IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 December 2023, accepted 18 December 2023, date of publication 19 December 2023,
date of current version 28 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3345225

RESEARCH ARTICLE

DeepMetaForge: A Deep Vision-Transformer Metadata-Fusion Network for Automatic Skin Lesion Classification

SIRAWICH VACHMANUS[1], THANAPON NORASET[1], WARITSARA PIYANONPONG[2], TEERAPONG RATTANANUKROM[2], AND SUPPAWONG TUAROB[1], (Member, IEEE)
[1]Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom 73170, Thailand
[2]Division of Dermatology, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok 10400, Thailand

Corresponding author: Suppawong Tuarob (suppawong.tua@mahidol.edu)

This work was supported by the Specific League Funds from Mahidol University.

ABSTRACT Skin cancer is a dangerous form of cancer that develops slowly in skin cells. Delays in diagnosing and treating these malignant skin conditions may have serious repercussions. Likewise, early skin cancer detection has been shown to improve treatment outcomes. This paper proposes DeepMetaForge, a deep-learning framework for skin cancer detection from metadata-accompanied images. The proposed framework utilizes BEiT, a vision transformer pre-trained as a masked image modeling task, as the image-encoding backbone. We further propose merging the encoded metadata with the derived visual characteristics while compressing the aggregate information simultaneously, simulating how photos with metadata are interpreted. The experiment results on four public datasets of dermoscopic and smartphone skin lesion images reveal that the best configuration of our proposed framework yields 87.1% macro-average F1 on average. The empirical scalability analysis further shows that the proposed framework can be implemented in a variety of machine-learning paradigms, including applications on low-resource devices and as services. The findings shed light on not only the possibility of implementing telemedicine solutions for skin cancer on a nationwide scale that could benefit those in need of quality healthcare but also open doors to many intelligent applications in medicine where images and metadata are collected together, such as disease detection from CT-scan images and patients' expression recognition from facial images.

INDEX TERMS Image-metadata fusion, deep learning, skin lesion classification.

I. INTRODUCTION

Skin cancer is among the most common cancerous diseases in many countries [1]. Early identification of skin cancer, while still not prevalent, is critical for improving treatment outcomes and may lead to lower mortality rates [2]. In medical practice, some benign and malignant conditions are difficult to distinguish from each other, as their dermatologic manifestations are very resembling and can be wrongly diagnosed as one another. For instance, skin-colored and pearly-rolled edge nodules on facial skin can be diagnosed clinically with cutaneous basal cell carcinoma (BCC) [3], but the other skin neoplasm can be a mimicker, such as SCC, amelanotic melanoma, and trichoepithelioma. Another example includes squamous cell carcinoma (SCC) [3], which is clinically similar to actinic keratosis (AK), amelanotic melanoma, BCC, warts, spindle cell tumor, traumatic wounds, or other benign tumors [4], while melanoma, the deadliest form of skin cancer [5], might be challenging to differentiate from seborrheic keratosis, melanocytic nevus, pigmented BCC, pigmented AK, lentigo, angiokeratoma, dermatofibroma, or some vascular abnormalities [6]. Therefore, due to the variety of treatments and prognoses, these dermatological disorders need actual diagnoses. Dermoscopy is a noninvasive, in vivo diagnostic procedure used largely to examine cutaneous lesions. Traditionally, such a procedure

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

has been utilized as a guide to distinguish between benign and malignant skin lesions before performing a skin biopsy to provide a definitive diagnosis.

The initial presentation of these skin conditions can cause anxiety in different aspects and encourage patients with suspected skin conditions to visit hospitals for professional diagnoses. These circumstances, causing the sheer influx of concerned patients, have posed bottleneck problems in many hospitals with insufficient dermatologists, causing hospital congestion that may lead to a decline in the quality of medical care [7]. Furthermore, studies have shown that patients who live far from major hospitals, particularly in rural regions or with poor incomes, identified long waits and long travel distances as barriers to receiving proper dermatologic care [8]. As a result, people in rural areas are more likely to disregard their health concerns until it is too late, owing to the expenses of communication and hospitalization, as well as their hectic routines [9], [10]. These issues behoove telemedicine solutions or other detection technologies that can facilitate diagnoses in patients with suspected skin disorders without requiring them to visit major hospitals. Such systems could potentially be used by non-dermatologists, especially primary care providers who work in rural areas, with guidance from artificial intelligence and remote professional dermatologists, allowing concerning patients in rural areas to receive mainstream treatment in a timely manner. Developing a telemedicine system for skin cancer detection, however, requires intelligent components that not only automatically and reliably distinguish malignant from benign skin diseases but also demand reasonable processing resources.

Previous literature has utilized deep learning technologies to produce classification models for skin cancer detection from images [11]. Furthermore, recent discoveries have shown that patients' metadata could provide additional useful information when appropriately integrated into the classifiers' training processes. The majority of the previously proposed metadata-fusing methods either concatenate the encoded metadata to the image embedding [12] or use metadata to guide visual feature extraction [13]. Although these techniques may be logical from the standpoint of the model, they fail to replicate the way in which people understand the meaning of pictures in conjunction with metadata. This is particularly relevant in the context of diagnosing skin lesions, where dermatologists rely on a combination of visual and metadata cues altogether rather than sequentially. Nevertheless, most of these approaches were only validated on a single dataset, limiting their evidence of generalizability and robustness against varying image quality and metadata compositions from diverse data sources. Furthermore, the preponderance of past research on metadata-fusing methodologies for skin lesion classification has merely examined the efficacy of their proposed models without evaluating their scalability in the context of system implementation. The ultimate goal of our research is to establish a system that can be accessible by both patients and healthcare providers throughout the country, particularly those whose physical access to major hospitals is hindered. Therefore, the realization aspects of the proposed algorithms are equally important.

In this paper, we propose DeepMetaForge, a visual transformer-based deep-learning framework for skin lesion classification using both images and patient metadata. The proposed framework utilizes the BEiT [14] backbone, which uses the self-attention mechanism to pre-train the model as a masked image modeling (MIM) task, inspired by the masked language modeling (MLM) task that was found successful for pre-training transformer-based language models [15]. To our knowledge, we are the first to evaluate the BEiT backbone on the skin lesion classification problem. Furthermore, we propose the Deep Metadata Fusion Module (DMFM) that combines the visual features extracted from BEiT and metadata encoded by a convolutional neural network (CNN) while simultaneously compressing and decompressing the amalgamated information, similar to the process of forging metal decoration onto a steel plate. We hypothesize that fusing the metadata while compressing the visual features allows the metadata to *blend* in with the image information more effectively. This process intuitively aligns with how humans perceive semantics from metadata-accompanied images, where metadata is interpreted simultaneously while digesting the image content rather than sequentially. The experiment results on four public datasets demonstrate that our proposed DeepMetaForge framework exhibits superior performance compared to the best image classification backbone and the state-of-the-art metadata-fusing approach by a large margin. In addition, the scalability analysis found that the BEiT backbone utilized in the proposed framework is scalable and could potentially be implemented in telemedicine applications where the framework can be run on both low-resource devices and as API services.

In future directions, the proposed framework could be generalized to research problems outside of imaging dermatology in which data consists of pictures and associated metadata, such as artwork interpretation and document figure categorization. Furthermore, the next generation of the BEiT backbone (BEiT-v3) uses Multiway Transformers to perform masked ''language'' modeling on images, texts, and image-text pairs, allowing such a backbone to be used in a variety of vision-language downstream tasks, including visual question answering, visual reasoning, and image captioning [16]. Such technologies could give rise to many innovative, intelligent medical innovations, such as the ability to explain models' decisions with natural language responses and to retrieve hand-written medical notes with natural language queries.

In summary, this research's key contributions are as follows:

- We propose a novel framework for skin lesion classification, DeepMetaForge, that uses BEiT as the backbone image encoder as well as a novel Deep Metadata

Fusion Module that combines the visual features with encoded metadata while compressing the amalgamated information. Such a novel concept is inspired by an intuition that humans comprehend metadata and images simultaneously while distilling a conclusion rather than sequentially.
- We empirically evaluate our proposed framework on four public skin lesion image datasets and compare the classification performance with stand-alone backbones and state-of-the-art metadata-fusing methods for skin lesion classification [12], [17], [18].
- We empirically investigate the scalability of the proposed framework by analyzing the trade-off between the efficacy and efficiency of the models. The insights shed light on the implementation aspect when extending the proposed framework in a real-world telemedicine system.
- We present parameter sensitivity analyses that impact the performance of the proposed framework, such as the impacts of different metadata modes, compression ratios, module components, and backbone configurations.

The rest of this article is organized as follows. Section II discusses the background of skin lesion classification from images and relevant approaches to addressing such a task. Section III explains the proposed DeepMetaForge framework, including the network architecture, the Deep Metadata Fusion module, the datasets used for validation, and the evaluation protocol. Section IV reports highlighted experiment results as well as relevant discussions. Finally, Section V concludes the paper.

II. BACKGROUND AND RELATED WORK

Automatic skin lesion classification has been a long-standing research problem in computational sciences and medicine [19]. The ability to automatically recognize cancerous skin conditions from images could prove helpful in building computer-aided systems for dermatologists. Furthermore, telemedicine applications could adopt such technologies to enable medical doctors to early diagnose patients who do not have access to hospitals with guided information from artificial intelligence. Such early-stage detection of malignant skin diseases could vastly increase the chances of successful treatment [20]. Various computer vision and machine learning techniques have been proposed to address skin lesion classification problems. The early techniques focused on extracting discriminative characteristics from images, including segmentation, lesion border, color, and other texture-based features [21]. These extracted features can then be used to train traditional machine-learning models such as Support Vector Machine (SVM) [22], neural networks [23], and CART [24].

Deep learning technologies have evolved in recent years to ease end-to-end training of machine learning models while reducing the need for human expertise to engineer features [25]. The availability of public datasets has expedited the growth of emerging deep-learning techniques by facilitating standard validation benchmarks. In many applications, including skin lesion classification, studies have reported deep learning approaches to perform superior to the traditional machine learning models trained with engineered features [26], [27]. In addition, the use of advanced deep learning techniques has facilitated the analysis and acquisition of knowledge from various kinds of images, therefore effectively tackling complex issues, including the segmentation of retinal layers [28] and the identification of objects in infrared thermal images [29], [30] where immune-based intelligent techniques can be used for diagnosis [31] and feature extraction [32] tasks. This section reviews recent deep-learning techniques applied to skin lesion classification problems. The first subsection discusses the techniques that utilize only image information. Then, since a novelty of our proposed network architecture is the ability to forge metadata onto image embedding, we also discuss relevant studies that utilize patients' metadata to enhance skin lesion classification in the second subsection.

A. SKIN LESION CLASSIFICATION USING ONLY IMAGES

Malignant skin lesion recognition from images can be framed as an image classification problem where conventional off-the-shelf pre-trained image embedding models can be directly applied.Jiahao et al. [33] evaluated the applicability of VGG-16, ResNet-50, and EfficientNet-B5 on the ISIC 2020 dataset and found EfficientNet-B5 to yield the best AUC-ROC. Similarly, Zhang and Wang [34] found DenseNet-201 to perform the best when compared with VGG-16 and ResNet-50 on the ISIC 2020 dataset.

