

## RESEARCH ARTICLE

# Skin Medical Image Captioning Using Multi-Label Classification and Siamese Network

YIHLON LIN<sup>1</sup>, KUIYOU LAI<sup>1</sup>, AND WENYU CHANG<sup>2,3</sup><sup>1</sup>Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Douliu, Yunlin 64002, Taiwan<sup>2</sup>Department of Dermatology, E-Da Cancer Hospital, I-Shou University, Kaohsiung 82445, Taiwan<sup>3</sup>School of Medicine for International Students, College of Medicine, I-Shou University, Kaohsiung 82445, Taiwan

Corresponding author: Wenyu Chang (changwenyu@gmail.com)

This work was supported by the National Science and Technology Council, Taiwan, under Grant MOST 110-2221-E-224-029.

**ABSTRACT** Image captioning is a process of automatically generating descriptive sentences for a given image. Text-to-image search is a form of search in which images are retrieved based on matching keywords and image features. We focus on the case in which multiple description sentences are generated for one image. In this study, we used four learning models: 1) a discriminator, which is a binary classifier that distinguishes skin from background using image segmentation; 2) an autoencoder; 3) a multiclass classification model combining the features from the discriminator and autoencoder and producing keyword labels; and 4) a Siamese network learning the textual similarity matching between colloquial description sentences of skin imaging pathology and keywords produced from the multi-class classifier. The experimental results show that the proposed method yields an accuracy of up to 99% for the testing data in terms of colloquial language of skin images. This study enabled users to read the skin. For teaching research on skin diagnosis, the proposed method can significantly relieve the shortage of training personnel and assist hospitals that lack resources for conducting case studies. The results of this study are expected to be feasible and can be applied in actual clinical teaching. For medical education in dermatology, the findings of this study contribute to the practical value of quantitative indicators and assessments for learning outcomes of medical students.

**INDEX TERMS** Fully convolutional network, image caption, discriminator, autoencoder, multi-label classification, Siamese network.

## I. INTRODUCTION

Dermatological diagnosis of skin diseases mainly relies on visual examination of skin lesions, followed by history taking, which involves hearing and verbal interactions between patients and physicians. The diagnostic process in dermatology is quite different from that in other specialties, as patients usually see and observe their own skin lesions before seeking medical advice. When patients visit physicians, who will perform a direct visual inspection, examine and interpret the skin lesions of the patients before inquiring about their detailed medical history. In recent years, many advanced dermatological applications of artificial intelligence have been developed for various purposes, such as aiding skin tumor classification, or diagnosing a certain disease, using clinical

images, dermatoscopic images, or mobile phone images. However, the fundamentals of preliminary and differential diagnoses are made mainly through information gained from visual inspection, which is contrary to other fields of clinical medicine. As a result, correct identification, analysis, and interpretation of skin lesions are essential, but also subjective, in performing dermatological diagnosis. This learning process requires repetition of long-term training and experiences during the education of medical students or trainees to achieve better accuracy. If visual inspection could be assisted by objective computer learning systems, this may help in simplifying the visual examination process, improving diagnostic accuracy during the visual and physical examination by trainees, and providing objective evidence on follow-up examination. Physical examination during dermatology training involves learning to read the patient, including vital signs and skin lesions, and inquiring about their medical history.

The associate editor coordinating the review of this manuscript and approving it for publication was Joewono Widjaja<sup>1</sup>.

However, this process is usually not systematic. Therefore, in this study, we raised the idea of the Platform for Learning to Read as a service mechanism for virtual teachers to help train students in the School of Medicine and trainees to improve their learning outcomes.

According to a popular textbook in dermatology on learning how to interpret skin lesions [1], skin reading is similar to reading text. On cutaneous examination, the skin lesion of each individual diagnosis has its own features. A complete description of these skin lesions is just like compositing a sentence or paragraph with words. A trained physician first recognizes the skin lesion type, identifies their shapes, colors, edges, and other features, and then further characterizes its arrangement and distribution. This process will further generate a complete paragraph. A fully described specific skin lesion by text may directly lead to the most possible diagnosis. Therefore, the premise of dermatological diagnosis is mandatory to understand the type of skin lesion, shape, color, arrangement, margination, distribution of lesions, consistency, etc. [1]. While learning to read skin diseases is a typical application of image captioning, we developed an artificial intelligence algorithm to simulate the special interpretation process in this study.

