A deep learning fusion network trained with clinical and high-frequency ultrasound images in the multi-classification of skin diseases in comparison with dermatologists: a prospective and multicenter study

An-Qi Zhu,[a,b,g,k] Qiao Wang,[a,b,c,k] Yi-Lei Shi,[d,k] Wei-Wei Ren,[a,b,c] Xu Cao,[d] Tian-Tian Ren,[e] Jing Wang,[f] Ya-Qin Zhang,[g] Yi-Kang Sun,[g] Xue-Wen Chen,[h] Yong-Xian Lai,[h] Na Ni,[h] Yu-Chong Chen,[h] Jing-Liang Hu,[d] Li-Chao Mou,[d] Yu-Jing Zhao,[a] Ye-Qiang Liu,[i] Li-Ping Sun,[b,c] Xiao-Xiang Zhu,[j,***] Hui-Xiong Xu,[g,*] and Le-Hang Guo,[a,b,c,**] China Alliance of Multi-Center Clinical Study for Ultrasound (Ultra-Chance)

[a]Department of Medical Ultrasound, Shanghai Skin Disease Hospital, School of Medicine, Tongji University, Shanghai, China
[b]Department of Medical Ultrasound, Shanghai Tenth People's Hospital, School of Medicine, Tongji University, Shanghai, China
[c]Shanghai Engineering Research Center of Ultrasound Diagnosis and Treatment, Shanghai, China
[d]MedAI Technology (Wuxi) Co., Ltd., Wuxi, China
[e]Department of Medical Ultrasound, Ma'anshan People's Hospital, Ma'anshan, China
[f]Department of Ultrasound, Jiading District Central Hospital Affiliated Shanghai University of Medicine & Health Sciences, Shanghai, China
[g]Department of Ultrasound, Zhongshan Hospital, Institute of Ultrasound in Medicine and Engineering, Fudan University, Shanghai, China
[h]Department of Dermatological Surgery, Shanghai Skin Disease Hospital, School of Medicine, Tongji University, Shanghai, China
[i]Department of Pathology, Shanghai Skin Disease Hospital, School of Medicine, Tongji University, Shanghai, China
[j]Chair of Data Science in Earth Observation, Technical University of Munich, Munich, Germany

eClinicalMedicine
2024;67: 102391

Published Online xxx
https://doi.org/10.1016/j.eclinm.2023.102391

Summary

Background Clinical appearance and high-frequency ultrasound (HFUS) are indispensable for diagnosing skin diseases by providing internal and external information. However, their complex combination brings challenges for primary care physicians and dermatologists. Thus, we developed a deep multimodal fusion network (DMFN) model combining analysis of clinical close-up and HFUS images for binary and multiclass classification in skin diseases.

Methods Between Jan 10, 2017, and Dec 31, 2020, the DMFN model was trained and validated using 1269 close-ups and 11,852 HFUS images from 1351 skin lesions. The monomodal convolutional neural network (CNN) model was trained and validated with the same close-up images for comparison. Subsequently, we did a prospective and multicenter study in China. Both CNN models were tested prospectively on 422 cases from 4 hospitals and compared with the results from human raters (general practitioners, general dermatologists, and dermatologists specialized in HFUS). The performance of binary classification (benign vs. malignant) and multiclass classification (the specific diagnoses of 17 types of skin diseases) measured by the area under the receiver operating characteristic curve (AUC) were evaluated. This study is registered with www.chictr.org.cn (ChiCTR2300074765).

Findings The performance of the DMFN model (AUC, 0.876) was superior to that of the monomodal CNN model (AUC, 0.697) in the binary classification ($P = 0.0063$), which was also better than that of the general practitioner (AUC, 0.651, $P = 0.0025$) and general dermatologists (AUC, 0.838; $P = 0.0038$). By integrating close-up and HFUS images, the DMFN model attained an almost identical performance in comparison to dermatologists (AUC, 0.876 vs. AUC, 0.891; $P = 0.0080$). For the multiclass classification, the DMFN model (AUC, 0.707) exhibited superior prediction performance compared with general dermatologists (AUC, 0.514; $P = 0.0043$) and dermatologists specialized in HFUS (AUC, 0.640; $P = 0.0083$), respectively. Compared to dermatologists specialized in HFUS, the DMFN model showed better or comparable performance in diagnosing 9 of the 17 skin diseases.

*Corresponding authors. Department of Ultrasound, Zhongshan Hospital, Institute of Ultrasound in Medicine and Engineering, Fudan University, Shanghai, 200032, China.

**Corresponding author. Shanghai Skin Disease Hospital, Department of Medical Ultrasound, Shanghai Tenth People's Hospital, Shanghai Engineering Research Center of Ultrasound Diagnosis and Treatment, School of Medicine, Tongji University, Shanghai, 200072, China.

***Corresponding author. Chair of Data Science in Earth Observation, Technical University of Munich, Munich, Germany.

E-mail addresses: xu.huixiong@zs-hospital.sh.cn (H.-X. Xu), gopp1314@hotmail.com (L.-H. Guo), xiaoxiang.zhu@tum.de (X.-X. Zhu).

[k]These authors contributed equally to this work.

Interpretation The DMFN model combining analysis of clinical close-up and HFUS images exhibited satisfactory performance in the binary and multiclass classification compared with the dermatologists. It may be a valuable tool for general dermatologists and primary care providers.

Funding This work was supported in part by the National Natural Science Foundation of China and the Clinical research project of Shanghai Skin Disease Hospital.

Keywords: Skin disease; Convolutional neural network; High-frequency ultrasound; Multi-classification

Research in context

Evidence before this study

We searched Google Scholar, PubMed, and Web of Science from database inception to Jun 30, 2023, using the combination of the following terms: "ultrasound" AND "skin disease" AND ("CNN" OR "convolutional neural network" OR "deep learning") AND ("clinical image" OR "close-up"). We found several studies focused on the application of convolutional neural networks (CNN) for skin classification using either high-frequency ultrasound (HFUS) or clinical images. However, few studies combined both image types to train CNNs for skin disease classification. Though it is proved that adding combined HFUS images can enhance the accuracy of skin disease diagnosis, HFUS examination may be relatively complex for most primary care physicians and dermatologists to monitor and accurately identify cancerous skin. To address this issue, it is necessary to develop an effective and convenient method to integrate images from both modalities to improve the diagnosis accuracy and popularization of HFUS examination.

Added value of this study

In this study, we developed a deep multimodal fusion network (DMFN) combining analysis of clinical close-up (external information) and HFUS images (internal information) for diagnosing skin diseases. Differing from previous studies limited to preselected skin diseases and binary classification, our DMFN model exhibited satisfactory performance in the binary and multiclass classification of a broad range of skin diseases compared with dermatologists.

Implications of all the available evidence

Our results indicated that the CNN-based DMFN model had promising performance in diagnosing skin diseases and improved the diagnostic performance of dermatologists who relied on visual inspection alone. Additionally, the DMFN model may be a feasible and potentially attractive method to inform primary care provider referral decisions effectively.

Introduction

Skin cancer, as the most common but complex malignancy in humans,[1,2] places significant burdens on healthcare services. It is primarily diagnosed through visual inspection, with dermoscopy commonly used. However, dermoscopy can only provide information about skin lesions' visible and external features, and a histopathological examination is usually required for confirmation. While the histopathological examination is considered the gold standard for diagnosing skin malignancies, this process is invasive, painful, and limited in sampling. To avoid unnecessary invasive procedures, non-invasive tools such as optical coherence tomography (OCT), reflectance confocal microscopy (RCM), and high-frequency ultrasound (HFUS) have been developed.[3–6] However, similar to dermoscopy, RCM and OCT may theoretically result in an inadequate assessment of lesions due to their limited depth penetration. In contrast, as one of the commonly used devices in clinical practice, HFUS has the potential to offer improved penetration capabilities, providing accurate quantification of tumor size and clear visualization of internal structures such as vessels and skin appendages.[7–9] Owing to its better visibility in the longitudinal direction, low cost, versatility, and real-time scanning,[5,10,11] HFUS has been increasingly used in dermatologic field for the initial differential diagnosis, surgical planning, and follow-up.[5,10–12]

Though it is proved that the addition of combined HFUS images can improve the accuracy of skin disease diagnosis,[5,12,13] HFUS examination may be relatively complex for most primary care physicians and dermatologists to monitor and accurately identify cancerous skin. To address this issue, it is necessary to develop an effective and convenient method to improve the diagnosis accuracy and popularization of HFUS examination.

