IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 July 2023, accepted 17 August 2023, date of publication 21 August 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3307014

RESEARCH ARTICLE

Collaborative Consultation Doctors Model: Unifying CNN and ViT for COVID-19 Diagnostic

TRONG-THUAN NGUYEN[1,2], (Student Member, IEEE),
TAM V. NGUYEN[3], (Senior Member, IEEE), AND MINH-TRIET TRAN[1,2,4], (Member, IEEE)

[1] Software Engineering Lab and Faculty of Information Technology, University of Science, Vietnam National University Ho Chi Minh City (VNU-HCM), Ho Chi Minh City 72711, Vietnam
[2] Vietnam National University, Ho Chi Minh City 71308, Vietnam
[3] Department of Computer Science, University of Dayton, Dayton, OH 45469, USA
[4] John von Neumann Institute, Vietnam National University Ho Chi Minh City (VNU-HCM), Ho Chi Minh City 71308, Vietnam

Corresponding author: Minh-Triet Tran (tmtriet@fit.hcmus.edu.vn)

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DS2020-42-01.

ABSTRACT The COVID-19 pandemic presents significant challenges due to its high transmissibility and mortality risk. Traditional diagnostic methods, such as RT-PCR, have limitations that hinder timely and accurate screening. In response, AI-powered computer-aided imaging analysis techniques have emerged as a promising alternative for COVID-19 diagnosis. In this paper, we propose a novel approach that combines the strengths of Convolutional Neural Network (CNN) and Vision Transformer (ViT) to enhance the performance of COVID-19 diagnosis models. CNN excels at capturing spatial features in medical images, while ViT leverages self-attention mechanisms inspired by human radiologists. Additionally, our approach draws inspiration from subclinical diagnosis, a collaborative process involving attending physicians and specialists, which has proven effective in achieving accurate and comprehensive diagnoses. To this end, we employ an early fusion strategy integrating CNN and ViT, then fed into a residual neural network. By fusing these complementary features, our approach achieves state-of-the-art performance in accurately identifying COVID-19 cases on two benchmark datasets: Chest X-ray and Clean-CC-CCII. This research has the potential to enable timely and accurate screening, aiding in the early detection and management of COVID-19 cases. Our findings contribute to the growing knowledge of AI-powered diagnostic techniques and demonstrate the potential for advanced imaging analysis methods to support medical professionals in combating the ongoing pandemic.

INDEX TERMS Medical image classification, transformer, convolutional neural network.

I. INTRODUCTION

The COVID-19 pandemic is widely recognized as a significant public health crisis due to its high transmissibility and mortality risks. Recent trends have shown a reversal in death disparities since the end of the first Omicron wave last March [1]. Especially, the virus remains a significant threat, claiming more than 400 lives per day nationwide and over 100 lives in Massachusetts last week [2]. Therefore, the recent resurgence of COVID-19 in some parts of the US necessitates continued monitoring and implementation of preventive measures to mitigate the spread of the virus. While widely utilized for COVID-19 detection, RT-PCR has certain limitations that hinder its effectiveness. These include time-consuming procedures, the possibility of false-negative results, limited availability of equipment, and stringent testing criteria. These factors can delay and impede the prompt and accurate screening of individuals potentially infected with the virus. In response to these challenges, there is a growing interest in leveraging AI-powered computer-aided imaging analysis [3], [4], [5] for COVID-19 diagnosis. By analyzing lung regions in Chest X-ray (CXR) images and CT scans, these AI-driven systems aim to provide an alternative and complementary approach to detecting COVID-19. Developing such machine-driven instruments is crucial to accurately and efficiently recognize COVID-19 in diagnostic imagery, helping to overcome the

The associate editor coordinating the review of this manuscript and approving it for publication was N. Ramesh Babu.

limitations of traditional diagnostic methods and enabling timely and reliable identification of infected individuals.

The Transformer architecture [6], initially developed for natural language processing, has found application in computer vision tasks due to its unique ability to capture long-range dependencies and global context information. This adaptation has given rise to Transformer-based models like the Vision Transformer (ViT) [7], which offer a compelling alternative to traditional Convolutional Neural Networks (CNN) in computer vision. In contrast to CNN, which relies on a hierarchical feature extraction process to combine lower-level features and construct higher-level representations, Transformer-based models leverage the self-attention mechanism to capture global dependencies among image patches or pixels. This self-attention mechanism enables them to effectively model the relationships between different parts of the image, regardless of their spatial proximity. By considering the entire image simultaneously, Transformer-based models can capture fine-grained details and exploit long-range dependencies, improving image classification tasks' performance. One notable advantage of Transformer-based models is their scalability, which is particularly valuable in complex and large-scale visual data applications. Transformer-based models excel in training on benchmark datasets [8], [9], enabling them to learn intricate patterns and generalize well to unseen examples. The ability of Transformer-based models to capture global context information and effectively handle large-scale visual data has opened up new possibilities in computer vision. Applications such as autonomous driving [10] and satellite imagery analysis [11] benefit from these models' capabilities, pushing the boundaries of what is achievable in computer vision tasks.

When attending physicians encounter challenging cases in subclinical diagnosis, they often seek consultation from specialists for further diagnosis discussion. This collaborative approach, driven by the expertise of multiple healthcare professionals, has proven effective in achieving accurate and comprehensive diagnoses. Motivated by this collaborative model, we explore the application of ViT and CNN with transfer learning for medical image analysis. On the one hand, the ViT model, analogous to a consulting doctor, brings unique strengths in capturing global contextual information and learning representations from visual data. On the other hand, the CNN model, resembling an attending physician, excels in capturing local patterns and extracting intricate details from medical images. Building upon this idea, we propose a novel hybrid model that synergistically integrates the strengths of both the CNN and ViT models in medical image classification. By combining the expertise of these two models, we aim to improve the accuracy and efficiency of diagnosing medical conditions based on visual information. This hybrid approach holds significant promise for medical image analysis, as it allows for a more comprehensive and accurate assessment of the underlying conditions. By leveraging the complementary strengths of the CNN and ViT models, we can enhance the overall diagnostic capabilities, leading to improved patient care and treatment outcomes.

In this paper, our contributions are threefold:

- We investigate the performance of state-of-the-art transformer-based visual classification models on the ChestXray [12] and Clean-CC-CCII [13] datasets, which comprise Chest X-ray and CT scan images, respectively.
- Inspired by the collaborative doctor consultation, We propose a novel hybrid model incorporating an early fusion strategy for combining transformer and CNN features, improving accuracy and efficiency in medical image analysis.
- We conduct extensive experiments. Our approach achieves impressive results, with accuracy rates of 98.86% and 95.62% observed on the ChestXray and Clean-CC-CCII datasets, respectively.

