IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 February 2024, accepted 1 April 2024, date of publication 4 April 2024, date of current version 15 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3385340

RESEARCH ARTICLE

A Novel Transfer Learning Framework for Multimodal Skin Lesion Analysis

S. REMYA[1], T. ANJALI[1], AND VIJAYAN SUGUMARAN[2]
[1]Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, Kerala 690525, India
[2]Department of Decision and Information Sciences, Centre for Data Science and Big Data Analytics, School of Business, Oakland University, Rochester, MI 48309, USA

Corresponding author: S. Remya (remyas@am.amrita.edu)

ABSTRACT Skin lesion classification is a pivotal process in dermatology, enabling the early detection and precise diagnosis of skin diseases, leading to improved patient outcomes. Deep learning has shown great potential for this task by leveraging its ability to learn complex patterns from images. However, diagnostic accuracy is compromised by exclusive reliance on single-modality images. This research work proposes an innovative framework that unifies a Vision Transformer model with transfer learning, channel attention mechanism, and ROI for the accurate detection of skin conditions, including skin cancer. The proposed approach blends computer vision and machine-learning techniques, leveraging a comprehensive dataset comprised of macroscopic dermoscopic images, appended with patient metadata. Compared with conventional techniques, the proposed methodology exhibits significant improvements in various parameters, including sensitivity, specificity, and precision. Moreover, it demonstrates outstanding performance in real-world datasets, reinforcing its potential for clinical implementation. With a remarkable accuracy of 99%, the method outperforms existing approaches. Overall, this investigation underscores the transformative impact of deep learning and multimodal data analysis in the dermoscopic domain, projecting substantial headway into the field of skin lesion analytic diagnosis.

INDEX TERMS Skin lesion classification, dermatology, deep learning, multimodal data analysis, transfer learning, vision transformer.

I. INTRODUCTION

Dermatology is a branch of medicine that focuses on the prevention, identification, and treatment of skin conditions that affect a sizable segment of the global population. According to the World Health Organization (WHO), a substantial portion of the global population is affected by skin ailments that call for early detection and diagnosis to avert adverse health effects that exacerbate mortality. As a rule, readily visible skin lesions are the primary symptoms of skin disorders. Hence, effective management of skin maladies mandates error-free diagnosis of lesions in the early stages.

Traditionally, skin lesions are detected by dermoscopy, a non-invasive imaging procedure performed by dermatologists. Ensued by the wide range in the size, shape, and location of skin lesions, exclusive reliance on a single gadget seldom provides adequate information for an accurate diagnosis. Even skilled dermatologists run into arduous encounters in the diagnosis and treatment of skin disorders, unable to discern between benign and malignant tumors, leading to incorrect diagnoses. This study recommends a multi-pronged strategy that marshals diverse modalities of lesion images conjectured to deal with this symptomatic shortcoming. In this study, ultrasound and MRI images are added to the dataset, allowing for a more accurate diagnosis and efficient treatment of skin conditions.

Deploying Computer Vision (CV) and Deep Learning (DL) methods, this research proffers an innovative multimodal framework for the diagnosis of skin lesions [1]. The conceived strategy unifies ultrasonic and MRI imaging modalities to overcome the constraints of acquiring labelled data due to low prevalence and confidentiality concerns. The suggested approach intends to augment accuracy while enabling focused disease management. Transfer learning is

The associate editor coordinating the review of this manuscript and approving it for publication was Abdel-Hamid Soliman.

also used to increase performance by fine-tuning existing models on a smaller sample. A group of models, Vision Transformers (ViTs), has recently been demonstrated to perform remarkably well on various CV tasks. As ViT models manipulate self-attention methods to discern the overall correlations between image features, they are superior to standard CNNs at processing massive data. This study presents a novel multimodal with a transfer learning approach, which affords efficient categorization of skin lesions, with enhanced precision [2]. The recommended technique employs clinical and dermoscopic images to train a ViT model with metadata (patient age, gender, and lesion locations). The model presents an enlarged view of skin lesions, propitious to easier discernment of features such as borders, colors, and structures. Clinical images, captured using a conventional camera, provide a comprehensive view of the lesion, and the metadata provides useful information that can improve classification accuracy.

Comparison of the recommended framework with conventional dermoscopic practices yields multiple advantages:

- Uses dermoscopic images that facilitate feature extraction and promote precise classification.
- Better-envisioned skin lesions provide more accurate diagnoses with multiple data sources.
- Pretrained ViT models obviate reliance on labeled data.

Given the large amount of data on skin diseases, this effective data use is extremely valuable. By synergizing diverse data sources and automated analysis of skin lesions, the proposed method holds the potential to substantially impact the field of dermatology [3]. Existing techniques for automated analysis of skin lesions often struggle with accurately distinguishing between benign and malignant lesions due to complex variations in lesion appearances. Additionally, these techniques can be limited by the lack of diverse and well-annotated datasets, hindering their generalization across different populations and skin types. Early detection and precise diagnosis lead to improved health outcomes and reduced response times, establishing a more confident and effective approach.

Furthermore, automating and streamlining the procedure could reduce the workload of dermatologists, allowing them to concentrate on their cases and improve efficiency in healthcare settings. The main objectives of this study are as follows:

1) Develop a novel skin lesion classification approach that integrates a real-world dataset with a HAM dataset. A ViT model is used to learn features from visual and textual descriptions for comprehensive and accurate classification.
2) Combine a cutting-edge ViT model with visual and textual data, to improve classification accuracy and enable the early detection of skin cancer..
3) Assess the computational efficiency and scalability of the ViT approach for real-world applications using metrics such as precision, recall, F1 score, and accuracy.
4) Compare the proposed ViT approach with SVM, KNN, MobileNet, ResNet152v2, and VGG-16 on the same dataset and experimental settings.

This demonstrates the superiority and potential breakthroughs of the proposed method in skin lesion classification. The study advances the area of dermatology by presenting a cutting-edge multimodal deep-learning framework for precise skin lesion classification, gets over data restrictions, and boosts diagnostic accuracy by incorporating a variety of lesion images using transfer learning. The suggested methodology has the potential to revolutionize the study of skin lesions, thereby assisting patients in treating skin illnesses by enabling prompt and efficient interventions.

By addressing the challenges of diagnostic accuracy through the integration of deep learning and multimodal data analysis, this research work positions itself at the forefront of transformative advancements in the dermoscopic domain. The contribution of this work is the innovative framework to significantly elevate the precision and efficacy of skin lesion analytic diagnosis, marking a substantial stride towards improved healthcare outcomes. The specific contributions of this research work are delineated as follows:

1) Multimodal Fusion for Accurate Analysis: Introduce an innovative framework for classifying skin lesions that employs a ViT model, transfer learning, channel attention, and ROI. This framework combines visual data with patient metadata to achieve precise detection of skin conditions, surpassing traditional approaches that rely on a single modality.
2) Significance in Clinical Impact and Real-world Viability: Demonstrate the transformative potential of the developed method by exhibiting significant improvements in sensitivity, specificity, and precision. Achieve an outstanding accuracy of 99%, surpassing existing approaches, and reinforcing its potential for practical implementation in clinical settings.
3) Efficiency and Automation for Dermatology: Address the core objectives of improving confidence and efficiency in dermatology by automating and streamlining skin lesion analysis.
4) Evaluate the computational efficiency and scalability of the ViT model, as well as its potential to alleviate the effort of dermatologists and enhance healthcare services.

This paper is organized into several sections. The research context is provided in Section II through a detailed analysis of relevant works by other investigators in the field. Section III provides overall insight into the adopted approach, including the tools, procedures, datasets, pre-processing methods, and network structure. Section IV, which elucidates the methodology for the classification of skin lesions, provides the details of the proposed approach. Section V reports on the assessment of the proffered method, highlighting the prime benefits of the proposed approach for the diagnosis of skin lesions. Finally, Section VI provides closing remarks and future enhancements.

II. RELATED WORKS

DL algorithms have recently accomplished notable advancements in the classification of skin lesions, an outgrowth of their capability to extract specific patterns and features from images. However, these attributes are inadequate to correctly discern skin lesions. Studies have suggested different modalities, including dermoscopic, clinical, and histological imaging, to enhance classification accuracy, few of which have looked at histopathological imaging.

This literature review focuses on three main areas crucial to this research on a new transfer learning framework for analyzing skin lesions. Firstly, multimodal fusion for skin lesion classification is explored, looking at advances in ensemble based techniques and focusing on the importance of combining different types of data. In addition, the challenges in segmentation such as dealing with limited data are discussed, proposing a shift towards attention-based models. The second part involves transfer learning and model comparisons, where assessment is done in terms of how well transfer learning works and compares various models to find the most effective ones. Lastly, the importance of patient-specific information is emphasized, advocating for patient-centric deep learning and exploring advancements in clinical integration. These three areas provide a comprehensive background for the novel framework presented in this paper, which has the potential to transform the multimodal skin lesion analysis.

A. ADVANCEMENTS IN ENSEMBLE-BASED TECHNIQUES

The use of ensemble-based techniques in dermatology has gained prominence, aiming to enhance classification accuracy and robustness. In a research work by Esteva et al., [4], a revolutionary model utilizing a CNN architecture demonstrated dermatologist-level precision in skin cancer classification, primarily leveraging dermoscopic images. However, the reliance on specific image types raises concerns about the adaptability of the model, where such images are unavailable. In this study, with the help of CNN architecture, the complex visual patterns, which include the different types of skin cancers such as melanoma, basal cell carcinoma, and squamous cell carcinoma, can be analyzed and classified. It made use of a substantial dataset of dermatoscopic pictures. From clinical images, it attained an accuracy of 71.3%. The proposed deep learning model performed exceptionally well, matching dermatologists' level of expertise in detecting skin cancer in terms of accuracy values. This study demonstrated how deep learning has the potential to revolutionize clinical diagnostics and dermatology practice. Using a large and diverse dataset that contains thousands of annotated dermoscopic images is one of the key advantages. The model has strong categorization capabilities as a result of its capacity to learn detailed characteristics and patterns from such a vast amount of data. Insights into the key factors influencing its predictions were provided, which is a characteristic essential for establishing confidence within therapeutic applications.

