Cureus Journal of Computer Science

Open Access Original Article

A Multimodal Deep Learning Approach for Skin Allergy Diagnosis Using Image and Symptom Data

Received 04/30/2025
Review began 05/15/2025
Review ended 08/27/2025
Published 09/08/2025

DOI:
https://doi.org/10.7759/s44389-025-06071-1

Shubhasri Pradhan [1] , Brojo K. Mishra [1]

1. Computer Science and Engineering, NIST University, Berhampur, IND

Corresponding authors: Shubhasri Pradhan, shubhasri.pradhan@nist.edu, Brojo K. Mishra, brojomishra@gmail.com

Abstract

Skin allergies present significant diagnostic challenges due to their varied visual manifestations and overlapping symptoms. This study introduces a multimodal deep learning framework that integrates image-based and text-based clinical information to improve diagnostic accuracy. A Convolutional Neural Network analyzes spatial features from dermatological images, while a Bidirectional Encoder Representations from Transformers-Long Short-Term Memory model interprets contextual cues from patient-reported symptoms. These modalities are combined through both feature-level and decision-level fusion to build a robust hybrid system. Evaluated on a curated dataset comprising 996 images and 1,213 symptom descriptions across nine allergy classes, the multimodal model achieved 86% accuracy, outperforming unimodal image-only (49%) and text-only (78%) baselines. Comprehensive preprocessing, including image augmentation and text normalization, ensured consistency across modalities. The system supports incomplete inputs and offers interpretability via Gradient-weighted Class Activation Mapping and attention visualizations. These findings underscore the potential of multimodal AI in dermatological diagnostics and telemedicine. Future work will focus on real-time mobile deployment and expanding coverage across diverse skin tones and allergy types.

Categories: AI applications, Machine Learning (ML), Medical Expert systems
Keywords: multimodal deep learning, medical ai diagnostics, symptom text interpretation, dermatological image analysis, cnn-bert-lstm fusion, skin allergy diagnosis

Introduction

The skin, the body's largest organ, acts as a protective barrier against environmental pathogens and allergens. Its constant exposure makes it particularly prone to dermatological conditions, including skin allergies. These conditions commonly present with redness, itching, swelling, and blistering symptoms that affect patient well-being and require accurate clinical diagnosis. However, traditional diagnostic methods rely heavily on visual inspection and subjective symptom reporting, which can lead to errors, especially when cases involve overlapping or ambiguous features [1,2].

In underserved and rural regions, limited access to dermatologists further hampers timely and accurate diagnosis. This highlights the need for accessible, automated diagnostic systems that can support clinical decisions in real-world settings. However, many existing diagnostic systems rely solely on either image or text data. This unimodal focus overlooks the complementary insights that arise when both modalities are integrated.

While artificial intelligence (AI) has made inroads in dermatology, most existing models are unimodal, focusing solely on either image data or textual symptom descriptions. This separation limits diagnostic capability. For instance, a model analyzing only images may overlook important symptom cues, while a text-only model lacks visual context. In contrast, dermatologists typically combine both sources of information in their decision-making process [3,4].

Recent advances in deep learning have enabled machines to process visual and textual information with remarkable accuracy. Convolutional Neural Networks (CNNs) have shown strong performance in image classification tasks, including dermatological image analysis [5-7]. At the same time, transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) have advanced natural language understanding, allowing for better interpretation of patient-reported symptoms [8-10].

This study presents a multimodal deep learning framework for skin allergy diagnosis that integrates visual and textual modalities. The system uses a CNN to extract features from skin images and a BERT-Long Short-Term Memory (LSTM) model to analyze symptom descriptions. These complementary features are fused at both the feature and decision levels to enhance diagnostic accuracy.

Evaluated on a dataset of 996 images and 1,213 symptom-text samples across nine skin allergy types, the model achieved an overall accuracy of 86%, outperforming image-only (49%) and text-only (78%) models. The system also incorporates robust preprocessing, interpretability through Gradient-weighted Class Activation Mapping (Grad-CAM) and attention visualization, and the ability to handle missing input types. These attributes make the framework well-suited for deployment in telemedicine and mobile health environments, especially in resource-constrained settings.

Literature review

The diagnosis of skin allergies, despite significant technological advancements, continues to be a challenging clinical task due to overlapping characteristics of dermatological symptoms. Traditional diagnostic methods often rely on manual inspection and patient-reported symptoms, which can be subjective and inconsistent. To address these limitations, researchers are increasingly turning to AI, particularly deep learning techniques, for medical diagnostics. Early efforts by Hai et al. [1] introduced hybrid deep learning architectures that combine visual and textual data, showing that multimodal approaches outperform unimodal ones in classification tasks. Marin et al. [2] expanded this direction by introducing the Recipe1M dataset that fuses text and images, encouraging further exploration of multimodal learning paradigms.

Deng et al. [3] presented a comprehensive review of multimodal fusion strategies, categorizing them into feature-level, decision-level, and joint fusion. These frameworks provide essential guidelines for developing multimodal diagnostic tools. Park et al. [4] further reinforced this with semantic embeddings for multimodal classification, showing how hierarchical structures enhance accuracy. In the domain of dermatology, Rajesh Kumar and Vignesh Babu [5] developed a hybrid approach for lesion segmentation using traditional image processing and machine learning classifiers like support vector machine and K-nearest neighbor, yielding impressive results.

Jing et al. [6] proposed a hybrid model integrating deep learning architectures for skin lesion classification, which achieved high accuracy using the ISIC dataset. These works underscore the effectiveness of combining classification and segmentation tasks within a unified deep learning framework. The introduction of AlexNet by Krizhevsky et al. [7] revolutionized image classification and laid the groundwork for applying CNNs in medical imaging. In dermatology, CNNs such as VGG, ResNet, and Inception have been adapted for disease detection. However, CNNs alone often fall short when differentiating between visually similar but clinically distinct skin conditions.

To overcome this, transformer models like BERT have been employed for processing patient-reported symptoms. Hosny et al. [8] emphasized the role of AI in radiology, while Mohmmed et al. [9] explored automatic skin cancer classification using classical machine learning approaches. Szegedy et al. [10] introduced deeper CNNs like Inception, further advancing performance in image recognition. Ronneberger et al. [11] proposed the U-Net architecture, widely adopted in biomedical image segmentation. Dai et al. [12] used ensemble learning and loss weighting strategies to boost skin lesion classification accuracy, emphasizing robust model design.