The remainder of this paper is organized as follows. In Section II, we briefly review methods for COVID-19 diagnosis. Then, we present our proposed method in Section III. Experiments and evaluation are discussed in Section IV. Finally, conclusions and future work are in Section V.

II. RELATED WORK

In recent years, there has been a surge in the development of various methods for diagnosing COVID-19, with the primary objective being the classification of medical cases into different categories, such as COVID-19, pneumonia, and normal. This section reviews relevant research on classification tasks according to the classes above.

A. CNN METHODS

CNN have significantly impacted medical imaging by leveraging their capacity to learn intricate and sophisticated representations through data-driven approaches. Jia et al. [14] proposed a method that addressed the challenge of gradient vanishing by dynamically combining features from various layers of MobileNet and ResNet. This approach effectively preserved important information throughout the network architecture, leading to improved performance. Song et al. [15] focused on refining the input data by removing boundary regions and filling in missing areas surrounding the lungs. They employed a modified ResNet50 architecture augmented with a feature pyramid network (FPN) module and a multi-layer perceptron (MLP) for prediction. Barzekar and Yu [16] proposed a novel CNN architecture called C-Net for the automated classification of biomedical images, specifically histopathological images for cancer diagnosis. The C-Net architecture consisted of multiple CNNs (Outer, Middle, and Inner) that worked together as feature extractors to classify images in terms of malignancy and benignancy. Aytaç et al. [17] presented a novel adaptive momentum optimizer for training CNN in medical image classification. The adaptive momentum dynamically adjusted the momentum rate based on error changes, eliminating the need for complex hyperparameter tuning. Musallam et al. [18] addressed the challenges of accurate diagnosis of brain diseases using a Computer-Aided Diagnosis (CAD) system for magnetic resonance imaging (MRI) images. It proposed a three-step preprocessing approach to enhance the quality of MRI images, along with a new CNN architecture designed specifically for diagnosing glioma, meningioma, pituitary, and normal images.

B. TRANSFORMER-BASED METHODS

Attention-based "Transformer" models have revolutionized deep learning by capturing long-range dependencies and learning powerful feature representations. ViT architectures have emerged as a considerable advancement, replacing convolutions with image patch sequences and achieving state-of-the-art performance across computer vision tasks. Sun et al. [19] proposed a novel pure transformer-based multi-view network for mammographic image classification. Their approach utilized a "cross-view attention block" structure to effectively fuse multi-view information, allowing for comprehensive analysis of the mammographic images. Additionally, they introduced a "classification token" mechanism to gather all relevant information for making accurate predictions in the final classification task. Xu et al. [20] introduced a transformer-based multi-modality deep learning framework designed to effectively fuse multiple sources of data for skin tumor analysis. Their approach incorporated clinical images, dermoscopic images, and clinical patient-wise metadata, enabling comprehensive and informative analysis for improved skin tumor diagnosis and classification. Almalik et al. [21] proposed a novel method called Self-Ensembling Vision Transformer (SEViT) to enhance the robustness of ViT against adversarial attacks. SEViT leveraged the resilience of initial blocks' feature representations to adversarial perturbations and combined multiple classifiers' predictions with the final ViT classifier to improve robustness. The proposed architecture was evaluated on chest X-ray and fundoscopy modalities, demonstrating its effectiveness in defending against various adversarial attacks in the gray-box setting.

C. HYBRID METHODS

The hybrid models used in medical image classification have emerged as promising approaches for improving the accuracy and robustness of diagnostic systems. Several studies have proposed different hybrid architectures to leverage the strengths of CNN and transformers. For instance, Kumar et al. [22] employed a hybrid CNN approach by integrating a ResNet 152 layer into the CNN architecture. Kumar et al. [23] introduced PHTrans, a model that combines transformers and CNNs in parallel to capture both global and local features and achieve superior segmentation performance. Zhou et al. [24] integrated DHRNet and a hybrid transformer to extract local and global features, allowing for exploring long-range dependencies. Rocha et al. [25] proposed a method called Hybrid CNN Ensemble (HCNNE) that combined features extracted by convolutional neural networks (CNN) and local binary patterns (LBP) for image classification. The method utilized an ensemble of multiple classifiers, where the Euclidean distance between LBP feature vectors and the confidence of CNN features classified by support vector machines were used as input to a multilayer perceptron classifier. Additionally, these features were used as input for other classifiers to create the final voting ensemble. Yuan et al. [26] presented CTCNet, a hybrid model that combines Swin Transformers and Residual CNNs, effectively blending complementary features using a cross-domain fusion block. MedViT [27] offerred a robust CNN-Transformer hybrid architecture specifically designed for medical image diagnosis. Their model overcomes concerns related to adversarial attacks and the reliability of deep medical diagnosis systems by leveraging the local feature extraction capabilities of CNN and the global connectivity of transformers. Additionally, they introduce an efficient convolution operation to mitigate computational complexity, and strategies to learn smoother decision boundaries and enhance model robustness. Jang and Hwang [28] proposed a three-dimensional Medical image classifier called a Multi-plane and Multi-slice Transformer (M3T) network for accurate classification of Alzheimer's disease (AD) in 3D MRI images. The M3T network combined 3D CNN, 2D CNN, and Transformer to leverage their strengths in representation learning and attention relationships. The 3D CNN captured local abnormalities using inductive bias, while the Transformer captured wider region abnormalities without inductive bias.

III. PROPOSED APPROACH

In this section, we introduce two important components of our approach: Deep Doctor and Deep Consultation. Deep Doctor represents our individual deep learning model, while Deep Consultation refers to the early fusion strategy. Additionally, we present the Deep Network Architecture, which serves as the classification model. These components are fundamental to our method and have significant implications for the overall framework.

A. DEEP DOCTOR

1) DenseNet-201

DenseNet-201 [29] is a deep CNN architecture that stands out for its dense connections between layers. These connections are designed to improve information flow and maximize feature reuse, leading to enhanced parameter efficiency and performance.