Most studies have proven convolutional neural networks (CNN) in skin lesion classification tasks achieved a diagnostic accuracy at or above the level of dermatologists.[14–17] Nevertheless, most CNN solutions in early studies were developed to diagnose skin lesions by only external information,[2,18,19] like clinical or dermoscopic images. However, CNNs for some new modalities (especially for HFUS) focused on internal information about skin lesions were insufficient. Moreover, the combining solution of external and internal information has not been developed.

On the other hand, the potential diagnoses of skin diseases are virtually extensive in clinical practice. However, most studies have focused only on selected specific diseases or the binary diagnostic classification (i.e., benign vs. malignant, nevi vs. melanomas, keratinocyte carcinomas vs. benign seborrheic keratosis).[20–23] While these studies were helpful, their limited scope was insufficient to deal with the challenge of diagnosing a broad spectrum of skin diseases. Therefore, a multiclass classification system that covers a wider range of skin diseases may be better suited for clinical needs.

Thus, we aimed to integrate clinical appearance (close-ups) with HFUS images by a CNN-based solution to improve dermatological referrals. We hypothesized that thoroughly combining external and internal information may yield a promising effect in diagnosing a broad spectrum of skin diseases, not only binary classification.

Methods

Ethics

This study was approved by the institutional review board of the ethics committees at all study centers (approval number: SSDH-IEC-SG-029-4.1). Informed consent was obtained for all participants. This study was registered through www.chictr.org.cn (ChiCTR2300074765) and reported according to the Checklist for Evaluation of Image-Based Artificial Intelligence Reports in Dermatology.[24]

Study design

This is a prospective and multicenter study. We first compared some candidate CNNs trained by clinical close-up and HFUS images and then selected the superior network based on the binary classification result. Simultaneously, human raters (general practitioners, general dermatologists, and dermatologists specialized in HFUS) were organized to diagnose the test cohorts. Their binary and multiclass classification results were compared with those of the selected network. Fig. 1 illustrates a comprehensive design of the study.

Image data sets

Between Jan 10, 2017, and Dec 31, 2020, a total of 1269 close-ups and 11,852 HFUS images of 1351 lesions were collected for training and validation cohorts at the Shanghai Skin Disease Hospital in Shanghai, China. From Jan 1, 2021 to Oct 31, 2022, consecutive patients with skin lesions were prospectively enrolled from Shanghai Skin Disease Hospital, Shanghai Tenth People's Hospital, Shanghai Jiading District Central Hospital in Shanghai, and the Ma'anshan People's Hospital in Anhui, China, as the test cohorts (887 close-ups and 2199 HFUS images of 422 lesions).

The clinical close-up images were taken by physicians using a camera (Nikon P510) and smartphones (Apple iPhone Xs, 12, XIAOMI Mi 12, and HUAWEI Mate 20). The photographic images of the training and validation cohort were taken under standardized conditions, ensuring a consistent image quality. However, the images of the test cohort were taken in different hospitals, leading to non-uniform backgrounds and lighting conditions.

For each hospital, the HFUS examinations were performed by experienced dermatologists who mastered dermatologic HFUS examinations. In detail, for each lesion, clear greyscale US images and color Doppler flow imaging (CDFI) images were obtained for further evaluation. The manufacturers of HFUS equipment applied in the study are presented in Supplementary Table S1.

The cases that fulfilled the following criteria were included: (1) availability of at least one clinical close-up image or one HFUS image, (2) the photographs and HFUS examinations were performed before any treatments or biopsies, (3) availability of an unequivocal histopathologic diagnosis. The exclusion criteria were as follows: (1) cases that were inadequate for clinical diagnosis, (2) cases with low image quality, (3) lesions covered by pen markings or tattoos in close-ups or obstructed by the posterior acoustic shadow of hyperkeratosis on HFUS. It was strictly prohibited to have any overlap between the datasets used for training, validation, and testing purposes.

Pathological ground truth

The diagnoses of skin diseases we enrolled were classified as benign or malignant based on histological pathology results.[25] Benign diseases included cysts, lipomas, nevus, benign keratosis-like lesions (BKL) including seborrheic keratosis (SK) and lichen planus-like keratosis, benign sebaceous neoplasms including sebaceous hyperplasia and sebaceous adenoma, haemangioma including angioma, cherry haemangioma, pyogenic granuloma and angiokeratoma, warts, and inflammation. Malignant diseases included skin cancers such as basal cell carcinoma (BCC), squamous cell carcinoma (SCC) including invasive SCC and keratoacanthoma, extramammary Paget's disease (EMPD), dermatofibrosarcoma protuberans (DFSP), malignant melanoma (MM), and precancerous lesions such as actinic keratosis (AK) and Bowen's disease (BD). Table 1 shows the frequencies of diagnoses in the training, validation, and test cohort.

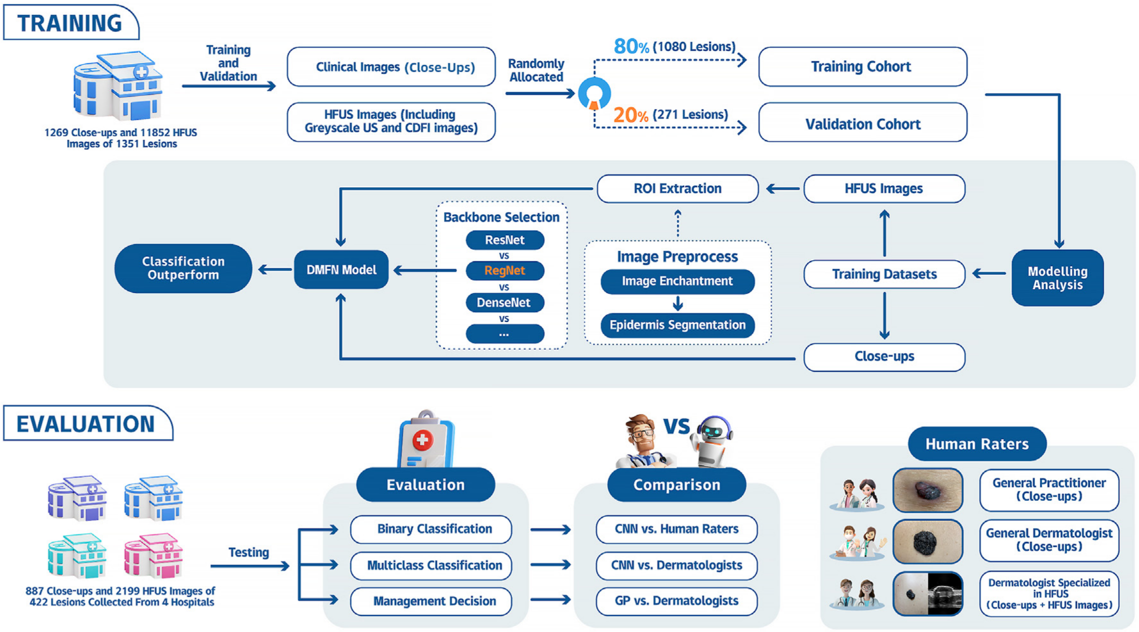

***Fig. 1:* Overview of the study design**. HFUS, high-frequency ultrasound; CDFI, color Doppler flow imaging; ROI, region of interest; DMFN, deep multimodal fusion network; CNN, convolutional neural network (including monomodal CNN model and DMFN model); GP, general practitioner.