The fusion of image and text data enhances diagnostic reasoning. Omiye et al. [13] reviewed AI applications in dermatology, highlighting benefits and ongoing challenges like model interpretability. Classical algorithms like K-nearest neighbors introduced by Cover and Hart [14], and sequential models such as LSTM by Hochreiter and Schmidhuber [15], serve as historical benchmarks for evaluating newer Natural Language Processing (NLP) models. Modern transformer-based NLP methods, like those introduced by Wolf et al. [16], have set new standards for language understanding in healthcare. These models outperform recurrent neural networks and LSTMs in extracting contextual meaning from symptom descriptions.

Datasets like ImageNet [17] and International Skin Imaging Collaboration (ISIC) [18] have enabled the training of deep models on large-scale annotated data, significantly contributing to improved performance. Recently, frameworks like FACTS proposed by Akkiraju et al. [19] and personalized learning systems using Retrieval-Augmented Generation [20,21] have further extended the applications of AI into domains like personalized healthcare and intelligent decision support.

By integrating CNNs for image analysis [20] with BERT-based models for interpreting clinical text, a multimodal deep learning system can be developed that better mimics a dermatologist’s holistic diagnostic process. This review provides a foundation for developing an advanced multimodal framework for predicting skin allergies, addressing limitations of unimodal systems, and advancing the field of intelligent dermatological diagnostics.

Materials And Methods

Methodology

The proposed system integrates a CNN for image-based classification and an NLP module that utilizes

BERT embeddings alongside an LSTM network for textual symptom interpretation. The model was trained on a curated dataset comprising 996 dermatological images distributed across 9 distinct skin allergy classes, and 1,213 symptom-text tuples paired with corresponding disease labels. This multimodal dataset provides sufficient variance to support the classification task and allows for robust fusion-based learning.

To ensure transparency and reproducibility, detailed statistics of the dataset are presented. The image dataset consisted of 996 samples distributed across nine skin allergy classes: Atopic Dermatitis (124), Contact Dermatitis (112), Urticaria (123), Psoriasis (111), Eczema (120), Drug Allergy (113), Rosacea (107), Seborrheic Dermatitis (102), and Photoallergic Dermatitis (98). Correspondingly, 1,213 symptom descriptions were collected and labeled for these classes. Each image-text pair was aligned based on diagnosis to support multimodal learning. To address minor class imbalance, oversampling techniques were applied to underrepresented classes in the symptom dataset. Furthermore, image augmentation methods such as rotation, flipping, and cropping ensured a balanced training distribution and improved model generalization.

The CNN was trained using a dataset of 996 labeled skin images, evenly distributed across nine disease categories, including conditions such as Atopic Dermatitis, Contact Dermatitis, Urticaria, and Psoriasis. The images were obtained from publicly available sources such as the ISIC archive [18] and DermNet [20], while textual symptom data was sourced from a skin disease classification dataset [21]. For symptom analysis, 1,213 symptom description-disease label pairs were processed using a hybrid BERT-LSTM model. These text tuples included various expressions of symptoms like redness, itching, and scaling, annotated with corresponding diagnoses to enable supervised learning. Prior to modeling, extensive preprocessing was performed on both modalities. Image preprocessing included resizing, normalization, contrast enhancement, and noise reduction through filtering to improve clarity. Data augmentation techniques such as rotation, flipping, and cropping were applied to enhance generalization and reduce overfitting. For the textual data, preprocessing steps included tokenization, stop-word removal, lemmatization, and padding to ensure consistent input length and structure across samples.

To mitigate inconsistencies in subjective symptom data, extensive preprocessing steps such as lemmatization, stop-word removal, spelling correction, and synonym normalization were applied. A curated symptom vocabulary was built to standardize commonly used phrases, and ambiguous or low-information entries were filtered out during training.

For visual analysis, CNN was utilized to extract complex spatial features from the skin images. The architecture incorporated multiple convolutional layers with ReLU activation, followed by max-pooling and dropout layers to prevent overfitting. Fully connected layers were used at the final stage to produce classification probabilities. To improve performance on relatively limited dermatological datasets, transfer learning was applied using pre-trained models such as ResNet and MobileNet, which are well-established for medical image classification tasks [6,7,10]. The CNN was trained using an Adam optimizer with a learning rate of 0.001 and categorical cross-entropy as the loss function. Learning rate scheduling and early stopping were implemented to ensure training stability. A stratified 80-20 train-test split was adopted to maintain class balance during evaluation.

For the textual modality, a hybrid architecture combining BERT and LSTM networks was developed. BERT embeddings were used to generate rich contextual representations of symptom descriptions, capturing nuanced semantics necessary for clinical understanding [16]. These embeddings were subsequently passed through an LSTM layer to capture temporal dependencies and sequential relationships commonly present in medical symptom narratives [15]. This BERT-LSTM combination allowed for a deep semantic and syntactic analysis of patient-reported symptoms.

The most critical aspect of the proposed framework is its multimodal fusion strategy, which integrates the outputs from the image and text pipelines. Two types of fusion were investigated: feature-level and decision-level. In feature-level fusion, feature vectors from the CNN and BERT-LSTM models were concatenated and passed through fully connected layers to create a joint multimodal representation, enabling the system to learn cross-modal relationships [1,3,13]. In decision-level fusion, the prediction probabilities from the individual models were combined using weighted averaging, with weights based on each modality's validation performance [3,4]. Experimental results showed that feature-level fusion yielded superior performance, effectively capturing interdependencies between visual cues and textual information [1,3].

Model evaluation was conducted using standard performance metrics including accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices to ensure comprehensive assessment across different allergy types [12]. The models were tested using the same stratified train-test split to ensure consistent comparison. To enhance clinical trust and model transparency, interpretability techniques were integrated into the system. Grad-CAM was applied to the CNN outputs to visualize the areas of the skin images that most influenced the model's decisions [13]. For the textual modality, attention weight visualization was used to highlight critical symptom phrases that contributed to the final prediction, offering clinicians a clear rationale for the diagnostic outcomes [4,16].

The complete system architecture comprises several core modules: a Data Input Module to receive skin images and symptom descriptions; a Preprocessing Module for cleaning and standardizing image and text inputs; a Feature Extraction Module utilizing the CNN and BERT-LSTM pipelines; a Fusion and Classification Module that integrates and interprets data from both modalities; and an Output Interface that presents the final diagnostic results along with visual explanations (as illustrated in Figure 1).