Image captioning is usually based on feature extraction and language model. Words in textbooks are representations of keywords of symptom features. Combination of several keywords may future represent the feature of particular signs.

It can be challenging to automatically evaluate the quality of image captions. Zhao et al. [2] proposed a multi-modal fusion architecture to generate descriptions of content images. Their model includes four sub-networks: (1) a convolutional neural network (CNN) for extracting image features, (2) an ATTssd model for extracting image attributes, (3) a language CNN model for modeling sentences, and (4) a recurrent neural network for word prediction. Based on the advances in image captioning research, image captioning approaches were classified in [3] into different categories. An attribute-assisted teacher-critical training strategy was introduced in [4] to facilitate the learning process of the captioning model. Sentence-level information has often been neglected. The learning of contrastive semantic similarity was considered in [5] for image captioning evaluation. To capture sentence-level representation, there are three progressive model structures: single-branch, dual-branch, and triple-branch models. In general, attention-based approaches often concentrate on individual visual features. However, Wang and Gu [6] focused on the relationship between image features. This provides important guidance for the generation of captions.

An increasing number of medical images have been adopted in the medical diagnosis process. An innovative learning-based framework for image captions was proposed in [7]. Their proposed method is efficient for automatically generating skin-imaging reports. In medical diagnostic systems that use deep learning techniques, boosting the performance of the learning model is usually achieved at the cost of diminishing explainability. Barata et al. [8] aimed to

address this major limitation and explainability of a skin cancer diagnostic system. The goal of this method is achieved at the cost of diminishing the explainability of medical diagnostic systems. This is because of syntactic complexity and long sentences. Using the Siamese neural network in that study, satisfactory results for biomedical text similarity evaluation could not be obtained. Li et al. [9] designed a self2self-attention model to solve syntactic complexity and long-sentence problems. In [10], the semantic representation of each sentence was used to generate its embedding. Such embeddings are then used for the external evaluation of the semantic parser.

An increasing number of medical images, including reading and writing reports, can be a big burden on physicians. To reduce the workload of physicians, the authors of [11], [12] proposed novel models to generate draft reports from related images. Considering the differences between patient and normal images, an X-ray image-captioning model was proposed in [11]. The proposed model includes a decoder to produce reports, which is realized using a transformer or long short-term memory (LSTM). As pointed out in [12], the occurrence of some medical terms is often observed in many different reports.

To strengthen the consistency of medical terms in the final report, Wang et al. [12] proposed a model that unifies template retrieval and sentence generation. For human-centric and remote-sensing image captioning, Yang et al. [13] proposed a novel Human-Centric Captioning Model (HCCM) to describe human behavior from a related image. Wang and Zhang [14] and Ye et al. [15] proposed a visual alignment attention model (VAA) and joint-training two-stage (JTTS) remote sensing images (RSI), respectively, to deal with remote sensing image captioning.

The remainder of this paper is organized as follows. In Section II, model selection is explained. In Section III, the material database and architecture of the proposed method are presented. We present our experimental setup and results in Section IV. The discussion and conclusions are presented in Section V.

## II. MODEL ARCHITECTURE

The applications of medical image captioning includes image and text encoding. In this section, we describe the architecture of the learning model used in this study. First, the discriminator and autoencoder are introduced. They were implemented using fully convolutional network (FCN) models. Next, a multi-class classification model was introduced. Finally, we describe a Siamese network.

### A. FCN MODEL OF AUTOENCODER WITH IMAGE SEGMENTATION

The network configurations of both the discriminator and autoencoder are shown in Fig. 1. The discriminator (indicated as (1) in Fig. 1) and autoencoder (indicated as (2) in Fig. 1) have similar network architectures, except for the output color channels. For the discriminator, the input was an original skin

image, and the output was a binary image within the skin region. The input and output of the autoencoder were the same. The loss function for training the autoencoder is the mean squared error (MSE). The purpose of the discriminator is to construct the features of the original image suitable for binary classification (i.e., skin or background), and that of the autoencoder is to construct the features of the original image suitable for reconstructing the original image. It should be noted that both the discriminator and autoencoder were implemented using the FCN models. The FCN model consists of feature encoding (i.e., encoder) and a decoding process (i.e., decoder), which mainly includes two types of blocks: a VGG Block and Inverse VGG Block. The former block is composed of multiple convolutional layers and one sub-sampling layer, whereas the latter is composed of multiple convolutional layers and one up-sampling layer. The details of each VGG block and the Inverse VGG block are shown in Fig. 1. The encoder (i.e., the feature extractor) of each model is the part from the input image to the output before the dropout layer, as shown in Fig. 1. After completing the training processes, the features produced by both encoders were concatenated to form the inputs of the multi-class classifier.

outputs were the probabilities of the input image belonging to each of the 21 keywords.

