

SkinDxFusion: An Integrated VGG19 and Vision Transformer Architecture for Dermatological Cancer Detection and GenAI-Based Reporting

Vaishnavi Rawoot¹, Dr.Gajendran Malshetty²

¹ Student, Department of Artificial Intelligence & Data Science, Sharnbasva University, Kalaburagi, Karnataka, India, vaishnavirawoot220@gmail.com

² Professor, Department of Artificial Intelligence & Data Science, Sharnbasva University, Kalaburagi, Karnataka, India, gajendra.malshetty@gmail.com

ABSTRACT

Skin cancer is one of the most common and rapidly increasing forms of cancer worldwide, making early and accurate diagnosis essential for improving patient outcomes. This paper presents SkinDxFusion, an intelligent dermatological analysis system that integrates a hybrid VGG19 and Vision Transformer (ViT) architecture with YOLOv11-based lesion detection and automated GenAI-driven medical report generation. The proposed hybrid model combines the local feature extraction capability of VGG19 with the global contextual learning ability of Vision Transformers, resulting in improved classification performance across multiple skin lesion categories. YOLOv11 is employed to accurately detect and localize lesion regions in both dermoscopic images and real-time video streams, enabling efficient clinical screening and practical deployment. To enhance usability, the system automatically generates structured dermatological reports containing predicted lesion type, confidence score, severity level, lesion characteristics, and clinical recommendations, reducing documentation effort and improving interpretability. The dataset consists of publicly available skin lesion images with appropriate preprocessing, normalization, augmentation, and class balancing techniques to improve model generalization. Experimental evaluation demonstrates that the proposed framework achieves superior accuracy, precision, and macro F1-score compared to standalone deep learning models while maintaining reliable detection performance. The developed web-based application provides a scalable, accessible, and decision-support solution for early skin cancer screening, assisting healthcare professionals in achieving faster, more consistent, and clinically meaningful dermatological diagnosis.

Keywords: Skin Cancer Detection, Deep Learning, VGG19, Vision Transformer, YOLOv11, Medical Image Analysis, Skin Lesion Classification, Automated Report Generation, Generative AI, Dermatology AI.

I. INTRODUCTION

Skin cancer has become one of the fastest-growing health challenges worldwide, with its incidence steadily increasing because of excessive ultraviolet radiation, environmental factors, and an aging population. Among different forms of skin cancer, melanoma is considered the most dangerous due to its high metastatic potential, making timely diagnosis essential for improving survival rates. Conventional diagnostic procedures primarily depend on visual examination, dermoscopy, and the experience of dermatologists. Although these methods are widely adopted, they are often affected by subjective interpretation, diagnostic variability, and limited accessibility in resource-constrained healthcare environments. Consequently, there is a growing demand for intelligent computer-aided diagnostic systems capable of providing accurate, reliable, and scalable support for dermatological screening. Recent progress in artificial intelligence, particularly deep learning, has significantly enhanced medical image analysis by enabling automatic extraction of complex visual patterns that are often difficult to identify through manual assessment alone [1]. Convolutional Neural Networks (CNNs) have achieved remarkable success in skin lesion classification because of their capability to learn hierarchical image representations, including edges, textures, shapes, and color distributions. Among various CNN architectures, VGG19 has demonstrated strong performance in capturing fine-grained local features from dermoscopic images. Nevertheless, CNN-based models primarily focus on local receptive fields and often fail to effectively model long-range spatial relationships that are crucial for differentiating lesions with similar visual characteristics [2]. To overcome this limitation, Vision Transformers (ViTs) have emerged as an effective alternative by utilizing self-attention mechanisms to capture global contextual information across the entire image. Although Vision Transformers provide superior contextual understanding, they generally require large-scale datasets and rich feature representations for optimal learning, limiting their standalone applicability in several medical imaging