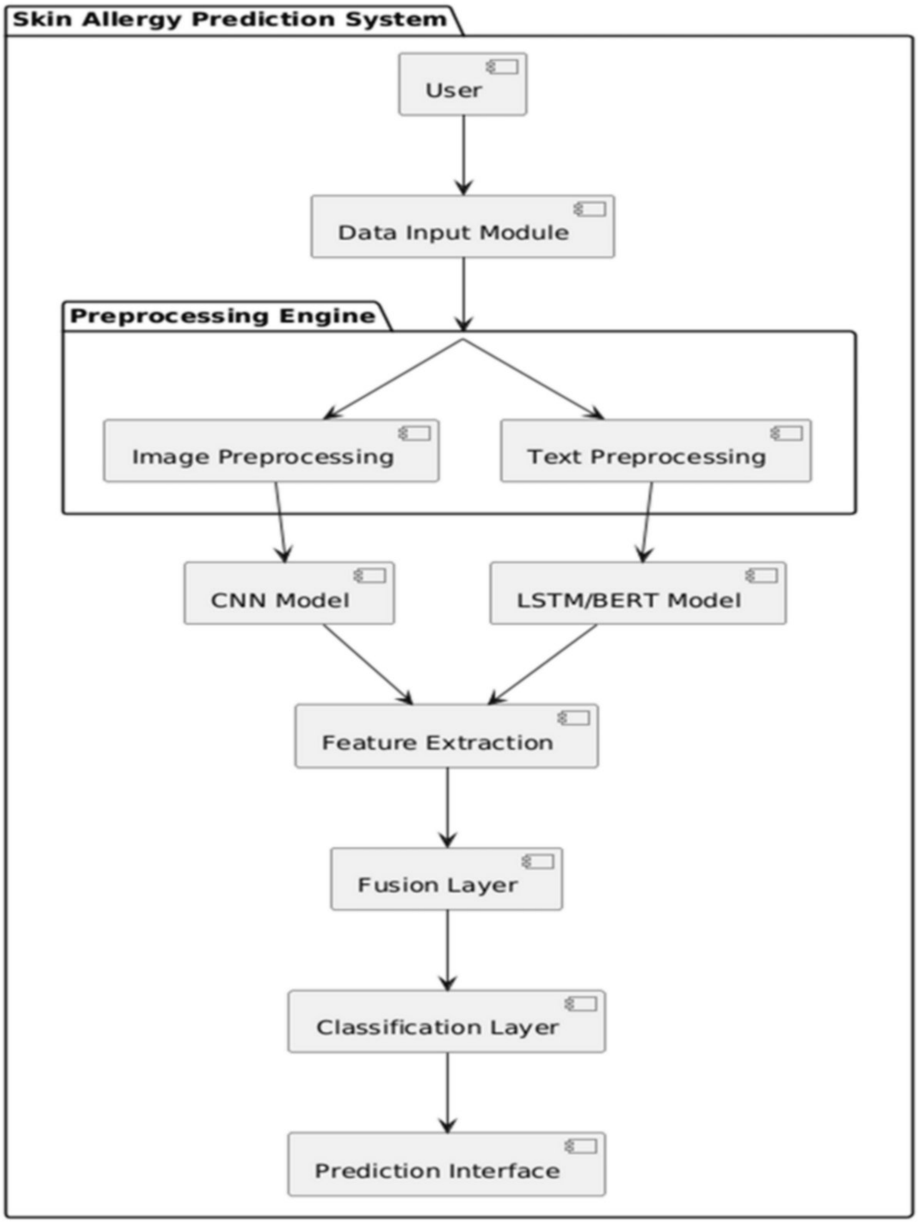

FIGURE 1: Framework of Skin Allergy Prediction Model

BERT, Bidirectional Encoder Representations from Transformers; CNN, Convolutional Neural Networks; LSTM, Long Short-Term Memory

This modular and interpretable design enables robust, multimodal prediction of skin allergies and represents a significant step forward in the development of intelligent tools for teledermatology and clinical decision support [5,13,19].

The CNN component includes three convolutional layers, each followed by ReLU activation and max-pooling layers. Dropout layers were applied after each pooling layer to reduce overfitting. The feature maps were flattened and passed to a dense layer before final classification.

The textual modality employs BERT-base for embedding generation (768-dimensional vectors), which are then passed to an LSTM layer with 128 hidden units and a dropout rate of 0.3. The LSTM output is fed into a fully connected dense layer followed by a softmax classifier for final prediction.

The fusion model concatenates the 512-dimensional CNN vector with the 128-dimensional LSTM output, and the combined 640-dimensional vector is passed through two dense layers (256 and 64 neurons, respectively) before classification.

Implementation

The implementation phase of the proposed hybrid skin allergy prediction system focuses on the practical realization of the conceptual design, integrating models for NLP and computer vision into a single multimodal diagnostic framework. The system was developed in Python, utilizing TensorFlow and Keras for building deep learning models, alongside supplementary libraries such as NumPy, Pandas, Matplotlib, and HuggingFace Transformers for data manipulation, visualization, and text processing [16].

A sequential Keras model was used to create CNN for the image classification component. Each of the three convolutional blocks in the CNN architecture consisted of a convolutional layer, a max pooling layer, and a ReLU activation function. To reduce overfitting, the final feature maps were flattened and passed through a dense layer with dropout regularization [7,10]. A carefully curated dataset of dermatological images was resized to 150 × 150 pixels and normalized to scale pixel values between 0 and 1 for model training [18,20]. Preprocessing was enhanced using the ImageDataGenerator class, applying data augmentation techniques such as random rotation, flipping, and brightness adjustments [6].

Simultaneously, the text analysis module was implemented to process symptom descriptions. Two approaches were adopted: an LSTM-based sequential model and a transformer-based BERT model. The LSTM model received tokenized and padded symptom inputs, converted into numerical sequences using pre-trained GloVe embeddings. It consisted of stacked LSTM layers with spatial dropout and dense layers culminating in a softmax classifier [15]. In contrast, the BERT model leveraged the MiniLM-L6-v2 variant from the Sentence Transformers library to generate context-rich embeddings [16]. The symptom descriptions were encoded into fixed-size vectors and passed to the LSTM model for classification. Robustness was ensured by merging multiple symptom datasets and applying data cleaning techniques like duplicate removal, text standardization, and noise filtering [4,17].