While pre-trained image models could be conveniently applied to the skin lesion datasets, the performance gaps still existed that called for the invention of more advanced network architectures. Zhang et al. [35] proposed to optimize the convolutional neural networks (CNN) with an improved version of the whale optimization algorithm [36] for skin cancer detection. Dermquest and DermIS were used as the benchmark datasets to compare their proposed method with an ordinary CNN, VGG-16, LIN, Inception-v3, and ResNet-50. Liu et al. [37] proposed incorporating doctors' perspectives when diagnosing skin cancer, including zooming, observing, and comparing into their proposed clinical-inspired network (CI-Net) using ResNet as the backbone. The evaluation was conducted on the ISIC 2016 - 2020 and PH2 datasets. Kaur et al. [38] developed LCNet, a novel end-to-end CNN-based framework that incorporates image resizing, oversampling, and augmentation specifically designed for melanoma skin lesion classification. Their method was tested on ISIC 2016, ISIC 2017, and ISIC 2020 datasets and was compared with conventional image classification backbones such as ResNet-18, Inception-v3, and AlexNet. Recently, Reis et al. [39] invented InSiNet, a deep convolutional approach for skin cancer detection and segmentation. They specifically raised the concern that typical

deep learning image encoding backbones can be extensive in size and consume high computational resources and proposed a new network with few parameters, resulting in a relatively lightweight model capable of cropping, segmenting lesions, and removing hair noises from the input images. Their method was evaluated on the ISIC 2018, 2019, and 2020 datasets, comparing against Inception-v3, DenseNet-201, ResNet-152v2, and EfficientNet-B0.

While deep learning techniques have been proposed to solve the skin lesion classification problem, a significant drawback of such approaches would be the lack of explainability when making predictions [40]. In the area of automatic skin cancer diagnosis, many attempts were made to explain the decisions from deep learning models. López-Labraca et al. [41] designed an interpretable system that allows CNN neurons to identify visual features and analyzes activation units to extract useful information for decision-making. Using their proposed system, they found that a small subset of channels was deemed relevant for a dermatologist compared to those of the baseline system. Recently, Rezk et al. [42] introduced an interpretable skin cancer diagnostic method that uses a skin lesion taxonomy to gradually acquire dermatological knowledge from a modified DNN architecture. These models were trained using clinical photographs that could be readily accessible to non-specialist healthcare professionals. The empirical investigations showed that the applied taxonomy is helpful for enhancing classification accuracy, comprehending the reasoning behind illness diagnosis, and identifying diagnostic mistakes. In addition, they used a sophisticated gradient-based class activation map approach that illustrates visual explanations when the model makes decisions.

The above-mentioned literature has investigated and proposed methods to improve skin lesion classification from images. Different studies focused on addressing challenges posed by the available skin lesion datasets, such as developing novel network architectures specific to the problem at hand, handling data imbalance, improving image quality, and enabling the models to provide valuable explanations when making decisions. However, all such methods only rely on skin lesion images as the sole information sources. Typically, during clinical dermoscopy, dermatologists also collect additional information about the patients, such as gender, age, and behavior (e.g., smoking, drinking, etc.), and about the diagnosed lesions, such as anatomical location, color, and bleeding. This research aims to investigate whether such metadata could help to improve the classification performance when incorporated into the network architecture while learning the image information. Therefore, the following subsection discusses relevant literature that attempts to use patients' metadata to improve skin lesion classification performance.

B. METHODS UTILIZING BOTH IMAGES AND METADATA

Oftentimes, images are accompanied by metadata for complemented information. Fusing different data sources to enhance model performance has long been investigated in computer vision [43], [44], [45], [46]. While traditional image classification models only require images as inputs, several studies have found that incorporating metadata during the training process could be helpful [47], [48], [49]. Integrating metadata into image classification could be as simple as concatenating metadata to the image features, training a dedicated learner with metadata and combining the probability outputs with those from image classifiers, or fusing metadata into the network architecture. Boutell and Luo [50] used a Bayesian network to incorporate camera metadata, such as brightness, flash, and subject distance, for scene image classification. Zhu et al. [51] extracted text lines from an image and combined them with the image content. Experimenting with an SVM classifier yielded an improvement from 81.3% to 90.1% in terms of accuracy. Yang et al. [47] enhanced theme classification using images from maps and their metadata such as name, title, keywords, and abstract. Langenberg et al. [48] used traffic lights' contextual metadata to assign each traffic light to its appropriate lane. Ellen et al. [49] employed geometric, geo-temporal, and hydrographic context metadata to improve plankton image classification. Lee et al. [52] proposed to combine deep learning models and traditional hand-crafted visual metadata features for biomedical image modality classification. Jony et al. [53], [54] fused image features and metadata to detect flooding in Flickr images.

In classifying skin lesions, as many public datasets contain patient information with lesion photos, the literature has developed approaches for exploiting such extra metadata to enhance classification performance. While a study reported that integrating patients' metadata, such as age, gender, and anatomical site, using a weighted average, concatenation-based, and squeeze-and-excitation (SE) approaches did not overall improve the skin lesion classification performance in a case study of 431 patients [55], some reported otherwise, especially those experimenting their proposed methods on larger datasets. Nunnari et al. [56] proposed concatenating probability from the image classifier with metadata when training traditional machine learning models. However, if neural networks are used as the primary classifier, then they proposed to concatenate one-hot encoded metadata with the image embedding layer. Their method yielded a 19.1% accuracy improvement on the ISIC 2019 dataset compared to classifiers trained only with image information. They also found that among the metadata attributes, age provided the most discriminative information, followed by body location and gender. Gessert et al. [12] proposed to encode metadata with two dense layers with dropout options and ReLU activation functions. The encoded metadata is then concatenated with the image embedding from the EfficientNet backbone. The choice of EfficientNet was particularly explored due to its ability to scale the model's width and depth according to the associated input size, leading to better classification results while using fewer parameters compared to other traditional image encoding

backbones. Their approach was evaluated on the ISIC 2019 dataset, yielding better performance than SENet-154, ResNext WSL, and EfficientNet classifiers trained only with image information. Ningrum et al. [17] focused on developing algorithms for melanoma detection from dermoscopic images using low-resource devices. Similar to Gessert et al. [12]'s work, they proposed to encode patients' metadata with a layer of the artificial neural network before concatenating it to the image embedding from a CNN model. Their method was evaluated on the ISIC 2019 dataset and was shown to be superior to using the CNN model alone. In addition to simply concatenating one-hot encoded metadata to image embedding, Li et al. [57] proposed to embed metadata with two neural networks using ReLU and Sigmoid as activation functions, then fused the embedded metadata with the image embedding using the multiplication function. While their method was shown to be superior to the concatenation-based approach, the improvement was marginal on the ISIC 2018 dataset. Furthermore, their concatenation baseline simply used the one-hot encoded metadata as-is without encoding it first, as done in their proposed multiplication-based method.

Another scheme of fusing metadata into visual information is using metadata to guide the image feature extraction without changing the dimension of visual feature maps, similar to the attention mechanism [58]. Pacheco and Krohling [13] proposed MetaBlock that uses an attention mechanism to supervise visual feature extraction. The block first encodes the patients' metadata and multiplies it with the image embedding where the hyperbolic tangent gate is used. These intermediate parameters are concatenated with another set of encoded metadata features, where the sigmoid gate is used to produce the final supervised visual features. Their method was evaluated on the ISIC 2019 and PAD-UFES-20 datasets. Similarly, Pundhir et al. [59] proposed a multiplication-based scheme to combine patients' metadata with encoded lesion images. However, their proposed Hyperpametric Meta Fusion Block multiples image embedding with encoded metadata using Leaky ReLU as the activation function then concatenates the intermediate parameters with another encoded metadata, after which the Swish [60] activation is used to produced supervised features. Their method was evaluated on the PAD-UFES-20 dataset, varying backbones such as MobileNet-v2, VGGNet-13, ResNet-50, EfficientNet-B4, and DenseNet-121. Cai et al. [61] proposed SLE (Soft Label Encoder), where metadata attributes are encoded as soft labels rather than binary values as done by the one-hot encoder. The soft labels are then used in a mutual attention decoder to produce the final embedding for classification with a fully connected layer. Their proposed method was evaluated on the ISIC 2018 dataset. Recently, Tang et al. [62] presented FusionM4Net, a multi-stage multi-modal learning algorithm for multi-label skin disease classification, which consists of two stages: first, learning feature information from clinical and dermoscopy pictures, and second, combining patient metadata and decision information from the two-modality images. Experiments conducted on the Seven-Point Checklist (SPC) dataset demonstrated that FusionM4Net was more accurate than any other existing state-of-the-art approach. A weakness of their approach is its dependence on two image sources (i.e., dermoscopic and clinical pictures), which may not be available in other datasets or practical scenarios.

While many studies have proposed various ways to merge patients' metadata into skin lesion classifiers' learning mechanisms, including concatenation-based, multiplication-based, and attention-based methods, the above-mentioned metadata-fusion approaches still fall short of analyses in several practical aspects. First, most of the proposed algorithms only validated their methods on one dataset, while generalizability must be evidenced by controlled experiments on multiple data sources with diverse characteristics. Second, it is our overarching intention to develop a telemedicine system accessible to healthcare practitioners and patients, especially those in suburban areas in developing countries where physical access to modern medical equipment is limited and experienced dermatologists are scarce. In doing so, the scalability aspect of the classification models must be explored to determine the appropriate deployment options to implement. In this paper, we propose a deep vision-transformer metadata-fusing framework for skin lesion classification. The novelty of our framework lies in two folds. First, we adopted BEiT (Bidirectional Encoder representation from Image Transformers), which uses a masked image modeling task to pre-train vision transformers [14]. Second, we propose a novel Deep Metadata Fusion Module (DMFM) to merge encoded metadata onto the image embedding during the compression, mimicking the process of ''forging'' decoration onto a metal piece, hence the name DeepMetaForge. Our proposed method is generalizable as evidenced by the experiments on four publicly available skin lesion datasets, comparing with both stand-alone image encoding backbones and the state-of-the-art metadata-fusing methods proposed by Gessert et al. [12] and Ningrum et al. [17]. Furthermore, we analyze the scalability of our proposed framework by studying the trade-off between classification performance and computational resource consumption.

III. METHODOLOGY

The novelty of our proposed DeepMetaForge framework is the use of BEiT backbone and the metadata forging mechanism. This section first discusses the proposed network architecture in detail. Then, the rest of this section walks through the experiment setup, including datasets, computational environments, and evaluation protocol.

A. NETWORK ARCHITECTURE

This section provides an overview of our network architecture, depicted in Fig. 1. Our network features a multi-modal base that can merge feature maps from various inputs. The

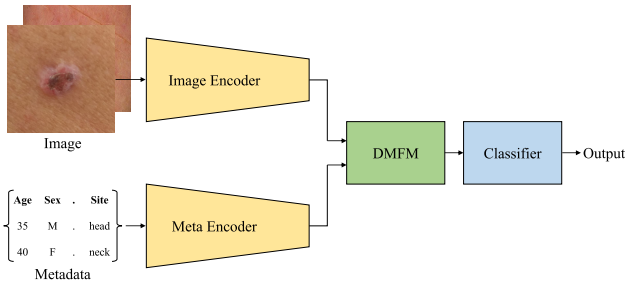

FIGURE 1. Overall DeepMetaForge network architecture.

network structure comprises three main parts: the image encoder, the meta encoder, and the combining module. The image encoder utilizes the state-of-the-art vision transformer backbone, BEiT [14], designed to extract high-level visual features from images. On the other hand, the metadata encoder employs a convolutional network structure to extract the feature map from the input data. To combine the feature maps from the image encoder and metadata encoder, we use a merging module called Deep Metadata Fusion Module (DMFM). This module effectively combines two feature maps to produce a more comprehensive and informative representation of the input. Overall, our network architecture is designed to take advantage of the strengths of both the image and metadata while effectively fusing them using a novel deep feature-level merging module.