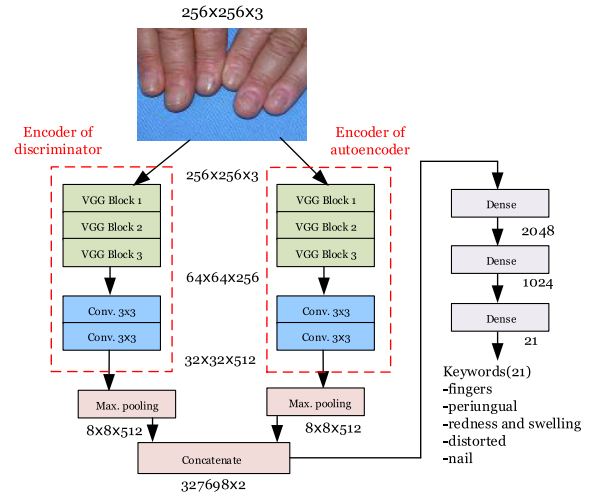


FIGURE 2. Flowchart of multi-class classifier.

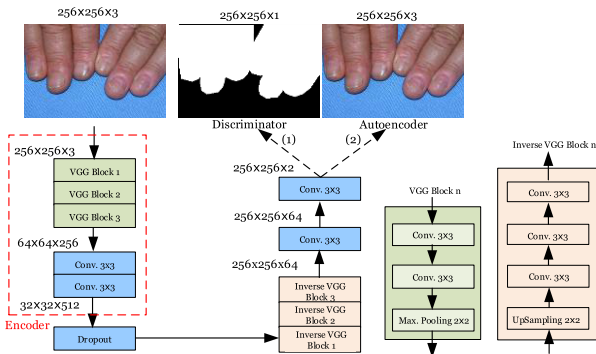


FIGURE 1. Architecture of discriminator and autoencoder.

### B. MULTI-CLASS CLASSIFICATION

The multiclass classification problem in this study is a multi-label multiclass classification problem. In this problem, each data instance may be associated with multiple labels, rather than just a single label. The predicted value of each output node represents the probability that the underlying input belongs to that class. The sum of the predicted values of all the output nodes need not be equal to one. The loss function for training the classifier is binary cross-entropy, and the activation functions of the output nodes are sigmoid functions.

The network configuration of the multi-class classifier is shown in Fig. 2. The input of the classifier was the input image, and the outputs were the associated keywords labeled by physicians. As shown in the figure, the features from the encoders of the discriminator and autoencoder were first extracted. After maximum pooling, these features were concatenated. Three dense layers were then stacked. The final

### C. TEXT FEATURE MATCHING USING SIAMESE NETWORK

Recently, few-shot learning becomes quite popular in the area of machine learning. Famous few-shot learning machines include Siamese networks, relation networks, prototypical networks, Gaussian prototypical network, matching networks. Specifically, Siamese networks can be used to judge whether two inputs belong to the same class. The Siamese network used in this study is shown in Fig. 3. The activation function of the output layer is the sigmoid function.

As mentioned above, the final outputs were the probabilities of the input image belonging to each of the 21 keywords. Usually, if the probability of belonging to a class is greater than 0.5, the input image is predicted to belong to that class, that is, the input image contains the corresponding keyword. According to our experimental results, the maximum number of keywords corresponding to an input image was less than 10. To speed up the training process, we set the input length of the “keywords” and “sentence description” in the Siamese network, as shown in Fig. 3, to 10.

Examples of the training data for the Siamese network are presented in Table 1. In the first column of the table, keywords of length 10 are treated as a whole. In the second column, each sentence was treated separately. If the keywords and the sentence description belong to the same class, it is labeled as “Y”, otherwise “N” in the third column. Therefore, the input of the Siamese network is a keyword-sentence pair. The final output of the Siamese network can identify the sentence descriptions from the pool of all sentences that are associated with the corresponding input image.

