10.1007/978-3-540-75286-8_38

[4] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, and P. Warier, "Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study," *Lancet*, vol. 392, no. 10162, pp. 2388–2396, Dec. 2018.

[5] A. Patel, S. C. van de Leemput, M. Prokop, B. Van Ginneken, and R. Manniesing, "Image level training and prediction: Intracranial hemorrhage identification in 3D non-contrast CT," *IEEE Access*, vol. 7, pp. 92355–92364, 2019.

[6] A. E. Flanders et al., "Construction of a machine learning dataset through collaboration: The RSNA 2019 brain CT hemorrhage challenge," *Radiol., Artif. Intell.*, vol. 2, no. 4, Jul. 2020, Art. no. e209002.

[7] M. Burduja, R. T. Ionescu, and N. Verga, "Accurate and efficient intracranial hemorrhage detection and subtype classification in 3D CT scans with convolutional and long short-term memory neural networks," *Sensors*, vol. 20, no. 19, p. 5611, Oct. 2020.

[8] M. DelRocini, C. Angelini, and G. Rasool, "Identification of abnormalities in head computerized tomography scans," in *Proc. IEEE SPMB*, Dec. 2020, pp. 1–4.

[9] G. Carannante, D. Dera, N. C. Bouaynaya, H. M. Fathallah-Shaykh, and G. Rasool, "SUPER-Net: Trustworthy medical image segmentation with uncertainty propagation in encoder–decoder networks," 2021, *arXiv:2111.05978*.

[10] D. Dera, G. Rasool, and N. Bouaynaya, "Extended variational inference for propagating uncertainty in convolutional neural networks," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2019, pp. 1–6.

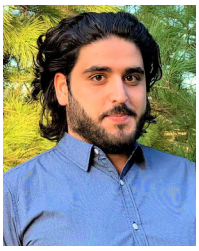
[11] A. Waqas, D. Dera, G. Rasool, N. C. Bouaynaya, and H. M. Fathallah-Shaykh, "Brain tumor segmentation and surveillance with deep artificial neural networks," in *Deep Learning for Biomedical Data Analysis*. Berlin, Germany: Springer, 2021, pp. 311–350.

[12] D. Dera, N. C. Bouaynaya, G. Rasool, R. Shterenberg, and H. M. Fathallah-Shaykh, "PremiUm-CNN: Propagating uncertainty towards robust convolutional neural networks," *IEEE Trans. Signal Process.*, vol. 69, pp. 4669–4684, 2021.