1) IMAGE ENCODER

Our proposed architecture uses the Bidirectional Encoder representation from Image Transformers (BEiT) [14] as the image encoder. It has been shown to outperform existing supervised pre-training models in masked image modeling tasks. The BEiT image encoder, illustrated in Fig. 2, tokenizes the original image into small visual tokens and randomly masks some image patches. These masked patches are then fed into the backbone transformer. In summary, the BEiT image encoder is a highly effective tool for pre-training vision transformers and an excellent choice for our network architecture.

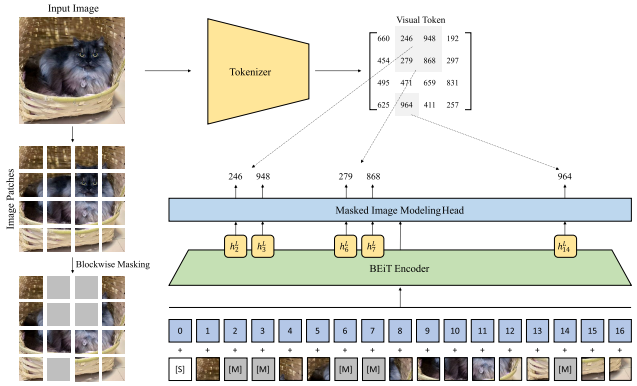

FIGURE 2. Illustration of the BEiT network structure, adapted from [14].

2) METADATA ENCODER

The metadata encoder in our architecture is based on a convolutional neural network (CNN), consisting of two sets of convolutional layers, each including a one-dimensional convolutional filter, a normalizer, and an activation function. The computation of the metadata encoder is described by (1), where ϕ represents the one-dimensional convolutional filter, ξ represents the one-dimensional batch normalizer, and ρ represents the Rectified Linear Unit (ReLU) activation function.

$$y = \rho[\xi\ (\phi\ (x))] \tag{1}$$

Let $y \in \mathbb{R}$ and $x \in \mathbb{R}$ denote the feature map from a set of convolutional layers and the input feature map. The metadata encoder is one of the critical components of our architecture, as it extracts the feature map from the input data. By utilizing a CNN structure, we can effectively extract relevant features from the input, which can be integrated with the feature map generated by the image encoder to create a comprehensive representation of the input data. The first CNN converts the input metadata into a feature map with 256 channels. The second converts the output of the first set into the same shape as the output of the image encoder. This allows us to merge the feature map generated by the meta encoder with the feature map generated by the image encoder using our proposed DMFM, resulting in a more comprehensive and informative representation of the input data.

3) DEEP METADATA FUSION MODULE (DMFM)

The Deep Metadata Fusion Module (DMFM) is a merging component in our architecture that fuses the feature maps generated by the image encoder and metadata encoder. This module is adapted from the Fused Module, introduced by Vachmanus et al. [63], to handle multi-modal visual information. However, the Fused module was designed for two-dimensional feature maps, while DMFM merges one-dimensional feature maps by removing the convolution with rate branch (cbr). Equation (2) defines the calculation of DMFM. Let $\hat{x} \in \mathbb{R}^c$ represent the concatenation result of the feature maps from the image encoder and metadata encoder, and $\hat{z} \in \mathbb{R}^{2c}$ denote the output feature map from DMFM, the asterisk symbol $(*)$ represents the concatenation operation, and z is calculated according to Equation (3).

$$\hat{z} = z * \hat{x} \tag{2}$$

$$z = \mu_2(\mu_1\ (\hat{x})) \tag{3}$$

The function $\mu(x)$ is described by Equation (4), where φ is the dropout operation with a 0.4 probability.

$$\mu(x) = \varphi[\rho(\phi\ (x))] \tag{4}$$

The result of μ_1 is a feature map with $\frac{\gamma}{c}$ channels, where γ denotes the compression ratio. The function μ_2 converts the feature map back to the input shape of the module. As a result, the feature map shape is roughly four times larger than the input shape. In our architecture, we use a compression ratio

of 8. The DMFM plays an essential role in our architecture, allowing us to merge the feature maps generated by the image encoder and meta encoder into a more comprehensive and informative representation of the input data.

The DMFM's operation is designed to imitate how a skin lesion image with metadata is comprehended. Specifically, one would understand the metadata and the associated skin lesion presented in the image altogether concomitantly while deducing a conclusion about whether it is a cancerous lesion. Our concept contradicts the popular concatenation-based method [12], where the simple concatenation of visual and metadata features are fed to the fully connected layer for classification, hence forcing the classifier to learn the two information pieces sequentially rather than simultaneously.

4) CLASSIFICATION LAYER

The classification layer of the network is tasked with reducing the channel of the fused feature map from the DMFM to a binary classification based on whether the skin lesion is benign or malignant. This layer comprises a fully connected layer, a one-dimensional batch normalizer, and a Rectified Linear Unit (ReLU) activation function. As previously described, the output of this layer is the ultimate output of the whole network structure and is vital for making an accurate prediction since it converts the DMFM-generated feature map into a binary classification. By applying a fully connected layer, the feature map's dimension is efficiently decreased to provide a reliable prediction of the skin condition.

The primary objective of the proposed network is to enable the detection of malignant skin lesions by binary classification rather than focusing on multiclass classification for more detailed identification of specific forms of skin cancer. This design choice is based on the following justifications. First, our primary objective is to use the model inside a telemedicine framework, with the primary purpose of screening individuals exhibiting symptoms of skin cancer to determine the need for further diagnosis by professional dermatologists. Hence, precise identification of malignant skin lesions is crucial. Framing the problem as a multiclass classification task, while providing more details about the predicted skin conditions, could not only be unnecessary for the objective but also result in poorer classification accuracy [64]. Furthermore, the primary contribution of this study is in the novel network architecture, which necessitates empirical validation for its suitability in accommodating diverse datasets with various types of skin conditions. Hence, to provide an equitable evaluation of the network's efficacy across various datasets, it is essential that the prediction task remains constant, hence enabling a comparative examination of the influence of input data on the accuracy of the model. Finally, framing the problem as a benign-vs-malignant classification task allows a fair comparison with several reputable methods addressing the skin lesion classification [12], [17], [18] that also addressed the problem using binary classification methods. Nevertheless, extending the proposed network for multiclass classification is trivial and can be achieved by modifying the output layers as needed.

TABLE 1. Statistics of the datasets used in this research.

Dataset	Camera	# Metadata Numeric Attributes	# Metadata Categorical Attributes	# Benign Samples	# Malignant Samples
PH2	Dermoscopy	0	13	203	76
SKINL2	Dermoscopy	1	3	160	40
PAD-UFES-20	Smartphone	3	17	405	1,089
ISIC 2020	Dermoscopy	1	2	31,954	575

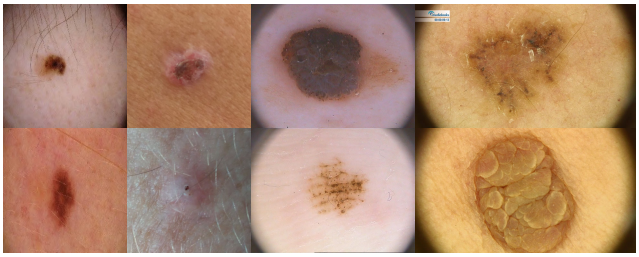

FIGURE 3. Example skin lesion images from each dataset. By column: ISIC 2020, PAD-UFES-20, PH2, and SKINL2. By row: Malignant and Benign.

B. DATASETS

We evaluate our proposed network architecture on four publicly available skin lesion datasets: ISIC 2020 [65], PAD-UFES-20 [66], SKINL2 [67], and PH2 [68]. The images in the PAD-UFES-20 dataset were taken with smartphones, while the others were with dermoscopy cameras. All of these datasets include both images and metadata, which provide complementary information for skin lesion classification.

The dataset contains a wide range of metadata, each tailored to specific categories. For example, the PH2 dataset contains the Histological Diagnosis, Asymmetry, Pigment Network, Dots/Globules, Streaks, Regression Areas, Blue-Whitish Veil, as well as color characteristics such as White, Red, Light-Brown, Dark-Brown, Blue-Gray, and Black. The SKINL2 dataset provides information on Gender, Age, Fototype, and Melanocytic attributes. The PAD-UFES-20 dataset includes data on smoking and drinking habits, age, pesticide exposure, gender, skin cancer history, cancer history, access to piped water and sewage systems, Fitzpatrick score, region, lesion diameter measurements, as well as factors like itchiness, growth, pain, changes, bleeding, elevation, and biopsy status. Lastly, the ISIC 2020 dataset encompasses information regarding gender, age approximation, anatomical site of the lesion, and diagnosis. These comprehensive datasets offer valuable insights for a wide range of dermatological and epidemiological research. A detailed description of each dataset is provided in Table 1. Example images from each dataset are depicted in Fig. 3.

Each dataset is divided into a training set, which comprises 70% of the data, a 10% validation set, and a testing set, which comprises the remaining 20%. To further enhance the reliability of our results, we employ 5-fold cross-validation

for all experiments conducted in this paper. During the training phase, we randomly split the training set into eight parts, with one part designated as a validation set for hyperparameter tuning. To improve the model's performance, we apply pre-processing techniques such as resizing all images to 224 × 224, with some tests conducted using 384, depending on the pre-trained models' architectures. Data augmentation techniques such as normalization, random flipping, and random rotation are performed to allow the model to handle a range of diverse inputs.

By validating our proposed architecture on these various datasets, we can show its efficacy in accurately classifying benign and malignant skin lesions in a wide range of scenarios. The use of these publicly available datasets is essential for demonstrating the practical generalizability of our proposed architecture to diverse sources of skin lesion images (i.e., smartphones and dermoscopy cameras) and different compositions of metadata details.

C. COMPUTATIONAL ENVIRONMENTS

During the training process, RGB images are resized to a height and width of 224 to meet the requirements of the backbone. The experiment is conducted on a single Linux machine with NVIDIA Geforce RTX 3090 GPU and Intel i7-12700K CPU with 32 GB of memory. The deep learning framework Pytorch is utilized to construct and train the network model. The network encoder is pre-trained on each original publication, while the other layers' parameters are randomly initialized. To re-adjust the weights of the training loss for each class, a cross-entropy function is used as Equation (5), where m represents the number of classes, p the predicted probability observation, and q the binary indicator (0 or 1).

$$Loss = -\frac{1}{m}\sum_{m} q\log(p) + (1-q)\log(1-p) \tag{5}$$

Momentum SGD is used as the optimization algorithm, whose parameter is set to 0.9. The initial learning rate ($rate_{ini}$) is set to 0.001 with a decay of 0.90 at every five epochs, and the learning rate (lr) is calculated by Equation (6), where $epochs$ denotes the global step epochs of training and dpe is the decay per epoch.

$$lr = rate_{ini} \times decay^{\frac{epochs}{dpe}} \tag{6}$$

D. EVALUATION METHODS

The network models are evaluated using a variety of metrics, including precision, recall, F1 score, accuracy, MCC, sensitivity, specificity, and NPV, to provide a comprehensive analysis of its effectiveness in skin lesion classification. Let the malignant class be referred to as the positive class and the benign class as the negative class. Precision measures the proportion of true positives among all predicted positives, while recall measures the proportion of true positives among all actual positives. The F1 score combines precision and recall to provide a balanced evaluation of the network's performance. In this research, we present precision, recall, and F1 of both the positive and negative classes. The macro-average F1 is simply the average of F1 scores from both classes, used to represent the overall classification efficacy. Accuracy measures the proportion of correct predictions among all predictions, while MCC, or Matthews correlation coefficient, considers true and false positives and negatives to evaluate the performance of the network. Additionally, sensitivity and specificity measure the proportion of true positive and true negative predictions, respectively, and NPV, or negative predictive value, measures the proportion of true negative predictions among all negative predictions.

