IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 March 2024, accepted 11 May 2024, date of publication 16 May 2024, date of current version 4 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3401777

RESEARCH ARTICLE

A Multimodal Transfer Learning Approach Using PubMedCLIP for Medical Image Classification

HONG N. DAO[1], **TUYEN NGUYEN**[1,2], **CHERUBIN MUGISHA**[1], **(Member, IEEE), AND INCHEON PAIK**[1], **(Senior Member, IEEE)**
[1]Department of Computer and Information Systems, The University of Aizu, Aizuwakamatsu, Fukushima 965-8580, Japan
[2]School of Computer Science, University of Technology Sydney, Ultimo, NSW 2007, Australia

Corresponding author: Incheon Paik (paikic@u-aizu.ac.jp)

ABSTRACT Medical image data often face the problem of data scarcity and costly annotation processes. To overcome this, our study introduces a novel transfer learning method for medical image classification. We present a multimodal learning framework that incorporates the pre-trained PubMedCLIP model and multimodal feature fusion. Prompts of different complexities are combined with images as inputs to the proposed model. Our findings demonstrate that this approach significantly enhances image classification tasks while reducing the burden of annotation costs. Our study underscores the potential of PubMedCLIP in revolutionizing medical image analysis through its prompt-based approach and showcases the value of multi-modality for training robust models in healthcare. Code is available at:https://github.com/HongJapan/MTL_prompt_medical.git.

INDEX TERMS Pre-trained model, medical image, classification task, contrastive language-image pre-training, feature fusion, multimodal model, prompt engineering.

I. INTRODUCTION

Deep learning (DL) is a powerful technique that facilitates significant advancements in medical image analysis [1], [2], [3]. However, training DL models can be challenging, especially when faced with limited data in the medical domain.

To address the issue of data scarcity, transfer learning (TL) has been introduced [4]. TL involves transferring pre-trained knowledge from a source task to a similar task. This approach not only reduces training time [5], but also proves beneficial when the target task lacks training data [6], [7]. TL can be applied using two main approaches: (*i*) utilizing a pre-trained network as a feature extractor and training a new classifier using the extracted features [8], [9], or (*ii*) fine-tuning the pre-trained network to suit the new task requirements [10].

In the medical domain, TL has been widely employed in medical image classification, addressing the limited availability of labeled medical image datasets. TL has been shown to enhance the performance of DL models for tasks such as breast cancer classification [11], [12], lung nodule classification [13], [14], and brain tumor classification [15], while reducing the need for extensive labeled data during training. However, despite its successes, applying TL to medical image classification remains challenging. Medical images possess unique characteristics which make it non-trivial to apply pre-trained models. Furthermore, previous studies on TL in medical image classification have primarily focused on a specific case (or a single dataset), for example digital pathology image analysis [16]. The transferability of pre-trained models across different medical image datasets and tasks requires further investigation.

One highly promising pre-trained model for transfer learning is the Contrastive Language-Image Pretraining (CLIP) model, introduced by OpenAI in 2021 [17]. CLIP stands as a state-of-the-art model that establishes associations between images and text through extensive training on a diverse collection of image-text pairs. However, it is worth noting that while the CLIP approach performs admirably in general data domains, it was initially trained on publicly available internet data. Consequently, it lacks domain-specific knowledge, particularly in specialized fields like medicine. To address this limitation, Eslami et al. introduced PubMedCLIP [18], a fine-tuned adaptation of CLIP tailored

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

for the medical domain. Their study revealed that leveraging the pre-trained PubMedCLIP features enhances visual question-answering (VQA) performance, surpassing current state-of-the-art baseline models.

In this work, we propose a model that takes advantage of PubMedCLIP's image and text feature representations. The robust visual-language representations allow our model to handle cases with limited training data. Experimental results demonstrate that the proposed multimodal model achieves excellent results in classifying medical images from different datasets. This paper is an extended version of our previous work [19]. Compared to [19], the main extensions are as follows. First, multiple prompts of different complexities are considered. Interestingly, it is shown that a richer prompt leads to much higher gains in classification accuracy. Second, a better feature fusion method is employed to further improve the performance. Third, two more datasets are used and more experiments are carried out, resulting in many insights into the behaviors of the model and reference methods.

The remainder of this paper is organized as follows. Section II presents related work on transfer learning and multimodal learning models. Section III describes the proposed approach and experimental setup. Extensive experimental results and discussions are provided in Section IV. Finally, conclusions are given in Section V

II. RELATED WORK

In this section, we review previous work related to TL in medical image classification, including multimodal models and the applications of pre-trained models.

A. TRANSFER LEARNING IN MEDICAL IMAGE CLASSIFICATION

Transfer learning has been employed in medical image classification to enhance model performance, particularly when training data is limited. This approach enables models to leverage knowledge of a pre-trained model learned on large datasets to improve the performance on smaller, domain-specific datasets. This saves time and costs, which is crucial in the medical imaging domain where datasets can be relatively small. Previous work related to TL in medical image classification can be categorized as follows. (*i*) Feature extraction: A common approach is to use a pre-trained model such as VGG [20], MobileNet [21], DenseNet [22], or EfficientNet [23] as the feature extractors and then train a classifier on top of the extracted features. This approach has been shown to improve classification accuracy in various cases [24], [25], [26]. (*ii*) Fine-tuning a pre-trained model: This approach involves adapting a pre-trained model specifically for the medical image classification task. The parameters of a pre-trained model are updated by training on the target dataset. Fine-tuning has proven to be effective in medical image classification tasks, such as colonoscopy frame classification [27], [28]. (*iii*) Multi-task learning: This approach involves training a model simultaneously on multiple related tasks. In medical image classification, multi-task learning has been used to improve the accuracy of models by leveraging the relationship between different medical imaging tasks [29], [30]. (*iv*) Domain adaptation: Domain adaptation in TL involves adapting a model trained on a source domain to a target domain with different distributions. In medical image classification, this approach has been used to address the problem of data imbalance and improve model performance on specific target domains [31]. TL has shown practicality in improving the performance of medical image classification models. However, these techniques result in high computational costs as discussed in [13] and [32]. Besides, not all pre-trained models that have been trained on large-scale natural image datasets perform optimally across all medical image modalities. For instance, a review paper by Morid et al. [33] highlighted that Inception models were commonly utilized in analyzing X-rays, endoscopic images, and ultrasound images, while GoogLeNet and AlexNet were frequently employed for MRI analysis. On the other hand, VGGNet models were mostly used in studying skin lesions, fundus images, and OCT (optical coherence tomography) data.

