MLFF-Net: a multi-model late feature fusion network for skin disease classification

Ajay Krishan Gairola[1,2], Vidit Kumar[2], Ashok Kumar Sahoo[1]
[1]Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, India
[2]Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India

Article Info

Article history:

Received Feb 26, 2024
Revised Apr 30, 2024
Accepted May 7, 2024

Keywords:

CNN
Deep learning
Feature extraction
Fusion
Skin disease
Vision transformer

ABSTRACT

Early diagnosis is paramount to preventing skin diseases and reducing mortality, given their global prevalence. Visual detection by experts using dermoscopy images has become the gold standard for detecting skin cancer. However, a significant challenge in skin cancer detection and classification lies in the similarity of appearance among skin disease lesions and the complexity of dermoscopic images. In response, we developed multi-model late feature fusion network (MLFF-Net), a multi-model late feature fusion network tailored for skin disease detection. Our approach begins with image pre-processing techniques to enhance image quality. We then employ a two-stream network comprising an enhanced densely linked network (DenseNet-121) and a vision transformer (ViTb16). We leverage shallow and deep feature fusion, late fusion, and an attention module to enhance the model's feature extraction efficiency. The subsequent feature fusion module constructs multi-receptive fields to capture disease information across various scales and uses generalized mean pooling (GeM) pooling to reduce the spatial dimensions of lesion characteristics. Finally, we implement and test our skin lesion categorization model, demonstrating its effectiveness. Despite the combination, convolutional neural network (CNN) outperforms ViT approaches, with our model enhancing the accuracy of the best model by 6.1%.

Corresponding Author:

Ajay Krishan Gairola
Department of Computer Science and Engineering, Graphic Era Deemed to be University
Dehradun, India
Email: ajaykgairola.cc@geu.ac.in

1. INTRODUCTION

Common proliferative diseases of the skin include skin lesions, which present in a wide variety and are clinically categorized as benign or malignant. In most cases, therapy is unnecessary for benign skin lesions [1] because they typically exhibit slow growth, do not invade nearby structures, and do not cause systemic effects. Malignant skin lesions, on the other hand, can invade neighboring tissues and metastasize to other organs, making prompt identification and treatment crucial. Therefore, developing a fast and precise approach for detecting skin cancer at an early stage is imperative. Artificial intelligence (AI) [2] has advanced to the point where it can now diagnose skin conditions using computer algorithms. This technology has the potential to significantly advance dermatology by aiding in the development of new diagnostic methods [3], [4]. To further enhance the accuracy of AI in diagnosis, dermatologists can contribute by providing the system with clinical data from their daily practice.

Khouloud *et al.* [5] introduced a deep-learning system for melanoma identification, combining a segmentation network and a classification network. The W-Net demonstrated superior accuracy and performance over conventional neural networks in both segmentation and classification tasks. Benyahia *et al.* [6] explored 24 AI classification methods and 17 pre-trained convolutional neural network (CNN) architectures for skin lesion classification. On the PH2 and ISIC2019 datasets, densely linked network (DenseNet-201) [7] coupled with a cubic support vector machine or fine k-nearest neighbors yielded the best results. Benyahia published her findings in the machine learning journal. Popescu *et al.* [8] integrated deep learning with crowdsourcing to enhance skin lesion classification. They developed a weight matrix by evaluating nine classification networks to improve prediction accuracy. Skin lesions were segmented using an optimized color feature (OCF), and a deep CNN was employed for classification, as proposed by Hasan *et al.* [9]. The hybrid technique aimed to enhance lesion contrast while minimizing artifacts. Features were extracted using the deep CNN-9 model before being merged with OCFs. Finally, a normal distribution-based high-ranking feature selection method was used to identify the most robust classification characteristics.

Here, we present a framework for skin disease categorization that leverages late feature fusion. Our primary contributions can be summarized as follows:

- We propose a multi-model late feature fusion network (MLFF-Net) designed for the multi-modal classification of skin diseases. This network integrates features from multiple modalities or sources, allowing for a more comprehensive representation of the data. The proposed representation is then utilized for the classification of several classes of skin diseases, enhancing the accuracy and robustness of the classification process.
- We demonstrate that late feature fusion methods can enhance the performance of our proposed multi-model MLFF-Net by utilizing improved versions of DenseNet-121 and vision transformer (ViTb16) models. These models are used to extract features from different modalities. By combining the input feature maps from these two models, we can capture and leverage the relationship between the two modalities, leading to improved classification performance.
- The feature fusion block (FFB) in our model is designed to create multiple receptive fields, allowing for the gathering of disease information at various scales. Following this, the generalized mean pooling (GeM) technique is employed to reduce the spatial dimensions of lesion features, enhancing the model’s ability to capture and process information efficiently.