tasks [3]. Recent research has therefore focused on hybrid deep learning architectures that combine the complementary strengths of CNNs and Vision Transformers. In these models, CNN layers are responsible for extracting detailed local features, whereas transformer modules learn global dependencies through attention mechanisms. This fusion enables more discriminative feature learning, resulting in improved classification accuracy, robustness, and generalization, particularly for visually similar benign and malignant skin lesions. Hybrid architectures have consequently emerged as a promising solution for enhancing automated skin lesion classification and reducing diagnostic inconsistencies in clinical practice [4]. Beyond lesion classification, precise localization of suspicious skin regions is equally important for practical clinical deployment. Modern object detection algorithms, particularly YOLOv11, provide rapid and highly accurate lesion detection by identifying affected regions with bounding boxes in both still images and live video streams. Its real-time inference capability makes YOLOv11 highly suitable for teledermatology, mobile healthcare, and automated clinical screening applications. Integrating lesion detection with classification creates a comprehensive diagnostic framework capable of simultaneously identifying lesion locations and determining their corresponding disease categories [5]. Interpretability has become another essential requirement for medical artificial intelligence systems. Presenting only prediction labels may not provide sufficient information for effective clinical decision-making. Therefore, automated report generation has gained significant importance by converting prediction results into structured clinical summaries. Such reports include lesion category, prediction confidence, severity assessment, lesion characteristics, and suggested clinical recommendations, thereby improving transparency, assisting healthcare professionals, and reducing documentation workload while promoting standardized reporting practices [6]. This work proposes SkinDxFusion, an integrated dermatological analysis framework that combines a hybrid VGG19-Vision Transformer classification model, YOLOv11-based real-time lesion detection, and Generative AI-driven automated report generation within a unified web-based platform. The proposed system supports classification of multiple benign and malignant skin lesion categories using publicly available datasets. Comprehensive preprocessing techniques, including image normalization, augmentation, and class balancing, are employed to improve model robustness and generalization. The effectiveness of the proposed framework is evaluated using standard performance metrics such as accuracy, macro F1-score, precision, recall, and Area Under the ROC Curve (AUC) to ensure reliable performance across all lesion categories [7]. Furthermore, the developed web application enables users to upload dermoscopic images, perform real-time lesion analysis, and receive structured diagnostic reports through an intuitive interface. This end-to-end system demonstrates the practical integration of artificial intelligence into dermatological workflows, providing an accessible and scalable solution for early skin cancer screening and clinical decision support. Although the proposed framework achieves promising performance, it is designed to assist healthcare professionals rather than replace expert clinical diagnosis. Future improvements will focus on expanding dataset diversity, enhancing model robustness, and conducting extensive clinical validation to support safe and reliable deployment in real-world healthcare settings [8].

II. RELATED WORKS

Recent advances in artificial intelligence have significantly transformed skin cancer detection and automated dermatological analysis, particularly through developments in deep learning for image classification, lesion detection, and automated clinical reporting. Earlier research mainly depended on conventional image processing techniques that extracted handcrafted features such as color distributions, texture descriptors, edge information, and shape characteristics. Although these approaches provided a foundation for computer-aided diagnosis, they often exhibited limited robustness when handling diverse lesion appearances, complex backgrounds, and varying imaging conditions. The introduction of deep learning, especially Convolutional Neural Networks (CNNs), marked a major breakthrough by enabling automatic feature learning directly from raw medical images, thereby achieving considerable improvements in diagnostic accuracy and reliability [9]. Pretrained architectures including VGG, ResNet, and Inception have demonstrated excellent performance on widely used benchmark datasets such as HAM10000 and ISIC, illustrating the effectiveness of transfer learning for dermatological image classification [10]. Despite these achievements, conventional CNN architectures primarily focus on local feature extraction and frequently struggle to capture long-range spatial dependencies that are important for distinguishing lesions with highly similar visual characteristics. To overcome these shortcomings, researchers have increasingly incorporated attention mechanisms and transformer-based architectures into medical image analysis. Vision Transformers (ViTs) utilize self-attention to establish relationships among distant image regions, enabling comprehensive global feature representation. Several studies have shown that integrating CNN feature extractors with transformer encoders results in hybrid architectures capable of simultaneously learning local texture information and global contextual dependencies [11]. Such hybrid models have demonstrated superior performance in skin lesion classification because dermatological images often contain irregular boundaries, scattered pigmentation