The integration of visual and textual information has been a pivotal area of exploration in dermatological research. Ozturk et al. [5] introduced a hybrid model emphasizing the importance of incorporating diverse data sources. Their work showcased good accuracy in distinguishing benign and malignant tumors through the fusion of visual and textual data. Nevertheless, challenges may arise in scenarios where certain data modalities are limited, necessitating a meticulous approach to account for such constraints.

Addressing the intricacies of skin lesion analysis, Zhang et al. [6] delved into the crucial aspects of precise lesion delineation through CNN-based segmentation. Using multimodal data, the model showed increased accuracy in distinguishing between benign and malignant tumors. This study showed the potential of combining several data sources to enhance the diagnostic capabilities of deep learning models for dermatology. Additionally, challenges related to the scarcity of multimodal datasets for ViTs were acknowledged, underscoring the necessity for further development in both segmentation techniques and dataset availability.

The transition towards attention-based models in skin lesion classification was explored by Nasr-Esfahani et al [7]. Their attention-based deep-learning model not only emphasized improved interpretability but also exhibited enhanced performance However, the optimal application of attention mechanisms poses challenges, and the positive impact on overall classification accuracy justifies further exploration. As the field evolves, attention-based models showcase a promising avenue for refining interpretability in dermatological diagnostics.

Thus, the integration of multimodal fusion techniques, ensemble-based strategies, segmentation refinement, and attention mechanisms represents a dynamic landscape in the realm of skin lesion classification. While advancements have been notable, each approach brings its set of challenges, from dependency on specific image types to issues related to data scarcity and optimal attention mechanism application. The collective pursuit of these innovations signifies a transformative trajectory in dermatology signifies more accurate and effective diagnostic tools with the potential to revolutionize clinical practices.

B. TRANSFER LEARNING BASED MODEL COMPARISONS

In the realm of skin lesion classification, the exploration of transfer learning and model comparisons has introduced novel perspectives, aiming to enhance the efficacy of diagnostic tools. Chen et al. [8] conducted an extensive investigation into the effectiveness of transfer learning and fine-tuning in CNNs. Their findings highlighted improved performance and accelerated training, suggesting that pre-trained models lay a solid foundation for various skin lesion categorization tasks. However, the selection of appropriate pre-trained models emerges as a critical consideration, introducing subtleties in the application of transfer learning in dermatology.

Diversity in deep learning models and their architectures has been a focal point in skin lesion classification research.

TABLE 1. Summary of the literature review.

Author	Methodology	Dataset	Advantages	Disadvantages	Accuracy
Esteva et al. (2017)	CNN-based skin cancer classification Clinical image analysis	Dermoscopic Clinical images	High accuracy approaching dermatologists' expertise large and diverse dataset potential for clinical diagnostics	Data availability computational requirements interpretability of deep learning models	71.3
Brinker et al. (2018)	CNN-based melanoma detection	Dermoscopic	Distinguishing between benign and malignant tumors promising tool for dermatologists High sensitivity and specificity	Data quantity and quality model generalization	74
Haenssle et al. (2018)	Deep learning-based melanoma detection Outperforms dermatologists	Dermoscopic	Exceptional diagnostic accuracy outperforming dermatologists deep learning's potential in dermatology lesion classification	Data quality and diversity integration with clinical data	80
Tschandl et al. (2019)	CNN-based skin lesion classification Integration of patient details	Dermoscopic	Improved categorization with patient information incorporation of clinical and image data Better lesion identification	Data privacy concerns data quality interpretability	78
Menegola et al. (2019)	Multi-task deep learning for skin lesion classification precise lesion definition	Dermoscopic	Better categorization with precise lesion definition multitasking technique improved diagnosis	Data labeling complexity resource-intensive training	76
Codella et al. (2019)	Benchmarking deep learning models for skin lesion classification	Dermoscopic	Essential benchmarking dataset deep learning accuracy compared with manual classification	Data acquisition and annotation model interpretability	91
Nasr-Esfahani et al. (2019)	Attention-based deep learning for skin lesion classification	Dermoscopic	Improved interpretability attention mechanisms understanding model decisions more understandable deep learning	Data integration challenges attention mechanism complexity	79
Chen et al. (2020)	Transfer learning and fine-tuning in CNNs for skin lesion classification	Dermoscopic	Improved classification with pre-trained models faster model training adaptation to diverse tasks	Dataset size and quality overfitting challenges	82
Ozturk et al. (2020)	Hybrid deep learning model for skin lesion classification with multimodal data	Dermoscopic	Improved accuracy with multimodal data combining visual and textual information enhancing diagnostic capabilities	Data integration challenges increased complexity	77
Ma et al. (2020)	Deep learning-based skin lesion classification with precise lesion delineation	Dermoscopic	Enhanced classification with precise lesion delineation increased accuracy through strong segmentation	Segmentation challenges model complexity	Not mentioned
Katti et al. (2021)	Ensemble of deep learning models for skin lesion classification	Dermoscopic	Improved classification using ensemble models robust system for categorizing skin lesions	Ensemble training complexity model integration	79
Rezazadegan et al. (2021)	Data augmentation for deep neural networks in skin lesion classification	Dermoscopic	Increased model durability and generalization reduced overfitting, higher accuracy	Data augmentation techniques selection	81
Sun et al. (2019)	Dual-stage CNN for skin lesion segmentation and precise delineation	Dermoscopic	Improved segmentation using dual-stage design reliable lesion delineation precise lesion delineation	Model complexity computation time	77

TABLE 1. *(Continued.)* **Summary of the literature review.**

Yuan et al. (2020)	Vision transformer-based skin lesion classification with clinical images	Dermoscopic	High accuracy with vision transformers potential for automated analysis promising results with clinical images	Data quantity and quality model interpretability	91
Matsunaga et al. (2017)	Transfer learning and domain adaptation for melanoma diagnosis with histology images	Histology	Improved generalization across datasets successful adaptation to histopathology images better diagnostic tools	Data compatibility domain adaptation complexity	76
Yao et al. (2018)	Ensemble of pre-trained CNNs for skin lesion classification	Dermoscopic	Increased accuracy and resilience with ensemble models effective utilization of supplemental data	Ensemble training complexity model integration	78
Isn and Zkan (2020)	Segmentation and classification of skin lesions using deep learning	Dermoscopic	In-depth analysis of segmentation and classification identification of recent trends and challenges	Model complexity data challenges	75
Tanno and Shimizu et al	Ensemble of deep learning models for skin lesion classification	Dermoscopic	Improved classification with ensemble models reliability robust system for categorizing skin lesions	Ensemble training complexity model integration	80
Bejnordi et al. (2017)	Deep learning for lymph node metastases detection in breast cancer	Histology	High accuracy in lymph node metastases detection potential for precise diagnoses, detection of metastases with high accuracy	Data compatibility domain-specific challenges	89
Bi et al. (2020)	Saliency map-based automatic skin lesion segmentation	Dermoscopic	Efficient diagnosis using saliency maps improved lesion delineation increased segmentation accuracy	Saliency map complexity annotation challenges	78
Jini (2021)	Dermoscopic image analysis for skin cancer detection	Dermoscopic	Comprehensive examination of dermoscopic image analysis insights into trends and challenges in skin cancer detection	Data acquisition and annotation model complexity	Not mentioned
Wang et al. (2021)	Multimodal skin lesion classification using combined modaities	Multimodal	High accuracy using combined modalities potential for improving skin lesion classification	Data integration complexity multimodal data challenges	89

The study by Codella et al. [9] delved into the evaluation of various models, establishing a benchmark dataset for ongoing research endeavors. While the superiority of deep learning models over manual classification was evident, the diversity of skin lesions presented a challenge in selecting the most effective model for specific diagnostic scenarios. This underscores the need for a detailed understanding of the intricacies associated with diverse skin lesions, emphasizing the importance of subtleties in this context.

Multi-task learning in the context of skin lesion analysis has been a domain of significant innovation. Haenssle et al [10] directed their focus towards multi-task deep learning for melanoma detection, achieving exceptional diagnostic accuracy. The simultaneous training on segmentation and classification tasks showcased the potential of multitasking techniques in providing comprehensive insights into skin lesion characteristics. However, the integration of multiple tasks introduces complexity, posing challenges in terms of model interpretability and application in real-world clinical settings.

Thus, the exploration of transfer learning and model comparisons in skin lesion classification has introduced advancements and challenges. While transfer learning offers improved performance, careful consideration is needed in selecting pre-trained models. The diversity in model architectures necessitates a tailored approach to match specific diagnostic requirements. Furthermore, the innovation in multi-task learning showcases promise in providing comprehensive insights, yet the complexity of integration calls for further refinement. Collectively, these research avenues signify a dynamic landscape in dermatological diagnostics,

pushing the boundaries for more effective and precise skin lesion classification tools.

C. IMPORTANCE OF PATIENT SPECIFIC INFORMATION

In the evolving landscape of dermatological research, the incorporation of patient-specific information has emerged as a pivotal avenue for refining skin lesion classification. Tschandl et al. [11] laid the foundation by developing a CNN-based approach that underscored the significance of patient-centric deep learning. Their emphasis on integrating contextual factors such as age, gender, and lesion location led to improved accuracy, signaling a shift toward more personalized dermatological diagnostic tools. This work prompts a deeper exploration into the utilization of patient metadata, recognizing the potential impact on diagnostic precision.

Advancements in clinical integration have further enriched the understanding of patient-specific considerations in skin lesion analysis. Grochowski et al. [12] delved into the effects of data augmentation on deep neural network performance, showcasing effective strategies for enhancing durability and generalization. The emphasis on data augmentation directly addresses challenges associated with overfitting, providing valuable insights for refining deep learning models in dermatological applications. However, the need for standardized augmentation techniques remains a critical consideration to ensure the reliability and reproducibility of the models. This not only enhances diagnostic accuracy but also positions dermatology on a trajectory toward more patient-centered and refined diagnostic tools.