In clinical practice, skin diseases exhibit a wide range of diversity and have a notable “Long Tail Effect”. That is, in statistics, although each subset may only contain a small amount of data, the cumulative amount increases significantly due to the large number of subsets. To ensure the precision of the model was not compromised

Classification	Diagnosis	Training cohort n = 1080	Validation cohort n = 271	Test cohort n = 422
Benign		486	123	235
	Cyst	136	34	49
	Lipoma	14	3	3
	Nevus	64	16	38
	BKL	74	19	17
	Benign sebaceous neoplasm	20	6	7
	Haemangioma	20	5	20
	Wart	33	9	10
	Inflammation	32	8	23
	Other[a]	93	23	68
Malignant (precancerous)		151	38	25
	AK	70	17	11
	BD	81	21	14
Malignant (cancer)		443	110	162
	BCC	216	54	65
	SCC	116	30	56
	MM	17	4	5
	EMPD	64	16	13
	DFSP	8	1	5
	Other[b]	22	5	18

Abbreviations: BKL, benign keratosis-like lesions; AK, actinic keratosis; BD, Bowen’s disease; BCC, basal cell carcinoma; SCC, squamous cell carcinoma; MM, malignant melanoma; EMPD, extramammary Paget’s disease; DFSP, dermatofibrosarcoma protuberans. The list of diagnoses of other[a] and other[b] is available in the Supplementary Table S2.

***Table 1:* Summary of diagnosis in the training, validation, and test cohorts.**

by the underrepresentation of specific skin diseases, we grouped rare benign and malignant diseases with small sample sizes (n < 8) into the groups of other^{a} (benign) and other^{b} (malignant), respectively. The list of other^{a} and other^{b} diagnoses is available in Supplementary Table S2.

Data preprocessing

In the HFUS image enhancement preprocessing module, firstly, we weakened the speckle noise of the HFUS image. There are many denoising techniques applied to improve the performance of skin disease diagnosis methods.[26] To avoid a loss of information, we used homomorphic filtering due to its effectiveness on our data types. Also, we applied histogram equalization rather than the intensity normalization methods, such as in the previous studies,[27,28] leading to increased computational costs. Then, we performed pyramid feature fusion based on spatio–temporal correlation and information complementarity on two augmented images. It is known that deep networks need huge data, and there are a lot of augmentation algorithms that have been applied to increase the reliability and robustness of the methods.[29–31] In this work, augmented images have been obtained carefully with these methods: homomorphic filtering algorithm, histogram equalization algorithm, and pyramid feature fusion algorithm (Figure S1). Finally, we introduced the skin segmentation module into the classification framework to improve the final classification accuracy. The details of parameters and methods are provided in Supplementary Method S1.

Deep multimodal fusion network model development

To help the classifier focus on the relevant features of the skin layer, we utilized the DeepLabv3+ network to segment the skin layer. Subsequently, based on prior knowledge and the segmentation mask, we cropped the skin areas and lesions from the initial HFUS image. This cropped image was then used for the subsequent classification task.

To integrate internal and external information, we designed a deep multimodal fusion network (DMFN) model that used skin lesions' appearance and HFUS information for deep feature fusion. The network input included three modal data of color clinical close-up images, greyscale US images, and CDFI images. Each modality data would go through DMFN for multi-layer feature extraction, and the features extracted by each layer of DMFN Block would be retained, fused, and finally classified into skin diseases. Meanwhile, we created a monomodal CNN model that utilized a similar training strategy to the DMFN model for comparison. The model was designed only for close-up images and could output binary or multiclass classification results. The workflow of CNN model development is presented in Figure S1.

Before building the DMFN and monomodal CNN model, we selected the appropriate backbone to test and compare a variety of CNNs (ResNet, DenseNet, EfficientNet, RegNet, etc.). For the test cohort, the RegNet network achieved superior performance than other networks, involving both monomodal and multimodal approaches ($P < 0.050$) (Figure S2 and Table S3). Given this reality, the RegNet network was selected as the backbone for building the DMFN and monomodal CNN model in our study.

Cross-entropy loss is a widely used loss function in classification tasks, which reflects the distance between the model prediction results and the true label of the data. Although hybrid loss functions have been used in some deep networks developed for skin disease diagnosis,[32] we used cross-entropy loss function due to its less computational complexity and efficiency with our datasets.

When performing a classification task, the output value was the predicted value of the category, which was converted into a probability value through the SoftMax function. The output category with the highest probability or top 3 was selected, and the model with the highest accuracy and recall on the validation dataset was saved. The top 1 prediction was used to evaluate the performance of the model. The detailed methods regarding the model development are provided in Supplementary Method S2.

Reader study

Human participants were divided into three groups: General practitioners (n = 3), general dermatologists (n = 4), and dermatologists specialized in HFUS (n = 3). All the participants had at least five years of experience. General practitioners and general dermatologists were presented with solely the close-up images, while dermatologists specialized in HFUS were presented with both close-up and HFUS images. General practitioners were supposed to indicate benign lesions or possible skin cancer and assess whether the patient required a referral. The other two types of dermatologists were asked to indicate the dichotomous diagnosis (benign vs. malignant), specific diagnoses, and management decisions (treatment/excision, follow-up, ignore/no action needed). The participants were blinded to the clinical information of patients.

As a result, we compared the diagnostic performance of CNN models and the participants to validate the value of CNN models in diagnosing skin diseases.

Heat map generation

To provide a better understanding of the prediction results obtained from the DMFN model, we generated heat maps using the gradient-weighted class activation mapping (Grad-CAM) method. The heat maps were produced by applying the packages pytorch-grad-cam 1.4.8 (https://github.com/jacobgil/pytorch-grad-cam).

Statistical analysis

Receiver operating characteristic (ROC) curves and the area under the ROC curves (AUC) were calculated to evaluate the diagnostic performance of the CNN models and human raters. The sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were also performed. Normal distribution was evaluated using the Kolmogorov–Smirnov test. Continuous variables with normal distribution were presented as mean ± standard deviation (SD). The comparison of AUCs was performed by the DeLong test.[33] McNemar's test was used to assess the differences in sensitivity and specificity. The t-test was used to analyze the accuracy differences among the different models and human raters. Results were considered statistically significant at $P < 0.050$. As for management decisions, the classification of "no referral", "ignore/no action needed", and "follow-up" for dermatologists were considered as true-negative for benign lesions, and "referral" and "treatment/excision" as true-positive for malignant lesions. Of note, "excision/treatment" and "follow-up examination" of AK were considered as true-positive due to its limited potential to progress to invasive carcinoma. Confusion matrices were used to indicate the diagnostic performance of specific skin diseases by CNN models and human raters. The software details and the reasons for selecting statistical methods are provided in Supplementary Method S3.

Role of the funding source

The funders played no role in the study design, data collection, data analysis, data interpretation, or writing of the manuscript. All authors approved the final manuscript for submission.

Results

Characteristics of patients and lesions

As shown in Table 2, a total of 1066 patients (mean age, 63.59 years ± 18.97; range, 4–93 years) with 1080 skin lesions were included in the training cohort, and 264 patients (mean age, 63.34 years ± 21.71; range, 35–81 years) with 271 lesions in the internal validation cohort. For the test cohort, a total of 422 skin lesions in 403 patients (mean age, 57.68 years ± 21.46; range, 3–97 years) were prospectively enrolled. The number of lesions in the head and neck was the highest among all cohorts. Detailed results of the anatomic location are also described (Table 2). After comparison, only three diseases in the other[a] category were found to be out-of-distribution in the test cohort (Table S2).

Diagnostic performance of CNN models

When evaluated on the test cohort, for binary classification, the AUC of the DMFN model combining analysis of HFUS and clinical close-up images was significantly higher than monomodal CNN model analyzing clinical close-ups alone (0.876; 95% confidence interval [CI], 0.843–0.911 vs. 0.697; 95% CI, 0.648–0.750; $P = 0.0063$) (Fig. 2, Table 3, and Table S4). Additionally, the DMFN model showed sensitivity and accuracy across the monomodal CNN model in the test cohort (Table 3). Both CNN models were better at diagnosing benign than malignant cases (Fig. 3). Nevertheless, the percentage of correct classifications of the DMFN model in malignant lesions was remarkably increased compared with the monomodal CNN model (87% vs. 57% correct predictions) (Fig. 3).