[13] S. Ahmed, I. E. Nielsen, A. Tripathi, S. Siddiqui, G. Rasool, and R. P. Ramachandran, "Transformers in time-series analysis: A tutorial," 2022, *arXiv:2205.01138*.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[16] Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, Jul. 2021.
[17] Y. Barhoumi and G. Rasool, "Scopeformer: N-CNN-ViT hybrid model for intracranial hemorrhage classification," 2021, *arXiv:2107.04575*.
[18] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[19] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.
[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[21] J. Broder and R. Preston, "Imaging the head and brain," in *Diagnostic Imaging for the Emergency Physician*. Philadelphia, PA, USA: Saunders, 2011, pp. 1–45.
[22] *RSNA Intracranial Hemorrhage Detection Challenge (2019)*. Radiological Society of North America. Accessed: Jul. 24, 2023. [Online]. Available: https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-intracranial-hemorrhage-detection-challenge-2019
[23] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.
[24] Z. Zhang and W. Zhang, "Pyramid medical transformer for medical image segmentation," 2021, *arXiv:2104.14702*.
[25] J. Muschelli, "Recommendations for processing head CT data," *Frontiers Neuroinform.*, vol. 13, p. 61, Sep. 2019, doi: 10.3389/fninf.2019.00061.
[26] M. Aubry and B. C. Russell, "Understanding deep features with computer-generated imagery," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2875–2883.
[27] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.
[28] W. Chen, Y. Liu, W. Wang, E. M. Bakker, T. Georgiou, P. Fieguth, L. Liu, and M. S. Lew, "Deep learning for instance retrieval: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7270–7292, Jun. 2023.
[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Institute of Electrical and Electronics Engineers Inc., 2015, doi: 10.1109/CVPR.2015.7298594.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[31] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[32] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.
[33] K. Simonyan and A. Zisserman A., "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[34] D. Haehn, J. Tompkin, and H. Pfister, "Evaluating 'graphical perception' with CNNs," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 1, pp. 641–650, Jan. 2019.
[35] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers and distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[38] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 630–645. [Online]. Available: https://doi.org/10.1007/978-3-319-46493-0_38.
[40] F. Chollet. (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras
[41] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," 2018, *arXiv:1811.12231*.
[42] X. Wang, T. Shen, S. Yang, J. Lan, Y. Xu, M. Wang, J. Zhang, and X. Han, "A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head CT scans," *NeuroImage, Clin.*, vol. 32, Jan. 2021, Art. no. 102785.
[43] M. Asif, M. A. Shah, H. A. Khattak, S. Mussadiq, E. Ahmed, E. A. Nasr, and H. T. Rauf, "Intracranial hemorrhage detection using parallel deep convolutional models and boosting mechanism," *Diagnostics*, vol. 13, no. 4, p. 652, Feb. 2023.
[44] World Health Organization. (2021). *The Top 10 Causes of Death*. [Online]. Available: https:/www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
[45] E. J. Benjamin et al., "Heart disease and stroke statistics—2019 update: A report from the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.
[46] S. S. Virani et al., "Heart disease and stroke statistics—2020 update: A report from the American Heart Association," *Circulation*, vol. 141, no. 9, pp. e139–e596, 2020.
[47] A. Winkler-Schwartz, J. Stein, and T. R. Marotta, "Hemorrhagic stroke," in *Stroke*. Cham, Switzerland: Springer, 2018, pp. 161–171.
[48] J. M. Wardlaw, G. Zoppo, T. Yamaguchi, and E. Berge, "Thrombolysis for acute ischaemic stroke," *Cochrane Database Syst. Rev.*, no. 3, 2003, doi: 10.1002/14651858.CD000213.
[49] S. R. Messe, S. E. Kasner, and J. A. Chalela, "CT evaluation of intracerebral hemorrhage," *Neuroimag. Clinics North Amer.*, vol. 16, no. 2, pp. 283–298, 2006.
[50] B. Deng, W. Zhu, X. Sun, Y. Xie, W. Dan, Y. Zhan, Y. Xia, X. Liang, J. Li, Q. Shi, and J. Jiang, "Development and validation of an automatic system for intracerebral hemorrhage medical text recognition and treatment plan output," *Frontiers Aging Neurosci.*, vol. 14, Apr. 2022, Art. no. 798132, doi: 10.3389/fnagi.2022.798132.
[51] Radiopaedia. *CT Head: An Approach*. Accessed: Mar. 13, 2023. [Online]. Available: https://radiopaedia.org/articles/ct-head-an-approach?lang=gb
[52] DCStang. *See Like a Radiologist With Systematic Windowing*. Accessed: Mar. 13, 2023. [Online]. Available: https://www.kaggle.com/dcstang/see-like-a-radiologist-with-systematic-windowing
[53] Reppic. *Gradient Sigmoid Windowing*. Accessed: Mar. 13, 2023. [Online]. Available: https://www.kaggle.com/reppic/gradient-sigmoid-windowing
[54] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3286–3295.
[55] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[56] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[57] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[58] C. Sun, F. Baradel, K. Murphy, and C. Schmid, "Learning video representations using contrastive bidirectional transformer," 2019, *arXiv:1906.05743*.
[59] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–14.
[60] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 13–23.
[61] L. Harold Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*.

[62] J. Schlemper, O. Oktay, L. Chen, J. Matthew, C. Knight, B. Kainz, B. Glocker, and D. Rueckert, "Attention-gated networks for improving ultrasound scan plane detection," in *Proc. Med. Image Anal.*, vol. 53, 2019, pp. 74–85.
[63] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnosing Alzheimer's disease from raw MRI scans using a deep learning model combining 3D CNN and self-attention mechanism," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3617–3628, 2021.
[64] Y. Barhoumi, "Efficient scopeformer: Towards scalable and rich feature extraction for intracranial hemorrhage detection using hybrid convolution and vision transformer networks," M.S. thesis, Rowan Univ., Glassboro, NJ, USA, 2023.
[65] H. Khan, P. M. Shah, M. A. Shah, S. U. Islam, and J. J. P. C. Rodrigues, "Cascading handcrafted features and convolutional neural network for IoT-enabled brain tumor segmentation," *Comput. Commun.*, vol. 153, pp. 196–207, Mar. 2020.
[66] H. Khan, N. C. Bouaynaya, and G. Rasool, "Adversarially robust continual learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Padua, Italy, Jul. 2022, pp. 1–8, doi: 10.1109/IJCNN55064.2022.9892970.
[67] J. M. Provenzale, "Imaging of traumatic brain injury: A review of the recent medical literature," *Amer. J. Roentgenol.*, vol. 186, no. 2, pp. 300–309, 2006.
[68] R. U. Kothari, T. Brott, J. P. Broderick, W. G. Barsan, L. R. Sauerbeck, M. Zuccarello, and J. Khoury, "The ABCs of measuring intracerebral hemorrhage volumes," *Stroke*, vol. 27, no. 8, pp. 1304–1305, Aug. 1996.
[69] M. Chawla, S. Sharma, J. Sivaswamy, and L. T. Kishore, "A method for automatic detection and classification of stroke from brain CT images," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2014, pp. 2639–2642.
[70] J. Sankar, M. C. Vanaja, and R. Venkatesh, "Computer-aided detection of intracranial hemorrhage from computed tomography images: A review," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 8, no. 3, pp. 363–380, 2013.



