**Methods**

**Ethical Approval and Informed Consent to Participate.** All studies related to human samples were approved by the ethics committee of the Second Xiangya Hospital of Central South University (No. 201930044). Each participant underwent testing after providing written informed consent, which agrees to the use of any accompanying images for data analysis and the publication of this study.

**Dataset.** Our study contains 386 validation cases recruited retrospectively based on the criteria that they had previously undergone a skin biopsy and were diagnosed with DLE, A-SCLE, P-SCLE, SLE, LP, EAC, Pso, DM, Vas, EM, Ecz or Ros from 2010 to 2021. To eliminate interference based on differences in skin type and skin condition on the experimental results, healthy controls with matched age were also included in the cohort (baseline characteristics shown in Extended Data Table 3-5). These cases mentioned above were screened from 884 cases treated at the Department of Dermatology in 25 institutions in China (the participating institutions are shown in Extended Data Table 1), and the screening criteria are presented in Extended Data Figure 4. For each case, we collected their medical history, which consisted of images and descriptions of clinical symptoms, laboratory test results, pathological examination records, medication history and diagnosis from clinicians. By the time skin biopsies were taken, all patients were clinically diagnosed and untreated. Two specialists of the Second Xiangya Hospital cooperated to classify each case into subsets of skin diseases. This classification was mainly based on the appearance of skin lesions, laboratory tests, pathological analysis, therapeutic effects and follow-up information.

The metadata of this study were classified into 3 major components: clinical skin images, multi-IHC images and 2019 EULAR/ACR scores. The clinical skin images are photographs of the skin lesion area. Each case contained one or more clinical skin images that conveyed the extent of skin lesions. Images were captured with various devices (smartphones or cameras). For each image, we ensure sufficient resolution and brightness for the deep learning system to recognize the image.

We employed the 2019 European League Against Rheumatism/American College of Rheumatology (EULAR/ACR) classification criteria to estimate the systemic involvement of the participants. The entry criterion requires a titer of antinuclear antibodies (ANAs) greater than 1:80 in HeP-2 cells or an equivalent positive test. As shown in Figure 2, the EULAR/ACR score mainly consists of 2 parts: clinical domains and immunology domains. The clinical domains include the constitutional domain (fever), hematological domain (leukopenia, thrombocytopenia, and autoimmune hemolysis), neuropsychiatric domain (delirium, psychosis, and seizure), mucocutaneous domain (nonscarring alopecia, oral ulcers, subacute cutaneous or discoid lupus, and acute cutaneous lupus), serosal domain (pleural or pericardial effusion and acute pericarditis), musculoskeletal domain (joint involvement), and renal domain (proteinuria>0.5 g/24 h or lupus nephritis diagnosed by renal biopsy). The immunological domains are formed by antiphospholipid antibody domains (anti-cardiolipin or anti-β2GP1 antibodies or lupus anticoagulant), complement protein domains (low C3 and/or low C4), and SLE-specific antibody domains (anti-dsDNA or anti-Smith antibody). Within each domain, only the highest weighted criterion was counted and applied to the total score, and the occurrence of a criterion on one occasion was sufficient.

The multi-IHC images are 20x images of skin specimens with combined staining of CD4-CD8-CD19-CD11b. Paraffin sections with a thickness of 3–5 µm were prepared from each patient and then subjected to dewaxing followed by antigen retrieval in buffer at pH 6.0, and 4 rounds of IHC staining were performed afterward. Sections were incubated with a DAPI solution before the slices were sealed and examined under a microscope. Antigen retrieval buffer, blocking buffer, anti-CD4 and anti-CD19 working solution, and HRP-conjugated secondary antibodies were obtained from the skin in situ immune cell detection kit provided by Guangzhou LBP Medicine Science & Technology Co., Ltd. The anti-CD11b working solution was purchased from Abcam (ab133357), and the anti-CD8 working solution was purchased from MXB (MAB-0021). Staining solutions were obtained from Akoya Biosciences.

**Multimodal Development.** The datasets were prepared for further running through two major processes: single-modal feature extraction and multimodal feature fusion. To help alleviate the selection bias, the dataset was randomly divided into a training dataset and a testing dataset at a ratio of 8:2. To ensure no overlap between the training dataset and testing dataset, an anonymous case ID was used. The testing dataset was used as a validation of training, and 10 repeated validation experiments were conducted. After a series of comparisons, the MMDLS was constructed with the best validation result.

**Single-modal feature extraction.** Compared with the traditional manual design of features based on medical experience, deep learning adopts a data-driven method to automatically extract image features, which can obtain deeper feature representations with stronger generalization ability 1. Therefore, in our study, we automatically extracted the features from each modality through convolutional neural networks in our multimodal data to effectively achieve information conversion and the fusion of each modality 2.