Concerning the specific diagnoses, the performance of the DMFN model was better than the monomodal CNN model (AUC, 0.707; 95% CI, 0.638–0.776 vs. AUC, 0.501; 95% CI, 0.452–0.551; $P = 0.0070$) (Fig. 2 and Table S4). The detailed confusion matrices of the CNN models revealed the percentage of correct prediction (Fig. 4). In general, the DMFN model tended to achieve higher accuracy than the monomodal CNN model for all diseases. The DMFN model could assist the monomodal CNN model in decreasing the possibility of misdiagnosing benign lesions as malignant and vice versa. For instance, 25% of the lipomas were incorrectly identified by the monomodal CNN model as malignant lesions (SCC) but were subsequently reclassified as benign by the DMFN model. Similarly, the DMFN model greatly reduced the likelihood that the monomodal CNN model might incorrectly diagnose some malignant conditions (including BD, MM, and DFSP) as benign.

Diagnostic performance of human raters for the binary classification

The general dermatologists performed significantly better than the general practitioners with only one close-up image at hand in the test cohort (AUC, 0.838; 95% CI, 0.798–0.882 vs. AUC, 0.651; 95% CI, 0.550–0.750; $P = 0.0083$). With additional information on HFUS, the diagnostic performance of dermatologists significantly improved to an AUC of 0.891 (95% CI, 0.857–0.921; $P = 0.0032$) (Fig. 2, Table 3, and Table S4). Furthermore, dermatologists with additional information on HFUS showed a superior sensitivity of 0.898 (95% CI, 0.857–0.937) when compared with general dermatologists (0.775; 95% CI, 0.717–0.830; $P = 0.00019$), while the specificity was not significantly improved (Table 3).

For benign and malignant lesions, the percentage of correct predictions was lowest in general practitioners. Adding HFUS information significantly improved dermatologists' accuracy in diagnosing malignancies (from 78% to 90%). However, there was no significant improvement in diagnosing benign lesions with HFUS information (Fig. 3).

Management decisions

Expectedly, the rate of correct management decisions was improved from general practitioners to

Characteristics	Training cohort n = 1080	Validation cohort n = 271	Test cohort n = 422
Patient demographics			
No. of unique individuals	1066	264	403
Age, y (mean ± SD)	63.59 ± 18.97	63.34 ± 21.71	57.68 ± 21.46
Sex (n, %)			
Male	576 (54.0%)	128 (48.5%)	205 (50.9%)
Female	490 (46.0%)	136 (51.5%)	198 (49.1%)
Lesion localization			
Head and neck	538	148	218
Scalp	73	19	24
Forehead	25	11	14
Temple	67	31	31
Periocular	45	13	18
Nose	96	21	31
Cheek	130	30	51
Lip & Chin	38	5	21
Others	64	18	28
Trunk	228	49	67
Extremities	212	53	108
Genitals and anus	102	21	29

Abbreviations: SD, standard deviation.

***Table 2:* The basic characteristics of patients and skin lesions.**

dermatologists with or without additional HFUS information (Table 3). The same observation was made in sensitivity and specificity. The AUC, sensitivity, and specificity of dermatologists with HFUS information for management decisions were 0.919 (95% CI, 0.889–0.944), 0.941 (95% CI, 0.903–0.973), and 0.915 (95% CI, 0.880–0.947), respectively. In comparison, with only one close-up image at hand, general dermatologists had a lower AUC (0.881; 95% CI, 0.846–0.913; P = 0.00089) and sensitivity (0.834; 95% CI, 0.788–0.890; $P = 0.00018$) but a higher specificity (0.953; 95% CI, 0.930–0.974; P = 0.049).

Diagnostic performance of CNN models vs. human raters

In the test cohort, for the binary classification task, we found that the DMFN model had a significantly higher AUC than the general practitioner (0.876; 95% CI,

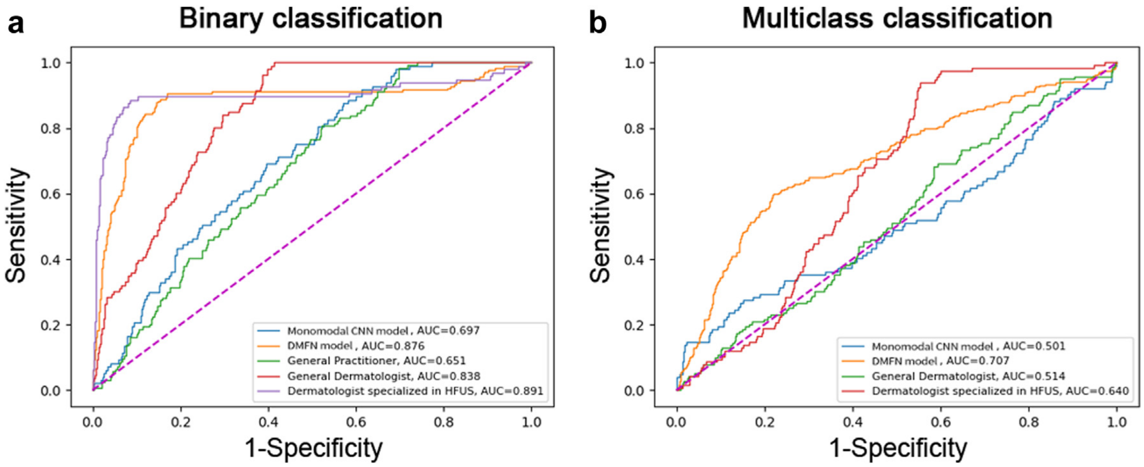

***Fig. 2:* ROC curves of monomodal CNN model, DMFN model, and human raters in the binary and multiclass classifications**. (a) ROC curves of monomodal CNN model, DMFN model, and human raters in the binary classification task. (b) ROC curves of monomodal CNN model, DMFN model, and human raters in the multiclass classification task. ROC, receiver operating characteristic; CNN, convolutional neural network; DMFN, deep multimodal fusion network; HFUS, high-frequency ultrasound; AUC, area under the ROC curve.

Ratings	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	PPV (95% CI)	NPV (95% CI)
Binary classification						
CNN model						
Monomodal CNN model	0.697 (0.648–0.750)	0.567 (0.507–0.637)	0.953 (0.919–0.975)	0.782 (0.746–0.818)	0.906 (0.848–0.947)	0.734 (0.683–0.781)
DMFN model	0.876 (0.843–0.911)	0.872 (0.824–0.917)	0.906 (0.870–0.942)	0.891 (0.863–0.917)	0.881 (0.834–0.923)	0.899 (0.860–0.935)
Human raters						
General practitioner	0.651 (0.550–0.750)	0.665 (0.564–0.764)	0.702 (0.616–0.791)	0.686 (0.591–0.775)	0.639 (0.539–0.737)	0.726 (0.601–0.816)
General dermatologist	0.838 (0.798–0.882)	0.775 (0.717–0.830)	0.940 (0.903–0.966)	0.867 (0.835–0.902)	0.912 (0.863–0.949)	0.840 (0.805–0.880)
Dermatologist specialized in HFUS	0.891 (0.857–0.921)	0.898 (0.857–0.937)	0.906 (0.868–0.945)	0.903 (0.874–0.931)	0.884 (0.836–0.929)	0.918 (0.886–0.954)
Management decision						
General practitioner	0.694 (0.634–0.743)	0.664 (0.598–0.729)	0.702 (0.622–0.782)	0.686 (0.645–0.826)	0.697 (0.651–0.748)	0.755 (0.706–0.797)
General dermatologist	0.881 (0.846–0.913)	0.834 (0.788–0.890)	0.953 (0.930–0.974)	0.900 (0.872–0.927)	0.934 (0.899–0.966)	0.878 (0.838–0.918)
Dermatologist specialized in HFUS	0.919 (0.889–0.944)	0.941 (0.903–0.973)	0.915 (0.880–0.947)	0.927 (0.899–0.948)	0.898 (0.861–0.936)	0.951 (0.921–0.978)

Abbreviations: CNN, convolutional neural network; DMFN, deep multimodal fusion network; HFUS, high-frequency ultrasound; AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; CI, confidence interval.