The architecture of DenseNet-201 consists of 201 layers, starting with a convolutional layer and a max pooling layer. The core building blocks of DenseNet are the dense blocks, which contain multiple convolutional layers. Within each dense block, the output feature maps of each convolutional layer are concatenated with the input feature maps

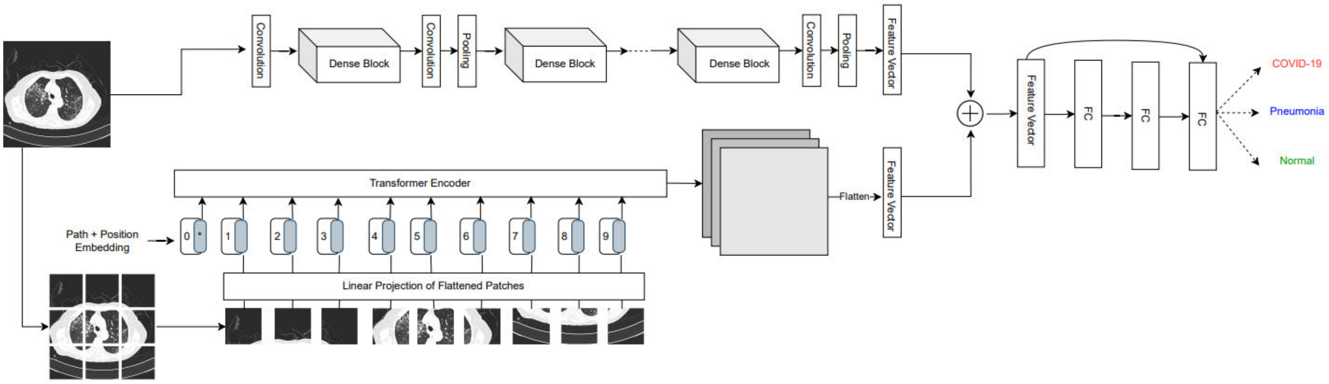


FIGURE 1. Our proposed approach. The method involves utilizing DenseNet-201 and T2T-ViT models for feature extraction, as shown in the above and below parts of the images, respectively. The features extracted from the early fusion are then fed into a feedforward network, which includes a skip connection.

of subsequent layers. This dense connectivity enables direct access to earlier layer features, facilitating effective information sharing and enabling the network to capture fine-grained details from earlier stages. The transition layers are introduced between dense blocks to manage computational complexity and control the growth of feature maps. Transition layers employ average pooling and 1 × 1 convolutional layers to reduce the dimensionality of feature maps. This reduction helps to compress and refine the learned representations while also promoting efficient computation. At the end of the DenseNet-201 network, a global average pooling layer is applied. This layer aggregates the feature maps by computing the average value of each channel across spatial dimensions, resulting in a global representation of the input image. This pooling operation helps to capture the most salient information from the entire image and facilitates robust feature extraction. Finally, the aggregated features are fed into a fully connected layer with a softmax activation function. This last layer maps the extracted features to predicted class probabilities, allowing the model to make predictions about the input image's class label.

Overall, DenseNet-201's architecture leverages dense connections to facilitate information flow and promote feature reuse. This design choice improves the network's ability to capture fine-grained details and effectively learn from the input data. The combination of dense connectivity, transition layers, and global average pooling enables DenseNet-201 to achieve powerful and accurate image classification performance.

2) VISION TRANSFORMER (ViT)

The ViT [7] is the first full-transformer model that directly applies the Transformer architecture to images, enhancing the spatial relationship among image pixels. The ViT model takes an input image and splits it into non-overlapping patches. These patches are then flattened and transformed into sequences of token embeddings.

The ViT model consists of a conventional transformer encoder, followed by a linear classification head. The transformer encoder incorporates multiple transformer blocks, each containing a self-attention mechanism. This self-attention mechanism allows the model to selectively attend to different patches in the input image, capturing the interdependencies and contextual relationships between them. The output of the final transformer block is passed to the linear classification head, which generates a probability distribution across the potential image classes. This allows the ViT model to classify the input image based on the learned representations and relationships captured by the transformer blocks. By directly applying the Transformer architecture to images, the ViT model overcomes the limitations of traditional CNNs by explicitly modeling long-range dependencies and capturing global context. This approach enhances the ability of the model to understand the spatial relationships among image pixels and effectively extract meaningful features for classification.

Overall, the ViT model offers a novel and powerful approach to image analysis, leveraging the strengths of the Transformer architecture to improve spatial understanding and achieve state-of-the-art performance in various image-based tasks.

3) TOKENS-TO-TOKEN ViT (T2T-ViT)

The T2T-ViT [30] architecture introduces a novel approach to image processing by employing a vision transformer model. It consists of two key components: the Tokens-to-Token (T2T) module and the T2T-ViT backbone, which work together to enhance the model's ability to extract meaningful features from images.

The T2T module plays a vital role in structuring the input image into tokens while capturing local structural information. It involves two main operations: Re-Structurization and Soft Split (SS). In the Re-Structurization step, the image is divided into fixed-size patches, treating each patch as an individual token. This process allows the model to capture local content present in the image. Subsequently, the SS operation further splits each patch into multiple sub-tokens, enabling the model to capture more detailed local information. This iterative sub-tokenization enhances the model's capacity to extract fine-grained features. In the final layer of the T2T module, a class token and a Sinusoidal Position Embedding (PE) are combined. The class token serves as a global representation, encapsulating comprehensive image information and capturing the overall content and context of the image. On the other hand, the Sinusoidal Position Embedding encodes the spatial position of each token within the image, providing the model with an understanding of the relative spatial relationships between tokens. By incorporating positional information, the model develops spatial awareness and can leverage this information during processing. The T2T-ViT backbone acts as the core architecture that utilizes the structured tokens generated by the T2T module. It consists of transformer layers, which process the tokens and extract meaningful representations. The transformer layers employ self-attention mechanisms to model the relationships between tokens, allowing the model to capture global context information. By combining both local and global information derived from the structured tokens, the T2T-ViT architecture enables the model to capture intricate details and comprehend contextual relationships within the image.

This unique integration of the T2T module and the T2T-ViT backbone empowers the T2T-ViT model to effectively leverage both local and global information from the image. By capturing fine-grained details and understanding contextual relationships, the model achieves improved performance across various image analysis tasks. This innovative architecture expands the capabilities of vision transformers, enabling them to process images with enhanced spatial awareness and capture rich visual information.

4) TRANSFORMER iN TRANSFORMER (TNT)

TNT [31] model is an architecture that extends the Transformer framework by introducing a Transformer block within each patch of an image. This allows for enhanced modeling of the spatial relationships among image pixels, leading to more detailed and accurate representations.

In the TNT model, the input image is first divided into non-overlapping patches. Each patch is then flattened and transformed into a sequence of token embeddings. These token embeddings serve as the input to the Transformer in the Transformer block. The Transformer in Transformer block consists of multiple sub-layers, including a self-attention mechanism and feed-forward neural networks, similar to the traditional Transformer. However, an additional sub-transformer in TNT is embedded within the Transformer block to capture finer details and interactions between tokens within each patch. The sub-transformer operates at a smaller scale, focusing on the interactions among visual words within a visual sentence. Visual sentences are created by subdividing each patch into smaller sub-patches called visual words. The sub-transformer independently calculates features and attention between visual words within each visual sentence. The shared network in TNT is responsible for computing the features and attention between visual words. This shared network ensures consistency and coherence across the visual words within a visual sentence, capturing relevant information and relationships. After processing the visual words within each visual sentence, the features are aggregated to represent the entire patch. This aggregation step combines the detailed features extracted from the sub-transformer with the global context captured by the conventional Transformer layers. The final output of the TNT model is obtained by applying a linear classification head to the features from the last Transformer block. This classification head generates a probability distribution across potential image classes, enabling the model to make predictions.