Recently, more advanced pre-trained models have been investigated (see Table 1). In [36], Ohata et al. considered 18 different image encoders in transfer learning for Path images. They showed that the best result of the experiment was provided by the DenseNet. In the research of Jimenez et al. [37] on breast tumor classification, DenseNet also demonstrated high accuracy in diagnosing benign and malignant tumors when compared with different pre-trained models. Similarly, Sharma et al. [39] employed DenseNet model with preprocessing techniques like normalization and data augmentation for Blood images. Meanwhile, in the study of Shaban et al. [34], they demonstrated that MobileNet exhibits superior performance, achieving the highest average accuracy compared to various classifiers on Path images. Also, Eroglu et al. [35] found that the highest accuracy was obtained with MobileNet features for Breast images. Kallipolitis et al. [38] utilized transfer learning with various pre-trained models on a dataset that is augmented by the Grad-CAM technique to highlight visual patterns relevant to each class. The experimental results showed that EfficientNet outperformed other models. In the study of Chola et al. [40], they employed EfficientNet as the backbone for Blood images which are pre-processed by image processing. In a comparison of different deep learning models for mamography breast images, Jafari et al. [41] demonstrated that among the individual models, EfficientNet consistently outperformed the others. Our study in [26] was the first to employ PubMedCLIP for medical image classification on various image types of MedMNIST dataset. However, the that solution is still unimodal, relying solely on image modality.

B. MULTIMODAL LEARNING

In recent years, there has been increasing interest in using both text and image data as input for medical

TABLE 1. Recent transfer learning studies on medical images.

Reference	Year	Pre-trained model	Image type	Note
[34]	2020	MobileNet	Path	Unimodal (Image)
[35]	2021	MobileNet	Breast	Unimodal (Image)
[36]	2021	DenseNet	Path	Unimodal (Image)
[37]	2021	DenseNet	Breast	Unimodal (Image)
[38]	2021	EfficientNet	Path	Unimodal (Image)
[39]	2022	DenseNet	Blood	Unimodal (Image)
[40]	2022	EfficientNet	Blood	Unimodal (Image)
[26]	2022	PubMedCLIP	Various image types of MedMNIST (Path; Pneumonia; Blood; Breast)	Unimodal (Image)
[41]	2023	EfficientNet	Breast	Unimodal (Image)
[19]	2023	PubMedCLIP	Breast	Multimodal (Image + text) No Prompt engineering
This paper	-	PubMedCLIP	Path; Blood; Breast	Multimodal (Image + text) With Prompt engineering

image analysis. Combining these two modalities allows for capturing both visual and semantic information, leading to improved accuracy and interpretability of classification results. Several recent studies have utilized medical reports to provide supervision information and learn multimodal representations by maximizing mutual information between the two input modalities [42], [43], [44]. Extracting labels from reports using natural language processing (NLP) has also been explored as a means to leverage information from the text [45], [46]. Transformer-based vision-and-language models are used for learning multimodal representations from image and associated reports, which outperform traditional CNN and RNN methods [47]. Attention mechanism have also been used to facilitate interactions between visual and semantic information [48].

Recently, Contrastive Language-Image Pretraining (CLIP) is an advanced pretrained model developed by OpenAI [17]. It applies contrastive learning with a huge dataset of 400 million image-text pairs obtained from the Internet. As a consequence of this multimodal training, CLIP can be used to find the text snippet that best represents a given image, or the most suitable image given a text query. One of the interesting advantages of CLIP is its ability to perform zero-shot learning [17]. Also, the high performance of CLIP features enables many new exciting applications, for example, pre-training model to address the challenge of limited labeled data [49], art classification [50], and image captioning [51].

In the medical domain, Eslami et al. [18] investigates the effectiveness of the pre-trained CLIP model for visual question answering (VQA) task. To tailor the CLIP model for applications in the medical field, the authors introduced the PubMedCLIP model by fine-tuning the original CLIP model. This approach employs pairs of medical images and associated text of various anatomical regions from the medical ROCO dataset [52].

In line with the new trend of using LMM in machine learning, our preliminary work [19] introduced the first multimodal transfer learning approach using PubMedCLIP, where text and image features are combined for classifying Breast images. In this paper, we present an extended solution with a new fusion method and prompt engineering. As a result, the proposed method can work with a small number of data samples and has good performance across different datasets.

III. METHODOLOGY

A. THE PROPOSED MULTIMODAL MODEL

As mentioned, our method aims to utilize the powerful multimodal representations of the PubMedCLIP. The method takes as input both an image and a description text. First, the image and text are encoded using PubMedCLIP, which produces a vector representation for each modality. These vector representations are then fed into a fusion module to produce a combined feature vector, which is used to predict a similarity score. Finally, the similarity scores are employed for classification.

The proposed model consists of three main stages: feature extraction, feature fusion, and class prediction. As shown in Figure 1, in the first stage, the image feature extraction component takes a medical image v as input and outputs an image feature vector $\vec{v}_i$. Similarly, the text feature extraction component will generate the text feature vector $\vec{q}_j$ for an input text description of image class q_j. For each pair of $(\vec{v}_i, \vec{q}_j)$, the feature fusion component produces a combined vector $\vec{Z}_{v_i,q_j}$ which is used to compute the similarity score between the image and the text.

We perform image feature extraction with two options, PubMedCLIP-RN50 and PubMedCLIP-ViT32. These two encoders are based on different technologies, namely CNN (PubMedCLIP-RN50) and Vision Transformer (PubMedCLIP-ViT32). This helps to see behaviors of CNN and Vision Transformer over different medical imaging modes in our study, including microscopic imaging and ultrasound scan imaging.

In our approach to effectively utilize image labels for model training, we draw inspiration from the methodology described in Radford et al.'s paper [17]. This approach acknowledges the importance of connecting text prompts with image content, a technique that has demonstrated enhanced performance compared to using simple labels alone [17]. In particular, it is shown that adding a simple word like "image" into the prompt can improve the performance. So, in this work, we consider a heuristic approach that