Table 3: **Diagnosis performance of different human participants and CNN models in the test cohort.**

0.843–0.911 vs 0.651; 95% CI, 0.550–0.750; $P = 0.0025$) and the general dermatologists (0.838; 95% CI, 0.798–0.882; $P = 0.0038$), but significantly lower than the dermatologists specialized in HFUS (0.891; 95% CI, 0.857–0.921; $P = 0.0080$). Similar observations were made for differences in sensitivity and the rate of correct classifications. Dermatologists outperformed the monomodal CNN model even with just close-up images of their availability (Fig. 2, Table 3, and Table S4).

When predicting benign lesions for the binary classification, the monomodal CNN model and the DMFN model performed at the same level as dermatologists (95% vs. 94% with close-up images and 91% vs. 91% with close-up and HFUS images, respectively). As for predicting malignant diseases, the DMFN model outperformed the dermatologists (87% vs. 78%) and general practitioners (87% vs. 66%) using only close-up images, but not dermatologists with additional HFUS information (87% vs. 90%) (Fig. 3). More specifically, in the multiclass classification, the DMFN model performed better on common malignant classes (such as BD, BCC, SCC, and DFSP) than the dermatologists evaluating close-up images only, but the diagnostic performance of the DMFN model in AK, BCC, and EMPD was not at the level of dermatologists in the same condition. Additionally, the DMFN model could correctly identify some malignant diseases (such as EMPD and MM) that dermatologists mistakenly label as benign. Except for benign sebaceous neoplasms, BKL, and lipoma, the DMFN model and dermatologists performed better or comparably to most benign lesions (Figs. 4 and 5).

Overall, for the classification of 17 types of skin diseases in the test cohort, the DMFN model achieved better performance than dermatologists specialized in HFUS (AUC, 0.707; 95% CI, 0.638–0.776 vs. AUC, 0.640; 95% CI, 0.608–0.675; $P = 0.0083$) (Fig. 2 and Table S4). However, the DMFN model's overall accuracy of 55% was close to that of dermatologists specialized in HFUS (59%, $P = 0.66$) (Figure S3). Specifically, the DMFN model attained an almost identical or better performance than dermatologists with close-up images at hand in AK, EMPD, nevi, BCC, BD, SCC, DFSP, warts, cysts, haemangioma, and inflammation. Compared to dermatologists with additional HFUS information, the DMFN model obtained satisfactory diagnostic performance for BD, DFSP, cysts, SCC, MM, warts, and nevi (Fig. 5).

The representative heat maps of CNN's analysis of some skin diseases are presented in Fig. 6.

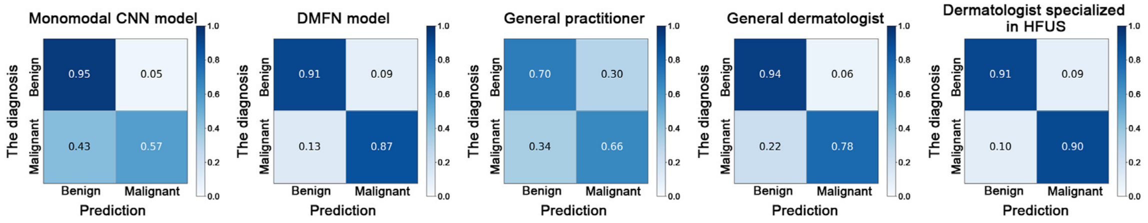

Fig. 3: **Confusion matrices of binary classification in the test cohort**. CNN, convolutional neural network; DMFN, deep multimodal fusion network; HFUS, high-frequency ultrasound.

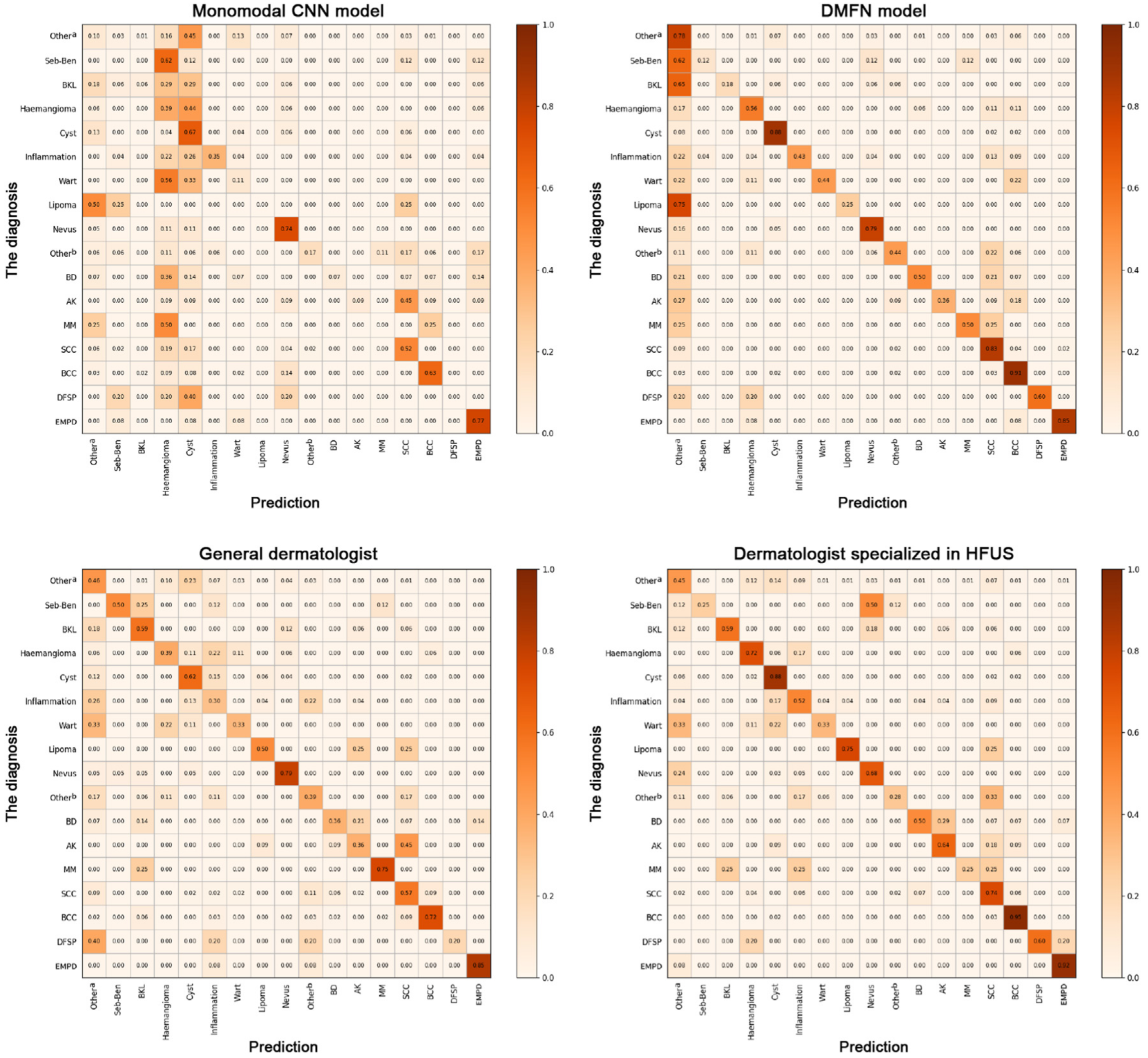

Fig. 4: **Confusion matrices of multiclass classification in the test cohort**. CNN, convolutional neural network; DMFN, deep multimodal fusion network; HFUS, high-frequency ultrasound; Seb-Ben, benign sebaceous neoplasms; BKL, benign keratosis-like lesions; BD, Bowen's disease; AK, actinic keratosis; MM, malignant melanoma; SCC, squamous cell carcinoma; BCC, basal cell carcinoma; DFSP, dermatofibrosarcoma protuberans; EMPD, extramammary Paget's disease. The list of diagnoses of other^a and other^b is available in the Supplementary Table S2.