patterns, and subtle structural variations. Furthermore, lightweight transformer variants have been introduced to reduce computational overhead while preserving classification accuracy, making them more suitable for practical healthcare applications and edge-device deployment [12]. Alongside image classification, lesion localization has become an equally important research direction. Object detection algorithms from the YOLO family have gained widespread acceptance due to their ability to perform fast and accurate real-time detection. The latest YOLOv11 architecture incorporates improved feature pyramid networks and optimized detection strategies that enhance localization performance even for low-contrast lesions and irregular lesion boundaries. Several studies have reported that combining YOLO-based detection with deep learning classifiers provides a complete diagnostic framework capable of both identifying lesion locations and determining disease categories, thereby improving clinical usability and supporting real-time dermatological screening applications [13]. The deployment of YOLOv11 in telemedicine and mobile healthcare platforms has further demonstrated its effectiveness for remote diagnosis and large-scale screening. Another major challenge addressed in previous research is class imbalance and dataset heterogeneity, both of which commonly affect medical image classification. Researchers have proposed various strategies including extensive data augmentation, synthetic image generation, oversampling techniques, and class-weighted loss functions to improve model robustness and ensure balanced learning across different lesion categories. Without appropriate balancing strategies, deep learning models tend to become biased toward majority classes, resulting in reduced sensitivity for rare but clinically significant conditions such as melanoma [14]. Advanced augmentation methods, including geometric transformations, brightness variation, color perturbation, and Generative Adversarial Networks (GANs), have been widely adopted to increase dataset diversity and minimize overfitting during model training. The interpretability of artificial intelligence systems has also emerged as an important research focus. Explainable AI techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) and attention visualization methods enable clinicians to understand the reasoning behind model predictions by highlighting diagnostically significant image regions. These visualization techniques improve transparency, increase clinician confidence, and demonstrate that the model focuses on medically relevant features instead of irrelevant background information [15]. The integration of explainability methods into hybrid CNN-transformer frameworks has further strengthened the reliability and clinical acceptance of automated diagnostic systems. In addition to image analysis, several recent studies have explored automated medical report generation to improve communication between AI systems and healthcare professionals. Encoder-decoder architectures and transformer-based language models have been employed to generate structured diagnostic reports directly from extracted visual features and classification outputs. These systems convert numerical predictions into clinically meaningful narratives that summarize lesion characteristics, confidence levels, disease severity, and suggested medical recommendations. Researchers have emphasized the importance of maintaining factual correctness, consistency, and clinical relevance while incorporating uncertainty estimation to avoid misleading conclusions [16]. More recently, the emergence of Large Language Models (LLMs) has significantly enhanced the fluency, completeness, and readability of automatically generated medical reports. Researchers have also investigated ensemble learning and multi-stage processing frameworks to further improve diagnostic performance. Ensemble approaches combine predictions from multiple deep learning models to enhance robustness and reduce prediction variance, whereas multi-stage pipelines sequentially integrate preprocessing, segmentation, classification, detection, explainability, and report generation into a unified workflow. These comprehensive systems have achieved competitive performance on benchmark datasets such as ISIC, demonstrating improved accuracy and reliability across diverse skin lesion categories [17]. Nevertheless, increased computational complexity and higher hardware requirements remain significant challenges for practical deployment, particularly in low-resource healthcare environments. Although remarkable progress has been achieved, several limitations continue to exist in current research. Most published studies depend on publicly available datasets that may not adequately represent variations in skin tone, lesion appearance, acquisition devices, and real-world clinical settings. In addition, differences in evaluation protocols and experimental methodologies make objective comparison between studies difficult. Most proposed systems have been validated under controlled laboratory conditions, while large-scale clinical evaluation remains relatively limited [18]. Addressing these issues requires collaborative efforts among researchers, dermatologists, and healthcare institutions to develop standardized datasets, establish common evaluation protocols, and conduct extensive real-world validation studies.

Overall, existing literature demonstrates a clear shift toward hybrid deep learning architectures, real-time detection frameworks such as YOLOv11, explainable artificial intelligence, and automated clinical report generation. These developments collectively improve the accuracy, interpretability, and scalability of computer-aided dermatological diagnosis. Continued advancements in these areas are expected to enhance early skin cancer detection, reduce diagnostic variability, and provide more effective decision-support systems for clinical dermatology practice [19].

III. DATASETS

The dataset employed in this research consists of a diverse collection of dermoscopic and clinical skin lesion images obtained from publicly available Kaggle repositories, providing broad coverage of different lesion categories and imaging conditions. It includes both malignant and benign skin lesions, such as melanoma, basal cell carcinoma, squamous cell carcinoma, actinic keratosis, nevus, vascular lesions, and other common dermatological disorders. The availability of multiple lesion classes enables the proposed model to learn discriminative features for accurate classification across various skin conditions. The collected images exhibit considerable diversity in terms of resolution, illumination, skin tone, background complexity, and image quality, thereby improving the model's ability to generalize to real-world clinical scenarios. Before training, all images undergo a comprehensive preprocessing pipeline that includes resizing to a uniform input resolution, pixel value normalization, and several data augmentation techniques such as image rotation, horizontal and vertical flipping, zooming, and color enhancement. These preprocessing operations increase the variability of the training data, reduce overfitting, and improve the robustness of the learned model. To address the inherent class imbalance commonly found in dermatological datasets, class-weighted loss functions and balanced sampling strategies are incorporated, ensuring that minority lesion categories receive adequate representation during training. The dataset is partitioned into training, validation, and testing subsets using stratified sampling to preserve the proportional distribution of each lesion category across all splits. In addition to the static image dataset, a Roboflow-based real-time dataset is utilized for training and evaluating the YOLOv11 detection module. This enables accurate localization and real-time detection of skin lesions in images and live video streams. The integration of both static and real-time datasets allows the proposed framework to achieve reliable lesion classification while supporting efficient deployment in practical dermatological screening and clinical environments.