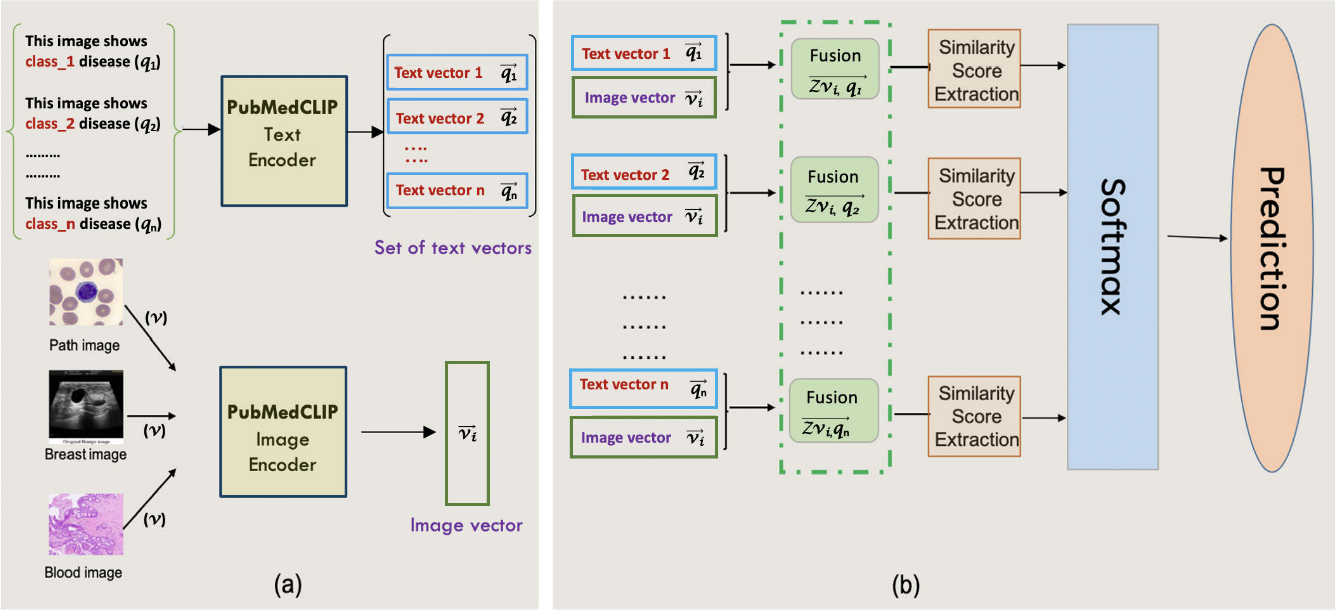

This image shows class_1 disease (q_1)
This image shows class_2 disease (q_2)
....
....
This image shows class_n disease (q_n)
PubMedCLIP Text Encoder
Text vector 1 $\vec{q_1}$
Text vector 2 $\vec{q_2}$
....
Text vector n $\vec{q_n}$
Set of text vectors
Path image
Breast image
Blood image
(v)
PubMedCLIP Image Encoder
$\vec{v_i}$
Image vector
(a)
Text vector 1 $\vec{q_1}$
Image vector $\vec{v_i}$
Fusion $\vec{Zv_i, q_1}$
Similarity Score Extraction
Text vector 2 $\vec{q_2}$
Fusion $\vec{Zv_i, q_2}$
Text vector n $\vec{q_n}$
Fusion $\vec{Zv_i, q_n}$
Softmax
Prediction
(b)

FIGURE 1. **Overview of our model. We feed the original image and label templates to the PubMedCLIP-text encoder and PubMedCLIP-Image encoder. Fusion technique is used to combine the two vectors. Finally, the softmax layer is added for classification the disease.**

gradually increases the contextual information in the prompt templates. The words we select for the prompts are commonly found in electrical health record (EHR), such as medical, image, disease, illness, symptom, sign, patient [53].

For each dataset, we have developed three distinct text prompt templates to guide the proposed model in the task of medical image classification. In addition to these prompts, we also include Prompt-0, which is simply the name of the label for each class. Specifically, the prompt templates are as follows.

1) Prompt-0: "{label}"
2) Prompt-1: "This image shows {label} disease".
3) Prompt-2: "In this medical image, there are indications of {label}".
4) Prompt-3: "Based on this medical image, it appears that the patient may be exhibiting signs or symptoms related to the {label} disease or illness".

As can be seen, these prompts offer varying levels of information, allowing the model to capture different aspects of the image. Specifically, Prompt-0 does not provide any additional context about the image, while more information is increasingly added to Prompt-1, Prompt-2, and Prompt-3. To facilitate this process, each dataset has a dictionary with descriptions of all the diseases present. These descriptions are encoded into text vectors, resulting in a set of text vectors specific to each dataset.

In the second stage, we combine the image and text features into a single feature vector using the feature fusion block. A straightforward approach for combining feature vectors is to multiply them element-wise. However, this method has limitations due to the poor interaction of the elements between the two vectors. Various fusion techniques have been developed to combine text and image feature vectors to maximize interactions. These approaches usually rely on the idea of making bilinear pooling computationally feasible. In this study, we employs the Multimodal Factorized Bilinear Pooling (MFB) method [54] for multimodal feature fusion because of its simplicity, ease of implementation, and a high convergence rate. MFB [54] is a pooling method that combines information from multiple modalities (e.g., image and text) by computing the outer product of their feature vectors and then factorizing the resulting matrix using a low-rank decomposition. This approach allows for efficient modeling of pairwise interactions between different modalities while reducing the feature dimensionality after pooling [55], [56]. A comparison of MFB with other fusion methods will be discussed in the next section.

In the third stage, class prediction is done based on combined feature vectors. Given a set of combined vectors $\{\vec{Z}_{v_i,q_j}\}$ for each pair of $(\vec{v}_i, \vec{q}_j)$, we employed a set of fully-connected layer blocks, each of which independently transforms $\vec{Z}_{v_i,q_j}$ to a scalar. These output scalar values will form the similarity scores between the image $\vec{v}_i$ and the text description $\vec{q}_j$. The blocks are denoted as Similarity Score Extraction modules in Figure 1b. Finally, a softmax layer normalizes the scores, yielding a probability distribution indicating the likelihood of the input image belonging to a description from the dictionary. The prediction is chosen by selecting the highest probability element from the distribution.

B. DATASETS

To conduct this research, we use three different medical datasets with different classes and imaging modes. The first is the Blood dataset, consisting of 17,092 microscopic peripheral blood cell images [11]. The images of this dataset are categorized into eight classes: neutrophils, eosinophils, basophils, lymphocytes, monocytes, immature granulocytes, erythroblasts, and platelets or thrombocytes. The second one is the Path dataset, containing 100,000 images of human colorectal cancer and healthy tissues [57]. The tissue images are organized into nine classes: adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM). The third is the Breast dataset containing 780 medical images of breast cancer using ultrasound scans [58]. The Breast dataset is organized into three classes: normal, benign, and malignant.

C. REFERENCE MODELS AND IMPLEMENTATION DETAILS

In order to evaluate the improvements of the proposed multimodal model with respect to previous multimodal and unimodal models, the following reference models are employed for our experiments.

- The multimodal model of [19], which is the preliminary version of our work. This model uses PubMedCLIP's image and text encoders without prompt engineering. Note that, in this model, we use only the Transformer-based encoder (PubMedCLIP-ViT32) because, as shown in [19] and [26], it is always better than the Resnet-based encoder. In the following, this model is denoted as PubMedCLIP-Multi.
- The unimodal model of [26] that only uses the image modality of PubMedCLIP. In the following, this model is denoted with two options PubMedCLIP-ViT32 and PubMedCLIP-RN50. Here, the image encoders of this unimodal model are exactly the same as those of the multimodal models.
- Three unimodal models using a popular pretrained model, namely DenseNet, MobileNet, or EfficientNet. As mentioned above, recent studies (e.g. [35], [36], [40]) just focus on a certain image type (e.g. Blood or Path), so their findings on the best pretrained model vary. In our evaluation, these models will be compared on the three datasets, using the same setting as the above unimodal and multimodal models.

To clearly see the performance differences of the models, our experiments use the same setup for all models. Especially, because we want to see the performances with a small amount of training data, no techniques of data augmentation and preprocessing are applied. The workflow of the unimodal models is shown in Figure 2, where the feature vector provided by a pre-trained model is input into a fully-connected layer for classification. For training of both multimodal and unimodal models, the learning rate is set to 1×10^{-3}, and the batch size is 16. All implementations are based on the PyTorch framework [59]. To obtain stable results, we repeat all experiments ten times and report the average scores over all experiment runs.

IV. EXPERIMENTS

In this section, we shows the performance comparison between the proposed model and the reference models on different datasets. We also perform extensive experiments with different fusion techniques, prompt templates, and different numbers of training samples.