Discussion

In the present study, for the first time, we developed a DMFN model based on both clinical close-up and HFUS images for evaluating a broad spectrum of skin lesions and compared its performance with human raters. Our results demonstrated that combining HFUS and close-up imaging modalities performed better than either modality alone. While close-ups depicted the surface aspects of the lesion, HFUS could supplement the internal characteristics of the lesion under the surface, revealing information about the tumor in both longitudinal and transverse planes. Therefore, for both dermatologists and CNN solutions, we suggested combining HFUS images with clinical examination for optimal results.

Since the 2017 landmark article that Esteva et al.[20] firstly introduced CNN to dermatology and showed expert-level performance, most subsequent studies focused on the classification of limited categories of some preselected skin diseases. For instance, Tschandl et al. reported the application of the CNN model in classifying pigmented and non-pigmented skin lesions, respectively.[18,34] A few studies investigated the performance of the dermoscopy-based CNN model in the diagnosis of melanomas.[21,22,35–40] In addition, several studies indicated that the CNN model could perform to a standard comparable to dermatologists in diagnosing inflammatory skin diseases and specific areas of skin, such as acral melanoma and onychomycosis,

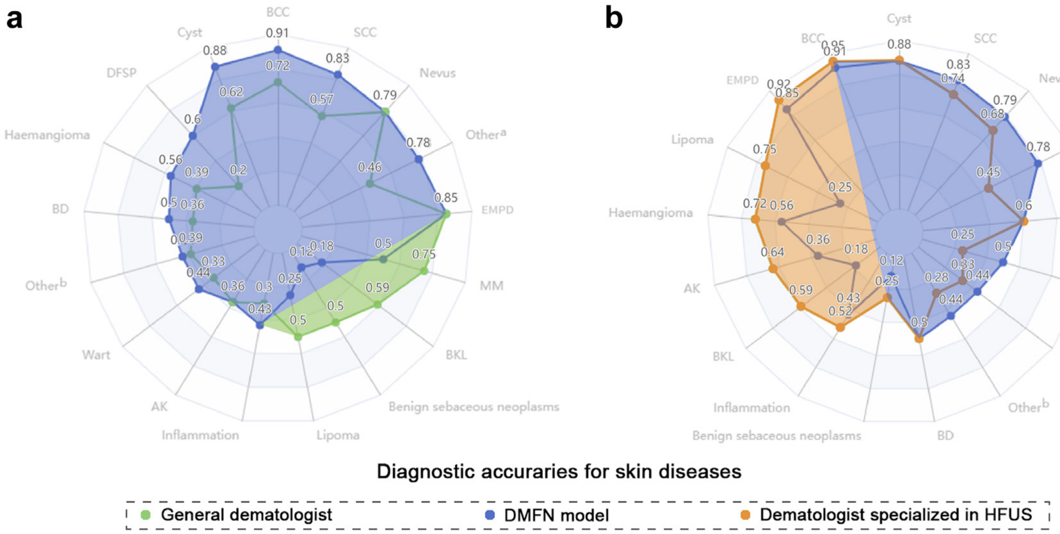

Fig. 5: **Radar charts comparing the performance of dermatologists and the DMFN model for each skin disease**. (a) A radar chart comparing the performance of general dermatologists and the DMFN model for each skin disease. (b) A radar chart comparing the performance of dermatologists specialized in HFUS and the DMFN model for each skin disease. In panel a, the green area indicated that the diagnostic accuracy of the human rater (dermatologist) was higher than that of the DMFN model, including MM, BKL, benign sebaceous neoplasms, and lipoma. On the contrary, in panel b, the orange area indicated that the diagnostic accuracy of the human rater (dermatologist + HFUS) was higher than that of the DMFN model, including BCC, EMPD, lipoma, haemangioma, AK, BKL, inflammation, and benign sebaceous neoplasms. The blue area in both panels indicated that the DMFN model's diagnostic accuracy was higher or comparable to that of two human raters for rest diseases. CNN, convolutional neural network; DMFN, deep multimodal fusion network; HFUS, high-frequency ultrasound; BKL, benign keratosis-like lesions; BD, Bowen's disease; AK, actinic keratosis; MM, malignant melanoma; SCC, squamous cell carcinoma; BCC, basal cell carcinoma; DFSP, dermatofibrosarcoma protuberans; EMPD, extramammary Paget's disease. The list of diagnoses of other[a] and other[b] is available in the Supplementary Table S2.

respectively.[23,41–43] However, using a limited set of diseases for training and testing may result in overestimating the performance of the CNN model for those specific diseases and lead to a lack of ability to generalize to other diseases. In the field of dermatology, healthcare professionals frequently come across patients presenting with skin conditions that can manifest in various possible diagnoses. As such, utilizing CNNs based on limited categories of preselected skin diseases could not be a feasible approach for clinical practice. Differing from previous studies, to be close to the real clinical situation, our multicenter study did not exclude any specific lesions. Based on this concept, the primary task of our study was a multi-classification problem, not just a simple benign-malignant dichotomy.

For the multiclass classification task, overall, the DMFN model, which analyzed both HFUS and clinical images, could classify lesions almost as accurately as expert raters. Specifically, the DMFN model could achieve comparable or better performance than dermatologists analyzing clinical and HFUS images in cysts, BD, DFSP, warts, nevus, SCC, and MM. However, it did not reach the accuracy of human raters in benign sebaceous neoplasms, BKL, and non-pigmented lesions with no obvious appearance changes such as lipoma. One possible reason for this was that these diseases were seldom biopsied, resulting in their infrequent occurrence in the training set. Additionally, important to mention to melanoma, that both the DMFN model and dermatologists tended to have lower diagnostic accuracy after reference to HFUS images. This is likely because some melanoma subtypes, such as lentigo maligna and acral lentiginous melanoma, were too thin to be effectively diagnosed by HFUS. Nevertheless, the DMFN model still had some advantages in the multi-classification diagnosis of melanoma. The DMFN model could reduce the probability of misdiagnosing melanoma as benign in comparison with dermatologists specialized in HFUS.

Since skin ultrasound is not popularized among dermatologists, our study suggested that the DMFN model could be a suitable tool to improve the diagnostic performance of dermatologists who have not mastered HFUS. Moreover, the DMFN models for multi-classification diagnosis presented a close or higher accuracy than dermatologists for most diseases (especially higher than dermatologists who only diagnose based on appearance). On the other hand, the multi-classification diagnostic results could also help dermatologists in the dichotomous diagnosis of benign and malignant.

With regard to binary classification tasks, the CNN models had comparable or better performance than