IV. METHODOLOGY

The proposed methodology presents a comprehensive workflow for automated skin lesion analysis, beginning with data collection and preprocessing, followed by feature extraction, lesion classification, real-time detection, and automated report generation. The framework integrates image preprocessing, a hybrid VGG19 and Vision Transformer (ViT) classification model, YOLOv11-based lesion localization, and Generative AI-driven report synthesis into a single end-to-end pipeline. This unified architecture ensures efficient processing of skin lesion images while delivering accurate classification, precise localization, and structured clinical reports. Figure 1 illustrates the overall architecture of the proposed system, depicting the sequential flow of data through each processing stage and demonstrating the interaction between the classification, detection, and report generation modules.

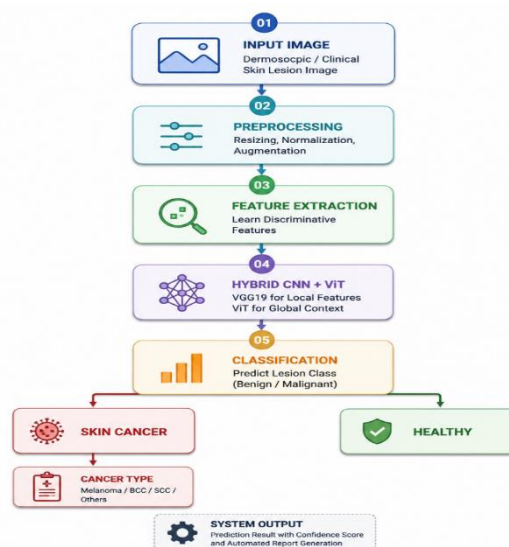


Fig. 1. Proposed SkinDxFusion system architecture illustrating the complete workflow from input image acquisition and preprocessing to hybrid CNN-ViT based skin lesion classification.

A. Data Collection

The data acquisition process involves collecting a comprehensive set of dermoscopic and clinical skin lesion images from publicly available repositories to ensure broad representation of different dermatological disorders. The dataset contains both benign and malignant lesion categories, allowing the proposed model to learn distinctive visual characteristics for accurate classification. Images are gathered under varying conditions, including differences in illumination, image resolution, background complexity, and skin tone diversity, thereby improving the robustness and generalization capability of the model. In addition, a Roboflow-based real-time dataset is incorporated for training and evaluating the YOLOv11 detection module, enabling reliable lesion localization in dynamic environments. Each image is carefully annotated and categorized to maintain data consistency and provide high-quality labels for supervised learning.

B. Data Preprocessing

Image preprocessing is performed to enhance data quality and prepare the dataset for effective model training. All input images are resized to a uniform resolution compatible with both the hybrid VGG19-Vision Transformer architecture and the YOLOv11 detection network. Pixel intensity values are normalized to improve numerical stability and accelerate model convergence. To increase dataset diversity and strengthen model generalization, various augmentation techniques including rotation, horizontal and vertical flipping, zooming, brightness adjustment, and color transformation are applied. Where necessary, image enhancement and noise reduction methods are also utilized to improve visual quality. Furthermore, class imbalance is mitigated using class-weighted loss functions and balanced sampling strategies, ensuring fair learning across all lesion categories while minimizing bias toward majority classes.

C. Feature Extraction Using Deep Learning

The feature extraction stage employs the VGG19 convolutional neural network to obtain rich spatial and textural representations from skin lesion images. Its deep hierarchical architecture effectively captures low-level visual cues such as edges, textures, and color variations, as well as higher-level semantic information including lesion shape, boundary irregularities, and structural patterns. The extracted feature maps are transformed into compact feature representations and subsequently forwarded to the Vision Transformer module. This integration preserves both detailed local information and broader contextual relationships, significantly improving the model's capability to distinguish between visually similar skin lesion categories.

