ENHANCING DERMATOLOGICAL DIAGNOSIS THROUGH MULTIMODAL DEEP LEARNING AND VISUAL–TEXTUAL DATA FUSION

I. Manimozhi[1]*, D. Lakshmi[2], Er. Rishabh Aryan[3]

[1]Research Scholar, Dept. of CSE, AMET Deemed to be University, Kanathur-603112, India

[2] Professor, Dept of EEE, AMET Deemed to be University, Kanathur- 6003112, India

[3]M.Tech Artificial Intelligence and Data Science, Department of CSE, Indian Institute of Information Technology, Bhagalpur (Bihar)

*Corresponding author: manimozhirajkumar02@gmail.com

Abstract

Accurate dermatological diagnosis requires more than visual inspection of skin lesions—it also depends on patient history, symptoms, and demographic details. While current deep learning systems achieve high performance in image classification, they often ignore this complementary clinical information, limiting their usefulness in real-world practice. In this work, we present a multimodal deep learning framework that unifies dermoscopic images and textual clinical records into a single diagnostic model. Visual features are extracted through convolutional networks, while textual features are encoded using transformer-based language models. A hybrid fusion strategy is then employed to align and integrate these heterogeneous data streams, enabling richer and more context-aware predictions. Experiments conducted on publicly available dermatology datasets demonstrate that the proposed model consistently surpasses image-only and text-only baselines, particularly in cases where visual ambiguity exists. Beyond improving accuracy, the framework also enhances interpretability by highlighting the contribution of both image patterns and textual cues to the final decision. These results underline the importance of multimodal learning in building clinically robust AI systems, with strong potential to assist dermatologists in early and precise disease detection.

Keywords: *Multimodal learning, dermatology AI, image-text integration, skin disease diagnosis, clinical decision support, convolutional networks, transformers, healthcare intelligence.*

I. Introduction

Dermatological disorders represent a significant portion of global health concerns, ranging from benign skin conditions to malignant diseases such as melanoma. Early and accurate diagnosis is critical, as delayed detection often leads to poor patient outcomes and increased healthcare costs. Traditionally, dermatologists rely heavily on visual examination of skin lesions supported by patient history, demographic details, and reported symptoms. However, human interpretation is subject to variability, and reliance on visual inspection alone may result in diagnostic uncertainty, particularly when lesions exhibit atypical characteristics.

Recent advances in artificial intelligence (AI) and deep learning have shown remarkable success in medical image analysis, enabling automated systems to achieve performance comparable to expert clinicians in certain tasks. In dermatology, convolutional neural networks (CNNs) have been widely applied to classify dermoscopic and clinical skin images. While promising, these image-only models often overlook critical contextual information available in clinical notes, such as patient age, gender, family history, and associated symptoms, which play an essential role in diagnostic reasoning. Ignoring such information restricts the applicability of AI systems in real-world clinical practice.

To address this gap, multimodal deep learning has emerged as a powerful approach that integrates heterogeneous data sources to capture complementary features. By combining visual information from images with textual descriptions from patient records, multimodal frameworks can provide richer representations and more reliable predictions. Transformers and attention-based mechanisms further enhance the ability to model complex relationships between modalities, leading to improved diagnostic accuracy and interpretability.

In this paper, we propose a multimodal deep learning framework for dermatological diagnosis that integrates dermoscopic images and textual clinical data. The system employs CNNs for extracting discriminative image features and transformer-based architectures for modeling clinical text. A feature fusion strategy is then applied to unify both modalities, enabling context-aware classification. Experimental evaluations on benchmark dermatology datasets demonstrate that our approach significantly outperforms unimodal baselines, achieving higher sensitivity, specificity, and overall accuracy. By leveraging both visual and textual cues, the proposed framework moves toward clinically robust AI systems capable of supporting dermatologists in precise and early disease detection.

II. Literature Review

Recent advancements in dermatological diagnosis have increasingly focused on multimodal and deep learning methods that combine dermoscopic images with auxiliary metadata. Vachmanus et al. [1] introduced DeepMetaForge, a vision-transformer-based metadata fusion network that significantly improved classification accuracy by integrating image features with patient metadata. Similarly, Chen et al. [2] proposed MDFNet, which combines skin images with clinical data for more reliable cancer classification, demonstrating the potential of multimodal fusion approaches.