### III. PROPOSED METHODS

The methods used in this study include data pre-processing, data augmentation and labeling, training discriminator and

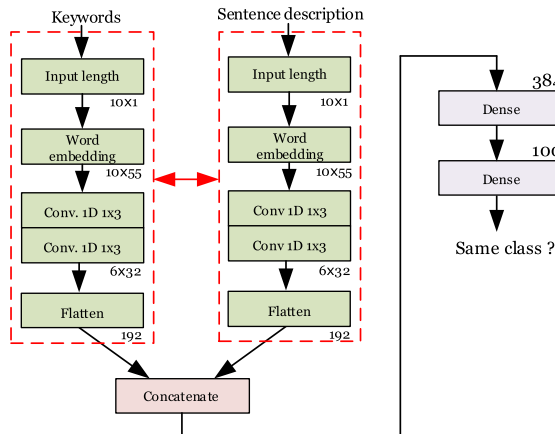


FIGURE 3. Flowchart of Siamese network.

TABLE 1. Examples of training data for siamese network.

Keywords	Sentence description	Same class
fingers	Finger	Y
periungual	feriungual redness and swelling	Y
redness and swelling	yellow pus appears under the cuticle	N
distorted	the nail plate is distorted	Y
... (10 items as a whole)	red scaly plaques with well-defined edges	N

autoencoder to extract features, and integration of the four model architectures.

A. MATERIAL DATABASE

In this study, training and testing datasets were collected from DermNet. previously known as DermNet NZ [16], which is an open and free dermatology resource. The database is supported by the New Zealand Dermatological Society and is contributed by numerous volunteer physicians globally, including dermatologists, health professionals, and students; therefore, the image database contains lesions of multiple ethnicities.

In this study, 62 important and easily identifiable images and corresponding sentence descriptions were selected for common skin imaging symptoms, such as herpes zoster, paronychia, and plaque psoriasis. The items and keywords were defined by Prof. Chang (a dermatologist at E-Da Hospital, Taiwan), as shown in Table 2. This table includes 21 keywords and four related items. Disease items had the highest number of keywords. For an image, a maximum of 10 keywords appeared. Images in the dataset included four different body places, four different severities, ten different skin diseases, and three colors. The number of occurrences for each keyword is listed in Table 3.

Table 4 shows that an image is associated with five keywords and five sentence descriptions. It should be noted that the image and sentence descriptions were obtained using DermNet. The purpose of this study is to add keywords

associated with the input image. Thus, the overall task of image captioning can be decomposed into a three-stage design. First, the features from the encoders of the discriminator and autoencoder are extracted after completing the training processes. Next, the multiclass classifier establishes the relationship between the input image and the keywords. Finally, the Siamese network builds the relationship between keywords from the classifier and sentence descriptions.

TABLE 2. The definition of keywords.

Items	Place	Severity	Disease	Color
Number	4	4	10	3
Keywords	-fingers -toes -periungual -nail	-redness and swelling -abscess -pus	-distorted -transverse ridges -skin cracks -grouped vesicles -erythematous papules -crusting -pustular -well-defined edges -scaly plaques -scale	-yellow -red -silvery white


TABLE 3. The number of occurrence of each keyword.

Disease Name	Keywords	Numbers
Herpes zoster	grouped vesicles	18
	erythematous papules	18
	crusting	15
	Pustular	5
	Paronychia	fingers
Paronychia	toes	2
	Periungual	17
	redness and swelling	14
	swelling	3
	distorted	10
	Nail	10
	transverse ridges	8
	yellow	10
	Pus	10
	Abscess	7
Plaque psoriasis	Red	22
	scaly plaques	22
	well-defined edges	22
	silvery white	16
	Scale	16
	skin cracks	1

B. DATA AUGMENTATION AND LABELING

We collected 62 images and 304 labeled description sentences. Each image contained more than one description sentence. To train the multi-label classification model, we augmented the original image by adjusting the flip, brightness, and contrast to obtain 7936 images and let the ratio of training datasets to test datasets be 8:2. For the Siamese network model, there were 304 data points. As shown in Table 5, the ratio of the training to testing datasets was 8:2. To evaluate

**TABLE 4. An example of image, keywords, and sentence description.**

Image	Keywords	Sentence description	Label
	-Fingers	fingers	1
	-periungual	periungual redness and	1
	-redness	swelling	
	and	yellow pus appears	0
	swelling	under the cuticle	
	-distorted	the nail plate is	1
-nail	distorted		
		red scaly plaques with	0
		well-defined edges	

the model effectively, we performed an overall test evaluation using only the test images.