D. Model Training and Classification

The classification framework utilizes a hybrid architecture that combines VGG19 with a Vision Transformer to exploit the advantages of both convolutional and attention-based learning. VGG19 extracts detailed local spatial features, while the Vision Transformer captures long-range dependencies and global contextual information using self-attention mechanisms. The fused feature representation is processed through fully connected layers to classify images into multiple skin lesion categories. During training, class-weighted optimization techniques are employed to address dataset imbalance and improve classification performance across all classes. The effectiveness of the proposed model is evaluated using standard performance metrics, including accuracy, precision, recall, macro F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC), ensuring comprehensive and reliable assessment.

E. Detection and Report Generation

The lesion detection component utilizes the YOLOv11 object detection model to accurately identify and localize suspicious skin lesions in both static images and real-time video streams. Its high-speed inference capability enables efficient deployment in practical clinical screening applications. The detected lesion regions are integrated with the classification results to produce comprehensive diagnostic information. Finally, a Generative AI-based automated reporting module converts the prediction results, confidence values, lesion characteristics, severity assessment, and relevant clinical information into structured dermatological reports. This automated reporting process enhances interpretability, improves clinical usability, supports informed decision-making, and significantly reduces the manual effort required for medical documentation.

V. EXPERIMENT

A. Evaluation Metrics

The effectiveness of the proposed skin lesion analysis framework is assessed using a comprehensive set of evaluation metrics to provide a reliable and unbiased performance analysis across all lesion categories. Overall classification accuracy is used to measure the percentage of correctly predicted samples. However, because medical datasets often suffer from class imbalance, accuracy alone does not adequately reflect model performance. Therefore, the Macro F1-score is adopted to provide equal importance to both majority and minority classes by balancing precision and recall. Precision evaluates the proportion of correctly classified positive predictions, whereas recall measures the model's capability to identify all relevant lesion instances, which is particularly important for detecting malignant conditions such as melanoma. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is also computed to evaluate the model's discriminative ability under different decision thresholds. Furthermore, a confusion matrix is analyzed to examine classification errors and identify confusion between visually similar skin lesion categories. The performance of the YOLOv11 detection module is evaluated using Mean Average Precision (mAP) at different Intersection over Union (IoU) thresholds, providing an accurate assessment of lesion localization performance. Collectively, these evaluation metrics offer a comprehensive analysis of the proposed system's classification accuracy, detection capability, robustness, and overall reliability for automated dermatological diagnosis.

B. Hyperparameters

The proposed hybrid CNN-Vision Transformer model is trained using carefully optimized hyperparameters to achieve stable convergence and high classification performance. All input images are resized to 224×224 pixels, satisfying the input requirements of both the VGG19 backbone and the Vision Transformer architecture. A batch size of 16 is selected to maintain an appropriate balance between computational efficiency and memory utilization. The model is initialized with a learning rate of 0.0001, enabling smooth and stable optimization throughout the training process. Training is performed for 20 to 50 epochs, with early stopping employed based on validation performance to prevent overfitting. The Adam optimizer is utilized because of its adaptive learning capability and faster convergence characteristics. To improve generalization, dropout regularization with a dropout rate of 0.3 is incorporated within the fully connected layers. For the YOLOv11 detection network, parameters including the confidence threshold and Intersection over Union (IoU) threshold are carefully tuned to maximize lesion detection accuracy while minimizing false-positive predictions. In addition, data augmentation settings such as image rotation, zooming, brightness variation, and flipping are applied to improve dataset diversity and strengthen model robustness. These hyperparameters are determined through extensive empirical experimentation to achieve optimal performance in both skin lesion classification and real-time detection tasks.

VI. RESULTS

The experimental evaluation demonstrates that the proposed hybrid framework outperforms individual deep learning architectures for skin lesion classification and detection. By integrating VGG19 with the Vision Transformer, the model effectively combines detailed local feature extraction with global contextual learning, resulting in improved classification performance across multiple dermatological categories. It accurately distinguishes visually similar lesions, particularly melanoma and nevus, while maintaining balanced performance on imbalanced datasets. Confusion matrix analysis shows reduced misclassification compared with conventional CNN models. The YOLOv11 detection module provides accurate real-time lesion localization in images and video streams, supporting practical clinical screening. Automated GenAI-based report generation produces structured diagnostic summaries containing predicted lesion type, confidence score, severity, and clinical recommendations. Overall, the proposed SkinDxFusion system delivers accurate, reliable, interpretable, and scalable performance for early skin cancer diagnosis applications.