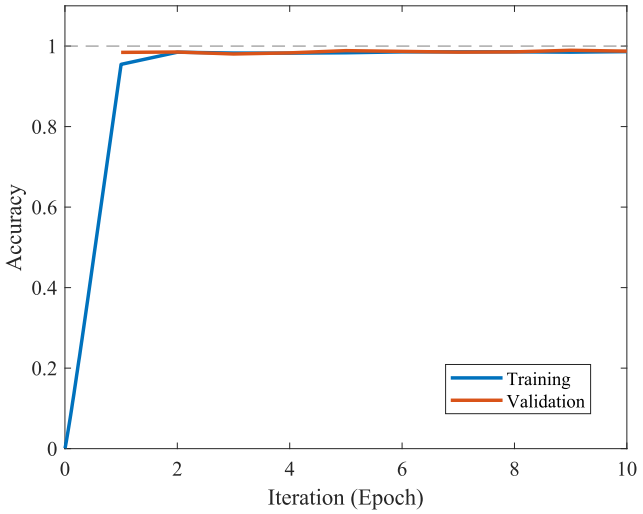

FIGURE 4. Comparison of the accuracy of training and validation across training iterations (epochs) using the BEiT-base-224 model on the ISIC 2020 dataset.

IV. EXPERIMENT, RESULTS, AND DISCUSSION

In this section, we evaluate the effectiveness of the proposed network structure in comparison to other existing approaches for skin lesion classification. Our evaluation is conducted on multiple publicly available datasets, as described in Section III-B. In addition to evaluating the overall performance of the proposed network, we also assess the construction and optimization of the DMFM, including the evaluation of individual components using ablation studies, as well as determining the optimal compression ratio for merging feature maps generated from different sources. By thoroughly evaluating the proposed approach, we can identify the strengths and weaknesses of the network and provide insights for future improvements in skin lesion classification.

During the training process, the primary goal is to identify the optimal training weights, which is achieved by closely monitoring the accuracy of the validation set. To address potential overfitting issues, a simultaneous check is kept on both the validation accuracy and the training accuracy. In Fig. 4, the relationship between accuracy and training iterations across epochs using the BEiT-base-224

model on the ISIC 2020 dataset is illustrated. Given the dataset's substantial size, it can be observed that the training accuracy stabilizes shortly after the first epoch. Both the training and validation accuracy values consistently remain at approximately 98-99%.

A. THE DMFM OPTIMIZATION

The DMFM is primarily determined by the compression ratio, denoted as γ, which controls the size of the feature map in the module. The feature map is compressed to a smaller shape during the merging process to facilitate the combination of the different feature maps. To determine the optimal compression ratio, we varied the value of γ according to powers of 2 to avoid any remainder during the separation process. Specifically, we tested compression ratios of 1, 2, 4, 8, 16, 32, and 64 in our experiments.

In this experiment, the performance of the DMFM was tested on the ISIC 2020 dataset, which is the largest dataset used in this research. The primary objective was to evaluate the DMFM's capability to merge feature maps generated from various sources, including both image and metadata, effectively. The evaluation metrics used were precision, recall, and F1 score for each class. By testing the DMFM on different datasets while varying the compression ratio, we could determine the optimal configuration for accurately classifying skin lesions.

TABLE 2. Comparison between different compression ratios on the ISIC 2020 dataset.

Comp. Ratio	Benign F1	Malignant F1	Macro Avg F1	Accuracy	MCC	Sensitivity	Specificity	NPV
1	0.997	0.799	0.898	0.993	0.800	0.770	0.997	0.996
2	0.995	0.629	0.812	0.991	0.636	0.623	0.998	0.993
4	0.996	0.666	0.831	0.992	0.667	0.671	0.998	0.994
8	***0.997***	***0.843***	***0.920***	***0.995***	***0.846***	***0.850***	0.997	***0.997***
16	0.996	0.735	0.866	0.992	0.748	0.694	0.997	0.995
32	0.995	0.559	0.777	0.990	0.602	0.510	***0.998***	0.991
64	0.996	0.710	0.853	0.992	0.726	0.638	0.998	0.994

The results of the experiment evaluating the performance of the DMFM on the ISIC2020 dataset are presented in Table 2. To ease analyses, the comparison of F1 scores is shown in Fig.5. Our primary objective was to determine the optimal compression ratio to effectively merge the feature maps generated from different sources, including both images and metadata, and accurately classify skin lesions as benign or malignant. The results show that the best compression ratio is 8, which produces the highest macro-average F1 score of 92.0%, roughly 18.4% higher than the lowest F1 score obtained in the experiment. In addition to the average F1 score, we also analyzed the precision and recall values for each class to evaluate the performance of the DMFM with varying compression ratios. The results show that the compression ratio of 8 generates the highest recall value for the malignant class, which is about 66.9% higher than the lowest recall value obtained in the experiment. Moreover, this compression ratio provides a well-balanced performance between precision and recall for both the benign and malignant classes, which is important in accurately classifying skin lesions. The results of this experiment demonstrate that the DMFM with a compression ratio of 8 is the most efficient configuration to merge metadata features into image data features for accurate skin lesion classification. By varying the compression ratio and evaluating the performance on different datasets, we can determine the optimal configuration for the DMFM to effectively forge feature maps generated from different sources. These findings contribute to the development of an effective and accurate system for skin lesion classification, which can aid in the early detection and treatment of skin cancer.

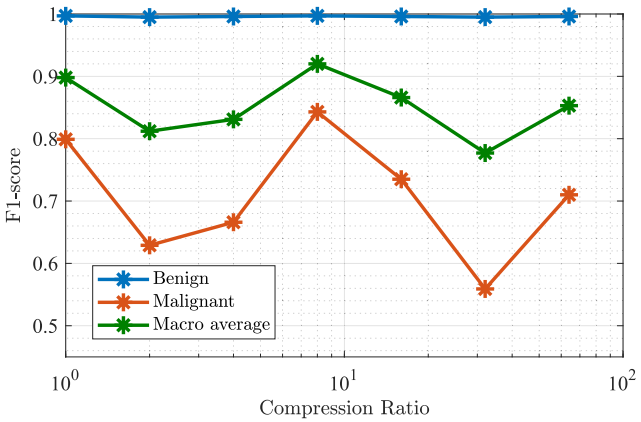

FIGURE 5. Comparison of F1 scores from different compression ratios. The blue, orange, and green lines represent the F1 scores of the benign class, malignant class, and macro average between the two classes, respectively.

B. ABLATION STUDIES ON DMFM

Ablation studies analyze the impact of individual components on a model's performance. In DMFM, they can provide insights into the compression branch and skip feature map. By removing one component at a time and analyzing performance, we can understand their importance and how they work together. This can guide the design of more effective models for skin lesion classification.

Table 3 displays the results of our experiment assessing the performance of the DMFM on the ISIC2020 dataset. To further investigate the impact of individual components on the overall performance of the model, we conducted an ablation study by removing the compression branch and the skip feature map, one component at a time. We then evaluated the performance on the same dataset to determine the importance of each component with respect to skin lesion classification performance. The results indicate that the combination of the compression branch and skip feature map yields the highest F1 score among all combinations. Notably, the proposed compression branch with the compression ratio (γ) can improve the efficiency of concatenation by 31.2%. Additionally, the precision of the malignant class was improved by approximately 1.8 times. These findings demonstrate the impact of both components on achieving accurate

TABLE 3. Comparison of the classification performance when using DMFM with different module components on the ISIC 2020 dataset. Δ F1 denotes the relative performance difference (improvement) of the best-performing method compared to each other approach.

Module Components		Benign	Malignant	Macro	Δ F1	Accuracy	MCC	Sensitivity	Specificity	NPV
Comp. Branch	Skip	F1	F1	Avg F1						
w/o	w	0.979	0.423	0.701	31.3%	0.960	0.456	0.751	0.964	0.995
w	w/o	0.991	0.341	0.666	38.1%	0.983	0.341	0.333	***0.999***	0.984
w	w	***0.997***	***0.843***	***0.920***	-	***0.995***	***0.846***	***0.850***	0.997	***0.997***

TABLE 4. Comparison of the classification performance when using DMFM with different feature map merging methods on the ISIC 2020 dataset. Δ F1 denotes the relative performance difference (improvement) of the best-performing method compared to each other approach.

Merging Components	Benign	Malignant	Macro	Δ F1	Accuracy	MCC	Sensitivity	Specificity	NPV
	F1	F1	Avg F1						
Summation	0.994	0.471	0.732	25.6%	0.988	0.471	0.439	***0.998***	0.989
Multiplication	0.994	0.608	0.801	14.9%	0.989	0.610	0.605	***0.998***	0.991
Concatenation	***0.997***	***0.843***	***0.920***	-	***0.995***	***0.846***	***0.850***	0.997	***0.997***

skin lesion classification and highlight the effectiveness of the proposed compression branch in improving the DMFM's performance.

Another crucial aspect in developing the DMFM architecture involves exploring various methods of merging feature maps. To achieve optimal skin lesion classification, another experiment was conducted with different merging operations for $\hat{x}$ in Equation (2), including summation, multiplication, and concatenation. These operations play a pivotal role in fusing the information from the compression branch and the skip feature map. Through rigorous testing and analysis, the aim is to determine which merging technique yields the best results, ultimately refining the DMFM architecture and advancing the field of skin lesion classification.

Table 4 presents the experimental results for various merging operations of DMFM on the ISIC2020 dataset. The findings indicate that employing the concatenation operation yields the highest F1 score compared to the other operations. This results in approximately a 25.6% improvement over the summation operation and a 14.9% increase over the multiplication operation. Consequently, the concatenation operation was the optimal choice for merging with DMFM.

C. IMAGE ENCODER BACKBONE COMPARISON

The proposed network architecture comprises two encoders that extract different feature maps: the metadata encoder and the image encoder. The metadata encoder uses a basic CNN without pre-trained weights, while the image encoder employs a large network structure that can extract or scope into multiple values in the image. To optimize training time and computational resources, the image encoder is initialized with pre-trained weights specific to our task. In this section, we conducted experiments to determine the optimal network backbone for the image encoder by testing three well-known network structures: the ResNext [69], EfficientNet [70], and BEiT [14] models. Our primary objective was to identify the most suitable model for accurately classifying skin lesions.

The results of our experiments, presented in Table 5, demonstrate that the ResNext model achieved an F1 score of approximately 85-86%, while EfficientNet ranges from 50-87% (depending on the model size), and BEiT achieves an F1 score of 90-92% on the ISIC 2020 dataset. The BEiT model outperforms the other models on the malignant class, resulting in the highest F1 score. Therefore, we conclude that the BEiT model is the most suitable image-encoding backbone for our proposed architecture. The high F1 score obtained using the BEiT model demonstrates its effectiveness in extracting image features and combining them with metadata to accurately classify skin lesions.

D. IMPACT OF METADATA INTEGRATION

Metadata, which includes patient information such as age, gender, anatomical location of the lesion, and histopathological information, is an essential factor that can provide additional information about a skin lesion that is not captured by image data alone. Incorporating metadata into skin lesion classification can be critical in accurately classifying skin lesions. For example, age can be a significant factor in identifying malignant melanoma, while the location of the lesion can provide insight into its potential malignancy. By combining metadata with image data, we can obtain a more comprehensive understanding of the lesion, leading to improved accuracy in classification. In our study, we aimed to investigate the impact of metadata on skin lesion classification by comparing the performance of our proposed model with and without metadata. We also tested the model using dummy metadata, where all data slots were set to zero, to evaluate the effect of metadata on model efficiency. By analyzing the results across different datasets, we aimed to gain insights into the importance of metadata in accurately classifying skin lesions and identify the optimal approach for incorporating metadata into the model.