**TABLE 5. Description of training and testing datasets.**

Multi-label classification model	Number
# of training dataset (80%)	7136
# of testing dataset(20%)	800
Total	7,936
Siamese network	Number
# of training dataset (80%)	243
# of testing dataset(20%)	61
Total	304

**TABLE 6. Description of datasets of various models.**

Stage	Stage 1	Stage 2	Stage 3
Model	Discriminator, autoencoder	Multi-label classifier	Siamese network
Training Dataset	7,136	7,136	243
Testing Dataset	800	800	61
Total	7,936	7,936	304(49/255)*

\*: Number of positive sample/number of negative sample

**TABLE 7. The definition of training parameters.**

Parameter	Number
Image Size	512×512
Max. length of keywords	21
Number of total keywords for colloquial sentence	43
Max. length of colloquial sentence	10
Number of keywords on an image	10
Dictionary size (including keywords and sentences)	51

**C. TRAINING SAMPLES**

What are the training datasets used? For the training dataset of the autoencoder, this is easy. The inputs and outputs were the same as those of the original training datasets. The input of the discriminator is the training input image, and the output is the corresponding image with binary values indicating the skin or background. As mentioned, the purpose of the discriminator is to construct the features of the original image suitable for binary classification (i.e., skin or background), and that of the autoencoder is to construct the features of the

original image suitable for reconstructing the original image. The main reasoning behind concatenating the features from discriminator and autoencoder is to retain the main features of the original input image while focusing on the skin region.

**TABLE 8. Performance of discriminator for skin image segmentation.**

	Accuracy	Skin Segmentation			
		Jl	DSC	Sensitivity	Specificity
Count	800	800	800	800	800
Mean	<b>0.9137</b>	<b>0.8710</b>	<b>0.9205</b>	<b>0.9467</b>	<b>0.6662</b>
Std	0.0478	0.2134	0.0540	0.0453	0.2170
Min	0.7957	0.0000	0.7334	0.7909	0.1714
Q1	0.8771	0.8419	0.8788	0.9161	0.5338
Q2	0.8998	0.9609	0.9152	0.9678	0.6381
Q3	0.9737	1.0000	0.9843	0.9810	0.8647

**TABLE 9. Performance of the skin autoencoder.**

	MSE	MAE
Count	800	800
Mean	<b>0.0429</b>	<b>0.1663</b>
Std	0.0221	0.0517
Min	0.0041	0.0518
Q1	0.0243	0.1250
Q2	0.0429	0.1662
Q3	0.0616	0.2089

**TABLE 10. Performance of the skin autoencoder.**

	Binary Accuracy	Accuracy
Count	800	800
Mean	<b>0.9951</b>	<b>0.9638</b>
Std	0.0347	0.1869
Min	0.5714	0.0000
Q1	1.0000	1.0000
Q2	1.0000	1.0000
Q3	1.0000	1.0000
Max	1.0000	1.0000

**TABLE 11. Testing dataset performance of proposed model.**

Model	multi-label classification	Siamese network
Misclassification error for testing	*29/800 (0.036)	23/800(0.028)





\*: Number of misclassification/number of testing datasets

**D. PROPOSED MODELS**

In this study, the skin disease description was divided into three stages: (1) In the first stage, the FCN model was used to establish skin segmentation (discriminator) and autoencoder model features. The purpose was to obtain a representative encoded feature representation of the input image that focused on the features of the skin region. (2) In the second stage, the two features from the discriminator and autoencoder are combined and fed into the multi-label keyword classification model to produce the keyword label



TABLE 12. The performances of proposed models.