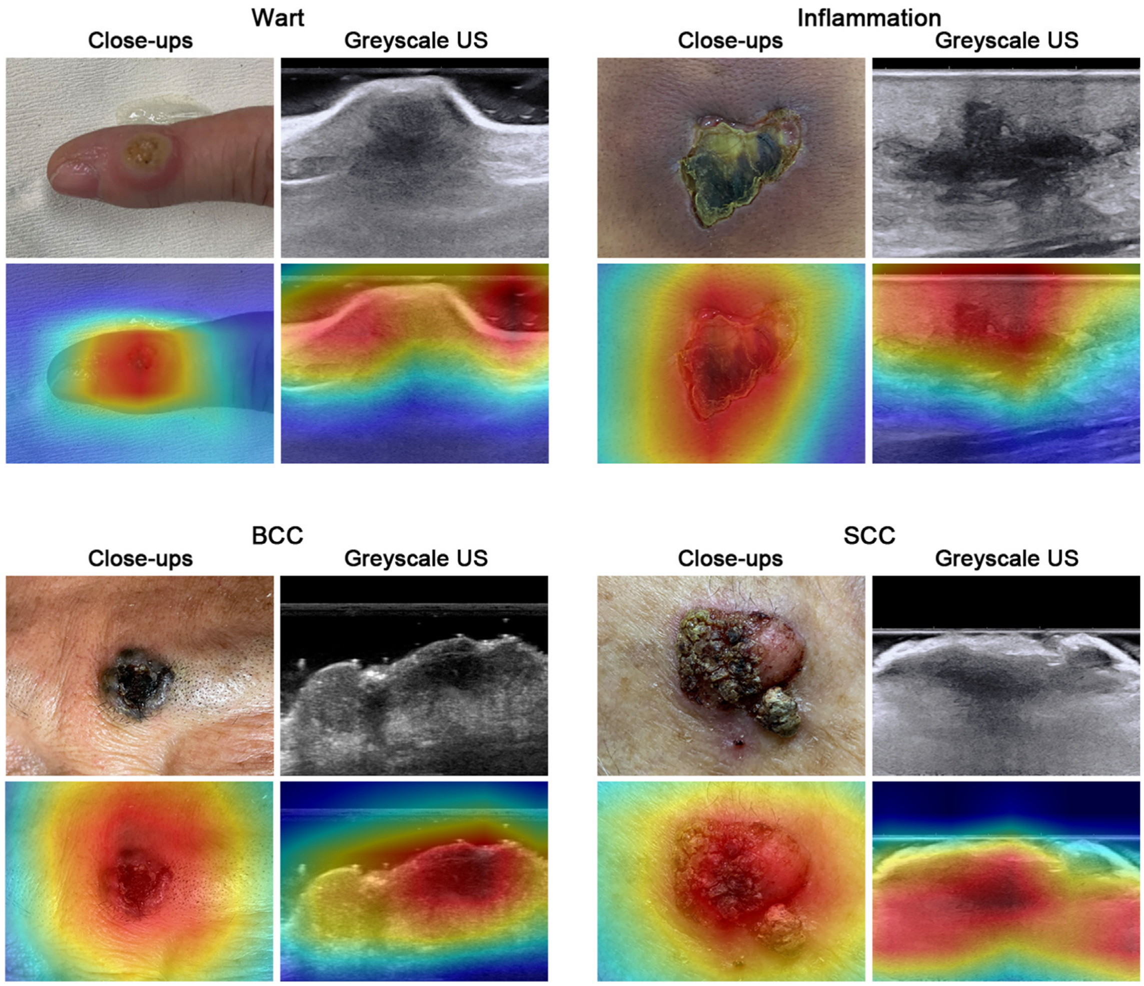

***Fig. 6:* The representative heat maps of the DMFN model's analysis using greyscale HFUS images in various skin diseases**. The varying color distributions reflected the most predictive regions that the DMFN model concentrates on distinct lesions. The areas highlighted in red on the heat map represented the informative features that played a crucial role during the prediction process. DMFN, deep multimodal fusion network; HFUS, high-frequency ultrasound; BCC, basal cell carcinoma; SCC, squamous cell carcinoma.

dermatologists. In detail, the DMFN model outperformed general practitioners for benign lesions and surpassed general dermatologists for malignant cases with only close-up images at hand. These findings indicated the potential of CNN models for further assisting the decision-making processes of dermatologists.

Given that the advantages of CNN-based classifiers do not rely on providing management decisions,[44] we only analyzed the management decisions for human participants. In clinical practice, dermatologists usually make management decisions rather than definitive classifications, whether it is simple dichotomous classifications (benign or malignant) or making specific diagnoses. Our data suggested that the diagnostic performance of the dermatologists' management decisions was improved over their definitive classifications in the same situation. The differences in sensitivity could translate into a closer clinical setting resulting in fewer missed malignant lesions for dermatologists. Yet, a prospective clinical study is still needed to evaluate whether dermatologists benefit from incorporating CNN classification into the decision-making process in a clinical real-life situation.

It is important for primary care providers to assess the necessity of referring a patient accurately and quickly to a dermatologist. In this regard, our present results revealed that both the monomodal CNN model and the DMFN model outperformed general practitioners in diagnostic performance. Given this reality, our CNN-based system may support primary care providers by providing valuable information regarding referral decisions and improved triaging.

There are also some limitations in our study. First, we did not provide CNNs with data other than HFUS and clinical close-up information for training, such as anatomical site, age, gender, skin tone, and lesion size or color variation. It is possible that the performance of our CNN-based system could be improved by adding above mentioned data. Second, only pathologically confirmed cases were enrolled. It may bring bias from the overrepresentation of malignant cases. Nevertheless, we believe that the advantages of an accurate diagnosis verified by pathology outweigh the disadvantages of verification bias. Third, in our study, we did not involve multiple levels of dermatologists, which may result in a generalized diagnosis. Thus, future studies could involve more individuals with diverse backgrounds to compare their diagnosis performance with our CNN. Fourth, it was better to evaluate our DMFN model on public skin datasets. Unfortunately, we did not find any database containing both clinical and HFUS images. Consequently, we had to evaluate the model's performance using our own dataset, which was limited to individuals of Chinese race and skin tone. This bias could affect CNN's generalization ability. Fifth, our exploratory study investigated multiple classification models (binary and multiclass DMFN) without multiplicity adjustment. Future studies should consider the multiplicity adjustment when applying multiple models simultaneously. Additionally, we used a 1:1 weight ratio for the two modalities (clinical close-ups and HFUS) in building our model because their actual weight was not known, given their varying values across different skin conditions. However, we believe the proportion of weights between the two modalities may float for different skin diseases. Further exploration is needed to determine the optimal weight ratio for each skin disease. Finally, we believe that using "human plus AI" is the most feasible form in clinical practice. Therefore, the impact of CNNs in assisting dermatologists in daily clinical practice could be evaluated in future in-depth studies. As a future work, the performance of the proposed method can be compared with the performance of a capsule neural network-based method since capsule networks have the ability to preserve spatial relationships of learned features, and therefore have been used recently for image classification tasks.[45–47]

In conclusion, we conducted a prospective study comparing dermatologists with CNN models that analyzed clinical close-up and HFUS images across a range of skin lesions. Our findings showed that the DMFN model, combining analysis of clinical close-ups (external information) and HFUS (internal information) images, performed better with clinical close-ups and HFUS images than with clinical close-ups alone. Additionally, the DMFN model can provide accurate binary classification and satisfactory multiclassification diagnoses for some diseases. Thus, our DMFN model may be a feasible and potentially attractive method to effectively inform primary care provider referral decisions.

Contributors

Conceptualization: HX.X., LH.G., XX.Z.

Data curation: AQ.Z., TT.R., J.W., YQ.Z.

Formal analysis: AQ.Z., X.C., JL.H., LC.M.

Funding acquisition: HX.X., LH.G., Q.W., YJ.Z.

Investigation: AQ.Z., Q.W., WW.R., X.C., YL.S., TT.R., J.W., YQ.Z., YK.S., XW.C., YX.L., N.N., YC.C.

Methodology: AQ.Z., Q.W., WW.R., X.C., YK.S., XW.C., YX.L., N.N., YC.C.

Accessing and verifying the underlying data: AQ.Z., LH.G., Q.W., X.C., YL.S., JL.H., LC.M.

Project administration: HX.X., LH.G., YL.S.

Resources: YL.S., YJ.Z., YX.L., YQ.L., LP.S.

Software: YL.S., X.C., JL.H., LC.M.

Supervision: YL.S., YJ.Z., YX.L., YQ.L.

Validation: WW.R., TT.R., J.W., YL.S., LC.M.

Visualization: AQ.Z., Q.W., X.C., YL.S., JL.H.

Writing-original draft: AQ.Z., X.C.

Writing-review & editing: HX.X., LH.G., XX.Z.

Data sharing statement

The data concerning patients is not publicly available due to privacy requirements but can be obtained from the corresponding author upon reasonable request approved by the institutional review board. The code for the model development is available online (https://github.com/cao1124/SkinLesionClassification).

Declaration of interests

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (81927801, 82151318, 82072092, 82001816) and the Clinical research project of Shanghai Skin Disease Hospital (lcfy2022-09).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2023.102391.

References

1 Stern RS. Prevalence of a history of skin cancer in 2007 results of an incidence-based model. *Arch Dermatol*. 2010;146(3):279–282.

2 Han SS, Moon IJ, Kim SH, et al. Assessment of deep neural networks for the diagnosis of benign and malignant skin neoplasms in comparison with dermatologists: a retrospective validation study. *PLoS Med*. 2020;17(11):e1003381.