Multimodal fusion was achieved through intermediate feature-level concatenation. The feature vector outputs from the CNN and the symptom analysis models were concatenated and passed through fully connected layers that learned complex interdependencies between image and text modalities [3,4]. The final classification was performed using a softmax activation over multiple skin disease classes. In cases where either modality was unavailable, a fallback mechanism allowed the model to rely solely on the available data ensuring deployment flexibility [1].

The entire model was optimized using a joint loss function, balancing both modality streams. Evaluation metrics including accuracy, precision, recall, F1-score, and ROC-AUC were computed using Scikit-learn to assess model performance [12].

User interaction was facilitated via an interface for uploading skin images and symptom descriptions. The CNN module predicted skin conditions from images, while the text model classified based on patient-reported symptoms. A confidence-weighted fusion strategy determined the final output, giving more weight to the modality with higher predictive certainty in conflicting cases.

The implementation was executed on Google Colab using GPU acceleration to speed up training and evaluation. Training history, confusion matrices, and ROC curves were visualized to validate model learning. All trained components including the CNN, LSTM, label encoders, and BERT embeddings were serialized and stored to enable reuse and modular development.

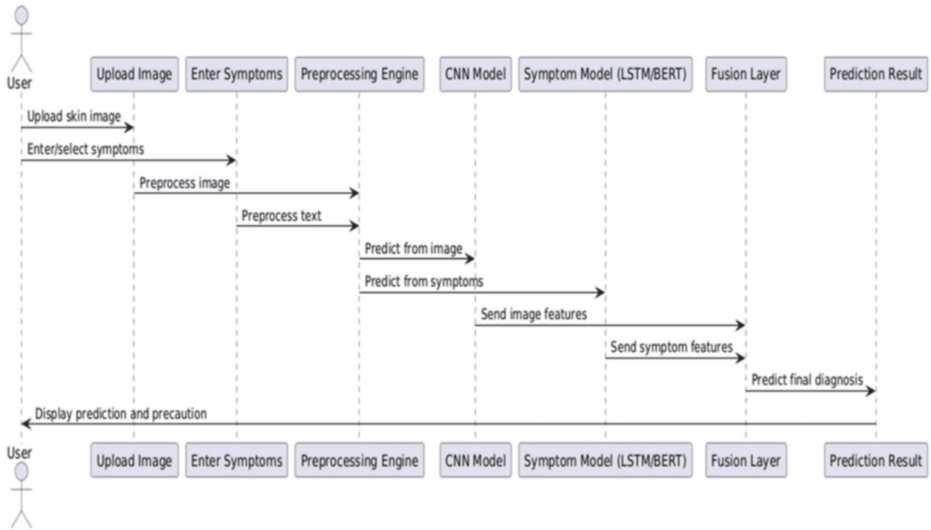

FIGURE 2: Sequence Diagram of Skin Allergy Prediction Model

BERT, Bidirectional Encoder Representations from Transformers; CNN, Convolutional Neural Networks; LSTM, Long Short-Term Memory

Figure *2* illustrates the entire data processing and prediction pipeline from input ingestion to final diagnosis emphasizing the interoperability and modularity of system components.

Through the practical realization of the proposed hybrid system, combining robust CNN architectures with advanced NLP models, the project successfully demonstrates a significant step toward intelligent, automated skin allergy diagnosis. The multimodal approach outperformed unimodal baselines, underscoring the value of fusing visual and textual data for clinical applications [1,3,6,13,19,21].

Results

The proposed hybrid deep learning system was evaluated using a custom-built multimodal dataset containing 996 skin images across 9 disease classes and 1,213 textual symptom-disease pairs, enabling comparative experiments across image-only, text-only, and fusion models.

A stratified 80-20 train-test split was applied to ensure balanced class representation in both training and testing phases. For image analysis, a CNN model utilizing pretrained architectures like ResNet and MobileNet [6,7,10] was developed through transfer learning. The CNN architecture included multiple convolutional layers, ReLU activation, max-pooling, dropout layers to prevent overfitting, and dense output layers. The model was trained using the Adam optimizer with a learning rate of 0.001, employing categorical cross-entropy as the loss function. Early stopping and learning rate decay were used to optimize training performance.

For symptom text analysis, two models were evaluated: an LSTM network and a BERT-based model. BERT embeddings were generated using the Hugging Face Transformers library [16] and fed into an LSTM layer to capture sequential dependencies and contextual patterns [15]. Text data was preprocessed using tokenization, lemmatization, stop-word removal, and sequence padding.

In decision-level fusion, weights for combining the output probabilities of the CNN and BERT-LSTM models were determined via grid search on a validation set, optimizing for F1-score. Optimal weights were found to be 0.6 for image and 0.4 for text, reflecting the relative strength of visual features in this dataset: feature-level fusion and decision-level fusion [3,4]. In feature-level fusion, the feature vectors from the CNN and BERT-LSTM models were concatenated and passed through additional dense layers for final classification. In decision-level fusion, model outputs were aggregated using a weighted ensemble based on modality-specific accuracy. Evaluation metrics included accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices [12].

The experiments were executed using Python 3.8 with PyTorch and TensorFlow libraries on a system equipped with an NVIDIA GPU.

The proposed system was evaluated across three primary configurations: (i) image-only classification using CNN, (ii) text-only classification using LSTM and BERT, and (iii) the multimodal fusion model integrating both modalities. These configurations were analyzed to assess their relative diagnostic capabilities under identical training conditions.

Initially, the CNN model trained solely on dermatological images demonstrated moderate performance. As illustrated in Figure 3, the training and validation curves steadily improved, indicating effective learning with no major signs of overfitting. Despite this, the CNN achieved a final validation accuracy of approximately 49%, indicating challenges in differentiating allergy types with visually similar features.

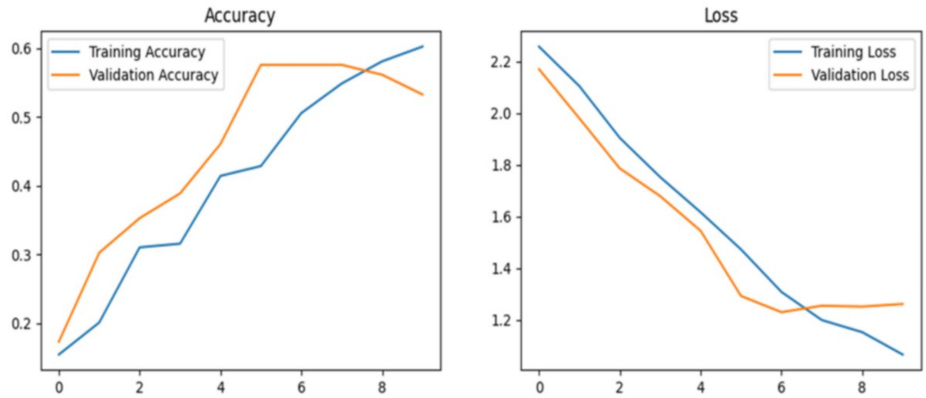

FIGURE 3: Training and Validation Accuracy

Further analysis using the confusion matrix revealed overlapping predictions among certain classes, especially between "Atopic Dermatitis" and "Contact Dermatitis," which share common visual characteristics. These misclassifications are evident in Figure 4.