Testing Images				
<b>Keywords (ground truth)</b>	<ul style="list-style-type: none"> <li>• silvery</li> <li>• white</li> <li>• scale</li> <li>• skin</li> </ul>	<ul style="list-style-type: none"> <li>• silvery</li> <li>• white</li> <li>• scale</li> <li>• skin</li> </ul>	<ul style="list-style-type: none"> <li>• fingers</li> <li>• periungual</li> <li>• redness and swelling</li> <li>• nail</li> <li>• distorted</li> <li>• transverse</li> <li>• ridges</li> <li>• yellow</li> <li>• pus</li> </ul>	<ul style="list-style-type: none"> <li>• abscess</li> <li>• red</li> <li>• scaly</li> <li>• plaques</li> <li>• well-defined</li> </ul>
<b>Predicted outcome of multi-label classifier</b>	<ul style="list-style-type: none"> <li>• silvery</li> <li>• white</li> <li>• scale</li> </ul>	<ul style="list-style-type: none"> <li>• abscess</li> <li>• red</li> <li>• scaly</li> <li>• plaques</li> <li>• well-defined</li> </ul>	<ul style="list-style-type: none"> <li>• fingers</li> <li>• periungual</li> <li>• redness and swelling</li> <li>• transverse</li> <li>• ridges</li> <li>• yellow</li> <li>• pus</li> </ul>	<ul style="list-style-type: none"> <li>• abscess</li> <li>• red</li> <li>• scaly</li> <li>• plaques</li> <li>• well-defined</li> <li>• edges</li> </ul>
<b>Status of the predicted outcome of multi-label classifier</b>	Partially correct	Partially correct	Partially correct	Partially correct
<b>Colloquial sentence description (ground truth)</b>	<ul style="list-style-type: none"> <li>• red scaly plaques with well-defined edges</li> </ul>	<ul style="list-style-type: none"> <li>• red scaly plaques with well-defined edges</li> </ul>	<ul style="list-style-type: none"> <li>• periungual redness and swelling</li> <li>• red</li> <li>• the nail plate is distorted with transverse ridges</li> <li>• yellow pus appears under the cuticle</li> </ul>	<ul style="list-style-type: none"> <li>• periungual redness and swelling</li> <li>• the scale is silvery white</li> <li>• toes</li> <li>• yellow pus appears under the cuticle</li> </ul>
<b>Predicted outcome of the Siamese network</b>	<ul style="list-style-type: none"> <li>• red scaly plaques with well-defined edges (0.631)*</li> </ul>	<ul style="list-style-type: none"> <li>• with redness and swelling (0.852)</li> <li>• the scale is silvery white (0.736)</li> <li>• toes (0.602)</li> <li>• yellow pus appears under the cuticle (0.657)</li> </ul>	<ul style="list-style-type: none"> <li>• periungual redness and swelling (1.0)</li> <li>• red (1.0)</li> <li>• the nail plate is distorted with transverse ridges (0.908)</li> <li>• yellow pus appears under the cuticle (1.0)</li> </ul>	<ul style="list-style-type: none"> <li>• with redness and swelling (0.826)</li> <li>• toes (0.622)</li> <li>• abscess (0.628)</li> <li>• yellow pus appears under the cuticle (0.989)</li> </ul>
<b>Correct classification</b>	Y	N	Y	N

\*: class probability.

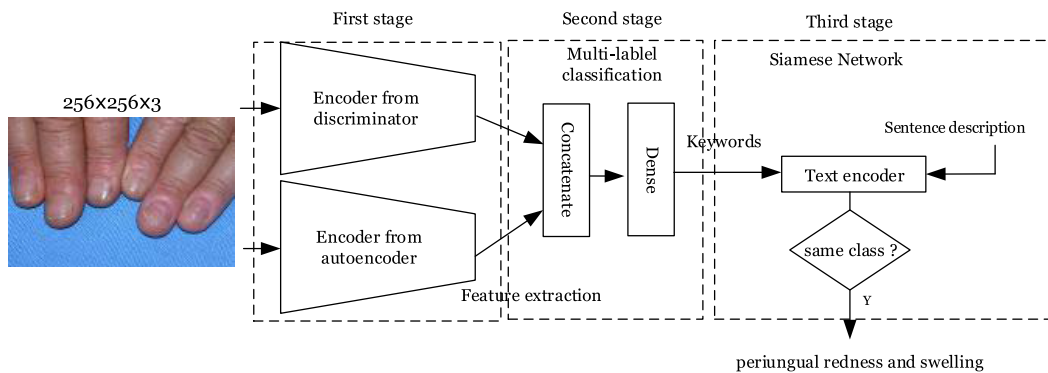


FIGURE 4. Flowchart of proposed methods.

set of the image. (3) In the third stage, the Siamese network was employed to obtain textual similarity matching

between keyword labels and colloquial sentence descriptions. The tasks in these three stages were integrated to

generate colloquial sentence descriptions for skin pathology. A flowchart of the process is shown in Fig. 4.

#### IV. EXPERIMENTAL RESULTS

The experiment mainly involved three model architectures: pre-trained discriminator and autoencoder, multi-label keyword classification, and Siamese network. We augmented the images to 7936 images. Randomly chosen 7139 data points were used to train the discriminator, autoencoder, and multi-label classifiers. The remaining 800 data points were used for testing purposes. For training the Siamese network, after pre-processing, 304 pieces of training data were used in total, including 49 positive samples and 255 negative samples. The related datasets are listed in Table 6.