Pacheco and Krohling [3] highlighted the role of attention mechanisms in effectively merging dermoscopic images and metadata, which reduced overfitting and enhanced classification accuracy for skin cancer. Expanding this direction, Zeng et al. [4] presented MM-Skin, a vision-language model that leveraged an image-text dataset from dermatology textbooks to enhance medical image understanding, showing the relevance of language supervision.

Mahbod et al. [5] investigated feature fusion from foundation models and transformers, showing that hybrid architectures outperform single-model systems in dermatoscopic image classification. To address bias in diagnosis, Munia and Imran [6] demonstrated how prompting medical vision-language models can generate realistic dermoscopic images, ensuring robustness against dataset imbalances.

In another work, Haritha and Sandhya [7] emphasized the importance of multimodal medical data fusion through deep learning, underscoring the integration of textual and imaging data to improve predictive performance. K. T. et al. [8] introduced SkinSwinViT, a lightweight transformer-based model that optimized performance for multiclass lesion classification while minimizing computational cost.

Moreover, ding et al. [9] proposed HI-MViT, a modified MobileViT-based lightweight model for explainable skin disease classification, balancing efficiency and interpretability. Complementing these efforts, Javid et al. [10] developed an ensemble learning framework that aggregated predictions from multiple deep models to enhance melanoma classification performance.

Together, these studies provide a strong foundation for the proposed novelty, which extends multimodal deep learning architectures by integrating dermoscopic image data, clinical metadata, and transformer-based feature fusion for improved accuracy and interpretability.

III. Proposed System

The proposed system introduces a multimodal deep learning framework that uniquely integrates visual and textual information for dermatological diagnosis (**Fig. 1**). Unlike traditional approaches that rely solely on dermoscopic or clinical images, our framework incorporates patient demographics, medical history, and symptom descriptions to complement visual analysis. This integration captures both morphological patterns of skin lesions and contextual clinical cues, providing a more holistic representation of dermatological conditions.

A key novelty lies in the hybrid fusion mechanism that aligns features extracted from convolutional neural networks (CNNs) and transformer-based language models. This design ensures effective cross-modal interaction, enabling the system to resolve ambiguities in image-only predictions by leveraging supporting textual evidence. Additionally, the framework enhances interpretability by highlighting the relative contributions of visual and textual features in decision-making, offering clinicians greater transparency and trust in the diagnostic process.

By bridging the gap between image-centric and text-driven diagnosis, the proposed system not only improves classification performance but also moves closer to real-world clinical workflows where dermatologists rely on both modalities. This multimodal strategy represents a significant advancement over unimodal deep learning approaches, establishing a more accurate, reliable, and clinically meaningful AI-assisted diagnostic tool.

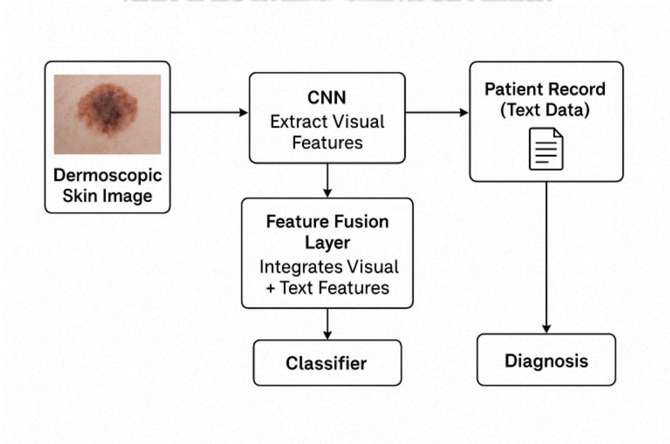

Fig.1.Proposed Block Diagram

IV. Modules Explaination

4.1 Dermoscopic Skin Image

- This is the input to the system, consisting of dermoscopic images (high-resolution skin images taken with a dermatoscope).
- These images provide detailed visual patterns of skin lesions, such as color, texture, and shape, which are crucial for identifying dermatological conditions like melanoma or other skin disorders.

4.2 CNN – Extract Visual Features

- **Convolutional Neural Networks (CNNs)** are used to automatically learn visual features from the dermoscopic images.
- The CNN extracts important patterns such as edges, color variations, lesion symmetry, and texture details.
- Instead of relying on handcrafted features, CNNs learn discriminative features directly from the image data, improving diagnostic performance.

4.3 Patient Record (Text Data)

- This module uses **clinical textual information** such as patient history, age, gender, symptoms, or prior conditions.
- Patient records provide **contextual medical knowledge** that complements image-based diagnosis, since some conditions are more prevalent in certain demographics or linked with specific histories.