A. EXPERIMENTAL SETTINGS

A key focus of our research was to examine how our model performs under conditions of limited training data. To achieve this, we gradually increase the number of training samples of each class. Specifically, we start with small numbers of training images per class, namely 10, 50, 100, and so on until eventually reaching 80% of the dataset. The images not used for training in each case are set aside for testing. We maintained the same setting for all evaluated models. The incremental increase in training data size enables us to explore the models' learning behaviors as they have access to more training samples. This provides valuable insights into the trade-off between training data volume and performance.

Our experiments evaluate the model's performance using accuracy as the primary metric to assess its ability to distinguish between various classes. The accuracy metric, represented by Equation 1, provides a comprehensive measure of the overall correctness of the model's predictions. The formula for accuracy metric is represented as follows:

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + F_N + T_N + F_P} \tag{1}$$

where T_P, T_N, F_P, F_N represent true positive, true negative, false positive, false negative, respectively.

B. EXPERIMENTAL RESULTS

1) FUSION TECHNIQUE COMPARISON

In the proposed model, to fuse the text and image vectors for prediction, we employed the MFB fusion technique. To show the benefit of this fusion technique, we compared this technique to two other popular fusion techniques, namely Multimodal Compact Bilinear Pooling(MCB) [60] and Multimodal Tucker Fusion (MUTAN) [61]. For simplicity, template Prompt-1 is used in this evaluation. In Figure 3, the performances of the proposed model using one of two vision backbones, PubMedCLIP-RN50 and PubMedCLIP-ViT32, together with the three fusion techniques are shown for the three datasets. For the Blood dataset, the results are shown in Figure 3(a), where both PubMedCLIP-RN50 and PubMedCLIP-ViT32 with MFB exhibit increasing accuracy as the number of shots is increased. When the number of shots exceeds 100, the curves reach high accuracy, around 90% for PubMedCLIP-ViT32 and around 85% for PubMedCLIP-RN50. However, when employing the MCB and MUTAN fusion techniques, the curves remain relatively flat, showing minimal improvement even when the number of shots is high. Moreover, the accuracies achieved by MCB and Mutan fusion techniques are significantly lower, approximately 70% for PubMedCLIP-ViT32 with Mutan, 58% for PubMedCLIP-RN50 with Mutan, 66% for PubMedCLIP-ViT32 with MCB, and 43% for PubMedCLIP-RN50 with MCB. With the Path dataset in

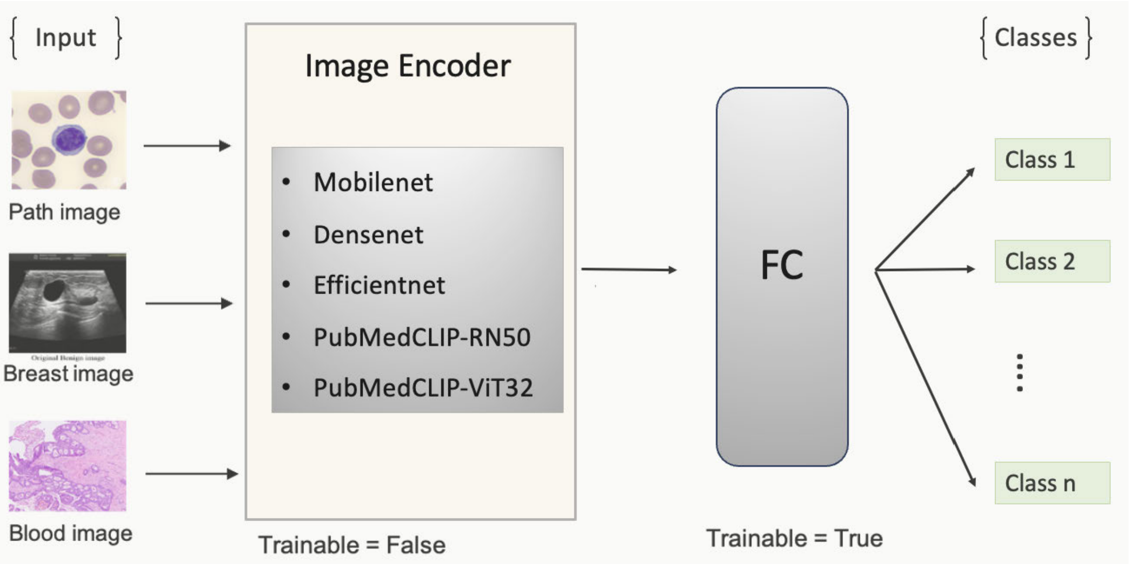
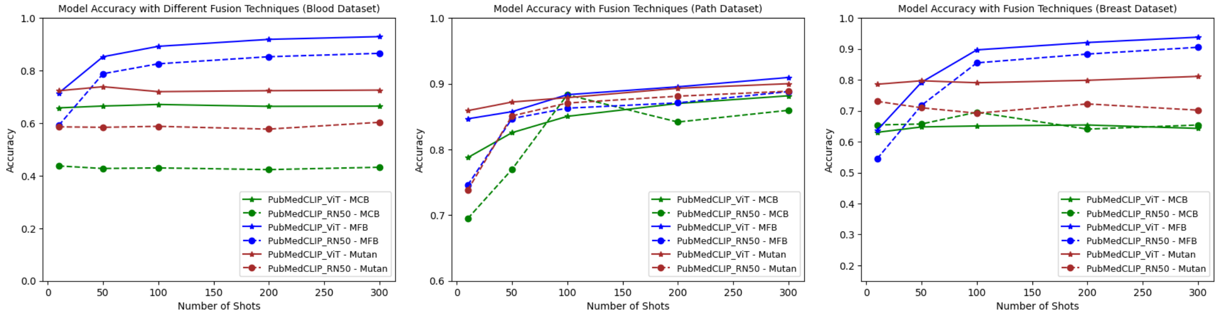


FIGURE 2. Unimodal model of transfer learning for medical image classification.



(a) Blood dataset (b) Path dataset (c) Breast dataset

FIGURE 3. Fusion technique comparison.

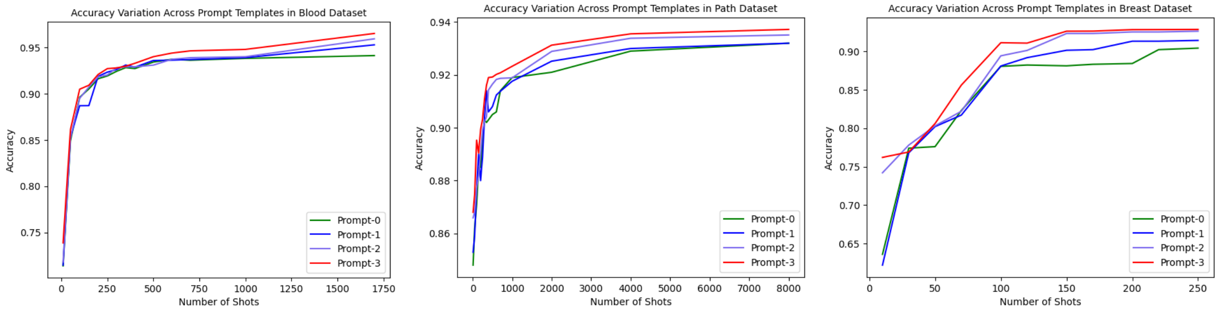
Figure 3(b), PubMedCLIP-ViT32 with MFB provides the highest curve among all the combinations. When the number of shots exceeds 200, the accuracy surpasses 90%. Besides, the MCB fusion technique provides highly unstable results. With the Breast dataset (Figure 3(c)), the behavior is similar to that in the Blood dataset. The MFB fusion technique demonstrates favorable results for both PubMedCLIP-RN50 and PubMedCLIP-ViT32, with increasing accuracy as the number of shots increased. However, the other fusion techniques show much lower results; the accuracy of Mutan with PubMedCLIP-ViT32 (PubMedCLIP-RN50) is consistently around 78% (70%). The MCB fusion technique results in about only 65% for both backbones. Among the three fusion techniques, MUTAN is only better than MFB at very small number of shots (e.g. 10 shots in Path and Breast datasets).