By incorporating a Transformer in a Transformer block, the TNT model enhances the ability to capture intricate spatial relationships among image pixels. It enables the extraction of fine-grained details and improves the overall representation of the input image. This approach has shown promising results in various computer vision tasks, demonstrating its potential for advancing the field of image analysis and understanding.

5) PoolFormer

PoolFormer [32] introduces the MetaFormer concept, which modifies the encoder part of the vision transformer architecture by replacing the token mixer with a PoolFormer module while keeping the other components unchanged. The PoolFormer module consists of a parallel set of multi-head self-attention layers that operate on patch embeddings, followed by position-wise feedforward networks.

In the PoolFormer module, each multi-head self-attention layer processes the patch embeddings, capturing the relationships between patches. These self-attention layer outputs are then pooled across the patch dimension using a pooling operation. This pooling operation aggregates the information from different patches and produces a fixed-size representation for each patch. The resulting patch representations obtained from the pooling operation are then passed through a set of transformer layers. These transformer layers further refine the visual representation by integrating local and global information within the patches. The transformer layers utilize self-attention mechanisms to model the dependencies between patches and capture contextual relationships.

By employing the PoolFormer module in the MetaFormer architecture, the encoder part of the vision transformer is transformed. The module effectively captures the relationships between patches and generates informative representations by pooling and integrating information from multiple patches. This modification enhances the vision transformer's ability to process images and extract meaningful visual features.

6) CONFORMER

The Conformer [33] architecture is a unique dual network structure that combines convolutional operations and self-attention mechanisms to leverage both local features and global representations. It consists of two main branches: a convolutional neural network (CNN) branch and a transformer branch.

The CNN branch focuses on capturing local contextual information by applying convolutional operations to the input data. Convolutional layers are adept at extracting spatial features and capturing local patterns within the data. This branch helps the Conformer model capture fine-grained details and local relationships. On the other hand, the transformer branch is responsible for capturing global contextual information by utilizing self-attention mechanisms. Self-attention allows the model to analyze the relationships between different elements in the input sequence and capture long-range dependencies. By incorporating self-attention, the transformer branch can capture global patterns and establish contextual relationships across the entire input. The Conformer architecture employs a stem module called the Feature Coupling Unit (FCU) to integrate local features and global representations effectively. The FCU progressively fuses the feature maps obtained from the CNN branch with the patch embeddings from the transformer branch. This interactive fusion process ensures that both local and global information are properly combined, enhancing the overall representation learning capability of the model. In terms of classification, the Conformer architecture utilizes the features obtained from the CNN branch and feeds them into one classifier. Simultaneously, it extracts the class token from the transformer branch and feeds it into another classifier. This dual classifier setup allows the model to leverage both local and global information for accurate classification.

By combining the strengths of the CNN and transformer branches, along with the integration provided by the FCU, the Conformer architecture can effectively capture both local and global contextual information. This makes it well-suited for various tasks that require a comprehensive understanding of the input data, such as image classification and sequence modeling.

B. DEEP CONSULTATION

We argue that DenseNet-201 and T2T-ViT have their own respective strengths. DenseNet-201 employs dense blocks that enable efficient learning of local features, allowing it to capture fine-grained details in the images. On the other hand, T2T-ViT utilizes self-attention mechanisms to model relationships between image patches, enabling it to capture long-range dependencies and understand the global context of the image. By fusing the features extracted from both models through early fusion, we obtain a more comprehensive representation of the input image. This combined representation incorporates both the detailed local features learned by DenseNet-201 and the holistic global context captured by T2T-ViT. The early fusion strategy not only facilitates the integration of these complementary features but also enhances the overall performance of the classification model in subsequent tasks. The resulting fused features provide a more robust and informative representation, leading to improved accuracy and effectiveness in medical image classification and analysis.

We consider the output tensor F_d of DenseNet-201, which represents the features extracted from the last fully connected layer. This tensor has a shape of (N, C_d), where N is the number of samples and C_d is the number of output features. Similarly, we denote the output tensor of T2T-ViT as F_t. Unlike DenseNet-201, T2T-ViT does not have a conventional fully connected layer. Instead, it incorporates a class token and a sinusoidal position embedding, which are concatenated with the final output of the Tokens-to-Token module. The resulting tensor F_t has a shape of (N, C_t), where C_t corresponds to the shape of the flattened feature maps generated by the Transformer encoder. To combine the information from DenseNet-201 and T2T-ViT, we concatenate the two tensors along the feature dimension. This fusion operation results in a new tensor F with a shape of $(N, C_d + C_t)$, which represents the fused feature. Mathematically, we define the early fusion as the concatenation of F_d and F_t:

$$F = [F_d, F_t] \in \mathbb{R}^{N \times (C_d + C_t)} \tag{1}$$

where [,] denotes the concatenation operation along the feature dimension.

C. DEEP NETWORK ARCHITECTURE

We introduce a feedforward neural network architecture for multi-class classification tasks in computer science and deep learning. Fig. 1 provides an overview of our proposed method. The architecture begins with an input layer that receives a 1D tensor containing the fused feature representation of the training data. This tensor is then passed through a fully connected layer with 1024 units, establishing connections between each unit and every element of the input tensor. The output of the fully connected layer is then processed by a rectified linear unit (`ReLU`) activation function, introducing non-linearity to the network and enabling it to learn complex patterns and representations. To enhance performance and prevent overfitting, batch normalization is applied, normalizing the input to each neuron within mini-batches by adjusting the mean and variance. To facilitate learning of residual mapping and identity mapping, another fully connected layer with 1024 units is utilized. The output of this layer is concatenated with the previous layer's output, forming a residual connection. This mechanism allows the network to capture the difference between the input and the desired output (residual mapping) while preserving important information from the input (identity mapping). The concatenated tensor resulting from the residual connection is then passed through another `ReLU` activation function, further enhancing the non-linear representations learned by the network. Additionally, a skip connection is incorporated by concatenating the output of this activation function with the input tensor, ensuring the preservation of original input information and facilitating better information flow through the network. Finally, the concatenated tensor undergoes further processing through a fully connected layer with a softmax activation function. This layer generates a probability distribution over the different classes, enabling the network to classify the input into the appropriate class. In summary, our proposed architecture includes input processing, feature extraction, non-linear transformations, residual connections, skip connections, and final classification. This approach aims to improve the network's ability to capture complex patterns and make accurate multi-class classifications.

IV. EXPERIMENTS AND EVALUATION

A. EXPERIMENTAL SETTINGS

1) DATASETS

We adopt two datasets, ChestXray and Clean-CC-CCII for the benchmark.

ChestXray [12] contains chest X-ray images of patients with COVID-19, pneumonia, and normal lungs. ChestXray dataset includes 6,432 X-ray images divided into two subsets. In particular, the training and testing set comprise 5,144 and 1,288 images, respectively.