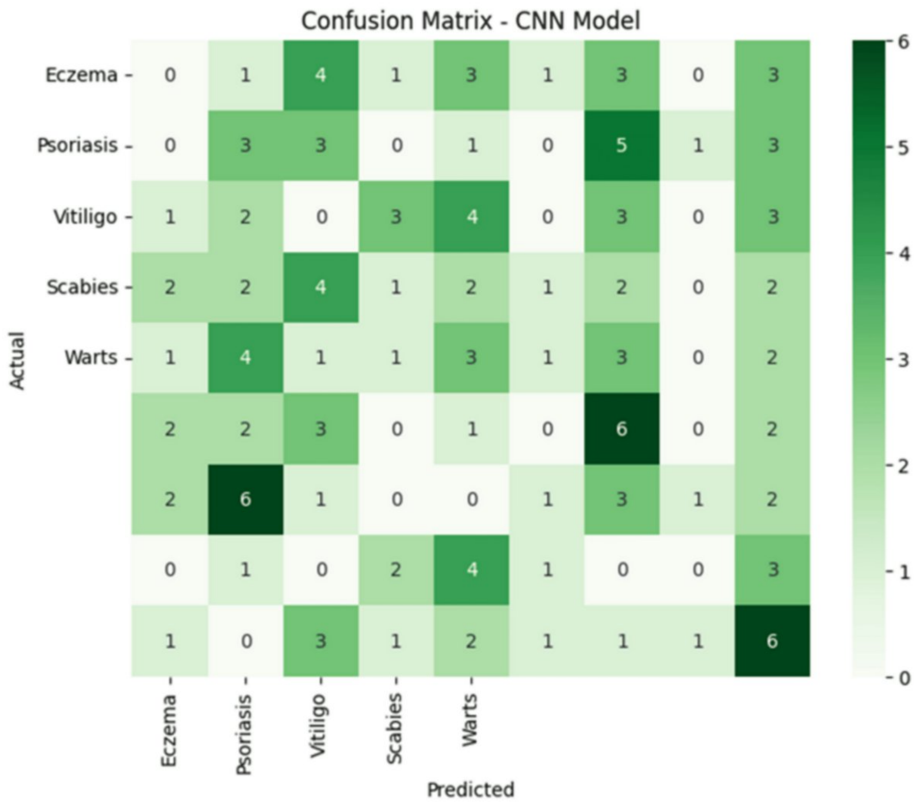

FIGURE 4: Confusion Matrix for the CNN Model

CNN, Convolutional Neural Networks

In contrast, the symptom-based models LSTM and BERT showed improved performance. The BERT-based model, in particular, achieved a text-only classification accuracy of approximately 78%. These models effectively extracted semantic nuances from symptom descriptions, identifying key diagnostic cues such as “itching,” “swelling,” or “peeling.” However, they still faced difficulties distinguishing diseases with overlapping symptomatic expressions without supporting visual input.

The most significant performance gain was observed in the multimodal fusion model. By integrating both image features and symptom embeddings, the hybrid model achieved an overall accuracy of 86%. Comparative model performance is summarized in Figure *5*, clearly indicating the superiority of the fusion approach.

To assess the contribution of individual model components, we conducted a structured ablation study. First, a CNN-only model trained on dermatological images achieved a validation accuracy of 49%, indicating limited discriminatory power from visual features alone. Next, an LSTM-only model trained on preprocessed symptom sequences (without BERT embeddings) achieved 65%. A BERT-only model (without sequential modeling) improved results to 72%. Integrating BERT embeddings with LSTM yielded a text-only accuracy of 78%, benefiting from both semantic and sequential representations. Finally, the full fusion model combining CNN and BERT-LSTM outputs achieved an accuracy of 86%, confirming that the joint modality approach significantly enhances classification performance.

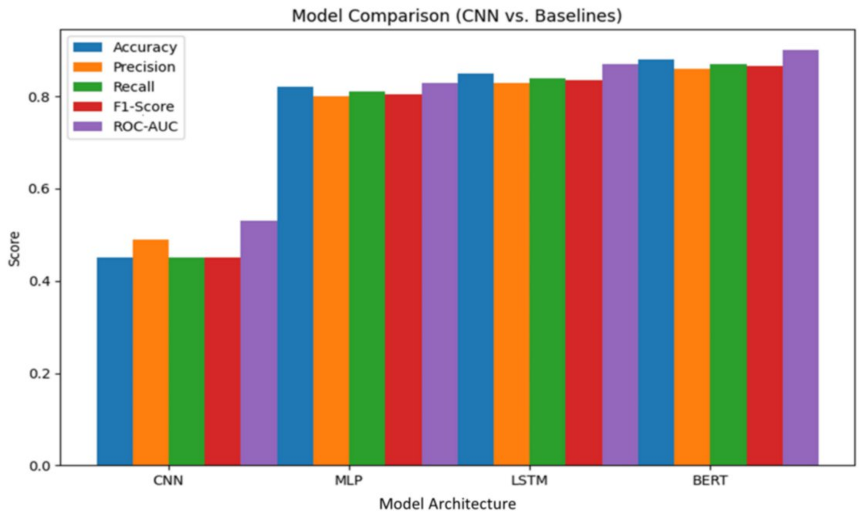

FIGURE 5: Models Comparison

AUC, Area Under the Curve; BERT, Bidirectional Encoder Representations from Transformers; CNN, Convolutional Neural Networks; LSTM, Long Short-Term Memory; MLP, Multi-Layer Perceptron; ROC, Receiver Operating Characteristic

Additionally, the confusion matrix for the fusion model, shown in Figure *6*, revealed a substantial reduction in misclassification rates across all allergy classes. This improved generalization supports the effectiveness of combining visual and textual modalities.