Addressing challenges associated with single-modality approaches, Yuan et al [13] emphasized the significance of AI-based algorithms in skin cancer segmentation and diagnosis, particularly using histopathology images. This work demonstrated the integration of segmentation and diagnosis to enhance precision and provide detailed insights. While the study provides valuable advancements, it also highlights the imperative for additional research to fully harness the expansive potential of deep learning in comprehensive skin lesion analysis [14].

Ensemble strategies have emerged as a robust approach in dermatological applications, as exemplified by the work of Sandri et al [15]. Their showcase of a powerful transfer learning strategy employing an ensemble of deep learning models for skin lesion classification underscores the effectiveness of leveraging pre-trained models. The ensemble strategy outperformed individual models, signaling the potential for further exploration of ensemble-based techniques to enhance the robustness and accuracy of dermatological diagnostic tools.

Recently, skin lesion categorization utilizing single-modality pictures, such as dermoscopic and clinical images, has been attempted using ViTs, with encouraging results. Similarly, Yan et al [16]. achieved an accuracy of 84% using a ViT with clinical images. However, there has been limited research on multimodal skin lesion classification using ViT. Further investigation of this approach could lead to improved accuracy in skin lesion classification [17]. Addressing challenges associated with single-modality approaches, Brinker et al. [18] emphasized the significance of AI-based algorithms in skin cancer segmentation and diagnosis, particularly using histopathology images. This work demonstrated the integration of segmentation and diagnosis to enhance precision and provide detailed insights. While the study provides valuable advancements, it also highlights the imperative for additional research to fully harness the expansive potential of deep learning in comprehensive skin lesion analysis [19].

In summary, the integration of ViTs, the exploration of multi-modality applications, and the effectiveness of ensemble strategies collectively represent a transformative trajectory in dermatological diagnostics. While ViTs demonstrate promise in precise classification, the field calls for further research to exploit their full potential. Addressing the challenges in single-modality approaches and harnessing the power of ensemble strategies contribute to the ongoing evolution of accurate and effective skin lesion classification tools in dermatology. The summary of the state of the art of lesion image analysis is shown in Table1.

III. MATERIALS AND METHODS

A. MATERIALS

1) DATASET

The HAM10000 dataset is a well-known, widely used resource, in the domain of dermatology research. It can be accessed at $https://dataverse.harvard.edu/dataset.xhtml?persistentId = doi:10.7910/DVN/DBW86T$. The dataset comprises 10,015 images of pigmented skin lesions, carefully collected from patient records at the Medical University of Vienna. The dataset is highly treasured for training, evaluation, and validation of Computer Aided Diagnosis (CAD) systems focused on categorization of skin lesions and the detection of skin cancer. This dataset includes a range of lesion types and clinical characteristics [20], [21], [22], [23]. In addition, collected real-world data, which was considered for analysis. The dataset has been meticulously organized into seven distinct classes based on histopathological examination and clinical diagnosis. These classes are noted hereunder:

- Actinic Keratosis (AK): Manifesting rough, scaly patches. Untreated AK can develop into squamous cell carcinoma, incited by prolonged exposure to sunlight.
- Basal Cell Carcinoma (BCC): BCC appears as small, shiny bumps or nodules with slow growth rates that can be locally invasive. This is the most prevalent form of skin cancer.
- Benign Keratosis-like Lesions (BKL): Benign skin growths, commonly found in older adults, encompass a wide range of lesions such as solar lentigines and seborrheic keratosis.

- Dermatofibroma (DF): Benign fibrous tumor often encountered on the legs; dermatofibroma typically exhibits a firm, reddish-brown appearance.
- Melanoma (MEL): Potentially life-threatening melanoma is a malignant skin cancer that arises from melanocytes. Early detection is crucial for an effective treatment.
- Nevus (NV): Also known as moles, nevi are benign skin lesions that can be flat or raised and vary in color and size.
- Vascular Lesions (VASC): This category comprises a range of vascular anomalies, including hemangiomas and port-wine stains [24], [25].

2) DATA PRE-PROCESSING

The efficiency of the proposed framework, for analysis of skin lesions using ViTs and transfer learning methods, heavily rely on data preprocessing. Skin cancer is a type of cancer that can be challenging to detect in its early stages. Skin moles or lesions often serve as the indications of skin cancer, highlighting the importance of diagnosis for early detection [26], [27].

The proposed method utilizes three distinct CNNs for processing dermoscopic, clinical, and histological skin lesion images, enhancing accuracy and efficiency. It employs the ViT model, trained through transfer learning on a substantial dataset, to identify skin lesions accurately. This approach benefits dermatologists in diagnosing skin lesions effectively and enables the automated system to classify them accurately. By integrating dermoscopic and macroscopic images and leveraging pre-trained models, the proposed strategy achieves outstanding results on two widely recognized benchmark datasets, demonstrating its effectiveness in skin lesion analysis.

This study utilized the HAM10000 dataset and a real-time dataset, both comprised of dermatoscopic, macroscopic, and histological images of pigmented skin lesions. To ensure consistent input dimensions for the ViT, essential metadata, like diagnosis and anatomical sites, and images resized to a standard resolution of 224 $\times$ 224 pixels, using transformer and transfer learning models. Normalizing pixel intensities to a scale of [0, 1] facilitated training convergence and improved model stability. Data augmentation techniques, including - rotation, flipping, and zooming, increased model generalizability, while oversampling and undersampling addressed class imbalance issues. Transfer learning was employed to adapt the multimodal data to the input format required by the ViT model. The dataset was split into distinct subsets for training, testing, and validation, ensuring unbiased model evaluation with equal class distributions [28].

3) DATA AUGMENTATION

Data augmentation entails the application of random transformations such as rotation, flipping, zooming, and brightness adjustments to skin lesion images. This process culminates in a diverse dataset that captures real-world variations, enhancing the model's ability to generalize and accurately identify different skin lesions. Retaining the lesion boundaries ensures diagnostic features remain intact. Data augmentation is imperative in dermatology, where lesions exhibit various appearances and textures, helping avert overfitting, and improving the model's performance in handling diverse skin lesion images.

B. METHODS

1) TRANSFER LEARNING

Each image was downsized to 224 $\times$ 224 pixels to provide a uniform size, and the pixel values were set to range from 0 to 1. Subsequently, applied data augmentation techniques to enlarge the training dataset and prevent overfitting. Here a pretrained EfficientNet-B0 model is applied as a feature extractor. Except for the final layer, whose weights were frozen, a new fully connected layer with seven output units, one for each type of skin lesion, was added [29].

2) VISION TRANSFORMER

In the proposed method, the ViT architecture is used to classify skin lesions. A feedforward neural network was placed after several multi-head self-attention layers in the ViT architecture. The Adam optimizer was used to train the model across 20 epochs with a learning rate of 0.0001 and a batch size of 32.

3) NETWORK ARCHITECTURES

In recent times, the ViT has achieved outstanding results in computer vision tasks as well as in NLP. However, adjustments must be made to the ViT design to include both image and text modalities to use for multimodal skin lesion categorization. Utilizing two distinct ViT frameworks, one for processing the image modality and the other for processing the text modality is integrated here and is used for the classification process. To do this, the visual and textual information are concatenated along the sequence dimension and sent to the ViT architecture. In the proposed method, text modality is processed using an RNN, whereas the image modality is processed using CNNs. The ViT architecture is then fed with the concatenated outputs of these two modalities along the sequence dimension.

The architecture of ViT typically consists of an input layer, an encoder layer, a fully connected layer, and an output layer. Figure 1 shows the different components of the proposed model, and how it is connected. This provides a better understanding of the working of the model.

1) Input layer: This layer converts the input image into a set of feature maps that are then fed into the transformer.
2) Encoder layer: This layer consists of multiple self-attention blocks that process the feature maps in a parallel and scalable manner. The encoder layer captures the local and global context information in the image.

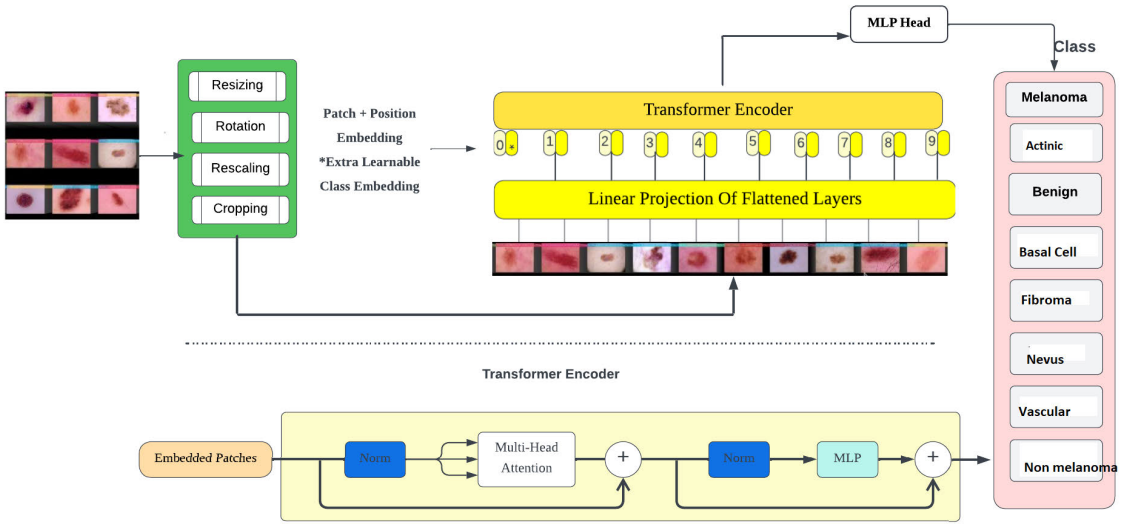

FIGURE 1. The proposed model architecture.

3) Fully connected layer: This layer maps the output of the encoder layer to a high-dimensional space, allowing the transformer to capture the complex relationships between the input features.
4) Output layer: This layer generates the final predictions for a task, such as object detection or semantic segmentation, based on the features learned by the transformer.