Inquiries by physicians on the medical history of patients or their communication about symptoms are colloquial sentences or keywords. Therefore, for medical students, the teaching and training of pathological explanations or evaluations uses simple colloquial descriptions. The number of key items labeled was 21 and the maximum length of the colloquial word dictionary was 43, as shown in Table 7.

As shown in Table 8, the mean accuracy, Jaccard index (JI), DICE (DSC), sensitivity, and specificity of the discriminator for image segmentation of 800 test images were 0.9137, 0.8710, 0.9205, 0.9467, and 0.6662, respectively. As shown in Table 9, the mean squared error (MSE) and mean absolute error (MAE) for 800 test images of the autoencoder model were 0.0429 and 0.1663, respectively. The experimental results show that features of both the discriminator and autoencoder yield good performances in binary classification (skin or background) and reconstruction of the original image.

The accuracy performance for 800 test images of the multi-label classifier is listed in Table 10. In calculating the “accuracy” in the second column of Table 10, the predicted keywords must match the given keywords exactly. In calculating “binary accuracy” in the first column of Table 10, any one of the predicted keywords is in the set of the given keywords. As shown in the table, the mean accuracy was 0.9638, and the mean binary accuracy was 0.9951. Both the accuracy performances were quite good.

The Siamese network training dataset had 304 samples, including 49 positive and 255 negative samples. The training and testing datasets contained 243 and 61 samples, respectively. The accuracy for the training dataset was 1.0 and that for the testing data was 0.8361.

To evaluate overall performance, 800 test images were used. The performances of the multi-label classifier and Siamese network are listed in Table 11. The misclassification rate for the multi-label classifier is 0.036, and that for the Siamese network is 0.028, which are all satisfactory. The experimental results show that even if there are some mistakes in predicted keyword generation, accurate colloquial sentence descriptions can still be obtained (e.g., from 29 reduced to 23).

Table 12 presents a comparison of the experimental outputs of the four input images. As shown in the table, although the predicted keywords do not exactly match the given ground truth keywords (only partially correct), correct colloquial sentence descriptions can still be obtained (two out of four).

#### V. CONCLUSION

In this study, a novel approach for skin medical image captioning is proposed. The main feature of this approach is that the overall task of image captioning is handled using a three-stage design. First, the features from the encoders of the discriminator and autoencoder are implemented using similar configurations of fully convolutional networks and are extracted after completing the training processes. Next, the multi-label classifier establishes the relationship between the input image and important keywords that contain key information about the image. Finally, the Siamese network builds a mapping between the keywords from the classifier and sentence descriptions. This approach has never been used before for skin image captioning and has shown promising results.

In this pilot study, we selected three typical and common skin diseases, paronychia, plaque psoriasis, and herpes zoster, from the DermNet website. The experimental results and performance showed that the proposed method is suitable for skin image captioning, especially for small datasets of medical images and informal sentences. The results of this study can provide an auxiliary platform for students at the School of Medicine to learn, read, and interpret skin lesions. We expect that the proposed approach will be extended to address other skin diseases. More importantly, this artificial intelligence platform with natural language development can serve as a cross-cultural and cross-national auxiliary platform without language barriers. In the future, matching networks with an attention mechanism [17] can be considered for measuring text similarity in the last step of the three-stage design.

#### REFERENCES

- [1] K. Wolff, L. Goldsmith, S. Katz, B. Gilchrest, A. S. Paller, and D. Leffell, *Fitzpatrick's Dermatology in General Medicine*, 8th ed. New York, NY, USA: McGraw-Hill, 2012.
- [2] D. Zhao, Z. Chang, and S. Guo, “A multimodal fusion approach for image captioning,” *Neurocomputing*, vol. 329, pp. 476–485, Feb. 2019.
- [3] S. Bai and S. An, “A survey on automatic image caption generation,” *Neurocomputing*, vol. 311, pp. 291–304, Oct. 2018.
- [4] Y. Huang, J. Chen, H. Ma, H. Ma, W. Ouyang, and C. Yu, “Attribute assisted teacher-critical training strategies for image captioning,” *Neurocomputing*, vol. 506, pp. 265–276, Sep. 2022.
- [5] C. Zeng, S. Kwong, T. Zhao, and H. Wang, “Contrastive semantic similarity learning for image captioning evaluation,” *Inf. Sci.*, vol. 609, pp. 913–930, Sep. 2022.
- [6] C. Wang and X. Gu, “Learning joint relationship attention network for image captioning,” *Exp. Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118474, doi: 10.1016/j.eswa.2022.118474.
- [7] F. Wu, H. Yang, L. Peng, Z. Lian, M. Li, G. Qu, S. Jiang, and Y. Han, “AGNet: Automatic generation network for skin imaging reports,” *Comput. Biol. Med.*, vol. 141, Feb. 2022, Art. no. 105037, doi: 10.1016/j.compbiomed.2021.105037.
- [8] C. Barata, M. E. Celebi, and J. S. Marques, “Explainable skin lesion diagnosis using taxonomies,” *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107413, doi: 10.1016/j.patcog.2020.107413.