In IHC images, the recognition of lupus subtypes mainly depends on the distribution and types of different immune cells; thus, extracting global and detailed information from IHC images is crucial. In general, deeper neural networks can effectively extract high-level semantic global features, whereas wider neural networks can capture fine-grained features at different scales. Therefore, to achieve comprehensive feature extraction and obtain the best performance of the model, it is necessary to balance the relationship between these factors effectively. To the best of our knowledge, the compound scaling method used in the EfficientNet model effectively solves this confusion 2. In addition, we compared the accuracy and complexity of the models under different compound scaling ratios by ablation experiments. Finally, we used the EfficientNet-B3 model to automatically extract the depth features from IHC images. The network is stacked with 16 modules, each of which uses convolution of different sizes to translate images and extract depth features through convolution operations.

For different lupus erythematosus subtypes, the difference in clinical images is not obvious. In particular, for subtypes with similar lesion areas, an accurate diagnosis is often difficult to determine in the clinic. Therefore, a deeper network is needed to extract more discriminative features. However, for clinical images with limited sample data, the gradient easily disappears because the network is too deep. The residual connection structure in the ResNet-18 model 3 combines the original features of the previous layer with the advanced abstract features, which effectively reduces the difficulty of training the deep network. Therefore, we used the more portable ResNet-18 model to extract the salient features of the skin lesion area. In addition, we compared a variety of popular network models in our multimodal data, and the experimental results show that ResNet-18 achieves better performance with clinical images.

In clinical practice, dermatologists not only evaluate the image morphological data but also make a comprehensive judgment based on clinical parameters. In our deep learning system, we combined clinical information with network features to provide complementary information for image features. We calculated the score for each patient based on the diagnostic criteria of the 2019 EULAR/ACR score 4. Each patient underwent 21 tests that assessed the patient’s involvement from different clinical perspectives, and the scores that met the criteria were added to provide a final aid diagnostic score for the patient. Then, we tried to magnify the differences among LE subtypes and other similar dermatoses by obtaining a higher resolution 2019 EULAR/ACR score. We separately incorporate the clinical 2019 EULAR/ACR scores of different magnification rates, including 1×, 50×, 100×, and 150×, where 1× indicates no expansion with a top-1 acc of 0.7916. The result increases with the continuous expansion of magnification rates. When the expansion rate reached 100×, the top1 acc reached the highest value of 0.8141. When increased to 150×, the top-1 acc decreased, demonstrating that the expansion reached saturation. Finally, we magnified and normalized the scores to obtain the shallow features of the clinical information modalities, which were then fed into the multimodal fusion module along with intermediate features of the other modality.

**Multimodal feature fusion.** As mentioned above, the data for each modality can be independently diagnosed by a dermatologist. However, dermatologists can make more evidence-based diagnoses when diagnostic information from different modalities is effectively combined. Combining rich data can effectively help doctors determine a diagnosis, and integrating data from multiple models can also benefit our network 5.

We first established an image feature extraction network composed of two encoders for clinical images and IHC images and learned a set of corresponding feature maps from each network structure. We considered using feature-level combinations to integrate the intermediate features generated by the two image modalities and to comprehensively assess information from different sources. We argued that the application of deep models in medical image analysis is often confronted with small sample learning problems, which can be effectively alleviated if image features are combined with clinical information 6. Therefore, the merged image features and clinical metadata features were combined to cooperatively train a classifier and achieve the final prediction of disease subtypes. In addition, we performed a comparison between single modal and multimodal approaches. The effective superposition of the modalities provided rich feature information and improved the performance of the model.

**Multimodal Training.** The enrolled patients were randomly divided into a training cohort and a test cohort at a ratio of 8:2, and the training cohort was used to optimize the parameters of the model. In addition, extensive care was taken to ensure that these patients were not split between the training and test cohorts. In the process of model training, we initialized the image features to extract the weight of the network from the model that has been pretrained for 1000 classification tasks on ILSVRC 2015 7. Then, we removed the final classification layer from the network and retrained it using our dataset to fine-tune the parameters of all layers.

Given that we focused on the contour position information formed by the aggregation of different types of immune cells and the overall spatial information of the skin lesion area in the clinical images, we did not use the patch method to crop the image before training. Instead, we resized the resolution of the two types of images to a specified size and sent them into the network directly. The whole training process includes forward propagation and backward propagation 8. First, the image was input into the network, and the encoder obtained the high-dimensional feature representation by extracting the hidden layer features in the image. Then, the classifier transformed the high-dimensional feature representation into the prediction results and input it into the network, and the errors of outputs and labels were propagated back to the network to update the model parameters. The cross-entropy 9 was chosen as the loss function, the learning rate was initialized at 0.00001, and an adaptive moment estimation (Adam 10) optimizer was applied to update the model parameters with a batch size of 5. As the network deepens, the learning rate gradually decreases but only to a minimum of 1×10-9.