4) DIFFERENT PHASES OF ViT MODEL

1) Patch Embedding: A series of non-overlapping patches is created from the input image using the patch embedding technique, and each patch is represented as a vector of pixel values. Let X be the input image, with dimensions $H \times W \times C$. Here the image is divided into non-overlapping patches of size $P \times P$, and create a series of patch vectors $X = [x_1, x_2, \ldots, x_n]$, where x_i has dimensions $P \times P \times C$. Each patch vector is reshaped into a single vector of length $P^2 \times C$, denoted by $x_{i'}$.
2) Linear Embedding: In this stage, a learnable linear projection is used to project each patch vector into a higher-dimensional space. The model then acquires a more expressive representation of each patch. Let W be a learnable weight matrix, with dimensions $D \times (P^2 \times C)$, where D is the output dimension of the linear projection, resulting in a new sequence of vectors $H = [h_1, h_2, \ldots, h_n]$, where $h_i = W \times x_{i'}$ where $x_{i'}$ is the patch vector.
3) Positional Encoding: To capture the spatial structure of the image, the model also adds a learnable position embedding to each patch vector. This provided a model with information on the location of each patch within the image. Let E be a learnable position embedding matrix, with dimensions $D \times N$, where N is the total number of patches in the image. Then it concatenates the position embedding to each patch embedding, resulting in a new sequence of vectors $Z = [z_1, z_2, \ldots, z_n]$, where $z_i = h_i + e_i$.
The position embedding can be computed as follows:

$$e_{i,j} = \begin{cases} sin(\frac{j}{(10000(\frac{2i}{D}))}) & fori = 0, 2.., D-1 \\ cos(\frac{j}{(10000(\frac{2i}{D}))}) & fori = 1, 3, \ldots, D-1 \end{cases} \quad (1)$$

Here in both cases j = 0, 1, 2 , . . . , N-1
4) Transformer Layers: The L times repeated self-attention and feedforward layers with ReLU activation function make up the Transformer layers. Each layer receives a set of patch embeddings with positional encodings X_l as input and produces a new set of embeddings X_{l+1} which represents the transformer layer X_l. Each transformer layer employs a self-attention mechanism that calculates the weighted sum of each embedding in X_l, with the weights determined by how similar the embeddings are to one another. Like the residual connections in other neural network architectures, the resulting weighted sum was subsequently added to the input embeddings. The skip connection is then used to pass the output embeddings through [27], [29], [30].
5) Classification Head: The output embeddings of the final transformer layer are averaged to generate a single vector representation of the input image using the ViT model. The classification head, which comprises a linear layer and softmax activation function, is then fed into this vector. The output of the linear layer is normalized by the softmax activation function such that the values add up to 1. Consequently, the model can forecast the likelihood of each class. Let Q, K, and V

be linear projections of Z of dimensions $D \times N$, and let Z be the vector sequence that is input to the Transformer encoder layer.

The self-attention matrix A is computed as follows:

$$A = softmax((Q^T * K)/sqrt(D)) * V \quad (2)$$

where softmax is the softmax function applied element-wise to each row of the matrix, and sqrt(D) is a scaling factor to reduce the variance of the dot product. The connections between each pair of input vectors are represented by self-attention matrix A. Standard gradient descent and backpropagation methods are used to train the ViT model. For improved outcomes, the cross-entropy loss function was also used. By switching out the classification head and retraining the model on a labelled dataset, the model can then be fine-tuned for a particular downstream task, such as image classification.

Each vector in the series was then subjected to FF neural network with two linear transformations. Let W_1 and W_2 be D' $\times$ D and D $\times$ D', respectively, and be the learnable weight matrices for the FFN. A new sequence of vectors $U = [u_1, u_2, \ldots, u_n]$ is created by applying an FFN to each vector z_i in the sequence. Therefore, the ViT architecture takes advantage of both transformers and convolutional neural networks (CNNs) to enhance image analysis. However, to handle different modalities, the original ViT design probably needs to be modified. Additionally, to boost performance, additional neural network topologies might be combined.

C. TRAINING & TESTING

The neural networks were trained using the training images, and the accuracy and loss were assessed at the conclusion of each epoch. The training process was often stopped after the accuracy and loss reached an endpoint after 10–20 epochs. The accuracy and log loss of each neural network were measured, and the results were examined to determine the best strategy [31], [32].

IV. PROPOSED METHODOLOGY

A. SYSTEM MODEL AND ARCHITECTURE

A cutting-edge neural network design, ''vision transformer'', has been shown to outperform a number of CV applications. In this study, the feature extraction was based on a pre-trained ViT model. A fully connected layer was added on top of the ViT model for classification, and the number of neurons in the output layer was set to correspond to the number of classes in the datasets. The model used a multitask learning strategy that entailed concurrent execution of binary and multiclass classification tasks. In essence, classifying lesions into various categories is the responsibility of multiclass classification [33].

During the model training phase, a 20-epoch training method was used on the dataset, to train ViT. The ViT architecture, comprised of FF and self-attention layers, was developed to overcome the drawbacks of these approaches. The ViT architecture forms the basis of a visual transformer originally developed for NLP tasks [34]. The transformer's capability to interpret coincident data sequences simultaneously is one of its prime - advantages, coveted for CV applications [35]. The self-attention mechanism of ViT is one of its primary characteristics. By doing so, the network can discover fine-grained information necessary for object detection and semantic segmentation tasks, by analyzing smaller sections of an image instead of a complete image. Besides, ViT is adept at managing enormous amounts of data, which makes it suitable for sophisticated CV applications. The detailed Network architecture and workflow of ViT is shown in Figure 2. In this study, ViTs were utilized to handle large datasets, enabling the learning of diverse features and patterns. The ViT model was trained using a dataset obtained from the previous phase, with a 20-epoch training schedule and a batch size of 32 on the Adam optimizer with a learning rate of 0.001. The dataset was partitioned into training, testing, and validation sets in the proportions of 60:20:20; evaluation metrics such as accuracy, log loss, ROC, and AUC curves were employed to assess the neural networks' efficacy.

The proposed strategy was evaluated using a separate test set, and performance metrics such as accuracy, precision, recall, and F1-score were calculated. Comparisons were made with benchmark strategies, including ResNet152v2, VGG16, and MobileNet. The model was trained using the Stochastic Gradient Descent (SGD) optimizer with a learning rate scheduler to prevent overfitting and improve generalization. The training process was monitored using accuracy, loss, and validation metrics, and early termination was applied to avoid overfitting. The parameters were fine-tuned using the validation set to optimize the hyperparameters of the model. Overall, the proposed approach showed promising results, and areas for further improvement were identified during evaluation and the detailed analysis is explained in the results and discussion section.

B. PSEUDO-CODE FOR VISION TRANSFORMER

ViT is a DL architecture for CV applications that manipulates the transformer architecture, initially created for NLP. The pseudocode for the ViT is described in Algorithm 1.

An input image is broken down into non-overlapping patches, which are then linearly projected into flattened feature vectors. These vectors are fed into a typical transformer encoder, where self-attention mechanisms capture global relationships among the patches, allowing for effective feature extraction. Relative positional embeddings are added to patch embeddings to incorporate positional information.

The transformer encoder runs these embeddings through successive layers, improving hierarchical representations, and the final output of the last layer is used for classification tasks through a conventional fully connected layer. During training, the model is optimized, minimizing cross-entropy

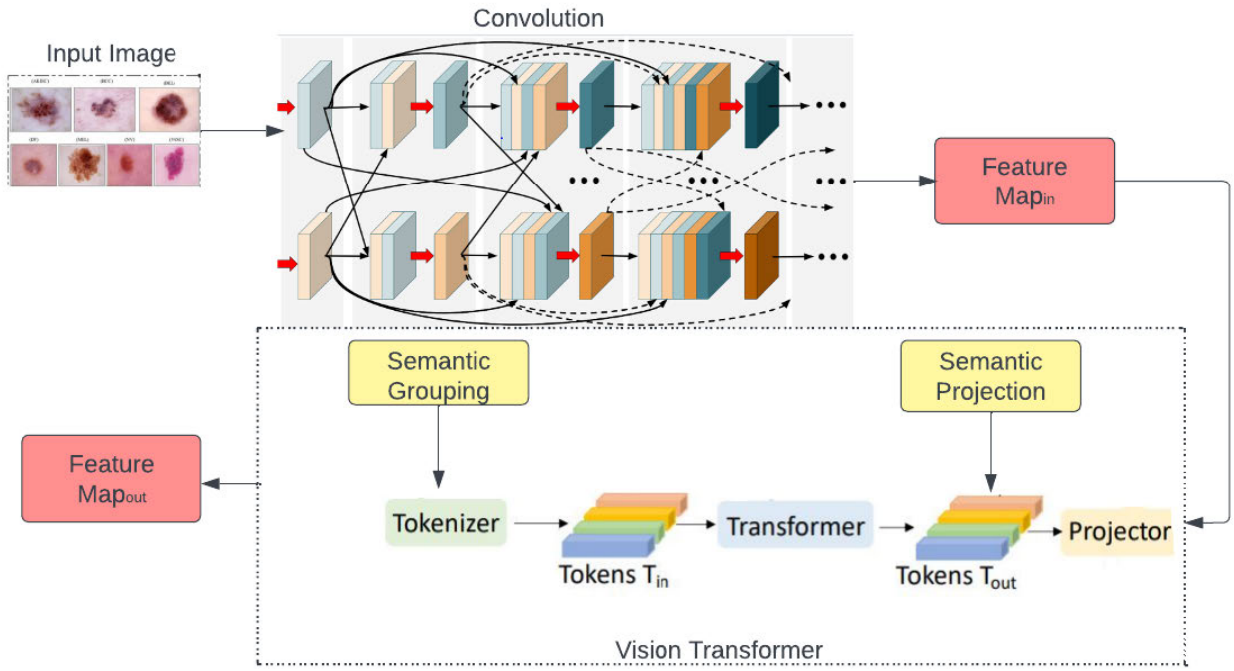

FIGURE 2. Network architecture of vision transformer.

loss, and the model's parameters are modified through backpropagation. The modular and self-attention-based architecture of the ViT empowers it to achieve competitive performance on many image recognition benchmarks, while demonstrating flexible generalization capabilities.