TABLE I

Performance Comparison of Deep Learning Models

Model	Accuracy	Precision	F1 Score
VGG19	85.2%	0.84	0.83
Vision Transformer (ViT)	87.6%	0.86	0.85
YOLOv11 (Detection)	89.1%	0.88	0.87
VGG19 + ViT (Proposed)	91.3%	0.90	0.89

A. Web Application Interface

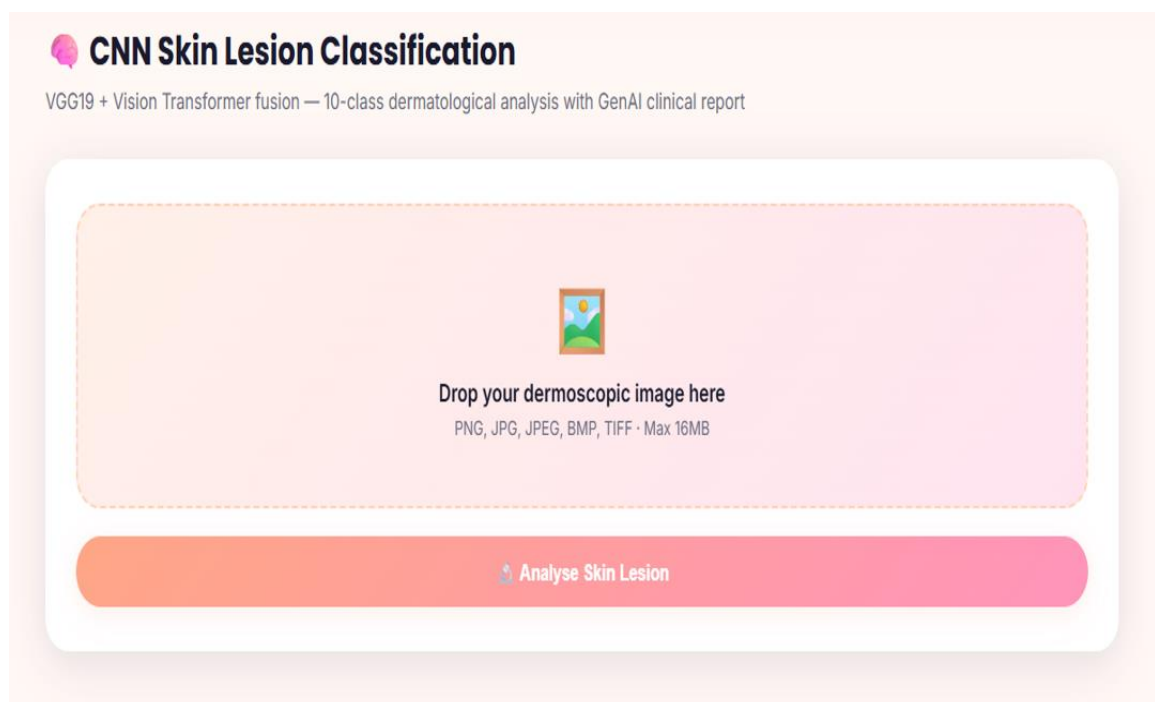


Fig. 2. Web interface of the proposed SkinDxFusion system, providing an intuitive platform for uploading dermoscopic images and performing AI-powered skin lesion analysis using the hybrid VGG19 and Vision Transformer model.