Clean-CC-CCII [13] comprises 340,190 slices/images for COVID-19 and normal encompassing both normal and common pneumonia cases. Clean-CC-CCII dataset is split into training and testing sets. In particular, the training contains 272,117 slices from 3,195 scans of 2,164 patients, and the testing set incorporates 68,073 slices from 798 scans of 534 patients.

2) EVALUATION METRICS

We use four metrics to evaluate the model performance:

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

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{4}$$

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{5}$$

True Positive (TP) and True Negative (TN) represent the number of correctly classified COVID-19 and non-COVID-19 scans, respectively. On the other hand, False Positives (FP) and False Negatives (FN) correspond to misclassified COVID-19 cases and non-COVID-19, respectively. The evaluation of ChestXray and Clean-CC-CCII datasets compromise both normal and pneumonia for the non-COVID-19 class. Additionally, the Accuracy is calculated as the macro-averaging value for all test data, and it serves as an evaluation metric of the model's overall performance.

B. IMPLEMENTATION DETAILS

We adopt a multi-GPU training strategy and employ the MMClassification Toolbox and PyTorch libraries for training our classification models. Initially, we train individual models using MMClassification, leveraging pre-trained DenseNet201 and T2T-ViT models as feature extractors. We enhance these models by adding fully connected layers with residual connections, batch normalization, dropout, and activation functions, as described in Section III-C. The models are compiled with an `Adam` optimizer and utilize the `Cross-Entropy` loss function. During training, we provide the training features and labels, specify the batch size and the number of epochs, and perform a 20% validation data split from the training set. The best model weights are automatically saved if they lead to improved validation accuracy at each epoch.

C. QUANTITATIVE ANALYSIS

1) EXPERIMENTAL RESULTS

We conduct an extensive analysis of the ChestXray dataset and the results, as shown in Table 1, including accuracy, precision, recall, and F1-score. These metrics serve as critical indicators of the model's effectiveness in classifying COVID-19, normal, and pneumonia cases. Among the models, the T2T model achieved an accuracy of 88.51%. While its precision of 93.89% suggests a high level of correctness in positive predictions, its relatively lower recall of 81.6% implies a significant number of missed positive cases. Consequently, the F1-score of 85.9% reveals a moderate overall performance. In contrast, the ViT model demonstrated superior performance with an accuracy of 92.24%. Its balanced precision of 91.59% and recall of 92.97% indicate its ability to correctly identify positive cases while minimizing false positives and false negatives. The F1-score of 92.24% reflects its strong overall performance. The DenseNet-201 model achieved a slightly higher accuracy of 93.18% compared to the ViT. It consistently exhibited stable precision, recall, and F1-score values of around 93%, highlighting its reliability in classifying the cases. The T2T-ViT model further improved the performance, reaching an accuracy of 95.42%. However, its relatively lower recall of 76.04% suggests the presence of a notable number of missed positive cases. Consequently, the F1-score of 81.59% indicates a comparatively lower overall performance compared to other models. On the other hand, the PoolFormer model showcased strong performance, attaining an accuracy of 97.13% and an impressive F1-score of 97.26%. Its high precision and recall values indicate its effectiveness in accurately identifying positive cases while maintaining a low rate of false positives and false negatives. Similarly, the Conformer model exhibited notable performance with an accuracy of 97.28% and an F1-score of 97.49%. Its high precision and recall values further reinforce its ability to classify cases accurately.

Our proposed method showcases outstanding performance across precision, recall, and F1-score metrics, providing strong evidence for the effectiveness of our early fusion approach in the ChestXray dataset. It outperforms all other methods with an exceptional accuracy of 98.86%. Furthermore, it exhibits high precision, recall, and F1-score values consistently around 98%, indicating its superior capability in accurately classifying COVID-19, normal, and pneumonia

TABLE 1. Experimental results on the ChestXray dataset. The best results for each evaluation are shown in red.

Method	Accuracy	Precision	Recall	F1-score
T2T	88.51	93.89	81.6	85.9
ViT	92.24	91.59	92.97	92.24
DenseNet-201	93.18	93.16	93.17	93.12
T2T-ViT	95.42	92.7	76.04	81.59
PoolFormer	97.13	96.96	97.57	97.26
Conformer	97.28	97.39	97.58	97.49
Ours	**98.86**	**98.16**	**98.22**	**98.19**

TABLE 2. Comparison of state-of-the-art methods on the ChestXray dataset. The best results for each evaluation are shown in red.

Method	Accuracy	Precision	Recall	F1-score
Late Consultation [12]	93.94	93.92	93.94	94.93
Early Consultation [12]	95.03	95.03	95.03	95.03
Ours	**98.86**	**98.16**	**98.22**	**98.19**

TABLE 3. Experimental results on the Clean-CC-CCII dataset. The best results for each evaluation are shown in red.

Method	Accuracy	Precision	Recall	F1-score
CovidNet [13]	88.69	90.48	88.08	89.26
COVID-AL [34]	86.60	N/A	N/A	N/A
Zhang *et al.* [35]	92.49	N/A	94.93	N/A
ViT [7]	93.22	92.77	91.95	92.33
RACNet [36]	95.33	N/A	N/A	N/A
Ours	**95.62**	**95.17**	**95.31**	**95.23**

cases. The comprehensive analysis presented in Table 1 corroborates the exceptional performance of our proposed model, as evidenced by its top-ranked accuracy, precision, recall, and F1-score metrics. These results affirm its efficacy in accurately classifying chest X-ray images in the ChestXray dataset.

2) COMPARISONS TO THE STATE-OF-THE-ART METHODS

We compare our proposed model with other state-of-the-art methods on the Chest X-ray dataset, as presented in Table 2 and Table 3 to provide a comprehensive evaluation. The Late Consultation and Early Consultation methods achieved accuracy scores of 93.94% and 95.03%, respectively, indicating their reasonable performance but falling slightly behind the top-performing models. While these models demonstrate solid performance, they exhibit lower recall scores, suggesting potential limitations in accurately identifying positive cases, which is critical in medical imaging applications. In contrast, our proposed method stands out by achieving the highest accuracy of 98.21% along with impressive precision, recall, and F1-score values. This indicates its superior performance in accurately classifying the different categories of chest X-ray images.

Moving beyond the ChestXray dataset, we also evaluated the performance of our method on the Clean-CC-CCII dataset. Our approach showcases outstanding performance, achieving an accuracy score of 95.62% along with high precision, recall, and F1-score values. Comparatively, the CovidNet method serves as a baseline, performing well with an accuracy score of 88.69%, precision of 90.48%, recall of 88.08%, and F1-score of 89.26%. However, the COVID-AL method exhibits lower performance metrics, and the ViT method demonstrates slightly lower performance metrics compared to the top-performing methods.