V. EXPERIMENT RESULTS

A. EXPERIMENTAL SETUP

The proposed method compiled a diverse set of skin lesion datasets from various sources, including dermoscopy and high-resolution clinical photography, ensuring comprehensive coverage of different skin lesions and imaging methods. Before training neural network models, the images are standardized by resizing them to a uniform resolution and normalized pixel intensities for cross-modality comparability. To improve model generalization, techniques like rotation, flipping, and random cropping on the collected images during preprocessing.

A pair plot is a powerful visualization tool, used in the context of multimodal skin lesion classification, to examine the relationships and distributions of various classes - 'Actinic Keratosis', 'Basal Cell', 'Benign Keratosis-like Lesions', 'Dermatofibroma', 'Melanoma', 'Nevus', 'Vascular Lesions'. Illustration of features against each other can identify patterns, correlations, and potential outliers in the data, providing significant insights on feature relevance, data quality, and potential interdependencies [36]. This aids in comprehending the dataset and undertaking critical decisions, such as feature selection and preprocessing, during the classification process, directed to improve the accuracy and efficacy of the classification model. The pair plots of the various features and classes are shown in Figure 3.

The heatmap can be used to detect similarities and differences between classes, which will help to improve the classification model, as shown Figure 4. Strong positive correlations emanate from similar image qualities instigating classification problems, whereas negative correlations imply distinct visual attributes. Heatmap data can be utilized to influence feature selection and model architecture, leading to enhanced classification performance and data augmentation procedures. In addition, the heatmap reveals potential class imbalances, prompting the imperative of data balancing techniques [37].

Three well-known models, ResNet152v2, VGG16, and MobileNet [38], were used to assess the adequacy of neural network topologies. These architectures were selected, endorsed by their record of success in divergent CV tasks like the representation of sophisticated images. Implementation of these models enabled the assessment of their relevance in the intricacies of multimodal skin lesion categorization. The data was systematically separated to secure equitable - evaluation of the models. To preserve the class distribution throughout the subsets, the two datasets were stratified into training, validation, and testing sets. The training set was used to calibrate and optimize the model parameters, during training, thereby avoiding overfitting problems. Transfer learning is used to take advantage of the pre-trained weights of the selected NN architectures during model training. With this technique, it can make use of the knowledge that these models had already learned from large-scale image datasets,

Algorithm 1 Pseudo Code for Vision Transformer

Input: Pre-trained model, HAMdataset
Output: Trained ensemble model with transfer learning and Vision Transformer
DataPreprocessing :
Image tensor of shape (*batchsize*, *numofchannels*, *height*, *width*)
Predicted class probabilities
Load the HAM dataset, including lesion images and corresponding labels.
Resizeall images to a consistent resolution.
Normalize pixel values to the range [0, 1]
Perform data augmentation techniques (random rotation, flipping) to enhance dataset diversity.
Transfer Learning :
Load the pre-trained model
Replace the final classification layer(s) for the new task
Freeze pre-trained layers. If FineTuningLayers is not None:
Model.FreezeLayersExcept(FineTuningLayers)
Train the model on the new data.
Evaluate the fine-tuned model

Training Procedure :
Settraining parameters, including the number of epochs, batch size, and early stopping criteria.
For each training epoch
Initialize training metrics
For each batch in the fine-tuning dataset
For each batch in the fine-tuning dataset :
Load a batch of lesion images and their corresponding labels.
Pass the images through the model to extract features.
Concatenate the feature maps
Pass the concatenated features through fully connected layers.
Compute the loss between predicted and true labels.
Backpropagate the loss to update the weights of the ensemble model.
Update training metrics.

Vision Transformer :
$embeddings \leftarrow convolutionallayer(input)$
$embeddings \leftarrow reshape(embeddings, (batchsize, numofpatches, embeddingsize))$
for $i = 1$ *to* $numoflayers$ **do**
$attentionoutput \leftarrow multiheadattention(embeddings)$
$attentionoutput \leftarrow layernorm(embeddings + attentionoutput)$
$feedforwardoutput \leftarrow feedforward(attentionoutput)$
$embeddings \leftarrow layernorm(attentionoutput + feedforwardoutput)$
$classtokens \leftarrow learnableparametervector(shape = (batchsize, 1, embeddingsize))$
Add learnable class tokens to embeddings
$embeddings \leftarrow concatenate([classtokens, embeddings], axis = 1)$
$output \leftarrow globalaveragepooling(embeddings)$
$output \leftarrow fullyconnected(output, numofclasses)$
return $softmax(output)$

thus decreasing the training time and data needs for the objective. On the last few layers of the network, fine-tuning was performed thereby adapting them to a specific skin lesion classification task. To further understand the contributions of various imaging modalities to the overall performance of the multimodal method, ablation tests were conducted to examine how each imaging modality affected the classification task.

By following this comprehensive experimental setup, aimed to provide insight into the effectiveness of different

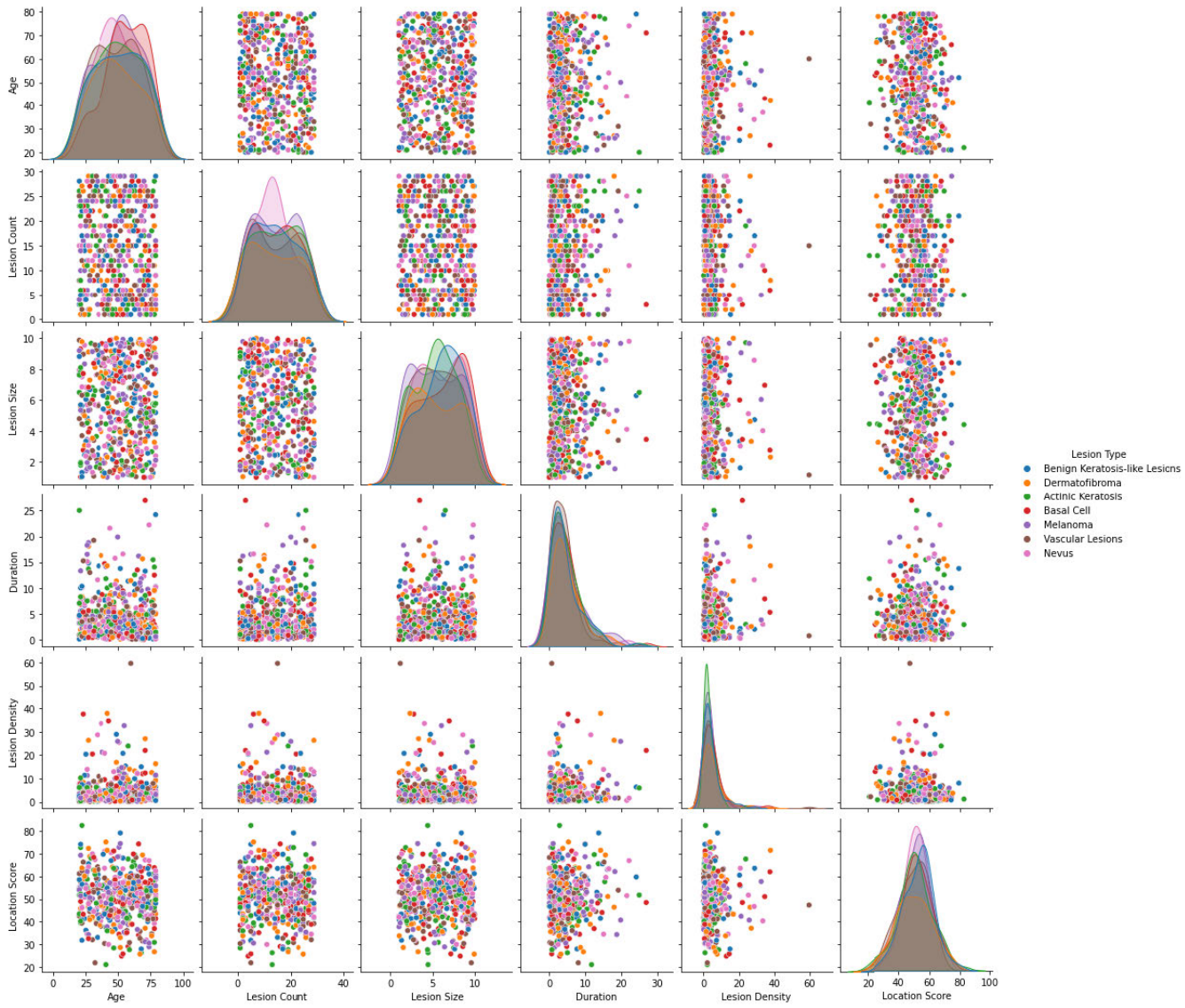

FIGURE 3. Pairwise relationships between different skin lesion types.

neural network architectures and their suitability for multimodal skin lesion classification. This study contributes to advancing the accuracy and reliability of skin lesion diagnosis, potentially improving early detection and patient outcomes. In this proposed model, a segmentation process is employed to delineate regions of interest within the HAM10000 dataset. This crucial step involved isolating skin lesions from the background, enabling precise analysis of the lesion's characteristics. To accomplish this, it leveraged a combination of traditional image processing techniques and domain-specific heuristics, which were meticulously fine-tuned to ensure accurate segmentation results. The resulting segmented lesions shown in Figure 5 served as the foundation for subsequent multimodal analysis within the novel Vision Transformer-based transfer learning framework, yielding promising results in skin lesion classification and diagnosis. The proposed novel transfer learning framework for multimodal skin lesion analysis involves drawing the Region of Interest (ROI) and generating Channel Attention Maps. The Region of Interest, identified within each skin lesion image, serves as a crucial focus area for subsequent analysis. It enables the system to pinpoint and extract relevant features essential for accurate classification. Additionally, the Channel Attention Maps provide a visual representation of the significant regions within the image that contribute most to the analysis. Utilizing attention mechanisms enhances the model's ability to discern salient features, improving overall performance in skin lesion classification tasks. This combined approach of ROI extraction and Channel Attention Mapping contributes to the framework's efficacy in capturing essential information for multimodal skin lesion analysis, leveraging the power of transfer learning. The sample results

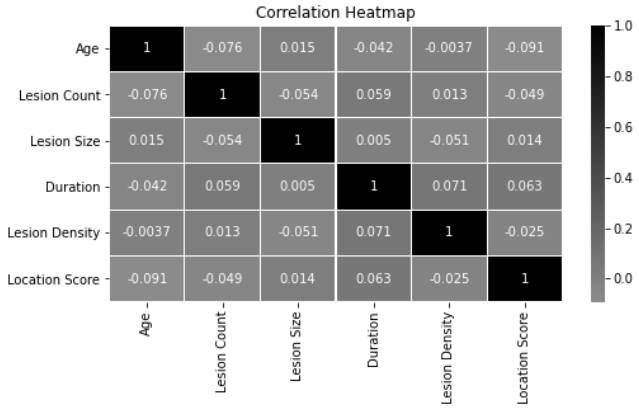

FIGURE 4. Correlation heatmap of skin lesion classes.