- [9] Z. G. Li and H. H. Y. C. Chen, "Biomedical text similarity evaluation using attention mechanism and Siamese neural network," *IEEE Access*, vol. 9, 2021, doi: [10.1109/access.2021.3099021](https://doi.org/10.1109/access.2021.3099021).
- [10] N. Bölücü, B. Can, and H. Artuner, "A Siamese neural network for learning semantically-informed sentence embeddings," *Exp. Syst. Appl.*, vol. 214, Mar. 2023, Art. no. 119103, doi: [10.1016/j.eswa.2022.119103](https://doi.org/10.1016/j.eswa.2022.119103).
- [11] H. Park, K. Kim, S. Park, and J. Choi, "Medical image captioning model to convey more details: Methodological comparison of feature difference generation," *IEEE Access*, vol. 9, pp. 150560–150568, 2021, doi: [10.1109/ACCESS.2021.3124564](https://doi.org/10.1109/ACCESS.2021.3124564).
- [12] F. Wang, X. Liang, L. Xu, and L. Lin, "Unifying relational sentence generation and retrieval for medical image report composition," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5015–5025, Jun. 2022, doi: [10.1109/TCYB.2020.3026098](https://doi.org/10.1109/TCYB.2020.3026098).
- [13] Z. Yang, P. Wang, T. Chu, and J. Yang, "Human-centric image captioning," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108545, doi: [10.1016/j.patcog.2022.108545](https://doi.org/10.1016/j.patcog.2022.108545).
- [14] Z. Zhang, W. Zhang, W. Diao, M. Yan, X. Gao, and X. Sun, "VAA: Visual aligning attention model for remote sensing image captioning," *IEEE Access*, vol. 7, pp. 137355–137364, 2019, doi: [10.1109/ACCESS.2019.2942154](https://doi.org/10.1109/ACCESS.2019.2942154).
- [15] X. Ye, S. Wang, Y. Gu, J. Wang, R. Wang, B. Hou, F. Giunchiglia, and L. Jiao, "A joint-training two-stage method for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, doi: [10.1109/TGRS.2022.3224244](https://doi.org/10.1109/TGRS.2022.3224244).
- [16] *DermNet NZ*. Accessed: Nov. 23, 2020. [Online]. Available: <https://dermnetnz.org/>
- [17] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching network for one-shot learning," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, 2017, pp. 1–9, doi: [10.48550/arXiv.1606.04080](https://doi.org/10.48550/arXiv.1606.04080).



**KUIYOU LAI** received the B.S. degree from the Department of Computer Science and Information Management, Providence University, Taichung, Taiwan, in 2020, and the M.S. degree from the National Yunlin University of Science and Technology, Yunlin, Taiwan, in 2022. His research interests include medical image processing and natural language processing.



**YIHLON LIN** received the B.S. degree in electrical engineering from the Department of Electronic Engineering, I-Shou University, Kaohsiung, Taiwan, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from the Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung, in 1999 and 2006, respectively. He is currently an Associate Professor with the Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Yunlin, Taiwan. His research interests include medical image processing, natural language processing, computer vision, and deep learning.



**WENYU CHANG** received the M.D. degree from the College of Medicine, National Taiwan University, Taiwan, in 2001, and the Ph.D. degree from the Graduate Institute of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan, in 2015. She is currently an Assistant Professor with the College of Medicine, I-Shou University, Kaohsiung, and the Director of the Department of Dermatology, E-Da Cancer Hospital, Kaohsiung. Her research interests include medical dermatology and deep learning in dermatological images.

...