In summary, our proposed model demonstrates exceptional performance on both the ChestXray and Clean-CC-CCII datasets, surpassing other state-of-the-art methods in terms of accuracy, precision, recall, and F1-score. This reinforces the effectiveness of our approach in accurately classifying COVID-19, normal, and pneumonia cases in various medical imaging datasets.

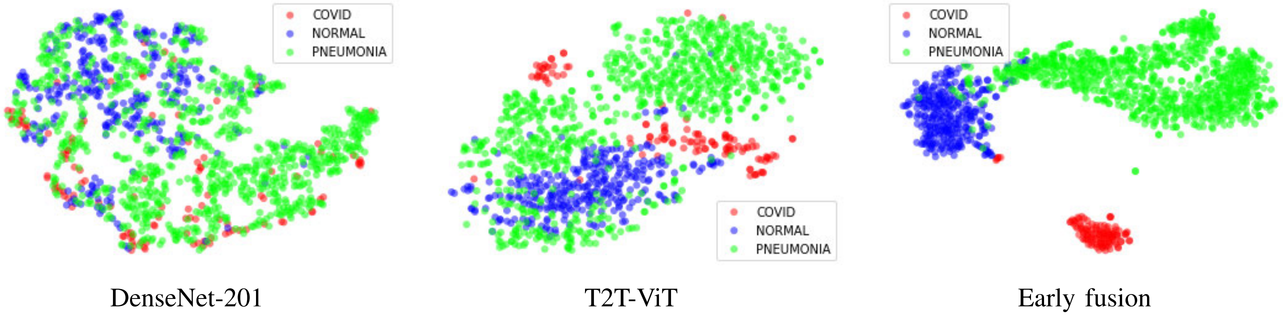
D. QUALITATIVE ANALYSIS

We utilize t-Distributed Stochastic Neighbor Embedding (t-SNE), a non-linear dimensionality reduction technique used to visualize high-dimensional data in a low-dimensional space while preserving the local and global structure of the data. As illustrated in Fig. 2, we reveal that DenseNet-201 generates clusters with no significant correlation, resulting in difficulty in visually distinguishing data points and possible confusion in interpretation. Although T2T-ViT refers to the ability to separate different categories in visually clear and easy to interpret, overlapping are still presented between COVID-19 and pneumonia classes. In contrast, the fused feature demonstrated a powerful distribution, which indicates well-separated, distinct clusters with minimal overlap. A powerful distribution in a cluster implies that the underlying features are highly informative and relevant for distinguishing between different groups or categories. Accordingly, fused features prove effective in the Linear SVM model and the feedforward neural network.

On the other hand, T2T-ViT demonstrates an improved ability to separate different categories in a visually clear and interpretable manner. However, despite its strengths, there are still regions of overlap between the COVID-19 and pneumonia classes. This suggests that T2T-ViT may face difficulties in fully disentangling the complex visual characteristics exhibited by these classes, possibly due to the similar radiological manifestations they can share.

In contrast, the fused feature representation exhibits a powerful distribution of clusters, showcasing distinct and well-separated groups with minimal overlap. This indicates that the fused features successfully capture highly informative and discriminative characteristics, enabling a more reliable differentiation between different groups or categories. Notably, the effectiveness of the fused features is evident in their integration into the Linear SVM model and the feedforward neural network, where they contribute to improved classification performance.

We continue with the analysis of chest X-ray images illustrated in Fig. 3. The proposed model exhibits noteworthy proficiency in accurately predicting COVID, normal, and



DenseNet-201 T2T-ViT Early fusion

FIGURE 2. t-SNE visualization of extracted features from individual models and fused features on the ChestXray testing set.

pneumonia classes, underscoring its discriminative capability among these categories. Nevertheless, it is essential to recognize that the model occasionally produces inaccurate predictions, specifically in misclassifying pneumonia and normal chest X-ray images. This observation implies that the model encounters difficulties in effectively capturing the subtle distinctions between these two classes, which could be attributed to overlapping radiographic features or variations in image quality.

Furthermore, when considering the diagnosis of COVID-19, normal, and pneumonia cases using CT scans, the level of confusion among these categories becomes more pronounced compared to chest X-ray images. This increased confusion arises from the shared visual characteristics observed in COVID-19 and pneumonia cases, such as ground-glass opacities, consolidation, and interstitial thickening, which pose challenges for accurate differentiation. In contrast, normal CT scans typically do not display any abnormal findings. The instances of confusion depicted in Fig. 4 provide valuable insights into the complexities encountered during the diagnostic process, underscoring the importance of further advancements to enhance the model's performance and reliability in accurately classifying such intricate scenarios.

In conclusion, the utilization of t-SNE visualization and in-depth analysis of the model's performance highlights the strengths and limitations of different approaches. The findings underscore the significance of advancing the fusion of features from diverse models to improve the accuracy and reliability of medical image analysis. Furthermore, they emphasize the ongoing need for refinement and augmentation of existing models to effectively address the challenges posed by complex imaging datasets in medical diagnosis.

E. ABLATION STUDY

The ablation study conducted in Table 4 aimed to investigate different combinations of hybrid CNN and Transformer features for the classification of chest X-ray images. The results provide valuable insights into the performance of these combinations and help inform the proposed approach, which replaces the SVM component with a residual neural network.

TABLE 4. Ablation study comparing fused features with SVM to the proposed method. The best results for each evaluation are shown in red.

	Accuracy	Precision	Recall	F1-score
Poolformer + HRNet	88.98	89.05	88.98	88.97
T2T + DenseNet201	89.67	89.92	89.67	89.69
Poolformer + DenseNet201	90.37	90.49	90.37	90.41
ViT + DenseNet201	94.10	94.10	94.10	94.10
Conformer + PoolFormer	96.04	96.02	96.04	96.03
Conformer + HRNet	96.04	96.02	96.04	96.03
Conformer + DenseNet201	96.12	96.10	96.12	96.11
T2T-ViT + DenseNet201	97.75	97.79	97.75	97.76
Ours	**98.21**	**98.21**	**98.21**	**98.21**

The combination of Poolformer and HRNet achieved an accuracy of 88.98%, demonstrating its effectiveness in correctly classifying the majority of test samples. It exhibited balanced precision, recall, and F1-score values, indicating a good trade-off between accurately identifying positive cases and minimizing false positives and false negatives. Similarly, the TnT and DenseNet201 combination achieved a slightly higher accuracy of 89.67% with comparable precision, recall, and F1-score values. Further improvement was observed with the Poolformer and DenseNet201 combination, which attained an accuracy of 90.37%. This combination showcased high precision, recall, and F1-score values, highlighting its effectiveness in accurately classifying COVID-19, normal, and pneumonia cases. The successful integration of Poolformer and DenseNet201 resulted in enhanced performance, leveraging their complementary features. The ViT and DenseNet201 combination achieved an even higher accuracy of 94.10%. It consistently demonstrated precision, recall, and F1-score values around 94.00%, indicating its robust performance in accurately classifying chest X-ray images. This combination effectively leveraged the powerful vision-based capabilities of the ViT model alongside the dense feature extraction of DenseNet201.