Our analysis revealed that incorporating metadata improved classification performance on most datasets for the average F1 score, as presented in Table 6. However,

TABLE 5. Classification performance of the DMFM with different image-encoding backbones on the ISIC 2020 dataset. Δ F1 denotes the relative performance difference (improvement) of the best-performing method compared to each other approach.

Backbone	Benign F1	Malignant F1	Macro Avg F1	Δ F1	Accuracy	MCC	Sensitivity	Specificity	NPV
ResNext 50	0.996	0.715	0.855	8.66%	0.992	0.731	0.685	0.9972	0.994
ResNext 101	0.995	0.728	0.861	7.89%	0.991	0.735	0.732	0.9955	0.995
EfficientNet b0	0.994	0.343	0.669	39.02%	0.987	0.343	0.355	0.9987	0.989
EfficientNet b1	0.991	0.084	0.537	72.92%	0.983	0.117	0.050	0.9995	0.983
EfficientNet b2	0.993	0.355	0.674	37.91%	0.986	0.397	0.270	0.9991	0.987
EfficientNet b3	0.991	0.026	0.509	82.74%	0.982	0.052	0.014	***0.9999***	0.983
EfficientNet b4	0.996	0.747	0.872	6.62%	0.992	0.751	0.706	0.9972	0.995
EfficientNet b5	0.995	0.575	0.785	18.41%	0.990	0.585	0.520	0.9982	0.991
EfficientNet b6	0.996	0.658	0.827	12.40%	0.992	0.660	0.656	0.9977	0.994
EfficientNet b7	0.995	0.561	0.778	19.44%	0.990	0.574	0.496	0.9986	0.991
BEiT-base-224	***0.997***	0.843	0.920	0.99%	***0.995***	0.846	0.850	0.9973	0.997
BEiT-base-384	0.997	***0.861***	***0.929***	-	0.995	***0.862***	***0.927***	0.9959	***0.999***
BEiT-large-224	0.997	0.833	0.915	1.57%	0.994	0.832	0.873	0.9961	0.998
BEiT-large-384	0.996	0.796	0.896	3.73%	0.993	0.794	0.816	0.9957	0.997

TABLE 6. Classification performance comparison between using only image information with the BEiT backbone (None), our proposed network with dummy metadata (Dummy), and our proposed network with actual metadata (Actual). Δ F1 denotes the relative performance difference (improvement) of the best-performing method compared to each other approach.

Dataset	Metadata	Benign F1	Malignant F1	Macro Avg F1	Δ F1	Accuracy	MCC	Sensitivity	Specificity	NPV
PH2	None	0.948	0.754	0.851	10.10%	0.915	0.728	0.700	0.969	0.930
	Dummy	0.558	0.343	0.451	107.94%	0.545	0.120	0.650	0.519	0.787
	Actual	***0.975***	***0.899***	***0.937***	-	***0.960***	***0.879***	***0.900***	***0.975***	***0.976***
SKINL2	None	0.915	0.749	0.832	4.72%	0.875	0.698	0.753	***0.920***	0.919
	Dummy	0.707	0.360	0.534	63.28%	0.630	0.121	0.382	0.723	0.738
	Actual	***0.926***	***0.817***	***0.871***	-	***0.896***	***0.755***	***0.830***	***0.920***	***0.938***
PAD-UFES-20	None	0.605	***0.880***	0.742	1.51%	***0.816***	0.502	***0.926***	0.521	***0.726***
	Dummy	0.338	0.445	0.392	92.31%	0.479	0.012	0.441	0.578	0.288
	Actual	***0.633***	0.874	***0.754***	-	0.813	***0.524***	0.889	***0.610***	0.695
ISIC 2020	None	0.864	0.118	0.491	87.32%	0.765	0.188	0.841	0.764	0.996
	Dummy	0.647	0.026	0.337	173.27%	0.534	-0.033	0.344	0.538	0.976
	Actual	***0.997***	***0.843***	***0.920***	-	***0.995***	***0.846***	***0.850***	***0.997***	***0.997***
Avgerage	None	0.833	0.625	0.729	19.39%	0.843	0.529	0.805	0.793	0.893
	Dummy	0.563	0.294	0.428	103.29%	0.547	0.055	0.454	0.589	0.697
	Actual	***0.883***	***0.858***	***0.871***	-	***0.916***	***0.751***	***0.867***	***0.876***	***0.901***

the degree of improvement varied depending on the dataset used. For instance, on the ISIC 2020 dataset, the use of metadata improved the performance by 87.3%, while on the PAD-UFES-20 dataset, the improvement was only 1.51%. Furthermore, we observed that the effect of metadata was more prominent in the efficient detection of malignant skin lesions. However, in the case of the PAD-UFES-20 dataset, the improvement from integrating metadata is only 1.51%, lower than those of other datasets. This may be due to the types of metadata that were not related to the diagnosis of the disease. In addition, PAD-UFES-20 is the only dataset whose images were taken by smartphones. Therefore, the reason for the small performance improvement could be due to the image characteristics themselves. Further investigation into the actual causes is needed to find a theoretical explanation for such a phenomenon.

In Fig. 6, we used GradCAM [71] to visualize the regions of interest (ROIs) of the network for image classification without metadata. The red areas in the bottom images represent the ROIs of the model. When comparing two samples, the network primarily focuses on the edges of the lesions. This visualization shows that using only image input cannot produce precise results, as the network's ROIs do not always correspond to the main information in the image. This can lead to misclassification of benign samples.

We also tested the model using dummy metadata, where all attributes' values are replaced with zeros, to evaluate the impact of metadata on model efficiency. Our results showed that the use of dummy metadata decreased the classification efficiency by about 41.2%, indicating that relevant and meaningful metadata is crucial for improving the performance of skin lesion classification. This also indicates the proposed network cannot tolerate missing metadata. This makes sense since the network was trained with actual metadata; therefore, the absence thereof during model evaluation could falsely guide the model's prediction.

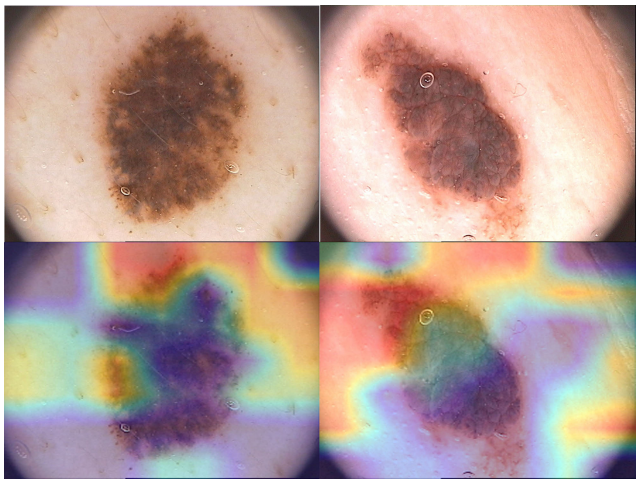

FIGURE 6. EfficientNet B7 ROI visualization of benign samples using the GradCAM technique: Top Row - Input Images, Bottom Row - ROI Visualization, Left Column - predicted as benign, Right Column - predicted as malignant.

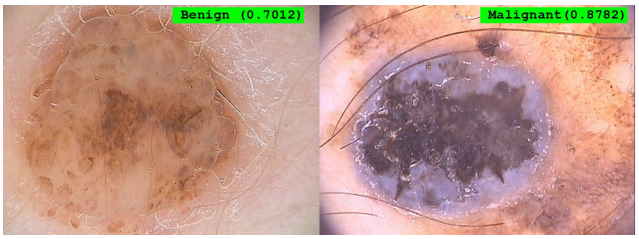

FIGURE 7. Prediction Outputs by BEiT with DMFM, with probability scores.

In conclusion, our study evaluated the impact of incorporating metadata in skin lesion classification using our proposed model. By comparing the model's performance with and without metadata across different datasets, we demonstrated that metadata could improve classification efficiency, especially for malignant detection. However, the effect of metadata on performance varies depending on the type of metadata and the dataset used. Therefore, it is crucial to carefully select and incorporate relevant metadata into the model to improve its efficiency in accurately classifying skin lesions. Further research is needed to determine the most appropriate types of metadata and the optimal approaches for incorporating them into skin lesion classification models.

E. NETWORKS COMPARISON

In the skin lesion classification domain, metadata has been shown to improve the overall classification performance in addition to using the image data alone. Various researchers have proposed new techniques for integrating metadata into their learning mechanisms and have continued to improve upon each other's work over time. In this study, we compared our proposed technique to three state-of-the-art methods, Jasil and Ulagamuthalvi [18] + DMF, Ningrum [17], and Gessert et al. [12], which also fuse metadata into their network architectures for skin lesion classification. The image encoding modules for these three baselines were reproduced as reported in their publications. Note that the skin lesion classification method proposed by Jasil and Ulagamuthalvi [18] does not inherently integrate image metadata into the network, making it difficult to compare with other metadata-fusing networks. Therefore, their proposed network was used as the image-encoder component in our proposed DMF network, hence referred to as *Jasil and Ulagamuthalvi + DMF*. By comparing the performance of these methods, we can gain insights into the strengths and weaknesses of each approach and identify areas for further improvement.

The results of our comparison study are presented in Table 7. By comparing the performance of the proposed DeepMetaForge (BEiT + DMF) network and three other published approaches (Jasil and Ulagamuthalvi [18], Ningrum et al. [17], Gessert et al. [12]), across all datasets, except for PH2, our DeepMetaForge network achieved the highest macro-average F1 score. As discussed earlier, the relevance and quality of metadata can have a significant impact on classification performance. This suggests that DMF is able to leverage metadata effectively to improve classification accuracy. Therefore, while the efficiency of only image classification for the malignant class is higher than that of the baseline networks (slightly higher than DMF), the DMF network outperformed Gessert et al. [12] by 34.29%, Ningrum et al. [17] by 75.15%, and Jasil and Ulagamuthalvi [18] by approximately 19.89% in terms of average F1 across all datasets.

It is worth noting that the Ningrum et al. method performs relatively inferior to other methods, despite their better performance reported in their publication [17]. One explanation could be that they only experimented on a subset of 1,200 images from the ISIC 2019 dataset with roughly 30% of positive samples and whose characteristics may be different from the whole dataset. Furthermore, our experiment, whose results are reported in Table 7, used the ISIC 2020 dataset, which is more comprehensive than the 2019 version while also presenting a severe data imbalance problem with only 1.78% positive samples.

In conclusion, our proposed DMF network demonstrates the highest efficiency compared to the other previously proposed metadata-fusing networks and establishes itself as the state-of-the-art method for image-metadata classification in skin lesion classification. Our results suggest that incorporating metadata into the classification system can improve the accuracy of the diagnosis, and the proposed network provides an effective approach for integrating metadata with image data for skin lesion classification.

F. EVALUATION OF VARIOUS NETWORK COMBINATIONS

While the proposed DeepMetaForge network features the BEiT backbone, this plug-and-play configuration could be easily changed to other image-encoding backbones, such

TABLE 7. Classification performance comparison between different metadata fusing networks (i.e., Ningrum et al., Gessert et al., Jasil and Ulagamuthalvi + DMF and our proposed DeepMetaForge network architecture (BEiT + DMF)) on the selected datasets. Δ F1 denotes the relative performance difference (improvement) of the best-performing method compared to each other approach.