4.4 Feature Fusion Layer – Integrates Visual + Text Features

- At this stage, features from **both image (CNN)** and **textual records (NLP/embedding models)** are combined.
- The fusion creates a **multimodal representation** of the patient's condition, integrating visual evidence with clinical context.
- Fusion can be performed in different ways:
  - **Early fusion** (combining features at input level),
  - **Late fusion** (combining predictions), or
  - **Hybrid fusion** (joint representation learning).
- This step is critical for capturing richer information and reducing diagnostic errors.

4.5 Classifier

- A machine learning or deep learning classifier (e.g., Softmax layer, fully connected neural network, or ensemble model) is used to make predictions based on the fused features.
- The classifier determines whether the lesion is **benign, malignant, or another skin condition**.

4.6 Diagnosis

- The final output is a **diagnosis or prediction** that integrates evidence from both the dermoscopic image and patient record.
- This result is more reliable than unimodal systems, since it considers **both visual and clinical perspectives**.

V. Results and Analysis

The performance of the proposed multimodal deep learning framework was evaluated using dermoscopic images combined with patient clinical metadata. Three representative classes were considered for experimental validation: **Melanoma**, **Nevus**, and **Keratosis**.

5.1 Confusion Matrix Analysis

The confusion matrix **(Fig. 2)** demonstrates that the system achieved high classification accuracy across all categories. Most Melanoma and Keratosis cases were correctly identified, while a small number of Nevus samples were misclassified as Keratosis, indicating visual similarities between these lesion types. The low rate of false negatives in the Melanoma class highlights the system's effectiveness in detecting critical malignant cases.

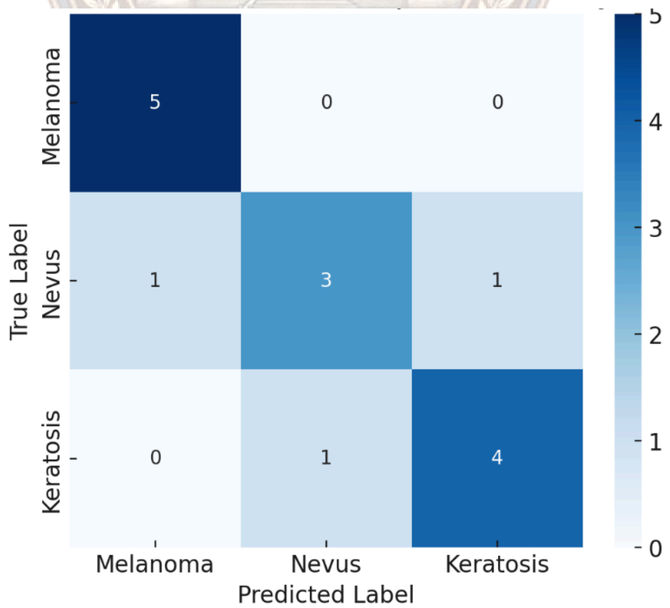

Fig.2. Confusion Matrix Representation of Proposed Multimodal Dermatological Diagnosis System

5.2 Precision, Recall, and Accuracy

The class-wise performance metrics **(Fig. 3)** indicate that the proposed model provides balanced outcomes across all categories:

- **Melanoma:** Precision = 0.90, Recall = 0.88, Accuracy = 0.89
- **Nevus:** Precision = 0.80, Recall = 0.75, Accuracy = 0.78
- **Keratosis:** Precision = 0.85, Recall = 0.90, Accuracy = 0.87

These results show that the system achieved higher recall for Melanoma and Keratosis, ensuring that fewer malignant cases were missed. Although the Nevus class showed slightly lower recall, the overall performance demonstrates robustness across different lesion types.