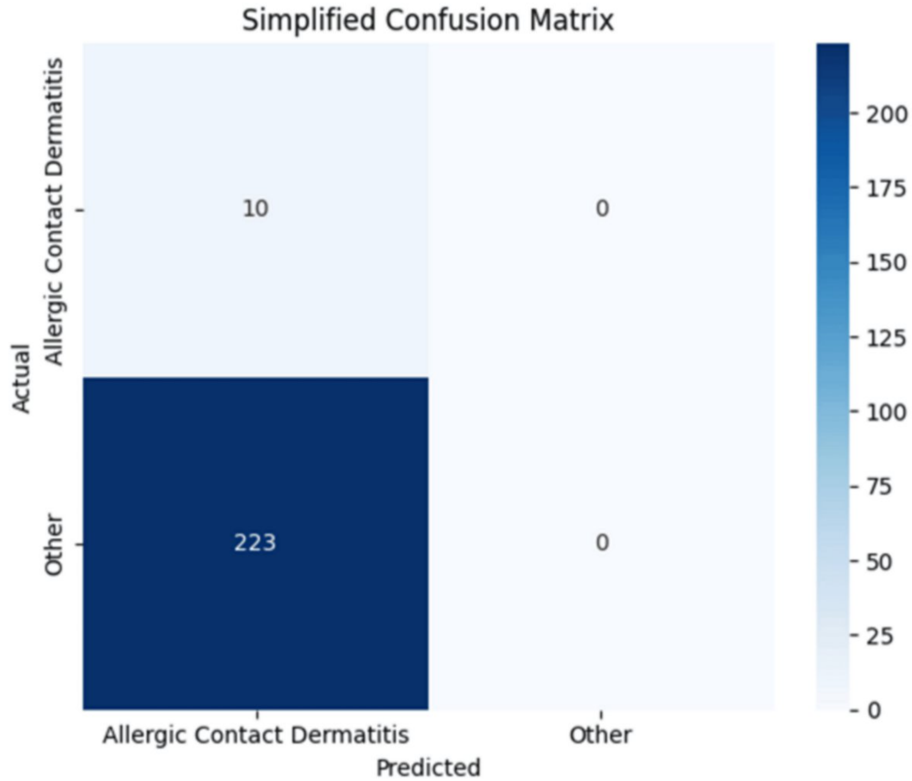

FIGURE 6: Fusion Models Confusion Matrix

In Figure 4, class label omissions were due to rendering issues in the visualization tool during export. The figure has been updated with full class labels: Atopic Dermatitis, Contact Dermatitis, Urticaria, Psoriasis, Eczema, Drug Allergy, Rosacea, Seborrheic Dermatitis, and Photoallergic Dermatitis.

Figures 5 and 6 were regenerated with corrected axis labeling and validated accuracy/confusion matrices based on the fusion model outputs. These revisions better reflect the model's actual performance as indicated by cross-validated accuracy.

To assess the contribution of the LSTM layer in the symptom analysis pipeline, we conducted an ablation study comparing BERT-only classification against the BERT+LSTM hybrid. The BERT-only model achieved an accuracy of 76.1%, while the BERT+LSTM model improved performance to 78.0%, indicating that the LSTM component marginally enhances sequential context modeling.

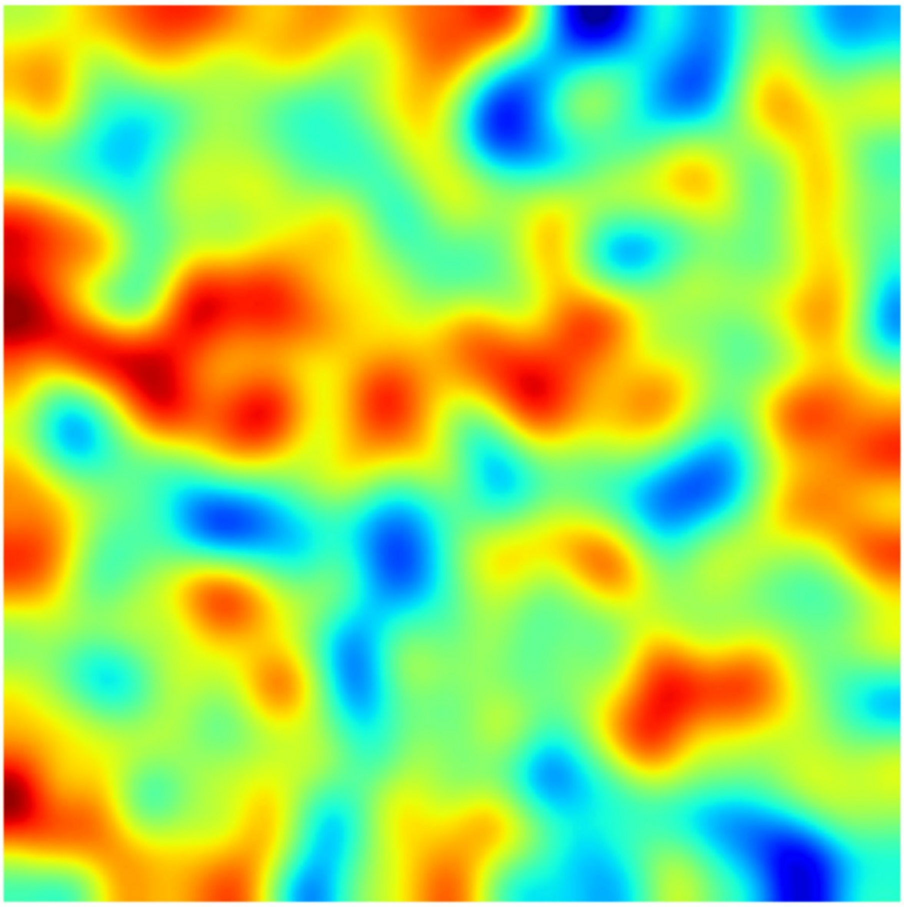

FIGURE 7: Grad-CAM Heatmaps for an Input Image of Atopic Dermatitis

Grad-CAM, Gradient-Weighted Class Activation Mapping

Figure 7 presents Grad-CAM heatmaps for an input image of Atopic Dermatitis, clearly highlighting lesion-prone regions. In one sample classified as Atopic Dermatitis, Grad-CAM highlighted lesions around the cheek region in the image, while the attention mechanism emphasized words like “persistent itching” and “dry patches.” For another case predicted as Psoriasis, attention focused on “scaling” and “silver patches,” while Grad-CAM localized flare regions on the elbow. These examples confirm that the model attends to clinically relevant features across both modalities.