Dataset	Network Architecture	Benign F1	Malignant F1	Macro Avg F1	Δ F1	Accuracy	MCC	Sensitivity	Specificity	NPV
PH2	Gessert (2020)	0.924	0.700	0.812	18.02%	0.880	0.635	0.725	0.919	**0.993**
	Ningrum (2021)	0.889	0.000	0.444	115.64%	0.800	0.000	0.000	**1.000**	0.800
	Jasil (2023) + DMF	**0.983**	**0.934**	**0.958**	-	**0.973**	**0.918**	**0.950**	0.979	0.988
	BEiT + DMF	0.975	0.899	0.937	2.27%	0.960	0.879	0.900	0.975	0.976
SKINL2	Gessert (2020)	0.798	0.359	0.579	50.60%	0.706	0.209	0.387	0.826	0.798
	Ningrum (2021)	0.848	0.200	0.524	66.36%	0.750	0.133	0.200	**0.933**	0.783
	Jasil (2023) + DMF	0.844	0.617	0.731	19.24%	0.781	0.480	0.632	0.837	0.859
	BEiT + DMF	**0.926**	**0.817**	**0.871**	-	**0.896**	**0.755**	**0.830**	0.920	**0.938**
PAD-UFES-20	Gessert (2020)	0.471	0.651	0.561	34.34%	0.611	0.172	0.601	0.560	0.434
	Ningrum (2021)	**0.848**	0.200	0.524	43.85%	0.750	0.133	0.200	**0.933**	**0.783**
	Jasil (2023) + DMF	0.613	0.827	0.720	4.66%	0.761	0.451	0.784	0.699	0.546
	BEiT + DMF	0.633	**0.874**	**0.754**	-	**0.813**	**0.524**	**0.889**	0.610	0.695
ISIC 2020	Gessert (2020)	0.956	0.327	0.642	43.44%	0.918	0.407	**0.927**	0.918	**0.999**
	Ningrum (2021)	0.992	0.000	0.496	85.47%	0.985	0.000	0.000	**1.000**	0.985
	Jasil (2023) + DMF	0.991	0.000	0.496	85.72%	0.982	0.000	0.000	**1.000**	0.982
	BEiT + DMF	**0.997**	**0.843**	**0.920**	-	**0.995**	**0.846**	0.850	0.997	0.997
Average	Gessert (2020)	0.787	0.509	0.648	34.29%	0.779	0.356	0.660	0.806	0.806
	Ningrum (2021)	**0.894**	0.100	0.497	75.15%	0.821	0.067	0.100	**0.967**	0.838
	Jasil (2023) + DMF	0.858	0.595	0.726	19.89%	0.874	0.462	0.591	0.879	0.844
	BEiT + DMF	0.883	**0.858**	**0.871**	-	**0.916**	**0.751**	**0.867**	0.876	**0.901**

as ResNext and EfficientNet, for portability. Figure 7 illustrates example prediction outputs. Therefore, this section provides a comprehensive evaluation of the proposed network architecture using different backbones, compared with the state-of-the-art Gessert et al. method [12] whose backbones are also varied for a fair comparison. Furthermore, the performance of selected image-encoding backbones alone is also reported for reference. Using different image encoders allows one to explore the effectiveness of different feature extraction methods in skin lesion classification. By varying the backbone, we can also evaluate the adaptability of each network architecture on different backbones that implement different architectures.

The evaluation of the proposed network model against the state-of-the-art approaches and only image classification was conducted using a variety of metrics, including F1, accuracy, and MCC. Δ F1 denotes the performance difference relative to the proposed DeepMetaForge network with the BEiT backbone. The results presented in Table 8, and the comparison of F1 score are shown in Fig. 8, show that, in most datasets, the proposed method outperformed the other baselines in many of the evaluation metrics. However, there were some evaluating metrics, such as sensitivity and specificity, in which the DeepMetaForge network with BEiT image encoder backbone did not achieve the highest score in the SKINL2 and PAD-UFES-20 datasets. It is important to note that these metrics focus only on each class of classification, and the difference between the best-performing method and the DMF with BEiT backbone is only marginal. On average, the proposed DMF network with BEiT outperforms the baselines and other configurations in all aspects. Specifically, the best configuration of the proposed network outperforms the best image-encoding backbone and metadata-fusing state-of-the-art methods by 19.39% and 8.49%, respectively.

This research provides valuable insights into the effectiveness of different network models and their ability to classify skin lesions accurately, which can have a significant impact on the development of more efficient and accurate diagnosis and treatment methods. Specifically, the experiment results on the four different datasets support our conjecture that fusing the metadata with visual features while compressing the fused information to extract the low-level representation is an effective approach to combining information from two different sources for skin lesion classification.

G. SCALABILITY ANALYSIS

Skin lesion classification models can be used in telemedicine applications to help dermatologists make informed decisions efficiently. These systems should be accessible via diverse platforms and available to patients and healthcare practitioners, especially those in rural areas. Evaluating the scalability of the proposed method is crucial to ensure its accessibility, usability, and reliability across different settings and scenarios.

In the previous part, we conducted an extensive evaluation and comparison of several skin lesion classification networks to determine the most suitable model for this task. We found that the proposed DeepMetaForge network with the BEiT image encoder backbone was the most efficient for the selected datasets. However, it is also essential to examine the trade-off between the models' efficacy and resource consumption in different scenarios to ensure its

TABLE 8. Classification performance comparison between different metadata fusing methods (i.e., None, Gessert et al., and our proposed DeepMetaForge network) and different backbone image encoders (i.e., ResNext50, EfficientNet-B7, and BEiT) on the selected datasets. Δ F1 denotes the relative performance difference (improvement) of the best-performing method compared to each other approach.

Dataset	Metadata-Fusing Method	Backbone	Benign F1	Malignant F1	Macro Avg F1	Δ F1	Accuracy	MCC	Sensitivity	Specificity	NPV
PH2	None	ResNext50	0.929	0.587	0.758	23.58%	0.880	0.586	0.475	0.981	0.885
		EfficientNet-B7	0.914	0.676	0.795	17.89%	0.865	0.601	0.700	0.906	0.925
		BEiT (base)	0.948	0.754	0.851	10.10%	0.915	0.728	0.700	0.969	0.930
	Gessert (2020)	ResNext50	0.962	0.769	0.865	8.30%	0.935	0.755	0.725	***0.988***	0.940
		EfficientNet-B7	0.924	0.700	0.812	15.41%	0.880	0.635	0.725	0.919	***0.993***
		BEiT (base)	0.969	0.871	0.920	1.86%	0.950	0.850	0.875	0.969	0.970
	DeepMetaForge	ResNext50	0.963	0.845	0.904	3.68%	0.940	0.810	0.825	0.969	0.957
		EfficientNet-B7	0.947	0.675	0.811	15.62%	0.910	0.646	0.650	0.975	0.923
		BEiT (base)	***0.975***	***0.899***	***0.937***	-	***0.960***	***0.879***	***0.900***	0.975	0.976
SKINL2	None	ResNext50	0.908	0.708	0.808	7.87%	0.860	0.642	0.634	0.945	0.875
		EfficientNet-B7	0.879	0.680	0.780	11.79%	0.824	0.559	0.684	0.877	0.881
		BEiT (base)	0.915	0.749	0.832	4.72%	0.875	0.698	0.753	0.920	0.919
	Gessert (2020)	ResNext50	0.842	0.129	0.486	79.35%	0.738	0.100	0.147	***0.960***	0.760
		EfficientNet-B7	0.798	0.359	0.579	50.60%	0.706	0.209	0.387	0.826	0.798
		BEiT (base)	0.919	0.800	0.859	1.41%	0.885	0.729	***0.843***	0.900	***0.941***
	DeepMetaForge	ResNext50	0.882	0.571	0.726	19.96%	0.818	0.512	0.518	0.930	0.847
		EfficientNet-B7	0.837	0.522	0.679	28.25%	0.767	0.426	0.476	0.875	0.823
		BEiT (base)	***0.926***	***0.817***	***0.871***	-	***0.896***	***0.755***	0.830	0.920	0.938
PAD-UFES-20	None	ResNext50	0.507	0.802	0.655	15.13%	0.718	0.313	0.784	0.538	0.483
		EfficientNet-B7	0.366	0.829	0.597	26.14%	0.731	0.227	0.895	0.289	0.511
		BEiT (base)	0.605	***0.880***	0.742	1.51%	***0.816***	0.502	***0.926***	0.521	***0.726***
	Gessert (2020)	ResNext50	0.349	0.677	0.513	46.86%	0.613	0.145	0.671	0.454	0.392
		EfficientNet-B7	0.471	0.651	0.561	34.34%	0.611	0.172	0.601	0.560	0.434
		BEiT (base)	0.607	0.851	0.729	3.32%	0.785	0.467	0.848	***0.617***	0.615
	DeepMetaForge	ResNext50	0.577	0.857	0.717	5.06%	0.787	0.439	0.880	0.536	0.629
		EfficientNet-B7	0.560	0.849	0.704	7.00%	0.775	0.415	0.868	0.526	0.608
		BEiT (base)	***0.633***	0.874	***0.754***	-	0.813	***0.524***	0.889	0.610	0.695
ISIC 2020	None	ResNext50	0.849	0.129	0.489	88.39%	0.743	0.191	0.824	0.740	0.996
		EfficientNet-B7	0.832	0.078	0.455	102.49%	0.716	0.098	0.598	0.718	0.989
		BEiT (base)	0.864	0.118	0.491	87.32%	0.765	0.188	0.841	0.764	0.996
	Gessert (2020)	ResNext50	0.951	0.393	0.672	36.91%	0.910	0.451	0.890	0.911	0.997
		EfficientNet-B7	0.956	0.327	0.642	43.44%	0.918	0.407	***0.927***	0.918	***0.999***
		BEiT (base)	0.979	0.423	0.701	31.27%	0.960	0.456	0.751	0.964	0.995
	DeepMetaForge	ResNext50	0.996	0.715	0.855	7.60%	0.992	0.731	0.685	0.997	0.994
		EfficientNet-B7	0.995	0.561	0.778	18.28%	0.990	0.574	0.496	***0.999***	0.991
		BEiT (base)	***0.997***	***0.843***	***0.920***	-	***0.995***	***0.846***	0.850	0.997	0.997
Avgerage	None	ResNext50	0.798	0.556	0.677	28.54%	0.800	0.433	0.679	0.801	0.810
		EfficientNet-B7	0.748	0.566	0.657	32.59%	0.784	0.372	0.719	0.698	0.826
		BEiT (base)	0.833	0.625	0.729	19.39%	0.843	0.529	0.805	0.793	0.893
	Gessert (2020)	ResNext50	0.776	0.492	0.634	37.29%	0.799	0.363	0.608	0.828	0.772
		EfficientNet-B7	0.787	0.509	0.648	34.29%	0.779	0.356	0.660	0.806	0.806
		BEiT (base)	0.868	0.736	0.802	8.49%	0.895	0.626	0.829	0.863	0.880
	DeepMetaForge	ResNext50	0.854	0.747	0.801	8.73%	0.884	0.623	0.727	0.858	0.857
		EfficientNet-B7	0.834	0.652	0.743	17.16%	0.860	0.515	0.622	0.844	0.836
		BEiT (base)	***0.883***	***0.858***	***0.871***	-	***0.916***	***0.751***	***0.867***	***0.876***	***0.901***

practical application in real-world solutions. Thus, in this part, we evaluate the proposed DeepMetaForge network on the ISIC 2020 dataset while varying different BEiT backbone models with different hyperparameter settings. With a thorough understanding of the model's efficacy-efficiency tradeoff, developers can choose the right model for specific applications.

We also analyzed the effect of the training dataset sizes on the model efficiency. We varied different training sizes, i.e., 10%, 30%, 50%, 70%, 90%, and 100%, and observed the performance on the test set. Fig. 9 plots the F1-score of the Malignant class and the macro-average F1-score as the function of training dataset size. The results show that training data size has a direct impact on performance, which begins to plateau when using over 80% of the training data. However, using the full dataset size still yields the optimal performance.