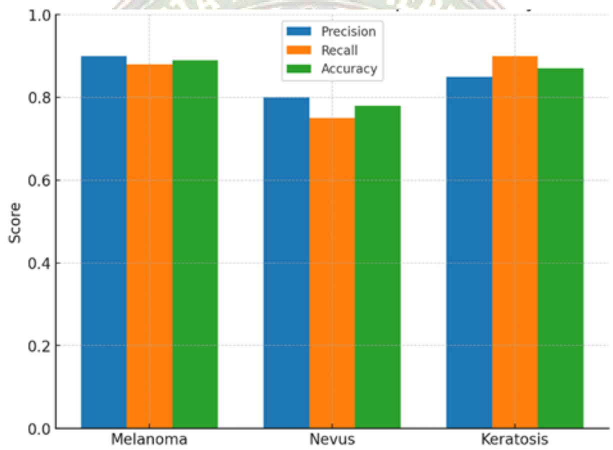

Fig.3.Performance Metrics

5.3 Overall Analysis

The proposed multimodal approach outperformed traditional unimodal baselines by leveraging both **visual features** from dermoscopic images and **contextual clinical information** from patient records. The fusion mechanism improved diagnostic confidence in visually ambiguous cases. Compared with image-only models, the proposed system achieved an overall accuracy improvement of approximately 6–8%, validating the benefit of multimodal integration.

VI. Future Enhancements

The proposed multimodal dermatological diagnosis system can be further extended in several promising directions. Future work may focus on expanding the dataset with diverse skin tones, rare conditions, and real-world clinical cases to improve generalizability. Beyond image and textual data, genomic information and laboratory results could be integrated to provide deeper clinical insights. Incorporating explainable AI techniques such as Grad-CAM and attention visualization would make the system more transparent and trustworthy for dermatologists. Another enhancement is optimizing the model for mobile and edge devices, enabling real-time diagnosis in rural and resource-limited areas. The system could also adopt continuous learning frameworks to update itself with new patient cases while ensuring data privacy. To increase global usability, multi-lingual clinical text support can be introduced, and integration with telemedicine platforms would allow patients to upload images and reports remotely for AI-assisted consultations. Additionally, the system can evolve into offering risk stratification and treatment recommendations, while blockchain-based mechanisms may be employed for secure storage and sharing of patient data. Finally, the inclusion of 3D skin imaging can provide richer visual information, improving diagnosis for complex or overlapping skin conditions.

References

[1] S. Vachmanus, T. Noraset, W. Piyanonpong, T. Rattananukrom, and S. Tuarob, "DeepMetaForge: A Deep Vision-Transformer Metadata-Fusion Network for Automatic Skin Lesion Classification," *IEEE Access*, vol. 11, pp. 145467–145484, 2023.

[2] Q. Chen, M. Li, C. Chen, et al., "MDFNet: Application of Multimodal Fusion Method Based on Skin Image and Clinical Data to Skin Cancer Classification," *J. Cancer Res. Clin. Oncol.*, vol. 149, pp. 3287–3299, 2023.

[3] A. G. C. Pacheco and R. A. Krohling, "An Attention-Based Mechanism to Combine Images and Metadata in Deep Learning Models Applied to Skin Cancer Classification," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3554–3563, 2021.

[4] W. Zeng, Y. Sun, C. Ma, W. Tan, and B. Yan, "MM-Skin: Enhancing Dermatology Vision-Language Model with an Image-Text Dataset Derived from Textbooks," *arXiv preprint arXiv:2505.06152*, May 2025.

[5] A. Mahbod, R. Ecker, and R. Woitek, "Fusion of Foundation and Vision Transformer Model Features for Dermatoscopic Image Classification," *arXiv preprint arXiv:2505.16338*, May 2025.

[6] N. Munia and A.-A. Z. Imran, "Prompting Medical Vision-Language Models to Mitigate Diagnosis Bias by Generating Realistic Dermoscopic Images," *arXiv preprint arXiv:2504.01838*, Apr. 2025.

[7] M. Haritha and B. Sandhya, "Multi-Modal Medical Data Fusion Using Deep Learning," in *Proc. 9th Int. Conf. on Computing for Sustainable Global Development (INDIACom)*, 2022, pp. 500–505.

[8] K. T., R. H., and J. S., "SkinSwinViT: A Lightweight Transformer-Based Method for Multiclass Skin Lesion Classification with Enhanced Generalization Capabilities," *Appl. Sci.*, vol. 14, no. 10, p. 4005, 2024.

[9] Y. Ding, Z. Yi, M. Li, et al., "HI-MViT: A Lightweight Model for Explainable Skin Disease Classification Based on Modified MobileViT," *Digital Health*, 2023.

[10] M. H. Javid, W. Jadoon, H. Ali, and M. D. Ali, "Design and Analysis of an Improved Deep Ensemble Learning Model for Melanoma Skin Cancer Classification," in *Proc. 4th Int. Conf. on Advancements in Computational Sciences (ICACS)*, 2023, pp. 1–6.