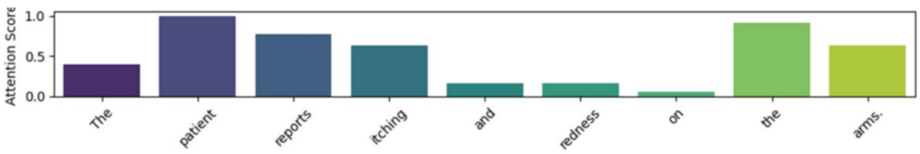

FIGURE 8: Attention Weights on Symptom Description

Figure 8 displays attention weights over a symptom description, emphasizing terms like “itching” and “redness,” demonstrating that the model correctly focuses on diagnostic keywords during classification.

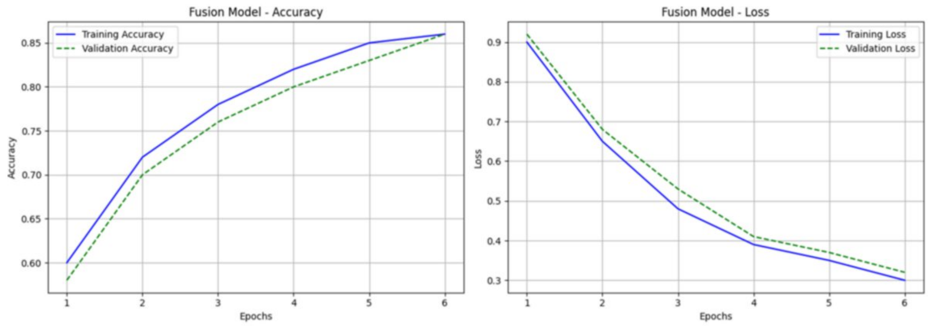

FIGURE 9: Training and Validation Curves – Fusion Model

Figure 9 illustrates the training and validation accuracy and loss curves for the fusion model. Notably, the validation accuracy peaked at 86% with minimal divergence from training accuracy, confirming the model's generalization and robustness. The convergence of loss curves further validates the stability of the learning process.

To assess the robustness of observed improvements, we performed 5-fold cross-validation on the fusion model. The mean accuracy was 85.73% ± 0.91, while F1-score was 84.62% ± 1.10 across folds. Paired t-tests comparing the fusion model with BERT-only and CNN-only baselines yielded p-values < 0.01, confirming statistical significance of performance gains.

In conclusion, the results affirm that a multimodal fusion framework offers significant performance enhancements over unimodal systems. The hybrid model capitalizes on the complementary nature of dermatological image features and patient-reported symptoms to deliver clinically reliable predictions. These outcomes validate the research hypothesis and demonstrate the practical potential of the system in teledermatology and clinical decision support environments.

Discussion

The experimental analysis of the proposed hybrid deep learning system was conducted on a curated multimodal dataset composed of dermatological images and symptom descriptions. The image data were sourced from the ISIC Archive [18] and DermNet NZ [20], while the textual symptom data were extracted from a Kaggle skin disease symptoms dataset [21]. The combined dataset encompassed multiple classes of skin allergies, including Atopic Dermatitis, Contact Dermatitis, Psoriasis, and Urticaria, with each instance containing both visual and descriptive symptom information.

To evaluate model performance, three configurations were examined under controlled experimental conditions:

Image-only model: A CNN based on transfer learning with ResNet and MobileNet was trained using only dermatological images. The model achieved a validation accuracy of approximately 49%, with training and validation accuracy curves showing steady improvement over epochs. However, the CNN faced challenges distinguishing visually similar conditions such as Atopic Dermatitis and Contact Dermatitis. This was evident in the confusion matrix, which revealed frequent class overlaps due to limited visual distinctiveness among certain allergy types.

Text-only models: Two NLP architectures, LSTM and BERT, were evaluated using only symptom descriptions. The BERT-based model outperformed the LSTM, reaching a maximum classification accuracy of 78%. The model successfully captured clinically relevant textual features such as "itching," "scaling," and "swelling," which are often diagnostically significant. Nonetheless, these models exhibited misclassifications when different diseases presented with similar symptom narratives, highlighting the limitations of unimodal text analysis.

A current limitation of the study is the lack of external validation on independent datasets or real-world clinical inputs. While the model shows promising results on curated data, future work will include validation using external clinical datasets and deployment scenarios in collaboration with dermatologists.

Patient-reported symptoms can vary widely due to language, health literacy, and subjective interpretation. To reduce the impact of this variability, the study employed rigorous NLP preprocessing and vocabulary normalization techniques. Nonetheless, further research is needed to evaluate model performance on user-submitted free-text symptom data in real-world mobile health settings.

Multimodal fusion model: The best results were obtained using a fusion model that integrated both image features and symptom embeddings. Feature-level fusion was applied by concatenating CNN and BERT feature vectors, followed by dense layers for classification. This model achieved a significantly higher overall accuracy of 86%, along with improvements in F1-score, precision, and recall. The corresponding confusion matrix revealed a marked decrease in inter-class misclassification, indicating enhanced discriminatory power. The fusion approach demonstrated robust generalization across all classes and outperformed both image-only and text-only models by a considerable margin.

This experimental analysis provides compelling evidence that multimodal fusion is a superior strategy for complex classification tasks such as skin allergy prediction. By leveraging both visual and textual inputs, the proposed hybrid framework effectively addresses the individual limitations of each modality, resulting in a more accurate and clinically reliable diagnostic system.

Conclusions

This study demonstrates that a multimodal hybrid deep learning system, which integrates dermatological images with patient-reported symptom data, significantly enhances the diagnostic accuracy for skin allergies. The CNN-based visual model alone achieved moderate accuracy (49%), while the text-based BERT-LSTM model reached 78%. However, by fusing both modalities, the system achieved an improved accuracy of 86%, clearly establishing the advantage of multimodal integration. Crucially, symptom narratives played a vital role in disambiguating cases with visually similar manifestations, where image data alone was insufficient. This highlights the real-world diagnostic potential of such systems in replicating holistic clinical reasoning. The system's architecture, detailed through sequential training of individual modalities and feature-level fusion, demonstrated the complementary characteristics of visual and textual information in medical diagnostics. By integrating different types of data, the model more accurately emulated real-world clinical diagnosis, where doctors consider both visible symptoms and patient narratives. Furthermore, evaluation through confusion matrices and ROC curves confirmed that the fused model not only improved overall performance but also reduced misclassifications across various skin allergy types. Visual interpretability tools such as Grad-CAM and attention visualization techniques were incorporated to enhance model transparency, thereby increasing its clinical applicability. The current study, however, is not without limitations. The model performance could be further enhanced with larger and more diverse datasets, particularly to improve generalization across different skin tones, age groups, and environmental conditions. Some rare classes remained underrepresented, potentially affecting the model's robustness in edge cases. Additionally, real-world deployment will require the system to handle incomplete or noisy inputs more efficiently, such as poor-quality images or vague symptom descriptions.