In this experiment, we compared the performance of different image encoder backbones by evaluating their parameter size, model size, training time, memory usage, and predicting time. The results, presented in Table 9, show that the larger size of the image encoder backbone yields a higher number of the model's parameters and also affects the model's physical size (i.e., storage space on HDD). The large version of BEiT is about three times larger than the base one. The training time per iteration with

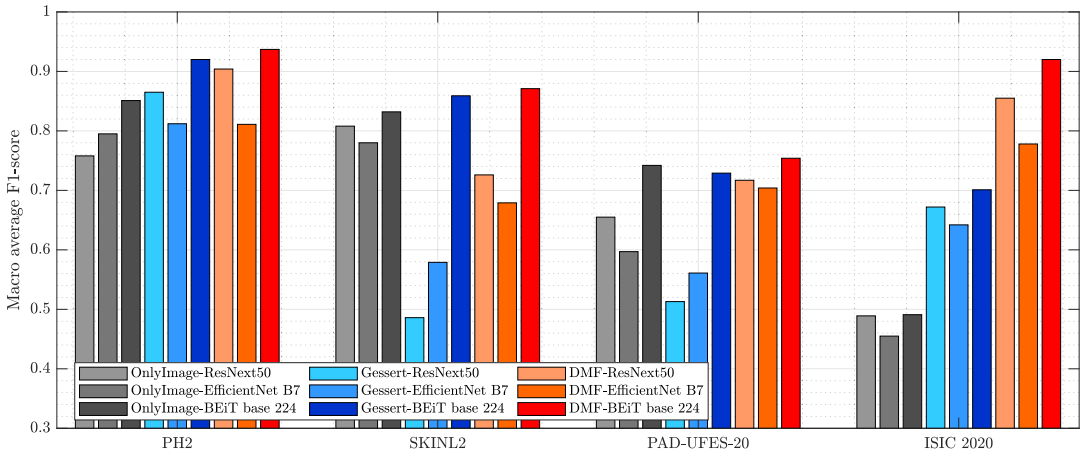

FIGURE 8. Comparison of macro-avg F1 scores from different network architectures and image encoding backbones on all datasets.

TABLE 9. Efficacy-efficiency tradeoff of the proposed DeepMetaForge network using different configurations of the BEiT backbones on the ISIC 2020 dataset.

Backbone	Parameter Size (M)	Model Size (MB)	Training Time/ Iteration (s)	Test Memory Usage (GB)	Average Testing Time (ms)	Macro-Avg F1	Accuracy	MCC
BEiT-base-224	9.31	389.0	0.079	2.07	4.1	0.920	0.995	0.846
BEiT-base-384	9.33	418.1	0.286	2.16	10.13	0.929	0.995	0.862
BEiT-large-224	31.1	1,265.5	0.257	2.87	10.91	0.915	0.994	0.833
BEiT-large-384	31.16	1,324.2	0.9	2.9	30.19	0.896	0.993	0.794

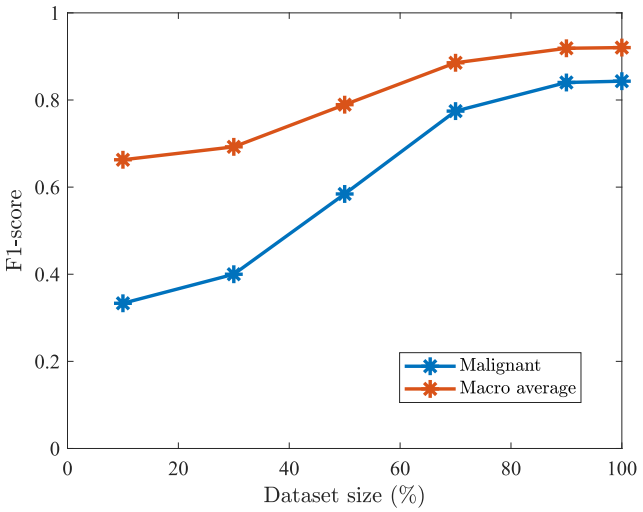

FIGURE 9. F1-score vs. dataset size on ISIC 2020 dataset using the proposed model.

a 16-batch-size of 384-input type took longer than the 224-input type. Additionally, the predicted memory reserve of the GPU for this type of network ranged between 2-3 GB. Due to the model's size, the larger size resulted in a longer predicting time, with the larger size being around three times longer than the base one, and the 384-input type being around three times longer than the 224-input type. Although the classification performance (macro-avg F1 and accuracy) and memory usage of these backbones do not differ much, we suggest using the BEiT-base-224 backbone for those who seek to adopt the proposed framework in a resource-limited environment, such as offline smartphone applications. Such a base model only consumes roughly 2GB of memory to operate, which can be accommodated by many modern smartphones. Once loaded in the memory, this base version also takes 4.1 ms to grade an input sample, which should be fast enough for real-time applications. However, if computation resources are not of critical concern, then the BEiT-base-384 version, which can encode larger images, hence yielding slightly better performance, is recommended. It is worth noting that the larger versions, such as BEiT-large-224 and BEiT-large-384, do not improve the classification efficacy but consume roughly 40% more memory to operate. Therefore, adopting such overkilling large models for skin lesion detection applications is not encouraged. Regardless, more investigation must be done to find ways to tweak these larger models to improve the classification performance.

H. LIMITATIONS

Although the proposed network architecture has demonstrated encouraging performance on the four chosen datasets of skin lesion images, the scope of the presented study primarily examined the performance of the proposed network, with limited emphasis on other essential processes

involved in deploying the model in practical systems. Such procedures include handling data imbalance, augmenting training data to enhance its quality, performing image segmentation to eliminate background noise, and applying image filtering to better certain characteristics of different skin conditions. Researchers and professionals interested in adopting the proposed techniques should thoroughly examine techniques for data preprocessing to further enhance the model's performance.

V. CONCLUSION AND FUTURE WORK

In this experiment, we proposed a novel network architecture, DeepMetaForge, for skin lesion classification, incorporating image and metadata information to improve classification accuracy. The proposed architecture features BEiT image-encoding backbone and the novel Deep Metadata Fusion Module (DMFM) that integrates visual and metadata features while blending them together simultaneously. We evaluated the performance of the DeepMetaForge network, along with other state-of-the-art approaches, on four datasets comprising skin lesion images taken from both dermoscopy and smartphone cameras. The results demonstrated that the proposed network with the BEiT image encoder backbone not only generalized well to different image sources and metadata compositions but also outperformed other networks in terms of F1, accuracy, and MCC, making it a suitable model for skin lesion classification when images and their metadata are available. A scalability analysis was conducted to investigate how the required computation resources would impact the classification performance of the proposed approach. This work can be extended by framing the problem as multiclass classification or object detection tasks, which can have significant implications for pre-screening skin lesion problems in remote areas. Future work can also focus on adapting the network to meet the specific needs of remote communities, ultimately improving public healthcare in underdeveloped and developing countries.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021.

[2] A. Stang and K. Jöckel, "Does skin cancer screening save lives? A detailed analysis of mortality time trends in Schleswig–Holstein and Germany," *Cancer*, vol. 122, no. 3, pp. 432–437, Feb. 2016.

[3] M. Zambrano-Román, J. R. Padilla-Gutiérrez, Y. Valle, J. F. Muñoz-Valle, and E. Valdés-Alvarado, "Non-melanoma skin cancer: A genetic update and future perspectives," *Cancers*, vol. 14, no. 10, p. 2371, May 2022.

[4] A. F. Jerant, J. T. Johnson, C. D. Sheridan, and T. J. Caffrey, "Early detection and treatment of skin cancer," *Amer. Family Physician*, vol. 62, no. 2, pp. 357–368, Jul. 2000.

[5] L. E. Davis, S. C. Shalin, and A. J. Tackett, "Current state of melanoma diagnosis and treatment," *Cancer Biol. Therapy*, vol. 20, no. 11, pp. 1366–1379, Nov. 2019.

[6] B. G. Goldstein and A. O. Goldstein, "Diagnosis and management of malignant melanoma," *Amer. Family Physician*, vol. 63, no. 7, pp. 1359–1369, 2001.

[7] D. J. Brailer, *A Theory of Congestion in General Hospitals*. Philadelphia, PA, USA: Univ. Pennsylvania, 1992.

[8] M. E. Cyr, D. Boucher, S. A. Korona, B. J. Guthrie, and J. C. Benneyan, "A mixed methods analysis of access barriers to dermatology care in a rural state," *J. Adv. Nursing*, vol. 77, no. 1, pp. 355–366, Jan. 2021.

[9] A. Wattanapisit and U. Saengow, "Patients' perspectives regarding hospital visits in the universal health coverage system of Thailand: A qualitative study," *Asia Pacific Family Med.*, vol. 17, no. 1, pp. 1–8, Dec. 2018.

[10] P. Jia, F. Wang, and I. M. Xierali, "Differential effects of distance decay on hospital inpatient visits among subpopulations in Florida, USA," *Environ. Monitor. Assessment*, vol. 191, no. S2, pp. 1–16, Jun. 2019.

[11] K. Hauser, A. Kurz, S. Haggenmüller, R. C. Maron, C. von Kalle, J. S. Utikal, F. Meier, S. Hobelsberger, F. F. Gellrich, and M. Sergon, "Explainable artificial intelligence in skin cancer recognition: A systematic review," *Eur. J. Cancer*, vol. 167, pp. 54–69, May 2022.

[12] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, Jan. 2020, Art. no. 100864.

[13] A. G. C. Pacheco and R. A. Krohling, "An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 9, pp. 3554–3563, Sep. 2021.

[14] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[15] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "AMMU: A survey of transformer-based biomedical pretrained language models," *J. Biomed. Informat.*, vol. 126, Feb. 2022, Art. no. 103982.

[16] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som, and F. Wei, "Image as a foreign language: BEiT pretraining for all vision and vision-language tasks," 2022, *arXiv:2208.10442*.

[17] D. N. A. Ningrum, S.-P. Yuan, W.-M. Kung, C.-C. Wu, I.-S. Tzeng, C.-Y. Huang, J. Y.-C. Li, and Y.-C. Wang, "Deep learning classifier with patient's metadata of dermoscopic images in malignant melanoma detection," *J. Multidisciplinary Healthcare*, vol. 14, pp. 877–885, Apr. 2021.

[18] S. P. G. Jasil and V. Ulagamuthalvi, "A hybrid CNN architecture for skin lesion classification using deep learning," *Soft Comput.*, Mar. 2023, doi: 10.1007/s00500-023-08035-w.

[19] O. Mehta, Z. Liao, M. Jenkinson, G. Carneiro, and J. Verjans, "Machine learning in medical imaging–clinical applications and challenges in computer vision," in *Artificial Intelligence in Medicine*, M. Raz, T. C. Nguyen, and E. Loh, Eds. Singapore: Springer, 2022, doi: 10.1007/978-981-19-1223-8_4.

[20] D. C. Araújo, A. A. Veloso, R. S. de Oliveira Filho, M.-N. Giraud, L. J. Raniero, L. M. Ferreira, and R. A. Bitar, "Finding reduced Raman spectroscopy fingerprint of skin samples for melanoma diagnosis through machine learning," *Artif. Intell. Med.*, vol. 120, Oct. 2021, Art. no. 102161.

[21] I. Maglogiannis and C. N. Doukas, "Overview of advanced computer vision systems for skin lesions characterization," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 5, pp. 721–733, Sep. 2009.

[22] M. d'Amico, M. Ferri, and I. Stanganelli, "Qualitative asymmetry measure for melanoma detection," in *Proc. 2nd IEEE Int. Symp. Biomed. Imag. Macro Nano*, Apr. 2004, pp. 1155–1158.