The combinations of Conformer with PoolFormer, Conformer with HRNet, and Conformer with DenseNet201 achieved high accuracies of 96.04%. 96.04% and 96.12%, respectively. These combinations showcased the effectiveness

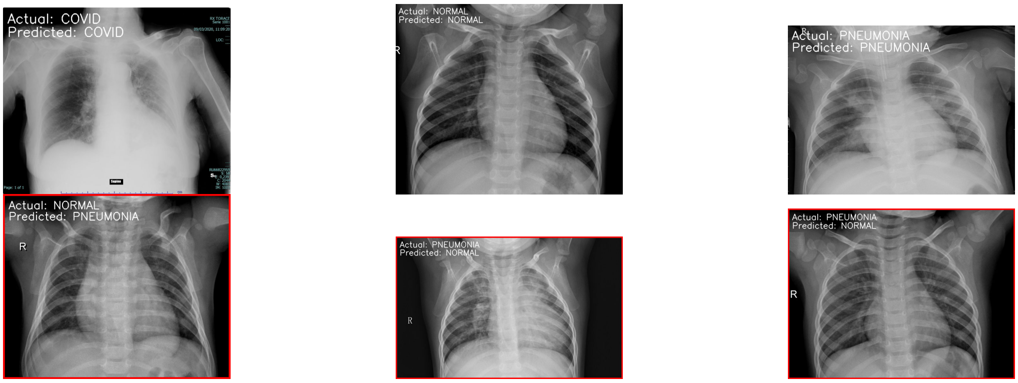


FIGURE 3. **Illustration of cases on the X-ray images predicted by our proposed method. The first row shows the visualizations of accurate predictions, while the second row displays the opposite.**

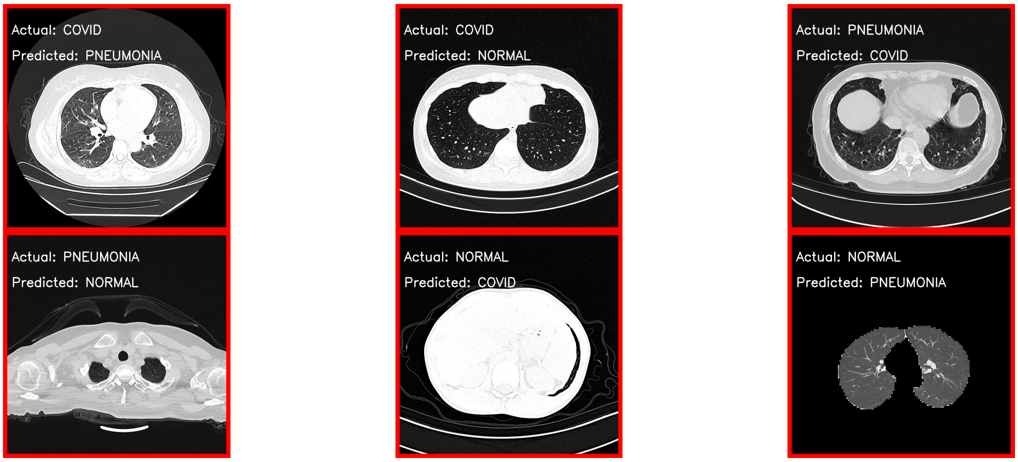


FIGURE 4. **Illustration of wrong predictions on CT scan images predicted by our proposed method.**

of integrating the Conformer model, known for capturing long-range dependencies, with different CNN-based models. The resulting models exhibited exceptional precision, recall, and F1-score values, underscoring their ability to classify chest X-ray images accurately.

The combination T2T-ViT of DenseNet201 demonstrated superior performance, achieving an accuracy of 97.75%. Its high precision, recall, and F1-score values showcased its effectiveness in correctly identifying positive cases while minimizing false positives and false negatives. This combination harnessed the strengths of both T2T-ViT and DenseNet201 models, resulting in significant improvements in classification performance. Based on the insights gained from the ablation study, the proposed approach outperformed all other combinations, achieving the highest accuracy, precision, recall, and F1-score of 98.21%. Replacing the SVM component with a residual neural network proves to be a successful modification, enhancing the model's classification performance.

V. CONCLUSION

In this paper, we introduce a novel approach that combines CNN and ViT techniques to enhance the performance of COVID-19 diagnosis models. Given the limitations of traditional diagnostic methods in the timely and accurate screening of potentially infected individuals during the COVID-19 pandemic, AI-powered computer-aided imaging analysis techniques have emerged as promising alternatives. By harnessing the spatial feature extraction capabilities of CNN and the self-attention mechanisms of ViT inspired by human radiologists, our proposed hybrid model achieves state-of-the-art performance in accurately identifying COVID-19 cases on two benchmark datasets, Chest X-ray and Clean-CC-CCII. The fusion of CNN and ViT features enables a more comprehensive analysis of medical images, facilitating improved diagnostic accuracy and aiding in efforts to combat the ongoing pandemic.

In the future, we will focus on further advancing the proposed hybrid approach for COVID-19 diagnosis. The effectiveness of the feature fusion strategy of CNN and ViT should be explored and optimized in different scenarios and datasets. Additionally, efforts should be made to expand the evaluation to larger and more diverse datasets to assess the generalizability of the hybrid model. Furthermore, incorporating additional clinical and demographic data into the model can potentially enhance its diagnostic capabilities and provide more valuable insights. By continued research and refinement of the proposed approach, we can contribute to the collective endeavors to mitigate the impact of the ongoing COVID-19 pandemic and improve the effectiveness of diagnostic methods in similar public health crises.

REFERENCES

[1] K. Lazar. *COVID is Still Killing People Every Day. But Its Main Victims Have Changed*. Accessed: Jul. 3, 2023. [Online]. Available: https://www.bostonglobe.com/2023/02/20/metro/covid-is-still-killing-people-every-day-its-main-victims-have-changed/

[2] J. Murphy. *COVID Hospitalizations: See the Latest Trend and Current Count*. Accessed: Jul. 3, 2023. [Online]. Available: https://www.nbcnews.com/datagraphics/covid-hospitalizations-see-latest-trend-current-count-rcna61053

[3] F. Azour and A. Boukerche, "An efficient transfer and ensemble learning based computer aided breast abnormality diagnosis system," *IEEE Access*, vol. 11, pp. 21199–21209, 2023.