In summary, based on the experiment results, the MFB fusion technique in general shows the best performance across the Blood, Path, and Breast datasets, for both PubMedCLIP-RN50 and PubMedCLIP-ViT32 backbones. In the following evaluations, we will exclusively present the results obtained using the MFB fusion technique.

2) PROMPT TEMPLATE EVALUATION

In this part, our evaluation involves testing each prompt template's performance as the number of training samples is increased from 10 samples per class up to 80% of the class. For simplicity, only PubMedCLIP-ViT32 is used the image encoder. The results presented in Table 2 highlight the different performances of the prompt templates (i.e. Prompt-0, Prompt-1, Prompt-2, Prompt-3). Futhermore, the results consistently demonstrate that Prompt-3 outperformed Prompt-0, Prompt-1 and Prompt-2 in all datasets. Especially, on the Path dataset, the performance of Prompt-3 quickly jumps to a high level after 500 shots. Meanwhile, on the Breast dataset, the performance of Prompt-3 saturates after 100 shots.

Additionally, the visualization in Figure 4 confirms the Prompt-3's consistent and superior performance. The results show that the performance of Prompt-0 is the lowest. More specifically, in Fig. 4, we can see that adding the words "image" and "disease" in Prompt-1 can help improve the performance on Blood and Breast datasets when the number of shots is high, and on Path dataset when the number of shots is medium (from 1500 shots to 5000 shots). Also, in general,



(a) Blood dataset (b) Path dataset (c) Breast dataset

FIGURE 4. Performance comparison of different prompt templates.

TABLE 2. Accuracy values for different prompts.

Dataset	No. of shots	Prompt-0	Prompt-1	Prompt-2	Prompt-3
Blood	10	0.714	0.715	0.718	0.739
	50	0.849	0.852	0.852	0.861
	100	0.896	0.887	0.895	0.905
	150	0.905	0.887	0.907	0.909
	200	0.916	0.919	0.918	0.921
	250	0.919	0.923	0.920	0.927
	300	0.924	0.926	0.928	0.928
	350	0.928	0.931	0.929	0.930
	400	0.927	0.928	0.929	**0.933**
	500	0.934	0.936	0.931	**0.939**
	600	0.937	0.936	0.937	**0.944**
	700	0.936	0.937	0.939	**0.946**
	1000	0.938	0.939	0.940	**0.948**
	80% of data	0.941	0.953	0.959	**0.965**
Path	10	0.848	0.853	0.866	0.868
	50	0.862	0.859	0.868	0.875
	100	0.871	0.877	0.877	0.895
	150	0.885	0.880	0.882	0.890
	200	0.885	0.890	0.890	0.903
	250	0.896	0.889	0.899	0.903
	300	0.903	0.897	0.904	**0.911**
	400	0.902	0.907	0.914	**0.919**
	500	0.903	0.908	0.906	**0.919**
	600	0.905	0.912	0.908	**0.921**
	700	0.906	0.913	0.912	**0.929**
	1000	0.914	0.917	0.913	**0.930**
	80% of data	0.932	0.932	0.935	**0.937**
Breast	10	0.636	0.622	0.742	0.762
	50	0.776	0.802	0.803	0.806
	100	0.883	0.891	0.894	**0.911**
	150	0.881	0.902	0.923	**0.926**
	200	0.884	0.913	**0.925**	**0.928**
	80% of data	0.904	0.913	**0.925**	**0.928**

Prompt-2 has better performance than Prompt-1 at most numbers of shots. This observation emphasizes the pivotal role of prompt engineering in the model's performance. The success of Prompt-3 can be attributed to its provision of richer contextual information, which better guides the model in associating image content with the corresponding medical condition. In our future work, we will further investigate the potential of leveraging more intricate and informative language constructs to enhance the performance of multimodal models in medical image classification. In the upcoming evaluation experiments, we will exclusively present results using Prompt-3 in the proposed model.

3) MODEL PERFORMANCE ACCURACY RESULTS

In this section, we compare the performances of the proposed model and reference models on the three datasets. The experimental results are given in Table 3. The performances of the models vary across the datasets. Here, we specifically explore the performances when the number of training samples (shots) gradually increases.

For the 10-shot learning scenario, we trained the models using ten images per class from each dataset and utilized the remaining images for testing. The results indicate that the proposed model (PubMedCLIP-ViT32) achieves the highest or second-highest accuracy across the three datasets. In the Path dataset, our model achieves the highest accuracy score among the models. However, all models perform poorly in the Breast dataset under the ten-shot learning setting.

Notably, PubMedCLIP-ViT32 exhibits superior performance compared to PubMedCLIP-RN50. So, in the following, the proposed model that employs PubMedCLIP-ViT32 is mostly referred to in the discussion.

For the 50-shot learning scenario, we increased the training data to 50 images per class. The results show that as the number of training images increases, the overall accuracy of the models improves. Our multimodal model achieves relatively high scores across all three datasets, with accuracy exceeding 80%. Notably, DenseNet and MobileNet perform well on the Blood and Path datasets but poorly on the Breast dataset.

Moving on to the 100-shot learning scenario, we fed 100 images per class into the models for training. The results indicate that our model's accuracy increases slower than MobileNet and DenseNet when transitioning from 50 to 100 training images per class in the Blood and Path datasets. Specifically, MobileNet achieves an accuracy of approximately 90% in the Blood and Path datasets, while DenseNet achieves a similar accuracy in the Path dataset. Nevertheless, our model performs well across all three datasets, with the accuracy surpassing 88%. Notably, in the Breast dataset, our model achieves an accuracy of over 92%, whereas other models fall below 80%.