B. CNN Detection

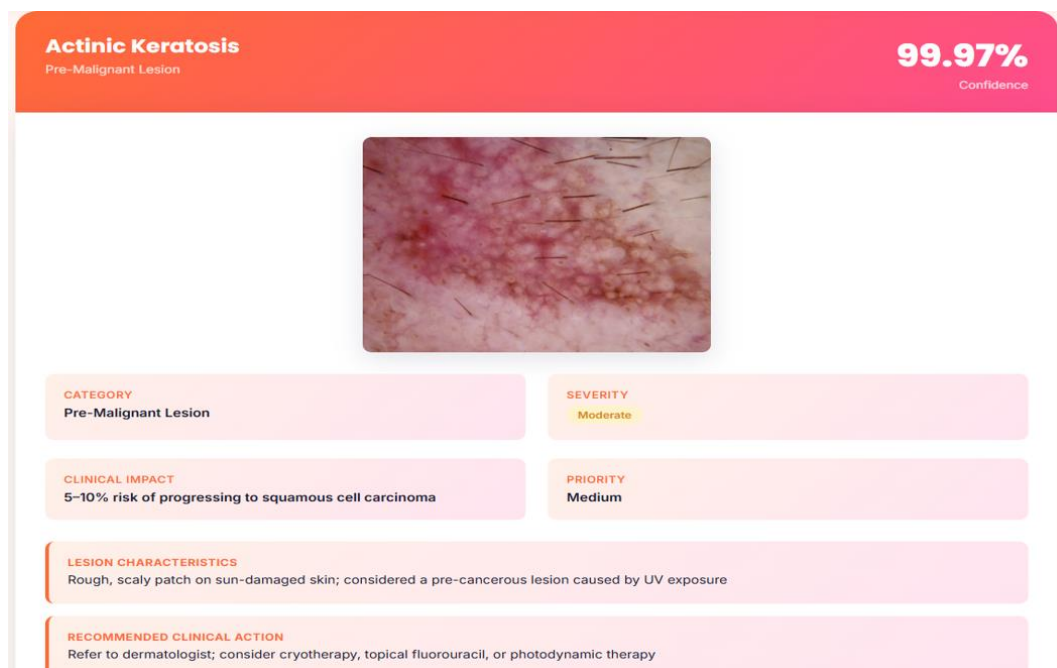


Fig. 3. CNN-based skin lesion classification interface presenting the predicted lesion type with its confidence score, clinical category, severity assessment, lesion characteristics, and recommended clinical action for comprehensive dermatological evaluation. *C.Yolov11 Interface*

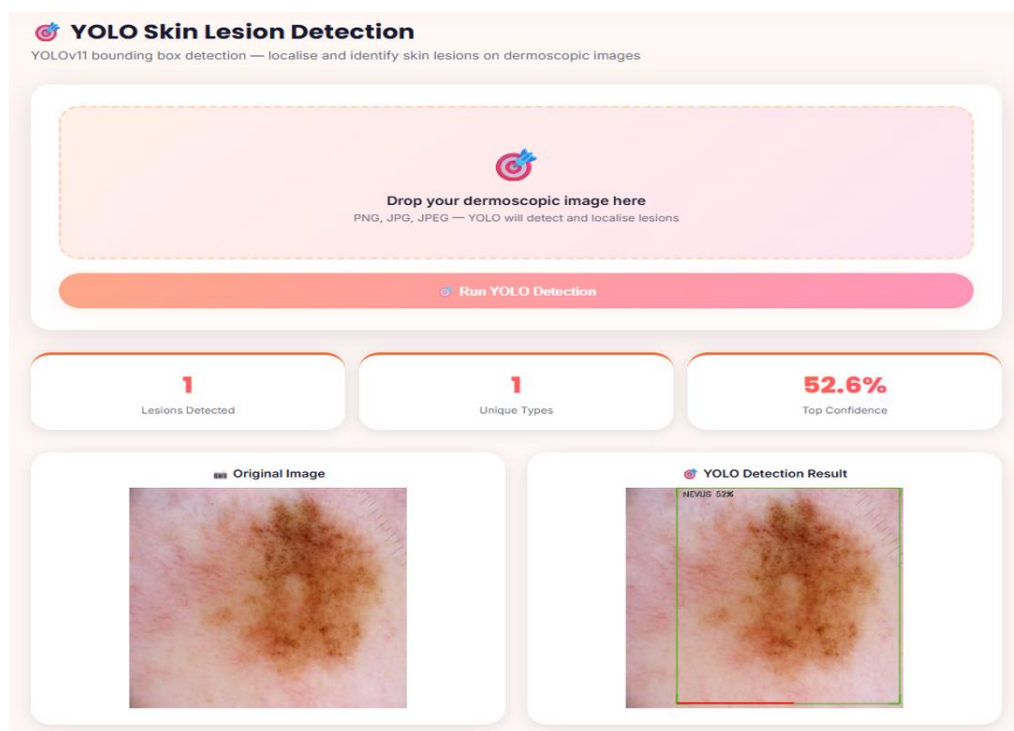


Fig. 4. YOLOv11-based skin lesion detection interface illustrating real-time localization of lesions through bounding box prediction, confidence score estimation, and detection statistics.