YASSINE BARHOUMI received the bachelor's degree in physics and chemistry from the University of Sciences, Tunis, Tunisia, and the first master's degree in physics of fluid dynamics and thermal transfers and the second master's degree in electrical and computer sciences engineering with a concentration in deep learning from Rowan University, where he is currently pursuing the degree. His master's thesis about Natural and Mixed Double Diffusive Convection of Nanofluids Inside Cavities and Onsets of Bifurcations. He is a graduate-level physics and engineering researcher. He was honored to study multiple aspects of physics and engineering during the undergraduate and graduate studies. He has extensively studied multiple domains in engineering and physics. During the ECE master's study, he extensively worked on multiple deep learning projects, including EMG signal detection using deep learning methods, hemorrhage detection using computer science deep learning methods, and brain tumor segmentation using computer science deep learning methods. He is also a Senior Machine Learning Engineer with MRIMath LLC, where he completed a four month internship in medical AI. He has applied AI and statistical methods to help build an MRI segmentation software with MRIMath LLC. He has studied multiple theoretical fluid dynamics graduate level courses and relevant courses from the ECE master's, such as machine learning, deep learning, reinforcement learning, concepts in AI: applications, transformers in NLP and computer vision, optimization and control, finite element analysis, applied mathematics, and complex analysis II. He is also interested in continual learning through MOOCs and bootcamps and a complete list of courses and MOOCs taken can be found on the LinkedIn website. His research assistantship at Rowan included assisting undergraduate clinics composed of mechanical, electrical, and machine learning teams, where he designed, developed, and built a set of robotic and prosthetic arms controlled by EMG signals via microcontrollers using novel machine learning and deep learning techniques. His duties included teaching machine learning and python basics and working side by side with the team through project completion and teaching with Circuits and Electronics Laboratories.

NIDHAL CARLA BOUAYNAYA (Member, IEEE) received the M.S. degree in pure mathematics and the Ph.D. degree in electrical and computer engineering (ECE) from the University of Illinois at Chicago. She is a Professor of ECE and the Director of the Rowan's Artificial Intelligence Laboratory (RAIL). She is currently the Associate Dean of research and graduate studies with the Henry M. Rowan College of Engineering. Previously, she was a Faculty Member with the University of Arkansas at Little Rock. She has coauthored more than 100 refereed journal articles, book chapters, and conference proceedings, such as IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Signal Processing Letters, *IEEE Signal Processing Magazine*, and *PLOS Medicine*. Her research is primarily funded by the National Science Foundation (NSF CCF, NSF ACI, NSF DUE, NSF I-Corps, NSF ECCS, NSF OAC, and NSF HRD), the National Institutes of Health (NIH), the U.S. Department of Education (USED), the New Jersey Department of Transportation (NJ DoT), the U.S. Department of Agriculture (USDA), the Federal Aviation Administration (FAA), Lockheed Martin Inc., and other industry. She is also interested in entrepreneurial endeavors. In 2017, she has Co-Founded and is a Chief Executive Officer (CEO) of MRIMath LLC, a start-up company, that uses artificial intelligence to improve patient oncology outcome and treatment response. MRIMath is funded by the NIH SBIR Program. Her research interests include big data analytics, machine learning, artificial intelligence, and mathematical optimization. She has won numerous Best Paper Awards, the most recent was at the 2019 IEEE International Workshop on Machine Learning for Signal Processing. She was also the winner of the Top Algorithm at the 2016 Multinomial Brain Tumor Segmentation Challenge (BRATS). She received numerous research and teaching awards, including the Rowan Research Achievement Award, in 2017, and the University of Arkansas at Little Rock Faculty Excellence Award in Research.

GHULAM RASOOL (Member, IEEE) received the B.S. degree in mechanical engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2000, the M.S. degree in computer engineering from the Center for Advanced Studies in Engineering (CASE), Pakistan, in 2010, and the Ph.D. degree in systems engineering from the University of Arkansas at Little Rock, in 2014. He is an Assistant Member with the Department of Machine Learning, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL, USA. He was a Postdoctoral Fellow with the Rehabilitation Institute of Chicago, Northwestern University, from 2014 to 2016. Before joining Moffitt, he was an Assistant Professor with the Department of Electrical and Computer Engineering, Rowan University. His current research focuses on building trustworthy multimodal machine learning and artificial intelligence model for cancer diagnosis, treatment planning, and risk assessment. His research efforts are currently funded by two National Science Foundation (NSF) Awards. Previously, his research was supported by the National Institute of Health (NIH), the U.S. Department of Education, NSF, the New Jersey Health Foundation (NJHF), Google, NVIDIA, and Lockheed Martin Inc. His work on Bayesian machine learning, has won the Best Student Award at the 2019 IEEE Machine Learning for Signal Processing Workshop.