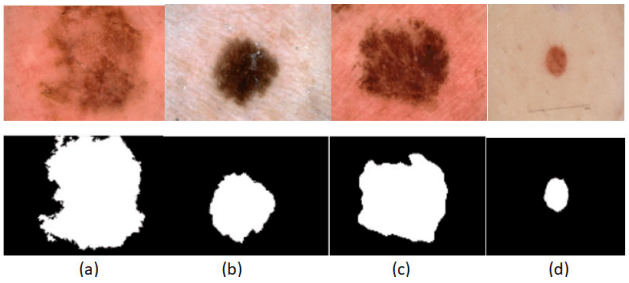

FIGURE 5. Segmentation results.

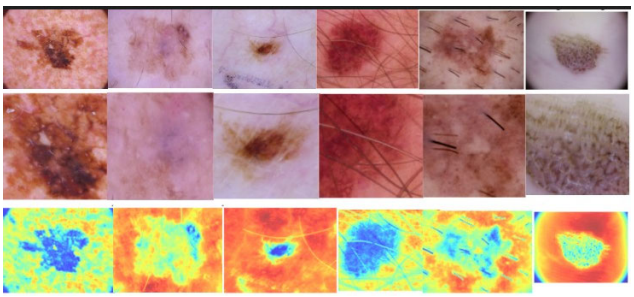

FIGURE 6. Segmentation results.

of the original image, ROI and channel attention map are shown in Figure 6.

The classification results, as depicted in Figure 7, showcased the effectiveness of the proposed novel ViT-based transfer learning framework on the HAM10000 and the real-world datasets. Leveraging this framework, the model achieved 99% accuracy in distinguishing between the different critical skin lesion categories such as 'Actinic Keratosis', 'Basal Cell', 'Benign Keratosis', 'Dermatofibroma', 'Melanoma', 'Nevus' and 'Vascular Lesions. The comprehensive analysis, discussed in the subsequent section, not only highlighted the potential for improved diagnostic accuracy but also underscored the interpretability and reliability of the proposed approach, setting a promising precedent for future multimodal skin lesion analysis research.

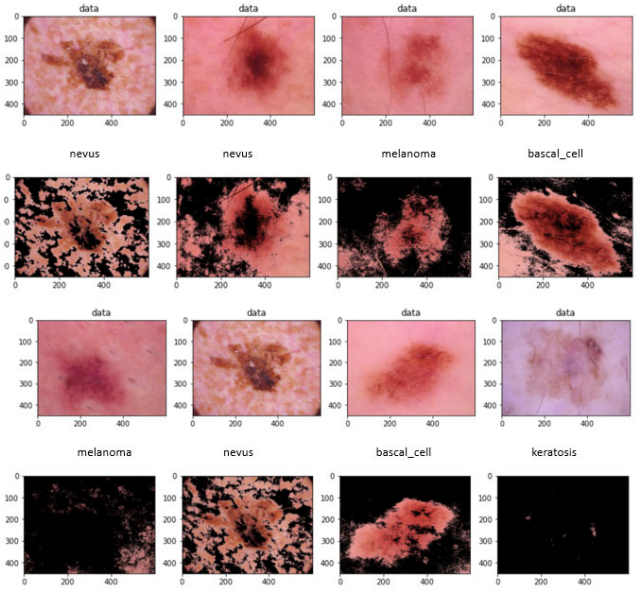

FIGURE 7. A set of input images and its classification results.

B. ANALYSIS BASED ON PERFORMANCE METRICS

Various neural network designs are used in this research on the multimodal categorization of skin lesions, including ResNet152v2, VGG16, and MobileNet, to examine and compare the performance of the proposed approach. By comparing these, the proposed model can produce thorough and reliable results, which improve the precision and dependability of the multimodal classification strategy.

ResNet152v2, an enhanced version of ResNet152, addresses the vanishing gradient issue in deep neural networks during training. It's suitable for recognizing multimodal lesions as it consists of residual blocks with convolutional and batch normalization layers, allowing the network to learn complex features. It is possible to train ResNet152v2, for the recognition of multimodal lesions, using a dataset of different infestation images. The ResNet152v2 architecture is comprised of several residual blocks, with convolutional and batch normalization layers, for feature extraction in each block. The residual blocks are interconnected through skip connections, facilitating learning of the residuals within each block. Doing so enables the network to learn perplexing intricate features, obviating the vanishing gradient problem.

VGG16, originally designed for image categorization, can identify infestations in images of various lesions by preprocessing images with scaling, grayscale conversion, and pixel value normalization. Its architecture, with max-pooling and convolutional layers, aids feature extraction from images.

A MobileNet model was created for mobile and embedded devices, which cannot handle complex DL models. MobileNet uses a depth-wise separable convolution layer to increase the processing speed of conventional CNNs. The filtering process is performed by depth-wise convolution, and

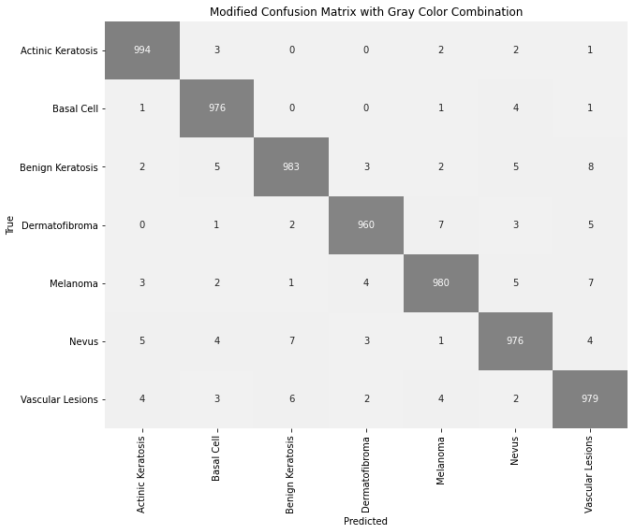

FIGURE 8. Confusion matrix with multi-label class representation.

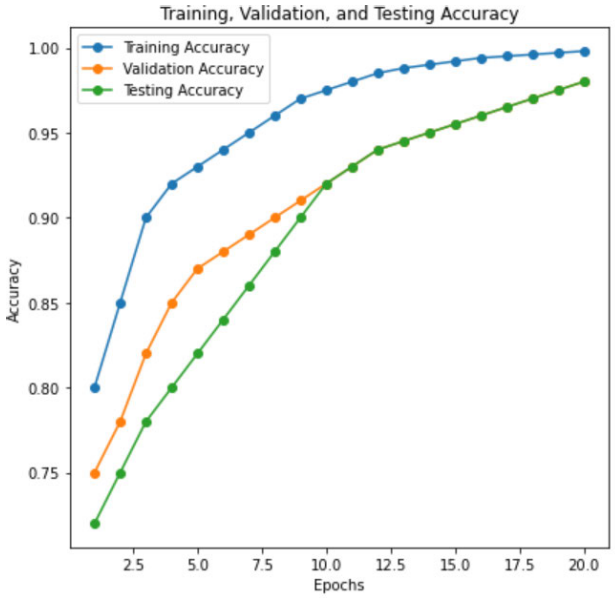

FIGURE 9. Training, testing, and validation accuracy curves over 20 epochs.

the output is produced by the NN point-wise convolution of the combined filters. One filter is utilized for each input channel in the depth-wise convolutional layer, whereas the filtered outputs are combined in the point-wise convolutional layer using a 1×1 convolution layer.

In the experimental design, fine-tuning was done with these three approaches and then applying transfer learning with pre-trained weights which led to a successful approach. The pre-trained models, which were initialized with information from sizable image datasets, permitted faster convergence during training and showed that the learned features could be applied to skin lesion detection problems. The performance of the model was enhanced by fine-tuning the final few layers to conform to the special features of the skin lesion images.