VII. CONCLUSION AND FUTURE SCOPE

In this research, a comprehensive skin lesion analysis framework, SkinDxFusion, has been developed by integrating a hybrid VGG19 and Vision Transformer (ViT) architecture with YOLOv11-based real-time lesion detection and Generative AI-powered automated report generation. The proposed framework effectively combines the local feature extraction capability of convolutional neural networks with the global contextual learning ability of transformer models, enabling accurate and robust classification across multiple dermatological categories. The integration of YOLOv11 further enhances the system by providing precise localization of lesion regions in both dermoscopic images and live video streams, making it suitable for real-time clinical screening and diagnostic applications. Comprehensive preprocessing techniques, including image normalization, augmentation, and class balancing, contribute to improved model generalization and reliable performance under diverse imaging conditions. Experimental evaluation demonstrates that the proposed hybrid framework achieves superior classification accuracy, precision, recall, Macro F1-score, and AUC compared with individual deep learning models while maintaining balanced performance across both benign and malignant lesion classes. The automated report generation module transforms prediction outputs into structured clinical reports containing lesion type, confidence score, severity level, lesion characteristics, and recommended clinical actions, thereby improving result interpretability and reducing the documentation burden on healthcare professionals. The developed web-based interface further enhances accessibility by providing an intuitive platform for image upload, lesion analysis, and report generation. Overall, the proposed framework serves as an efficient, scalable, and reliable decision-support system for early skin cancer detection, assisting dermatologists in improving diagnostic consistency and facilitating timely clinical intervention. Future work will focus on improving the robustness, scalability, and clinical applicability of the proposed framework by incorporating larger and more diverse datasets that represent different skin tones, lesion variations, imaging devices, and acquisition conditions. The integration of multimodal clinical information, including patient history, demographic details, genetic information, and laboratory records, may further improve diagnostic accuracy and personalized healthcare recommendations. Future research can also investigate lightweight transformer architectures, advanced ensemble learning methods, and efficient optimization techniques to reduce computational complexity while maintaining high predictive performance. Furthermore, incorporating Explainable Artificial Intelligence techniques such as Grad-CAM, attention visualization, and feature attribution methods will improve transparency and increase clinician confidence in AI-assisted diagnosis. Deployment on cloud, mobile, and edge computing platforms will enable accessible real-time screening in remote and resource-limited healthcare settings. Finally, extensive clinical validation through collaboration with dermatologists and medical institutions will be essential to ensure the safety, reliability, and widespread adoption of the proposed system in real-world healthcare environments.

REFERENCES

- [1] Y. Wu, X. Chen, and L. Zhou, "Skin Cancer Classification With Deep Learning," IEEE Access, 2022.
- [2] M. Azeem, S. Hussain, and F. Shah, "SkinLesNet: Classification of Skin Lesions and Detection," IEEE Access, 2023.
- [3] D. P. Yadav, R. Singh, and P. Kumar, "Dual-scale Lightweight Cross Attention Transformer for Skin Lesion Classification," IEEE Trans. Emerging Topics Comput. Intell., 2024.
- [4] N. Ahmad, H. Tariq, and R. Nawaz, "A Novel Framework of Multiclass Skin Lesion Recognition," Computers in Biology and Medicine, 2023.
- [5] K. Behara, S. R. Das, and A. Kumar, "An Improved Skin Lesion Classification Using a Hybrid CNN-based Approach," Biomedical Signal Processing and Control, 2024.
- [6] B. Shetty, M. Khadilkar, and S. Patil, "Skin Lesion Classification of Dermoscopic Images Using Deep Learning," Healthcare Analytics, 2022.
- [7] N. Gessert, M. Nielsen, and R. Werner, "Skin Lesion Classification Using Ensembles of Multi-scale CNNs," ISIC Challenge / IEEE Workshop, 2020.
- [8] F. Wu, L. Zhang, and J. Huang, "Automatic Generation Network for Skin Imaging Reports," IEEE Access, 2022.
- [9] P. A. Vieira, R. Morais, and J. Santos, "The Automated Generation of Medical Reports from Medical Images," Applied Sciences, 2024.
- [10] V. Ravi, K. Suresh, and D. Prasad, "Integrating Vision Transformers for Advanced Skin Lesion Evaluation," Journal of Imaging, 2024.
- [11] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," Nature, 2017.

-
- [12] N. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at ISIC," IEEE Int. Symp. Biomedical Imaging, 2018.
- [13] J. Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," IEEE Conf. Computer Vision and Pattern Recognition, 2016.
- [14] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.
- [15] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," ICCV, 2017.
- [16] T. Chen et al., "A Simple Framework for Contrastive Learning of Visual Representations," ICML, 2020.
- [17] K. He et al., "Deep Residual Learning for Image Recognition," CVPR, 2016.
- [18] O. Ronneberger et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation," MICCAI, 2015.
- [19] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," ICCV, 2021.