**Multimodal Evaluation.** We analyzed the performance of the multimodal model in obtaining diagnostic results by calculating the top-1 accuracy, sensitivity, specificity, F1-score, and other evaluation indicators. Because 13 types of lupus erythematosus and similar subtypes are diagnosed, the proportions of true positives (TPs) and true negatives (TNs) are seriously out of balance in the calculation of evaluation indicators for each disease. This phenomenon collectively results in indicators, such as specificity, that are not sensitive to model changes. Therefore, in the multiclassification task, sensitivity and top-1 accuracy are more important to verify the effectiveness of the method 11.

In addition to numerical results, we also provide some visual results to evaluate the method performance 12. Extended Data Fig. 2 shows the confusion matrix of the single-modal and the mutual fusion of multimodal approaches. Element (i, j) of each confusion matrix represents the empirical probability of predicting class j when the ground truth is class i. Light blue represents a low percentage, and deep blue represents a high percentage. Extended Data Fig. 3 shows the t-SNE plot of clusters for the 13 categories. The t-SNE plot helps us visualize the clustering of higher-dimensional feature vectors, and different colored point clouds represent the different disease categories. As shown in these two visualization results, the classification results of most subtypes are significantly improved with the effective fusion of multiple modalities, especially DLE, P-SCLE, A-SCLE, and SLE subtypes. These diseases often coexist and transform into each other, thus making them challenging to visually diagnose.

**Comparison to clinicians and pathologists.** To compare the performance of MMDLS with clinicians and pathologists, we retrospectively collected the diagnosis of each case. Clinical diagnosis is made from the initial medical record at the outpatient department and is mostly based on features of skin lesions and clinical manifestations. The pathological diagnosis is collected from pathological reports that rely highly on H&E staining, the gold standard for diagnosis. A diagnosis was made by two senior dermatologists who reached a consensus based on all images, medical history and follow-up data. The gold standard is used as a true label, and diagnoses from clinicians, pathologists and MMDLS were used as predicted labels. Here, for better classification, we introduced predicted labels, including undefined diagnosis (UD), other skin diseases that do not belong to the 13 categories (OSD), and uncertain lupus subtype (ULE), to the clinical and pathological diagnosis. Regarding the consistency between the predicted label and true label, indices, such as specificity, sensitivity, precision and F1 score, were chosen to compare the performances of clinicians, pathologists and MMDLS. Additionally, confusion matrix visualization was used for a more intuitive comparison (Figure 5).

**Hardware and software.** All experiments were implemented on an ASUS high-performance computing cluster, and we perform training using NVIDIA GeForce 1080 graphical processing units (GPUs) with 16 TB of SSD local storage. Each model was trained on a single GPU. We used the PyTorch 13 frame (version 1.1.0) to load data and implement the algorithm CaseViewer to access the original IHC images. Clinical images were obtained using a mobile phone or digital camera.

**References**

1. Liu, Y.*, et al.* A deep learning system for differential diagnosis of skin diseases. *Nat Med* **26**, 900-908 (2020).

2. Tan, M. & Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. (2019).

3. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *IEEE* (2016).

4. Aringer, M.*, et al.* 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. *Ann Rheum Dis* **78**, 1151-1159 (2019).

5. Esteva, A.*, et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118 (2017).

6. Xie, Y., Zhang, J., Xia, Y., Fulham, M. & Zhang, Y. Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. *Information Fusion*, 102-110 (2018).

7. Russakovsky, O.*, et al.* ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* **115**, 211-252 (2015).

8. Zheng, X., Yao, Z., Huang, Y., Yu, Y. & Zhou, J. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. *Nature Communications* **11**(2020).

9. Boer, P., Kroese, D.P., Mannor, S. & Rubinstein, R.Y. A Tutorial on the Cross-Entropy Method. *Annals of Operations Research* **134**, 19-67 (2005).

10. Kingma, D. & Ba, J. Adam: A Method for Stochastic Optimization. *Computer Science* (2014).

11. Wu, H.*, et al.* A deep learning-based smartphone platform for cutaneous lupus erythematosus classification assistance: Simplifying the diagnosis of complicated diseases. *Journal of the American Academy of Dermatology* (2021).

12. Hollon, T.C.*, et al.* Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nature medicine* **26**, 52-58 (2020).

13. Paszke, A.*, et al.* Automatic differentiation in PyTorch. (2017).