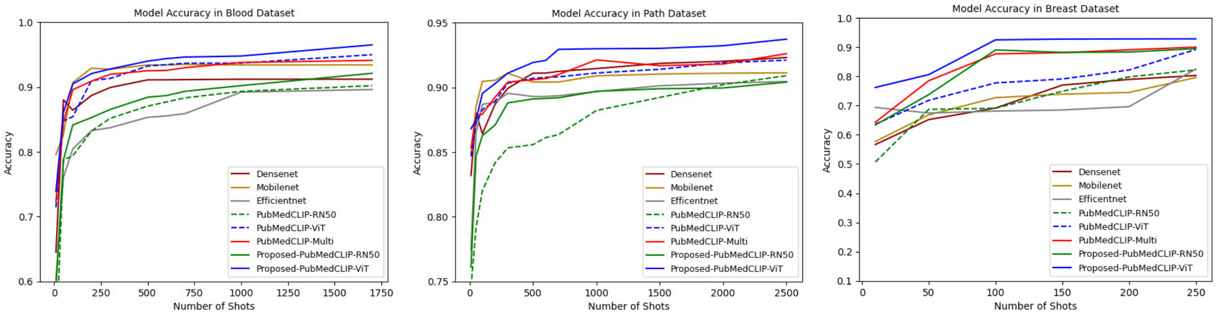
Further increasing the training data to 200 images per class, our model demonstrates outstanding performance across all three datasets. It achieves an accuracy of 92.1% in the Blood dataset, 90.3% in the Path dataset, and 92.8% in the Breast dataset, comparable to those of MobileNet. Compared to DenseNet, our model performs better by approximately 3% in the Blood dataset, 14% in the Breast dataset, and slightly

TABLE 3. Model performance.

Few shot	Pre-trained model	Blood dataset	Path dataset	Breast dataset
10-Shots	DenseNet	0.646	0.832	0.567
	MobileNet	0.795	0.849	0.577
	EfficientNet	0.603	0.774	0.694
	PubMedCLIP-RN50	0.497	0.746	0.507
	PubMedCLIP-ViT32	0.714	0.846	0.636
	PubMedCLIP-Multi	0.723	0.847	0.643
	Proposed-PubMedCLIP-RN50	0.691	0.761	0.634
	Proposed-PubMedCLIP-ViT32	0.739	0.858	0.762
50-Shots	DenseNet	0.880	0.889	0.652
	MobileNet	0.826	0.885	0.668
	EfficientNet	0.761	0.865	0.675
	PubMedCLIP-RN50	0.722	0.791	0.687
	PubMedCLIP-ViT32	0.847	0.873	0.778
	PubMedCLIP-Multi	0.852	0.872	0.785
	Proposed-PubMedCLIP-RN50	0.787	0.847	0.737
	Proposed-PubMedCLIP-ViT32	0.861	0.868	0.806
100-Shots	DenseNet	0.864	**0.902**	0.692
	MobileNet	**0.907**	**0.904**	0.727
	EfficientNet	0.804	0.869	0.681
	PubMedCLIP-RN50	0.794	0.821	0.691
	PubMedCLIP-ViT32	0.854	0.883	0.777
	PubMedCLIP-Multi	0.887	0.878	0.877
	Proposed-PubMedCLIP-RN50	0.841	0.862	0.890
	Proposed-PubMedCLIP-ViT32	**0.905**	0.895	**0.927**
200-Shots	DenseNet	0.887	**0.905**	0.789
	MobileNet	**0.929**	**0.905**	0.745
	EfficientNet	0.833	0.888	0.696
	PubMedCLIP-RN50	0.832	0.842	0.749
	PubMedCLIP-ViT32	0.910	0.889	0.822
	PubMedCLIP-Multi	0.911	0.892	0.891
	Proposed-PubMedCLIP-RN50	0.853	0.871	0.883
	Proposed-PubMedCLIP-ViT32	**0.921**	**0.903**	**0.928**
300-Shots	DenseNet	0.899	0.907	-
	MobileNet	**0.927**	**0.910**	-
	EfficientNet	0.837	0.895	-
	PubMedCLIP-RN50	0.851	0.853	-
	PubMedCLIP-ViT32	0.913	0.903	-
	PubMedCLIP-Multi	0.919	0.902	-
	Proposed-PubMedCLIP-RN50	0.865	0.888	-
	Proposed-PubMedCLIP-ViT32	**0.927**	**0.911**	-
500-Shots	DenseNet	0.911	0.908	-
	MobileNet	**0.933**	0.904	-
	EfficientNet	0.853	0.892	-
	PubMedCLIP-RN50	0.870	0.855	-
	PubMedCLIP-ViT32	**0.932**	0.907	-
	PubMedCLIP-Multi	0.924	0.911	-
	Proposed-PubMedCLIP-RN50	0.884	0.891	-
	Proposed-PubMedCLIP-ViT32	**0.939**	**0.919**	-
80% dataset	DenseNet	0.926	0.918	0.803
	MobileNet	**0.939**	0.915	0.796
	EfficientNet	0.873	0.912	0.825
	PubMedCLIP-RN50	0.902	0.895	0.822
	PubMedCLIP-ViT32	**0.949**	0.917	0.892
	PubMedCLIP-Multi	0.938	0.919	0.90
	Proposed-PubMedCLIP-RN50	0.921	0.918	0.892
	Proposed-PubMedCLIP-ViT32	**0.965**	**0.937**	**0.928**

lower by 0.2% in the Path dataset. When we increase the training data to 300 images per class, our model excels across all the datasets.

The dependence of model performances on the number of training samples and datasets can be seen more clearly in Figure 5. With the Blood dataset (Figure 5(a)), our model initially obtained the second-highest accuracy at 200 shots, trailing behind MobileNet. However, from 300 shots onward, the proposed model outperformed all other models. With the Path dataset, initially the proposed model again performs worse than MobileNet. However, at 500 shots, the result of MobileNet is lower than the proposed model. Especially with the Breast dataset (Figure 5(c)), the proposed model consistently achieved the highest accuracy across all numbers of shots. Meanwhile, all other models, including MobileNet, have much lower performances on this dataset. It can be concluded that the proposed model can consistently achieve good results across different datasets.



(a) Blood dataset (b) Path dataset (c) Breast dataset

FIGURE 5. Performance of the models on each dataset.

TABLE 4. Ablation study's settings and results.

Case	New Fusion	New Prompt	Dataset		
			Blood	Path	Breast
Case-1	-	-	0.941	0.919	0.898
Case-2	✓	-	0.953	0.934	0.915
Case-3	-	✓	0.945	0.923	0.911
Case-4	✓	✓	**0.965**	**0.937**	**0.928**

Regarding the multimodal model PubMedCLIP-Multi, its performances on Path and Blood datasets are comparable to the unimodal PubMedCLIP-ViT32 (Figure 5(a) and (b)); however, on Breast dataset, it is much better than PubMedLCIP-ViT32 and other unimodal models (Figure 5(c)). Among the unimodal models, PubMedCLIP-ViT32 is, in general, the best one over all three datasets, except at some small numbers of shots. Meanwhile, the performances of DenseNet, MobileNet, and EfficientNet vary across the datasets. Moreover, the proposed model's results are consistently highest over a wide range of the number of shots. This shows the promising capabilities of both multimodal and unimodal solutions based on PubMedCLIP, thanks to its very large scale.

C. ABLATION STUDY

In this part, we investigate the contributions of the two new components of the proposed model, including the new fusion and the new best prompt (i.e. Prompt-3). So, the comparison includes the following cases:

- Case-1: No new components (i.e. our preliminary model in [19])
- Case-2: Using the new fusion only.
- Case-3: Using the new prompt only.
- Case-4: Using the new fusion and the new prompt (i.e. the proposed model)

Here, for simplicity, we also employ only PubMedCLIP-ViT32, which is the best encoder for image modality. The accuracy results of the above four cases when training data is 80% of a dataset are shown in Table.4. It can be seen that the gains by the new fusion can only be up to 1.7%. Meanwhile, the gains by the new prompt are up to 1.3% and lower than the gains by the new fusion. When both new fusion and new prompt are used, the gains are 2.4%, 1.8%, and 3% on the Blood, Path, Breast datasets, respectively. These results mean that each new component can improve the performance, and when they are combined, the joint improvement is higher than individual improvements. So, the two new components are complementary to each other, and both are beneficial for the high performance of the proposed model.