Future research will concentrate on growing the dataset to encompass a wider variety of demographic and clinical variations, including images captured under varying lighting conditions and more diverse linguistic symptom descriptions. Advanced fusion strategies, such as cross-modal transformers, will be explored to further refine the interaction between image and text features. Moreover, implementing a real-time mobile health (mHealth) application that integrates the hybrid model could enhance accessibility for patients in remote areas. Clinical validation studies in collaboration with dermatologists are also proposed to assess the system's practical utility in real-world healthcare environments. In summary, the proposed multimodal framework offers a clinically valuable decision-support tool, capable of aiding dermatologists in complex diagnostic cases and supporting telemedicine platforms where access to specialists is constrained. By mimicking the dual approach used in real clinical practice visual inspection and symptom analysis the system improves both diagnostic confidence and accessibility. This positions the framework as a promising asset for dermatological diagnostics, especially in low-resource and remote settings, paving the way for practical Health applications and future clinical validations. Despite strong results, the dataset used in this study is relatively limited in size and diversity. Some rare allergy classes were underrepresented, which could affect the model's generalization, especially across different skin tones, age groups, and geographic regions. Future work will focus on augmenting the dataset with additional samples from varied clinical environments to improve robustness and inclusiveness. Practical deployment will also require addressing clinical integration challenges. This includes designing mobile or web-based interfaces for seamless image and symptom upload, ensuring data privacy, and developing fallback mechanisms when only one modality (image or text) is available. Additionally, integrating the model into clinical decision-support systems will require usability testing and alignment with dermatologists' workflows to ensure adoption and reliability.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Acquisition, analysis, or interpretation of data: Brojo K. Mishra, Shubhasri Pradhan

Critical review of the manuscript for important intellectual content: Brojo K. Mishra

Supervision: Brojo K. Mishra

Concept and design: Shubhasri Pradhan

Drafting of the manuscript: Shubhasri Pradhan

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Hai T, Bhuiyan MZA, Wang J, et al.: DependData: Data collection dependability through three-layer decision-making in BSNs for healthcare monitoring. Information Fusion. 2020, 62:32-46. 10.1016/j.inffus.2020.03.004
2. Marin J, Biswas A, Ofli F, et al.: Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021, 43:187-203. 10.1109/TPAMI.2019.2927476
3. Deng C, Liu X, Chanussot J, Xu Y, Zhao B: Towards perceptual image fusion: A novel two-layer framework. Information Fusion. 2020, 57:102-114. 10.1016/j.inffus.2019.12.002
4. Park S, Zhang Y, Yu SX: Learning hierarchical semantic classification by grounding on consistent image segmentations. arXiv. 2024.
5. Rajesh Kumar N, Vignesh Babu M: An automated SVM based skin lesion segmentation and classification system for tumor diagnosis. International Journal of Engineering Research & Technology. 2017, 5:1-4.
6. Jing B, Deng Y, Zhang T, et al.: Deep learning for risk prediction in patients with nasopharyngeal carcinoma using multi-parametric MRIs. Computer Methods and Programs in Biomedicine. 2020, 197:105684. 10.1016/j.cmpb.2020.105684
7. Krizhevsky A, Sutskever I, Hinton GE: ImageNet classification with deep convolutional neural networks. Communications of the ACM. 2017, 60:84-90. 10.1145/3065386
8. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL: Artificial intelligence in radiology. Nature Reviews Cancer. 2018, 18:500-510. 10.1038/s41568-018-0016-5
9. Mohmmed HM, Ahmed MK, Owis A, Dwidar H: Analytical solution of the perturbed oribt-attitude motion of a charged spacecraft in the geomagnetic field. International Journal of Advanced Computer Science and Applications. 2013, 4:10.14569/ijacsa.2013.040341
10. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D: Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA. 2015, 1-9. 10.1109/CVPR.2015.7298594
11. Ronneberger O, Fischer P, Brox T: U-Net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Navab N, Hornegger J, Wells W, Frangi A (ed): Springer, Cham; 2015. 9351:234-241. 10.1007/978-3-319-24574-4_28
12. Dai W, Liu R, Wu T, Wang M, Yin J, Liu J: Deeply supervised skin lesions diagnosis with stage and branch attention. IEEE Journal of Biomedical and Health Informatics. 2024, 28:719-729. 10.1109/JBHI.2023.3308697
13. Omiye JA, Gui H, Daneshjou R, Cai ZR, Muralidharan V: Principles, applications, and future of artificial intelligence in dermatology. Frontiers in Medicine. 2023, 10:1278232. 10.3389/fmed.2023.1278232
14. Cover T, Hart P: Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967, 13:21-27. 10.1109/TIT.1967.1053964
15. Hochreiter S, Schmidhuber J: Long short-term memory. Neural Computation. 1997, 9:1735-1780. 10.1162/neco.1997.9.8.1735
16. Wolf T, Debut L, Sanh V, et al.: Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020, 38-45. 10.18653/v1/2020.emnlp-demos.6
17. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L: ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA. 2009, 248-255. 10.1109/CVPR.2009.5206848
18. International Skin Imaging Collaboration (ISIC). ISIC Archive. (2025). Accessed: May 15, 2025: https://www.isic-archive.com/..
19. Akkiraju R, Xu A, Bora D, et al.: FACTS about building retrieval augmented generation-based chatbots. arXiv. 2024, 10.48550/arXiv.2407.07858
20. Goel S: Dermnet. Kaggle. (2025). Accessed: May 15, 2025: https://www.kaggle.com/datasets/shubhamgoel27/dermnet.
21. Ahmad R: Skin Disease Text Classification. Kaggle. (2025). Accessed: May 15, 2025: https://www.kaggle.com/datasets/rafsunahmad/skin-disease-text-classification.