3 MacFarlane D, Shah K, Wysong A, Wortsman X, Humphreys TR. The role of imaging in the management of patients with non-melanoma skin cancer diagnostic modalities and applications. *J Am Acad Dermatol*. 2017;76(4):579–588.

4 Malvehy J, Pellacani G. Dermoscopy, confocal microscopy and other non-invasive tools for the diagnosis of non-melanoma skin cancers and other skin conditions. *Acta Derm Venereol*. 2017;97:22–30.

5 Wortsman X. Common applications of dermatologic sonography. *J Ultrasound Med*. 2012;31(1):97–111.

6 Alfageme F, Wortsman X, Catalano O, et al. European federation of societies for ultrasound in medicine and biology (EFSUMB) position statement on dermatologic ultrasound. *Ultraschall der Med*. 2021;42(1):39–47.

7 Barcaui ED, Carvalho ACP, Lopes FPPL, Pineiro-Maceira J, Barcaui CB. High frequency ultrasound with color Doppler in dermatology. *An Bras Dermatol*. 2016;91(3):262–273.

8 Czajkowska J, Badura P, Korzekwa S, Platkowska-Szczerek A. Deep learning approach to skin layers segmentation in inflammatory dermatoses. *Ultrasonics*. 2021;114:106412.

9 Levy J, Barrett DL, Harris N, Jeong JJ, Yang XF, Chen SC. High-frequency ultrasound in clinical dermatology: a review. *Ultrasound J*. 2021;13(1):24.

10 Catalano O, Roldan FA, Varelli C, Bard R, Corvino A, Wortsman X. Skin cancer: findings and role of high-resolution ultrasound. *J Ultrasound*. 2019;22(4):423–431.

11 Kleinerman R, Whang TB, Bard RL, Marmur ES. Ultrasound in dermatology: principles and applications. *J Am Acad Dermatol*. 2012;67(3):478–487.

12 Wortsman X, Wortsman J. Clinical usefulness of variable-frequency ultrasound in localized lesions of the skin. *J Am Acad Dermatol*. 2010;62(2):247–256.

13 Dinnes J, Bamber J, Chuchu N, et al. High-frequency ultrasound for diagnosing skin cancer in adults. *Cochrane Database Syst Rev*. 2018;12:CD013188.

14 Göçeri E. Convolutional neural network based desktop applications to classify dermatological diseases. In: *2020 IEEE 4th international conference on image processing, applications and systems (IPAS)*. IEEE; 2020:138–143.

15 Göçeri E, Karakas AA. Comparative evaluations of CNN based networks for skin lesion classification. In: *The 14th international conference on computer graphics. Visualization, computer vision and image processing (CGVCVIP)*. 2020:1–6.

16 Göçeri E. Impact of deep learning and smartphone technologies in dermatology: automated diagnosis. In: *2020 Tenth international Conference on image processing theory, Tools and applications (IPTA)*. IEEE; 2020:1–6.

17 Goceri E. Automated skin cancer detection: where we are and the way to the future. In: *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE; 2021:48–51.

18 Tschandl P, Rosendahl C, Akay BN, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatol*. 2019;155(1):58–65.

19 MacLellan AN, Price EL, Publicover-Brouwer P, et al. The use of noninvasive imaging techniques in the diagnosis of melanoma: a prospective diagnostic accuracy study. *J Am Acad Dermatol*. 2021;85(2):353–359.

20 Esteva A, Kuprel B, Novoa RA. Dermatologist-level classification of skin cancer with deep neural networks. *Oncologie*. 2017;19(11-12):407–408.

21 Winkler JK, Fink C, Toberer F, et al. Association between durgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol*. 2019;155(10):1135–1141.

22 Marchetti MA, Codella NCF, Dusza SW, et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *J Am Acad Dermatol*. 2018;78(2):270–277.e1. https://doi.org/10.1016/j.jaad.2017.08.016.

23 Yu C, Yang S, Kim W, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. *PLoS One*. 2018;13(3):e0193321.

24 Daneshjou R, Barata C, Betz-Stablein B, et al. Checklist for evaluation of image-based artificial intelligence reports in dermatology: CLEAR derm consensus guidelines from the international skin imaging collaboration artificial intelligence working group. *JAMA Dermatol*. 2022;158(1):90–96.

25 World Health Organization. International statistical classification of diseases and related health problems (ICD). Available at: https://www.who.int/standards/classifications/classification-of-diseases. Accessed July 11, 2021.

26 Goceri E. Evaluation of denoising techniques to remove speckle and Gaussian noise from dermoscopy images. *Comput Biol Med*. 2023;152:106474.

27 Göçeri E. Intensity normalization in brain MR images using spatially varying distribution matching. In: *The 11th international conference on computer graphics, visualization, computer vision and image processing (CGVCVIP)*. 2017:300–304.

28 Göçeri E. Fully automated and adaptive intensity normalization using statistical features for brain MR images. *Celal Bayar Üniv Fen Bilim Derg*. 2018;14(1):125–134.

29 Göçeri E. Image augmentation for deep learning based lesion classification from skin images. In: *2020 IEEE 4th international conference on image processing, applications and systems (IPAS)*. IEEE; 2020:144–148.

30 Göçeri E. Medical image data augmentation: techniques, comparisons and interpretations. *Artif Intell Rev*. 2023;56(11):12561–12605.

31 Göçeri E. Comparison of the impacts of dermoscopy image augmentation methods on skin cancer classification and a new augmentation method with wavelet packets. *Int J Imag Syst Tech*. 2023;33(5):1727–1744.

32 Göçeri E. An application for automated diagnosis of facial dermatological diseases. *İzmir Katip Çelebi Üniversitesi Sağlık Bilimleri Fakültesi Dergisi*. 2021;6(3):91–99.

33 Delong ER, Delong DM, Clarkepearson DI. Comparing the areas under 2 or more correlated receiver operating characteristic curves - a nonparametric approach. *Biometrics*. 1988;44(3):837–845.

34 Tschandl P, Codella N, Akay BN, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol*. 2019;20(7):938–947.

35 Brinker TJ, Hekler A, Enk AH, et al. Deep neural networks are superior to dermatologists in melanoma image classification. *Eur J Cancer*. 2019;119:11–17.

36 Brinker TJ, Hekler A, Enk AH, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*. 2019;111:148–154.

37 Brinker TJ, Hekler A, Enk AH, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur J Cancer*. 2019;113:47–54.

38 Fink C, Blum A, Buhl T, et al. Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas. *J Eur Acad Dermatol Venereol*. 2020;34(6):1355–1361.

39 Winkler JK, Sies K, Fink C, et al. Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition. *Eur J Cancer*. 2021;145:146–154.

40 Winkler JK, Tschandl P, Toberer F, et al. Monitoring patients at risk for melanoma: may convolutional neural networks replace the strategy of sequential digital dermoscopy? *Eur J Cancer*. 2022;160:180–188.

41 Han SS, Park GH, Lim W, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: automatic construction of onychomycosis datasets by region-based convolutional deep neural network. *PLoS One*. 2018;13(1):e0191493.

42 Han B, Jhaveri RH, Wang H, Qiao D, Du J. Application of robust zero-watermarking scheme based on federated learning for securing the healthcare data. *IEEE J Biomed Health Inform*. 2023;27(2):804–813.

43 Schielein MC, Christl J, Sitaru S, et al. Outlier detection in dermatology: performance of different convolutional neural networks for binary classification of inflammatory skin diseases. *J Eur Acad Dermatol Venereol*. 2023;37(5):1071–1079.

44 Cook DA, Sherbino J, Durning SJ. Management reasoning: beyond the diagnosis. *JAMA*. 2018;319(22):2267–2268.

45 Göçeri E. Analysis of capsule neural networks for image classification. In: *The 15th international conference on computer graphics, visualization, computer vision and image processing (CVGCVIP)*. 2021:1–6.

46 Göçeri E. Capsule neural networks in classification of skin lesions. In: *The 15th international conference on computer graphics, visualization, computer vision and image processing (CVGCVIP)*. 2021:29–36.

47 Göçeri E. Classification of skin cancer using adjustable and fully convolutional capsule layers. *Biomed Signal Process Control*. 2023;85:104949.