D. DISCUSSIONS

The above results demonstrate the capabilities of the proposed model, which outperforms reference models in two aspects:

- The superior performances are consistent across three different image types. Whereas previous studies just focus on a certain type (e.g., either Blood, Path, or Breast).
- The behavior is also consistent over a wide range of the number of shots. It should be noted that existing studies mostly try to enlarge the amount of training data (e.g. by various data augmentation techniques) to improve the performance.

The advantages of the proposed model can be attributed to the robustness (or generalizability) of the large-scale and multimodal nature of the pre-trained PubMedCLIP model, together with prompt engineering and feature fusion.

It should be noted that the image encoder in the proposed model is the same (i.e., unmodified) as those used in unimodal models (using either PubMedCLIP-RN50 or PubMedCLIP-ViT32). However, thanks to the processing of both image input and text input, the proposed multimodal model always outperforms the corresponding unimodal model. This is an interesting benefit of large multimodal models like PubMedCLIP.

In addition, the experiments show that PubMedCLIP-ViT32 always performs better than

PubMedCLIP-RN50 in both unimodal and multimodal cases. On the Blood dataset, the unimodal model using PubMedCLIP-ViT32 is only worse than the multimodal model using PubMedCLIP-ViT32, which is even better than all other unimodal and multimodal models. This means the vision transformer technology is more effective than CNN in this classification task.

Our results also emphasize the importance of text prompt engineering to enhance a model's performance. In our study, adding more medical context into the prompt template helps the model understand more about the image that the model needs to classify. The improved performance when incorporating such keywords into the prompt can be attributed to the unique capabilities of the PubMedCLIP model, which is a fine-tuned version of CLIP tailored for medical applications. PubMedCLIP has been trained with a huge amount of images and associated text. A text prompt can be considered as a context input into the multimodal model. It seems that when appropriate words are provided in the prompt, the context will be clearer to the model, and thus, the performance at the output will be higher. So, it is important to empower the model with a richer context rather than a simple label or short description.

Furthermore, our model's robustness in image classification accuracy is fortified by fusing feature vectors of image and text inputs. This fusion of image and text vectors, coupled with an extensive text vector dictionary, equips our model to tackle a broad spectrum of medical conditions, ensuring consistent high accuracy across diverse image classification tasks. This multifaceted solution has been shown to be beneficial in medical image classification, with limited training data and adaptability across various datasets.

V. CONCLUSION AND FUTURE WORK

In this work, we have investigated the capability of transfer learning based on PubMedCLIP for medical image classification. We proposed a multimodal model that harnesses text prompts and images to achieve high accuracy even with limited training data, surpassing the performance of traditional transfer learning models. The advantages of the proposed model could be attributed to the multimodal pretrained backbones, prompt engineering, and feature fusion. Especially, the effective use of prompt templates in our model highlights its potential for various image classification domains. For future work, we will extend this approach by enhancing prompts through developing automated or context-aware prompts, which may improve the model's performance across diverse domains. Additionally, we will further evaluate the adaptability of the proposed model to various medical subfields and exploring cross-domain applications.

REFERENCES

[1] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *J. Pathol. Informat.*, vol. 7, no. 1, p. 29, Jan. 2016.

[2] H. Sharma, N. Zerbe, I. Klempert, O. Hellwich, and P. Hufnagl, "Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology," *Computerized Med. Imag. Graph.*, vol. 61, pp. 2–13, Nov. 2017.

[3] D. S. W. Ting et al., "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, p. 2211, Dec. 2017.

[4] R. Ribani and M. Marengoni, "A survey of transfer learning for convolutional neural networks," in *Proc. 32nd SIBGRAPI Conf. Graph., Patterns Images Tuts.*, Oct. 2019, pp. 47–57.

[5] D. Sarkar, R. Bali, and T. Ghosh, *Hands-On Transfer Learning With Python: Implement Advanced Deep Learning and Neural Network Models Using TensorFlow and Keras*. Birmingham, U.K.: Packt Publishing Ltd, 2018.

[6] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–18.

[7] O. Hadad, R. Bakalo, R. Ben-Ari, S. Hashoul, and G. Amit, "Classification of breast lesions using cross-modal deep learning," in *Proc. IEEE 14th Int. Symp. Biomed. Imag.*, Apr. 2017, pp. 109–112.

[8] S. Saxena, S. Shukla, and M. Gyanchandani, "Pre-trained convolutional neural networks as feature extractors for diagnosis of breast cancer using histopathology," *Int. J. Imag. Syst. Technol.*, vol. 30, no. 3, pp. 577–591, Sep. 2020.

[9] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *Proc. Int. Conf. Image Anal. Recognit.*, 2018, pp. 737–744.

[10] A. Mahbod, I. Ellinger, R. Ecker, Ö. Smedby, and C. Wang, "Breast cancer histological image classification using fine-tuned deep network fusion," in *Proc. 15th Int. Conf*, 2018, pp. 754–762.

[11] A. Acevedo, S. Alférez, A. Merino, L. Puigví, and J. Rodellar, "Recognition of peripheral blood cell images using convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 180, Oct. 2019, Art. no. 105020.

[12] Y. Wang, E. J. Choi, Y. Choi, H. Zhang, G. Y. Jin, and S.-B. Ko, "Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning," *Ultrasound Med. Biol.*, vol. 46, no. 5, pp. 1119–1132, May 2020.

[13] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[14] G. Liang and L. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," *Comput. Methods Programs Biomed.*, vol. 187, Apr. 2020, Art. no. 104964.

[15] A. Tiwari, S. Srivastava, and M. Pant, "Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019," *Pattern Recognit. Lett.*, vol. 131, pp. 244–260, Mar. 2020.

[16] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis," *Comput. Biol. Med.*, vol. 128, Jan. 2021, Art. no. 104129.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[18] S. Eslami, G. de Melo, and C. Meinel, "Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain?" 2021, *arXiv:2112.13906*.

[19] H. N. Dao, T. Nguyen, C. Mugisha, and I. Paik, "A multimodal transfer learning approach for medical image classification," in *Proc. IEEE Int. Conf. Consum. Electronics-Asia (ICCE-Asia)*, Oct. 2023, pp. 1–18.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[22] 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.

[23] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[24] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan, and A. Mittal, "Pneumonia detection using CNN based feature extraction," in *Proc. IEEE Int. Conf. Electr., Comput. Commun. Technol. (ICECCT)*, Feb. 2019, pp. 1–7.

[25] T. Kaur and T. K. Gandhi, "Automated brain image classification based on VGG-16 and transfer learning," in *Proc. Int. Conf. Inf. Technol. (ICIT)*, Dec. 2019, pp. 94–98.