[23] S. E. Umbaugh, R. H. Moss, and W. V. Stoecker, "Applying artificial intelligence to the identification of variegated coloring in skin tumors," *IEEE Eng. Med. Biol. Mag.*, vol. 10, no. 4, pp. 57–62, Dec. 1991.

[24] M. Wiltgen, A. Gerger, and J. Smolle, "Tissue counter analysis of benign common nevi and malignant melanoma," *Int. J. Med. Informat.*, vol. 69, no. 1, pp. 17–28, Jan. 2003.

[25] A. Iosifidis and A. Tefas, "Deep learning for visual content analysis," *Signal Process., Image Commun.*, vol. 83, Apr. 2020, Art. no. 115806.

[26] T.-C. Pham, C.-M. Luong, M. Visani, and V.-D. Hoang, "Deep CNN and data augmentation for skin lesion classification," in *Proc. Asian Conf. Intell. Inf. Database Syst.*, Dong Hoi City, Vietnam, Mar. 2018, pp. 573–582.

[27] H. K. Gajera, D. R. Nayak, and M. A. Zaveri, "A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104186.

[28] L. Ngo, J. Cha, and J.-H. Han, "Deep neural network regression for automated retinal layer segmentation in optical coherence tomography images," *IEEE Trans. Image Process.*, vol. 29, pp. 303–312, 2020.

[29] X. Yu, Z. Zhou, Q. Gao, D. Li, and K. Ríha, "Infrared image segmentation using growing immune field and clone threshold," *Infr. Phys. Technol.*, vol. 88, pp. 184–193, Jan. 2018.

[30] X. Yu and X. Tian, "A fault detection algorithm for pipeline insulation layer based on immune neural network," *Int. J. Pressure Vessels Piping*, vol. 196, Apr. 2022, Art. no. 104611.

[31] X. Yu, Y. Lu, and Q. Gao, "Pipeline image diagnosis algorithm based on neural immune ensemble learning," *Int. J. Pressure Vessels Piping*, vol. 189, Feb. 2021, Art. no. 104249.

[32] Z. Zhou, B. Zhang, and X. Yu, "Immune coordination deep network for hand heat trace extraction," *Infr. Phys. Technol.*, vol. 127, Dec. 2022, Art. no. 104400.

[33] W. jiahao, J. Xingguang, W. Yuan, Z. Luo, and Z. Yu, "Deep neural network for melanoma classification in dermoscopic images," in *Proc. IEEE Int. Conf. Consum. Electron. Comput. Eng. (ICCECE)*, Jan. 2021, pp. 666–669.

[34] Y. Zhang and C. Wang, "SIIM-ISIC melanoma classification with DenseNet," in *Proc. IEEE 2nd Int. Conf. Big Data, Artif. Intell. Internet Things Eng. (ICBAIE)*, Mar. 2021, pp. 14–17.

[35] N. Zhang, Y.-X. Cai, Y.-Y. Wang, Y.-T. Tian, X.-L. Wang, and B. Badami, "Skin cancer diagnosis based on optimized convolutional neural network," *Artif. Intell. Med.*, vol. 102, Jan. 2020, Art. no. 101756.

[36] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Adv. Eng. Softw.*, vol. 95, pp. 51–67, Feb. 2016.

[37] Z. Liu, R. Xiong, and T. Jiang, "CI-Net: Clinical-inspired network for automated skin lesion recognition," *IEEE Trans. Med. Imag.*, vol. 42, no. 3, pp. 619–632, Mar. 2023.

[38] R. Kaur, H. GholamHosseini, R. Sinha, and M. Lindén, "Melanoma classification using a novel deep convolutional neural network with dermoscopic images," *Sensors*, vol. 22, no. 3, p. 1134, Feb. 2022.

[39] H. C. Reis, V. Turk, K. Khoshelham, and S. Kaya, "InSiNet: A deep convolutional approach to skin cancer detection and segmentation," *Med. Biol. Eng. Comput.*, vol. 60, no. 3, pp. 643–662, Mar. 2022.

[40] P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. Andía, C. Tejos, C. Prieto, and D. Capurro, "A survey on deep learning and explainability for automatic report generation from medical images," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–40, Jan. 2022.

[41] J. López-Labraca, I. González-Díaz, F. Díaz-de-María, and A. Fueyo-Casado, "An interpretable CNN-based CAD system for skin lesion diagnosis," *Artif. Intell. Med.*, vol. 132, Oct. 2022, Art. no. 102370.

[42] E. Rezk, M. Eltorki, and W. El-Dakhakhni, "Interpretable skin cancer classification based on incremental domain knowledge learning," *J. Healthcare Informat. Res.*, vol. 7, no. 1, pp. 59–83, Mar. 2023.

[43] M. B. Alatise and G. P. Hancke, "A review on challenges of autonomous mobile robot and sensor fusion methods," *IEEE Access*, vol. 8, pp. 39830–39846, 2020.

[44] H. Dhayne, R. Haque, R. Kilany, and Y. Taher, "In search of big medical data integration solutions—A comprehensive survey," *IEEE Access*, vol. 7, pp. 91265–91290, 2019.

[45] L. Leng, M. Li, C. Kim, and X. Bi, "Dual-source discrimination power analysis for multi-instance contactless palmprint recognition," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 333–354, Jan. 2017.

[46] L. Leng and J. Zhang, "PalmHash code vs. PalmPhasor code," *Neurocomputing*, vol. 108, pp. 1–12, May 2013.

[47] Z. Yang, Z. Gui, H. Wu, and W. Li, "A latent feature-based multimodality fusion method for theme classification on web map service," *IEEE Access*, vol. 8, pp. 25299–25309, 2020.

[48] T. Langenberg, T. Lüddecke, and F. Wörgötter, "Deep metadata fusion for traffic light to lane assignment," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 973–980, Apr. 2019.

[49] J. S. Ellen, C. A. Graff, and M. D. Ohman, "Improving plankton image classification using context metadata," *Limnol. Oceanography, Methods*, vol. 17, no. 8, pp. 439–461, Aug. 2019.

[50] M. Boutell and J. Luo, "Photo classification by integrating image content and camera metadata," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2004, pp. 901–904.

[51] Q. Zhu, M.-C. Yeh, and K.-T. Cheng, "Multimodal fusion using learned text concepts for image categorization," in *Proc. 14th ACM Int. Conf. Multimedia*, Oct. 2006, pp. 211–220.

[52] S. L. Lee, M. R. Zare, and H. Müller, "Late fusion of deep learning and handcrafted visual features for biomedical image modality classification," *IET Image Process.*, vol. 13, no. 2, pp. 382–391, Feb. 2019.

[53] R. I. Jony, A. Woodley, and D. Perrin, "Flood detection in social media images using visual features and metadata," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Dec. 2019, pp. 1–8.

[54] R. I. Jony, A. Woodley, and D. Perrin, "Fusing visual features and metadata to detect flooding in Flickr images," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2020, pp. 1–8.

[55] J. Höhn, "Combining CNN-based histologic whole slide image analysis and patient data to improve skin cancer classification," *Eur. J. Cancer*, vol. 149, pp. 94–101, May 2021.

[56] F. Nunnari, C. Bhuvaneshwara, A. O. Ezema, and D. Sonntag, "A study on the fusion of pixels and patient metadata in CNN-based classification of skin lesion images," in *Proc. Int. Cross-Domain Conf. Mach. Learn. Knowl. Extraction*, Dublin, Ireland, Aug. 2020, pp. 191–208.

[57] W. Li, J. Zhuang, R. Wang, J. Zhang, and W.-S. Zheng, "Fusing metadata and dermoscopy images for skin disease diagnosis," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1996–2000.

[58] M.-H. Guo, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, pp. 331–368, Mar. 2022.

[59] A. Pundhir, S. Dadhich, A. Agarwal, and B. Raman, "Towards improved skin lesion classification using metadata supervision," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 4313–4320.

[60] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*.

[61] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang, "A multimodal transformer to fuse images and metadata for skin disease classification," *Vis. Comput.*, vol. 39, no. 7, pp. 2781–2793, Jul. 2023.

[62] P. Tang, X. Yan, Y. Nan, S. Xiang, S. Krammer, and T. Lasser, "FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102307.

[63] S. Vachmanus, A. A. Ravankar, T. Emaru, and Y. Kobayashi, "Multi-modal sensor fusion-based semantic segmentation for snow driving scenarios," *IEEE Sensors J.*, vol. 21, no. 15, pp. 16839–16851, Aug. 2021.

[64] T. J. D. Berstad, M. Riegler, H. Espeland, T. de Lange, P. H. Smedsrud, K. Pogorelov, H. Kvale Stensland, and P. Halvorsen, "Tradeoffs using binary and multiclass neural network classification for medical multidisease detection," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2018, pp. 1–8.

[65] V. Rotemberg, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci. Data*, vol. 8, no. 1, p. 34, Jan. 2021.

[66] A. G. C. Pacheco, G. R. Lima, A. S. Salomão, B. Krohling, I. P. Biral, G. G. de Angelo, F. C. R. Alves Jr., J. G. M. Esgario, A. C. Simora, P. B. C. Castro, F. B. Rodrigues, P. H. L. Frasson, R. A. Krohling, H. Knidel, M. C. S. Santos, R. B. do Espírito Santo, T. L. S. G. Macedo, T. R. P. Canuto, and L. F. S. de Barros, "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data Brief*, vol. 32, Oct. 2020, Art. no. 106221.

[67] S. M. M. de Faria, J. N. Filipe, P. M. M. Pereira, L. M. N. Tavora, P. A. A. Assuncao, M. O. Santos, R. Fonseca-Pinto, F. Santiago, V. Dominguez, and M. Henrique, "Light field image dataset of skin lesions," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 3905–3908.

[68] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH2–A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 5437–5440.

[69] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995.

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

[71] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," 2016, *arXiv:1610.02391*.

SIRAWICH VACHMANUS received the M.E. and Ph.D. degrees from Hokkaido University, Japan, in 2019 and 2022, respectively. He is currently a Lecturer with the Faculty of Information and Communication Technology, Mahidol University, Thailand. His research interests include computer vision, deep learning, machine learning, sensor fusion, and artificial intelligence for robots.

THANAPON NORASET received the B.Sc. degree from the Faculty of Information and Communication Technology, in 2007, and the Ph.D. degree in computer science from Northwestern University, USA, in 2017. He is currently a Faculty Member with the Faculty of Information and Communication Technology, Mahidol University, Thailand. His research interests include natural language processing and machine learning.

WARITSARA PIYANONPONG received the M.D. degree from the Faculty of Medicine, Thammasat University, and the Diploma degree in family medicine from Trang Hospital. Currently, she is a Dermatology Resident with the Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Thailand. Her research interest includes dermatology research, which also includes the role of artificial intelligence in dermatology.

TEERAPONG RATTANANUKROM received the M.D. degree from the Medical School, Khon Kaen University, the M.Sc. degree in dermatology from the University of Hertfordshire, U.K., and the Diploma degree from the Thai Board of Dermatology, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Thailand. Currently, he is a Clinical Instructor with the Division of Dermatology, Department of Medicine, Faculty of Medicine, Ramathibodi Hospital, Mahidol University. His research interests include the treatment of cutaneous lymphoma, melanoma, non-melanoma skin cancers, and surgical techniques in Mohs surgery.

SUPPAWONG TUAROB (Member, IEEE) received the Ph.D. degree in computer science and engineering and the M.S. degree in industrial engineering from The Pennsylvania State University and the B.S.E. and M.S.E. degrees in computer science and engineering from the University of Michigan-Ann Arbor. Currently, he is an Associate Professor in computer science with the Faculty of Information and Communication Technology, Mahidol University, Thailand. His research interests include data mining in large-scale scholarly, social media, healthcare domains, and applications of intelligent technologies for social good.