[4] G. Moon, S. Kim, W. Kim, Y. Kim, Y. Jeong, and H.-S. Choi, "Computer aided facial bone fracture diagnosis (CA-FBFD) system based on object detection model," *IEEE Access*, vol. 10, pp. 79061–79070, 2022.

[5] J. Onno, F. Ahmad Khan, A. Daftary, and P.-M. David, "Artificial intelligence-based computer aided detection (AI-CAD) in the fight against tuberculosis: Effects of moving health technologies in global health," *Social Sci. Med.*, vol. 327, Jun. 2023, Art. no. 115949.

[6] 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.*, vol. 30, 2017, pp. 1–12.

[7] 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*.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision–ECCV 2014*. Cham, Switzerland: Springer, 2014, pp. 740–755.

[10] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7073–7083.

[11] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision transformers for land-cover segmentation and classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1421–1430.

[12] K. A. Phung, T. T. Nguyen, N. Wangad, S. Baraheem, N. D. Vo, and K. Nguyen, "Disease recognition in X-ray images with doctor consultation-inspired model," *J. Imag.*, vol. 8, no. 12, p. 323, Dec. 2022.

[13] X. He, S. Wang, X. Chu, S. Shi, J. Tang, X. Liu, C. Yan, J. Zhang, and G. Ding, "Automated model design and benchmarking of deep learning models for COVID-19 detection with chest ct scans," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 6, pp. 4821–4829.

[14] G. Jia, H.-K. Lam, and Y. Xu, "Classification of COVID-19 chest X-ray and CT images using a type of dynamic CNN modification method," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104425.

[15] Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, J. Chen, R. Wang, H. Zhao, Y. Chong, J. Shen, Y. Zha, and Y. Yang, "Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2775–2780, Nov. 2021.

[16] H. Barzekar and Z. Yu, "C-Net: A reliable convolutional neural network for biomedical image classification," *Expert Syst. Appl.*, vol. 187, Jan. 2022, Art. no. 116003.

[17] U. C. Aytaç, A. Güneş, and N. Ajlouni, "A novel adaptive momentum method for medical image classification using convolutional neural network," *BMC Med. Imag.*, vol. 22, no. 1, pp. 1–12, Dec. 2022.

[18] A. S. Musallam, A. S. Sherif, and M. K. Hussein, "A new convolutional neural network architecture for automatic detection of brain tumors in magnetic resonance imaging images," *IEEE Access*, vol. 10, pp. 2775–2782, 2022.

[19] Z. Sun, H. Jiang, L. Ma, Z. Yu, and H. Xu, "Transformer based multi-view network for mammographic image classification," in *Proc. 25th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Singapore: Springer, 2022, pp. 46–54.

[20] J. Xu, Y. Gao, W. Liu, K. Huang, S. Zhao, L. Lu, X. Wang, X.-S. Hua, Y. Wang, and X. Chen, "RemixFormer: A transformer model for precision skin tumor differential diagnosis via multi-modal imaging and non-imaging data," in *Proc. 25th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 624–633.

[21] F. Almalik, M. Yaqub, and K. Nandakumar, "Self-ensembling vision transformer (SEViT) for robust medical image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 376–386.

[22] K. S. Ananda Kumar, A. Y. Prasad, and J. Metan, "A hybrid deep CNN-Cov-19-Res-Net transfer learning architype for an enhanced brain tumor detection and classification scheme in medical image processing," *Biomed. Signal Process. Control*, vol. 76, Jul. 2022, Art. no. 103631.

[23] W. Liu, T. Tian, W. Xu, H. Yang, X. Pan, S. Yan, and L. Wang, "PHTrans: Parallelly aggregating global and local representations for medical image segmentation," in *Proc. 25th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 235–244.

[24] Q. Zhou, S. Ye, M. Wen, Z. Huang, M. Ding, and X. Zhang, "Multi-modal medical image fusion based on densely-connected high-resolution CNN and hybrid transformer," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21741–21761, Dec. 2022.

[25] M. M. M. Rocha, G. Landini, and J. B. Florindo, "Medical image classification using a combination of features from convolutional neural networks," *Multimedia Tools Appl.*, vol. 82, no. 13, pp. 19299–19322, May 2023.

[26] F. Yuan, Z. Zhang, and Z. Fang, "An effective CNN and transformer complementary network for medical image segmentation," *Pattern Recognit.*, vol. 136, Apr. 2023, Art. no. 109228.

[27] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: A robust vision transformer for generalized medical image classification," *Comput. Biol. Med.*, vol. 157, May 2023, Art. no. 106791.

[28] J. Jang and D. Hwang, "M3T: Three-dimensional medical image classifier using multi-plane and multi-slice transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20686–20697.

[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[30] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 538–547.

[31] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 15908–15919.

[32] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10809–10819.

[33] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 357–366.

[34] X. Wu, C. Chen, M. Zhong, J. Wang, and J. Shi, "COVID-AL: The diagnosis of COVID-19 with deep active learning," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101913.

[35] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, and K. Wang, "Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423–1433, Jun. 2020.

[36] A. Arsenos, D. Kollias, and S. Kollias, "A large imaging database and novel deep neural architecture for COVID-19 diagnosis," in *Proc. IEEE 14th Image, Video, Multidimensional Signal Process. Workshop (IVMSP)*, Jun. 2022, pp. 1–5.

TRONG-THUAN NGUYEN (Student Member, IEEE) received the B.S. degree in data science from the University of Information Technology, Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam, in 2022. He is currently a Researcher with the Software Engineering Laboratory (SELab), University of Science, VNU-HCM. His research interests include machine learning and computer vision.

TAM V. NGUYEN (Senior Member, IEEE) received the Ph.D. degree from the National University of Singapore, in 2013. He was a Research Scientist and a Principal Investigator with the ARTIC Research Centre, Singapore Polytechnic. He was also an Adjunct Lecturer with the National University of Singapore. In 2016, he joined the University of Dayton, Dayton, OH, USA, as an Assistant Professor, where he is currently an Associate Professor with the Department of Computer Science. His research interests include computer vision, applied deep learning, multimedia content analysis, and mixed reality.

MINH-TRIET TRAN (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Science, Vietnam National University Ho Chi Minh City (VNU-HCM), in 2001, 2005, and 2009, respectively. In 2001, he joined the University of Science. He was a Visiting Scholar with the National Institutes of Informatics (NII), Japan, from 2008 to 2010, and the University of Illinois at Urbana–Champaign (UIUC), from 2015 to 2016. He is currently the Vice President of the University of Science, VNU-HCM, and the Director of the John Von Neumann Institute, VNU-HCM. His research interests include computer vision and human–computer interaction, cryptography and security, and software engineering. He is a Membership Development and Student Activities Coordinator of IEEE Vietnam. He is also a member of the Advisory Council for Artificial Intelligence Development of Ho Chi Minh City and the Vice Chairperson of the Vietnam Information Security Association (VNISA), South Branch.