[26] H. N. Dao, T. N. Quang, and I. Paik, "Transfer learning for medical image classification on multiple datasets using PubMedCLIP," in *Proc. IEEE Int. Conf. Consum. Electronics-Asia (ICCE-Asia)*, Oct. 2022, pp. 1–4.

[27] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[28] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4761–4772.

[29] J. Li, G. Zhao, Y. Tao, P. Zhai, H. Chen, H. He, and T. Cai, "Multi-task contrastive learning for automatic CT and X-ray diagnosis of COVID-19," *Pattern Recognit.*, vol. 114, Jun. 2021, Art. no. 107848.

[30] J. Gan, L. Xiang, Y. Zhai, C. Mai, G. He, J. Zeng, Z. Bai, R. Donida Labati, V. Piuri, and F. Scotti, "2M BeautyNet: Facial beauty prediction based on multi-task transfer learning," *IEEE Access*, vol. 8, pp. 20245–20256, 2020.

[31] P. Zhang, J. Li, Y. Wang, and J. Pan, "Domain adaptation for medical image segmentation: A meta-learning method," *J. Imag.*, vol. 7, no. 2, p. 31, Feb. 2021.

[32] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018.

[33] M. A. Morid, A. Borjali, and G. Del Fiol, "A scoping review of transfer learning research on medical image analysis using ImageNet," *Comput. Biol. Med.*, vol. 128, Jan. 2021, Art. no. 104115.

[34] M. Shaban, R. Awan, M. M. Fraz, A. Azam, Y.-W. Tsang, D. Snead, and N. M. Rajpoot, "Context-aware convolutional neural network for grading of colorectal cancer histology images," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2395–2405, Jul. 2020.

[35] Y. Eroğlu, M. Yildirim, and A. Çinar, "Convolutional neural networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mRMR," *Comput. Biol. Med.*, vol. 133, Jun. 2021, Art. no. 104407.

[36] E. F. Ohata, J. V. S. D. Chagas, G. M. Bezerra, M. M. Hassan, V. H. C. de Albuquerque, and P. P. R. Filho, "A novel transfer learning approach for the classification of histological images of colorectal cancer," *J. Supercomput.*, vol. 77, no. 9, pp. 9494–9519, Sep. 2021.

[37] Y. Jiménez Gaona, M. J. Rodriguez-Alvarez, H. Espino-Morato, D. Castillo Malla, and V. Lakshminarayanan, "Densenet for breast tumor classification in mammographic images," in *Proc. Int. Conf. Bioeng. Biomed. Signal Image Process.*, 2021, pp. 166–176.

[38] A. Kallipolitis, K. Revelos, and I. Maglogiannis, "Ensembling EfficientNets for the classification and interpretation of histopathology images," *Algorithms*, vol. 14, no. 10, p. 278, Sep. 2021.

[39] S. Sharma, S. Gupta, D. Gupta, S. Juneja, P. Gupta, G. Dhiman, and S. Kautish, "Deep learning model for the automatic classification of white blood cells," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–13, Jan. 2022.

[40] C. Chola, A. Y. Muaad, M. B. Bin Heyat, J. V. B. Benifa, W. R. Naji, K. Hemachandran, N. F. Mahmoud, N. A. Samee, M. A. Al-Antari, Y. M. Kadah, and T.-S. Kim, "BCNet: A deep learning computer-aided diagnosis framework for human peripheral blood cell identification," *Diagnostics*, vol. 12, no. 11, p. 2815, Nov. 2022.

[41] Z. Jafari and E. Karami, "Breast cancer detection in mammography images: A CNN-based approach with feature selection," *Information*, vol. 14, no. 7, p. 410, Jul. 2023.

[42] T.-M. Harry Hsu, W.-H. Weng, W. Boag, M. McDermott, and P. Szolovits, "Unsupervised multimodal representation learning across medical images and reports," 2018, *arXiv:1811.08615*.

[43] G. Chauhan, R. Liao, W. Wells, J. Andreas, X. Wang, S. Berkowitz, S. Horng, P. Szolovits, and P. Golland, "Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent (MICCAI)*, 2020, pp. 529–539.

[44] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3922–3931.

[45] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 590–597.

[46] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019, *arXiv:1901.07042*.

[47] Y. Li, H. Wang, and Y. Luo, "A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 1999–2004.

[48] Z. Zhang, P. Chen, X. Shi, and L. Yang, "Text-guided neural network training for image recognition in natural scenes and medicine," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1733–1745, May 2021.

[49] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "SLIP: Self-supervision meets language-image pre-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 529–544.

[50] M. V. Conde and K. Turgutlu, "CLIP-Art: Contrastive pre-training for fine-grained art classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3951–3955.

[51] M. Barraco, M. Cornia, S. Cascianelli, L. Baraldi, and R. Cucchiara, "The unreasonable effectiveness of CLIP features for image captioning: An experimental analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4661–4669.

[52] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology objects in context (ROCO): A multimodal image dataset," in *Proc. 7th Joint Int. Workshop*, Sep. 2018, pp. 180–189.

[53] H. N. Dao and I. Paik, "Patient similarity using electronic health records and self-supervised learning," in *Proc. IEEE 16th Int. Symp.*, Dec. 2023, pp. 1–15.

[54] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1839–1848.

[55] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020.

[56] D. Sharma, S. Purushotham, and C. K. Reddy, "MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain," *Sci. Rep.*, vol. 11, no. 1, p. 19826, Oct. 2021.

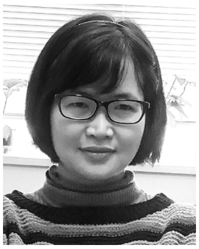
[57] J. N. Kather, N. Halama, and A. Marx, "100,000 histological images of human colorectal cancer and healthy tissue (v0.1)", Zenodo, 2018, doi: 10.5281/zenodo.1214456.

[58] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, Feb. 2020, Art. no. 104863.

[59] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[60] A. Fukui, D. Huk Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," 2016, *arXiv:1606.01847*.

[61] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2631–2639.



HONG N. DAO received the M.B.A. degree from Chungnam National University, South Korea. She is currently pursuing the Ph.D. degree with The University of Aizu. Her research interests include data analytics for economics and big data of smart city.

TUYEN NGUYEN received the bachelor's degree in computer science from The University of Aizu, in 2022. He is currently pursuing the Ph.D. degree with the University of Technology Sydney, Australia. His research interests include image processing, computer vision, and quantum computing.

CHERUBIN MUGISHA (Member, IEEE) received the bachelor's degree in computer science from Université Lumiére de Bujumbura, Burundi, in 2013, and the M.Sc. degree in computer science from The University of Aizu, Japan, in 2020, where he is currently pursuing the Ph.D. degree. His research interests include algorithms for multimodal machine learning methods integrating NLP, structured and unstructured data, and its application to medical data.

INCHEON PAIK (Senior Member, IEEE) received the M.E. and Ph.D. degrees in electronic engineering from Korea University in 1987 and 1992, respectively. He is currently a Professor with the University of Aizu, Japan. His research interests include deep learning applications, ethical LLMs, machine learning, big data science, and semantic web services. He is a member of the IEICE, IEIE, and IPSJ.