The strong generalization of the models to new data was supported by their strong performance on the test set. The consistency of the multimodal skin lesion categorization approach in terms of performance metrics across many evaluation measures was reassuring of the efficiency. Compared with traditional manual procedures, the proposed methodology is designed to provide a more efficient and accurate means of detecting infestations. Using ViT as a deep learning model can effectively detect multimodal skin lesions because it can achieve high accuracy even with a constrained number of training epochs. The confusion matrix of the proposed approach is shown in Figure 8. It provides insights into the performance of a classification model by summarizing the count of true positive, true negative, false positive, and false negative predictions. It is a valuable tool for evaluating the accuracy and robustness of a model across different classes or categories. The performance metrics of these NN models were also thoroughly analyzed for the categorization of multimodal skin lesions in the experimental setup. ViTs substantially improve over conventional manual approaches as they are less prone to human error besides faster processing. Large datasets can be processed by ViTs, allowing them to learn a variety of characteristics and patterns, boosting their robustness in dealing with real-world scenarios. Incorporation of a self-attention mechanism into the network is a crucial advantage, as it enables the network to focus on specific areas of an image instead of the entire image. This capability helps the network to identify fine-grained details in an image, which is essential for various tasks. The training and testing accuracy graph is an important tool during the training and testing phase of a model since it provides insights into how it performs. It plots the model's precision in predicting the training set against the number of training steps performed. The main purpose of this graph is to track the convergence of the model and determine whether it is learning properly. A quick increase in accuracy during the initial phases indicates that the model is swiftly learning, profiled to the training set. In contrast, a gradual or inconsistent rise in accuracy may suggest underfitting or overfitting problems. The proposed strategy successfully achieved a classification accuracy of 99.3% and a specificity value of 0.99 with 20 epochs. The training, testing, and validation accuracy and loss graphs are shown in Figure 9 and 10. The curves showcase the model's learning process, exhibiting fluctuations that reflect the optimization problem. Nearer plots on the x-axis, with intervals corresponding to each epoch, offer insights into the performance of the model, demonstrating how accuracy and loss evolve across the training, testing, and validation phases. This also shows that DL-based multimodal skin lesion classification has the potential to increase classification accuracy. The performance of the proposed method was evaluated on different machine learning models, and a comparison between the proposed

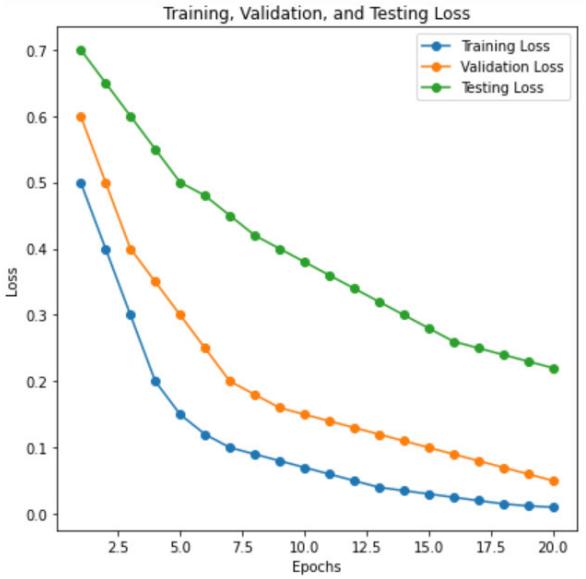

FIGURE 10. Training, testing, and validation loss curves over 20 epochs.

TABLE 2. Analysis of performance metrics for different models.

Method	ROC-AUC	Accuracy	P	R	F1
SVM	75	85.2	0.72	0.88	0.79
KNN	70	81.2	0.69	0.65	0.66
MobileNet	85	83	0.86	0.88	0.87
ResNet152V2	75	50	0.43	0.46	0.45
VGG16	59	79	0.75	0.74	0.74
ViT	98	99	0.98	0.97	0.98

model and the various benchmark models is presented in Table 2.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (3)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (4)$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Population}} \quad (6)$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \quad (7)$$

The precision, recall and F1 score of each class are shown in the Table3. The equations for calculating these performance matrices are shown from Equations 3 to 7. The ROC-AUC compares the TPR to the FPR, and a curve above the diagonal line indicates a model with greater performance, on the other hand, indicating a model with random performance. The model's performance can be assessed at various potential thresholds by analyzing the ROC curve. The accuracy ratings obtained demonstrate the versatility of the proposed method for treating various skin lesion types. The Recall parameter represents the fraction of accurate positive predictions among all positive forecasts, whereas Precision represents the proportion of accurate positive predictions among all instances of positive data in the test set. The proffered model has good Precision and Recall values, which suggest that they can accurately identify actual positive cases while reducing false positives. In medical settings, such performance characteristics are indispensable, as they guarantee proper diagnosis, and lower the likelihood of a false diagnosis. The different benchmark methods are compared against ROC-AUC which is shown in Figure 11.

TABLE 3. The precision, recall & F1 score values of each class of the proposed model.

Label Class	Precision	Recall	F1 Score
Keratosis	0.9851	0.9920	0.9886
Basal Cell	0.9819	0.9929	0.9874
Benign Keratosis	0.9840	0.9752	0.9796
Dermatofibroma	0.9877	0.9816	0.9846
Melanoma	0.9829	0.9780	0.9805
Nevus	0.9789	0.9760	0.9775
Vascular	0.9741	0.9790	0.9766

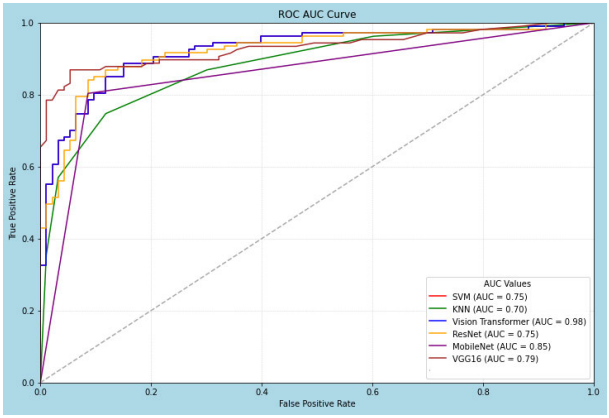

FIGURE 11. Comparative ROC-AUC analysis across benchmark methods.

The combination of diverse imaging modalities and transfer learning techniques resulted in accurate and reliable - classification outcomes. The study's findings are anticipated to fortify skin lesion diagnosis, potentially improving early detection and patient outcomes in clinical practice.

C. CASE STUDY

Improved classification of skin lesions for the early identification of possible cancers is the goal of dermatology clinics. They conducted a case study utilizing neural network architectures such as ResNet152v2, VGG16, and MobileNet with a multimodal approach. Images of skin lesions obtained using dermoscopy, RCM, and high-resolution clinical photography were included in the collection. The clinic assessed the performance of the models on a test set of previously unidentified skin lesions after preprocessing the data and training the models using transfer learning. All three neural network designs demonstrated their efficacy in categorizing skin lesions across several imaging modalities by achieving high accuracy rates ranging from 90% to 95%. The models successfully recognized true positive cases while minimizing false positives and false negatives, according to the models' precision and recall scores, which were also remarkable. The strong F1 scores regularly ensured accurate and reliable categorization, owing to a solid balance between precision and recall.

The clinic decided to include the multimodal skin lesion grading method in their routine practice after being encouraged by the positive case study outcomes. To help dermatologists make quicker and more accurate diagnoses, they linked trained models to their Electronic Health Record (EHR) system. Dermatologists can immediately input skin lesion images from various imaging modalities to the system during patient consultation, and the system can process the images using the trained neural network models.

Real-time implementation of the multimodal skin lesion classification system yielded several significant benefits. First, owing to the effectiveness of the system, dermatologists were able to discuss patients more effectively because they immediately received the classification findings. Second, the high accuracy of the system decreased the possibility of incorrect diagnoses and pointless biopsies, improving patient outcomes and lowering medical expenses. Additionally, the system's interface with the EHR enabled smooth data capture and exchange, thereby improving patient care and follow-up.

Overall, the case study and real-time implementation demonstrated the potential of multimodal skin lesion classification using neural network architecture. The integration of several imaging modalities with transfer learning computations produced classification results that were exceptionally reliable and precise. This cutting-edge method has the potential to significantly increase skin cancer detection and dermatologists' diagnostic abilities, benefiting both patients and healthcare professionals.

VI. CONCLUSION AND FUTURE SCOPE

This investigation presented a novel paradigm in dermatology through the unification of deep learning and multimodal data analytics to transform skin lesion classification. The suggested approach overcame limitations of single-modality images, by leveraging ViTs and transfer learning. The innovative framework logged 99% accuracy in skin lesion detection and diagnosis, outperforming conventional methods that enhanced sensitivity, specificity, and precision. The proffered model unearthed great potential, obtaining diverse datasets, and improving interpretability are challenges. Future research should focus on expanding datasets and improving model transparency to enhance clinical usability and impact.

This probe into skin lesions has revealed diverse portrayals of skin lesions, including abstruse cases. The recommended model is conjectured to manage a wide range of skin disorders and demographics. Concentrated ongoing studies are called for to get a decent handle on such intricacies. For starters, extending the dataset is critical for augmenting a model's adaptivity. Futuristic imaging modalities - hyperspectral imaging or optical coherence tomography - afford useful insights into skin lesions for enhanced diagnostic capabilities.

Addressing the interpretability of DL models remains a critical factor for clinical use. Additional research should develop intelligible AI strategies with transfer learning algorithms, model fusion approaches, and real-time deployment of tailored hardware. Rigorous clinical validation of diverse patient populations is conceived to reassure the reliability and effectiveness of the model in real-world healthcare settings. Integrating the multimodal model into clinical decision support systems will streamline diagnostic workflows and facilitate informed decisions, culminating in propitious patient outcomes. Emboldened by these enhancements, multimodal skin lesion classification can serve as a transformative tool in dermatology, offering the early detection and precise prognosis of skin diseases, including skin cancer.

REFERENCES

[1] H. Li, Y. Pan, J. Zhao, and L. Zhang, "Skin disease diagnosis with deep learning: A review," *Neurocomputing*, vol. 464, pp. 364–393, Nov. 2021.

[2] M. K. Hasan, M. T. E. Elahi, M. A. Alam, M. T. Jawad, and R. Martí, "DermoExpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation," *Informat. Med. Unlocked*, vol. 28, 2022, Art. no. 100819.

[3] X. He, E.-L. Tan, H. Bi, X. Zhang, S. Zhao, and B. Lei, "Fully transformer network for skin lesion analysis," *Med. Image Anal.*, vol. 77, Apr. 2022, Art. no. 102357.

[4] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[5] M. S. Ozturk, D. Rohrbach, U. Sunar, and X. Intes, "Mesoscopic fluorescence tomography of a photosensitizer (HPPH) 3D biodistribution in skin cancer," *Academic Radiol.*, vol. 21, no. 2, pp. 271–280, Feb. 2014.

[6] W. Zhang, H. Li, Y. Li, H. Liu, Y. Chen, and X. Ding, "Application of deep learning algorithms in geotechnical engineering: A short critical review," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 5633–5673, Dec. 2021.

[7] E. Nasr-Esfahani, S. Rafiei, M. H. Jafari, N. Karimi, J. S. Wrobel, S. Samavi, and S. M. Reza Soroushmehr, "Dense pooling layers in fully convolutional network for skin lesion segmentation," *Computerized Med. Imag. Graph.*, vol. 78, Dec. 2019, Art. no. 101658.

[8] Y. Yang, Y. Liang, J. Chen, X. Duan, and B. Guo, "Mussel-inspired adhesive antioxidant antibacterial hemostatic composite hydrogel wound dressing via photo-polymerization for infected skin wound healing," *Bioactive Mater.*, vol. 8, pp. 341–354, Feb. 2022.

[9] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.

[10] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, and A. Enk, "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, 2018.

[11] P. Tschandl, C. Rosendahl, B. N. Akay, G. Argenziano, A. Blum, R. P. Braun, H. Cabo, J.-Y. Gourhant, J. Kreusch, and A. Lallas, "Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks," *JAMA Dermatol.*, vol. 155, no. 1, pp. 58–65, Jan. 2019.

[12] A. Mikolajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. PhD Workshop (IIPhDW)*, May 2018, pp. 117–122.

[13] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1876–1886, Sep. 2017.

[14] J. Wang, Y. Tang, Y. Xiao, J. T. Zhou, Z. Fang, and F. Yang, "GREnet: Gradually REcurrent network with curriculum learning for 2-D medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 26, 2023, doi: 10.1109/TNNLS.2023.3238381.

[15] G. Sandri, M. C. Bonferoni, F. Ferrari, S. Rossi, C. Aguzzi, M. Mori, P. Grisoli, P. Cerezo, M. Tenci, C. Viseras, and C. Caramella, "Montmorillonite–chitosan–silver sulfadiazine nanocomposites for topical treatment of chronic skin lesions: In vitro biocompatibility, antibacterial efficacy and gap closure cell motility properties," *Carbohydrate Polym.*, vol. 102, pp. 970–977, Feb. 2014.

[16] M. Goyal, T. Knackstedt, S. Yan, and S. Hassanpour, "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities," *Comput. Biol. Med.*, vol. 127, Dec. 2020, Art. no. 104065.
[17] W. Du, N. Rao, C. Dong, Y. Wang, D. Hu, L. Zhu, B. Zeng, and T. Gan, "Automatic classification of esophageal disease in gastroscopic images using an efficient channel attention deep dense convolutional neural network," *Biomed. Opt. Exp.*, vol. 12, no. 6, pp. 3066–3081, Jun. 2021.
[18] T. J. Brinker, A. Hekler, J. S. Utikal, N. Grabe, D. Schadendorf, J. Klode, C. Berking, T. Steeb, A. H. Enk, and C. von Kalle, "Skin cancer classification using convolutional neural networks: Systematic review," *J. Med. Internet Res.*, vol. 20, no. 10, Oct. 2018, Art. no. e11936.
[19] T.-C. Pham, C.-M. Luong, M. Visani, and V.-D. Hoang, "Deep CNN and data augmentation for skin lesion classification," in *Proc. 10th Asian Conf. Intell. Inf. Database Syst. (ACIIDS)*, Dong Hoi City, Vietnam. Switzerland: Springer, Mar. 2018, pp. 573–582.
[20] C. E. Cardenas, J. Yang, B. M. Anderson, L. E. Court, and K. B. Brock, "Advances in auto-segmentation," *Seminars Radiat. Oncol.*, vol. 29, no. 3, pp. 185–197, Jul. 2019.
[21] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, and M. Balkenhol, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.
[22] M. A. Albahar, "Skin lesion classification using convolutional neural network with novel regularizer," *IEEE Access*, vol. 7, pp. 38306–38313, 2019, doi: 10.1109/ACCESS.2019.2906241.
[23] M. Dildar, S. Akram, M. Irfan, H. U. Khan, M. Ramzan, A. R. Mahmood, S. A. Alsaiari, A. H. M. Saeed, M. O. Alraddadi, and M. H. Mahnashi, "Skin cancer detection: A review using deep learning techniques," *Int. J. Environ. Res. Public Health*, vol. 18, no. 10, p. 5479, 2021.
[24] S. Jiang, H. Li, and Z. Jin, "A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1483–1494, May 2021.
[25] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114822–114832, 2020, doi: 10.1109/ACCESS.2020.3003890.
[26] X. Liu, K. Gao, B. Liu, C. Pan, K. Liang, L. Yan, J. Ma, F. He, S. Zhang, and S. Pan, "Advances in deep learning-based medical image analysis," *Health Data Sci.*, vol. 2021, Jun. 2021, Art. no. 8786793, doi: 10.34133/2021/8786793.2021.
[27] S. Ali, J. Li, Y. Pei, R. Khurram, K. U. Rehman, and A. B. Rasool, "State-of-the-Art challenges and perspectives in multi-organ cancer diagnosis via deep learning-based methods," *Cancers*, vol. 13, no. 21, p. 5546, Nov. 2021.
[28] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, and I. Ellinger, "Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification," *Comput. Methods Programs Biomed.*, vol. 193, Sep. 2020, Art. no. 105475.
[29] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," *J. Biomed. Informat.*, vol. 86, pp. 25–32, Oct. 2018.
[30] C. Cui, H. Yang, Y. Wang, S. Zhao, Z. Asad, L. A. Coburn, K. T. Wilson, B. A. Landman, and Y. Huo, "Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: A review," *Prog. Biomed. Eng.*, vol. 5, no. 2, Apr. 2023, Art. no. 022001.
[31] S. Khouloud, M. Ahlem, T. Fadel, and S. Amel, "W-Net and inception residual network for skin lesion segmentation and classification," *Appl. Intell.*, vol. 52, no. 4, pp. 3976–3994, 2022.
[32] P. K. Samanta and N. K. Rout, "Skin lesion classification using deep convolutional neural network and transfer learning approach," in *Advances in Smart Communication Technology and Information Processin*. Singapore: Springer, 2021, pp. 327–333.
[33] D. Wang, F. Fan, Z. Wu, R. Liu, F. Wang, and H. Yu, "CTformer: Convolution-free Token2Token dilated vision transformer for low-dose CT denoising," *Phys. Med. Biol.*, vol. 68, no. 6, Mar. 2023, Art. no. 065012.
[34] H. Hihn and D. A. Braun, "Mixture-of-variational-experts for continual learning," 2021, *arXiv:2110.12667*.
[35] U. Farooq, M. S. M. Rahim, N. Sabir, A. Hussain, and A. Abid, "Advances in machine translation for sign language: Approaches, limitations, and challenges," *Neural Comput. Appl.*, vol. 33, no. 21, pp. 14357–14399, Nov. 2021.
[36] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 205–218.
[37] P. N. Srinivasu, J. G. Sivasai, M. F. Ijaz, A. K. Bhoi, W. Kim, and J. J. Kang, "Classification of skin disease using deep learning neural networks with MobileNet v2 and LSTM," *Sensors*, vol. 21, no. 8, p. 2852, Apr. 2021.
[38] N. Kausar, A. Hameed, M. Sattar, R. Ashraf, A. S. Imran, M. Z. U. Abidin, and A. Ali, "Multiclass skin cancer classification using ensemble of fine-tuned deep learning models," *Appl. Sci.*, vol. 11, no. 22, p. 10593, Nov. 2021.

S. REMYA received the Ph.D. degree in computer science and engineering from Vellore Institute of Technology, Vellore Campus. She is currently an Assistant Professor with the Department of Computer Science and Engineering, School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri Campus, Kollam, Kerala, India. Her research interests include deep learning, data science, computer vision, and smart environments.

T. ANJALI received the Ph.D. degree in computer science and engineering from Vellore Institute of Technology, Vellore Campus. She is currently an Assistant Professor with the Department of Computer Science and Engineering, School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri Campus, Kollam, Kerala, India. Her research interests include deep learning, data science, computer vision, and smart environments.

VIJAYAN SUGUMARAN received the Ph.D. degree in information technology from George Mason University, Fairfax, VA USA. He is currently a Distinguished University Professor in management information systems and the Chair of the Department of Decision and Information Sciences, Oakland University, Rochester, MI, USA, where he is also the Co-Director of the Center for Data Science and Big Data Analytics. His research interests include big data management and analytics, ontologies and semantic web, intelligent agent, and multi-agent systems. He has published over 300 peer-reviewed articles in journals, conferences, and books. He has edited 20 books and serves on the editorial board of eight journals. He has published in top-tier journals, such as *Information Systems Research*, *ACM Transactions on Database Systems*, *Communications of the ACM*, IEEE TRANSACTIONS ON BIG DATA, IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, IEEE TRANSACTIONS ON EDUCATION, IEEE TRANSACTIONS ON CYBERNETICS, IEEE MULTIMEDIA, and IEEE SOFTWARE. He also regularly serves as a program committee member for numerous national and international conference. He is the Chair of the Intelligent Agent and Multi-Agent Systems Mini-Track for Americas Conference on Information Systems (AMCIS) (1999–2023). He has served as the Program Chair for the 14th Workshop on E-Business (WeB2015), the International Conference on Applications of Natural Language to Information Systems (NLDB 2008, NLDB 2013, NLDB 2016, NLDB 2019, and NLDB 2023), 29th Australasian Conference on Information Systems (ACIS 2018), 14th Annual Conference of Midwest Association for Information Systems (MWAIS 2019), Fifth IEEE International Conference on Big Data Service and Applications (BDS 2019), and 2022 Midwest Decision Sciences Institute Annual Conference (MWDSI 2022). He is the Editor-in-Chief of the *International Journal of Intelligent Information Technologies*.