2. PROPOSED METHOD

To enhance the accuracy of skin lesion classification, we have developed a MLFF model. Our model is built upon the ViTb16 network and an upgraded version of the DenseNet-121 network. By merging the output features of these two networks and employing multi-receptive field approaches, we achieve multi-granularity and multi-scale global features. The late feature fusion module plays a crucial role in enhancing the model’s ability to distinguish between healthy and unhealthy areas. Our network architecture consists of a two-stream network, a multi-classification module, and the late feature fusion module. The proposed model’s structure is illustrated in Figure 1. The details are mentioned below:

2.1. Data pre-processing

To achieve the intended results and enhance the quality of the images, pre-processing is necessary. Figure 1(a) shows that images could have distracting elements like hair, air bubbles, and noises. Image preparation encompasses:

- Image resizing: the proposed CNN reduces the input image to 224×224 dimensions to accommodate the variety of image sizes. Despite this scaling down, the data from the photograph remains unaltered.
- Image standardization: to normalize the data, we first convert the input image to a collection of pixels with a range of 0 to 1. Splitting each pixel by 255 normalizes the image pixel values, which range from 0 to 255, to a range of pixels from 0 to 1.

2.2. The enhanced DenseNet block’s structure

DenseNet [7] is a renowned approach in image categorization, known for its dense connection method shown in Figure 1(b). In DenseNet, each layer receives feature maps from all preceding layers and passes its own feature maps to all subsequent layers, promoting feature reuse and alleviating the vanishing gradient problem. The DenseNet-201 model, with additional network layers, has been fine-tuned, demonstrating the effectiveness of dense connections in various image classification tasks. Zhou [10] demonstrated the robustness of DenseNet by addressing the vanishing gradient problem. Attention mechanisms, such as the one in SENet [11], offer several benefits in network models, including improved classification performance, enhanced data extraction from images, and the ability to focus on specific regions of interest. SENet’s squeeze-excitation block, designed to improve model representation and capture channel interactions, implements this mechanism. However, adding a squeeze-excitation block to DenseNet’s internal module did not enhance classification performance on a private dataset of acne-like skin disorders. This limitation may stem from SENet's channel reduction method, which does not accurately replicate the connection between input weight vectors and the model. To overcome this limitation, Wang *et al.* [12] introduced efficient channel attention (ECA), which uses 1D convolution to identify channel interactions. ECA includes an efficient excitation module for cross-channel stimulation and a squeeze module for global spatial data, simplifying the concept and eliminating the need for channels to indirectly correspond with their k-nearest neighbors.

2.3. The architecture of the ViT

The ViT architecture begins by dividing the input image into non-overlapping patches of a predefined size. These patches are then augmented with position embeddings, and the resulting tokens are inputted into a transformer encoder. Finally, an MLP head layer is used for image classification. Figure 1(c) provides a visual representation of this basic framework. The transformer network in ViT consists of two main modules: The multi-layer perceptron (MLP) and the multi-head self-attention (MSA). Each module incorporates layer normalization (LN), and both benefit from the residual structure. The formulation for each layer of the transformer can be expressed as (2):

$$O_L = MSA\big(LN(O_{L-1})\big) + O_{L-1} \tag{1}$$

in which O_L stands for the L-th layer's output. As a last point, the classifier in this study is fed the first token from the last layer embedded O_L. The image's multi-scale features are stored in O_L, and a classifier is applied to it for classification.

Feature fusion block, the uneven distribution of affected regions in skin lesion images requires us to define the range of recovered attributes using the size of the convolutional kernel. These regions often exhibit varying diameters and poor continuity. Larger convolutional kernels are better suited for extracting global characteristics from pathological images, while smaller kernels excel at capturing local features. However, relying solely on smaller kernels may necessitate a deeper network to ensure a sufficiently large output feature mapping, potentially leading to overfitting. On the other hand, using larger convolutional kernels can be inefficient as they do not consider local information. Furthermore, stacking larger kernels can increase computational complexity and reduce model efficiency.

2.4. The architecture of the ECA block

In the context of these shared features, we have developed a new block structure for DenseNet by incorporating the ECA block. As inputs, layer L takes in all of the feature mappings that preceded it. The notation $[i\_0, i\_1, \dots, i\_(L-1)]$ stands for the merging of feature mappings produced in layers $0, 1, \dots, L-1$. This combination operation involves several steps, including batch normalization (BN), rectified linear units (ReLUs), pooling, convolution, and an ECA block denoted as $i\_0, i\_1, \dots, N\_L$. Figure 1(d) depicts the construction of both the enhanced block and layer, which are both made up of numerous improved layers connected by dense connections. Our enhanced layer, from top to bottom, includes the following steps: an ECA block, Conv 1×1, BN, ReLU, Conv 3×3, BN, ReLU, and Conv 5×5.

$$i\_L = N\_L\ ([i\_0, i\_1, \dots, i\_(L-1)]) \tag{2}$$

2.5. The FFB

Our goal in creating this feature fusion module is to collect data on a wider range of skin lesions by capturing their multi receptive fields. As depicted in the FFB of Figure 1(e), the multi receptive fields are made up of numerous convolution layers with different kernel sizes. To capture more diseased areas, a multi receptive field must encompass a wider area of skin lesions. Each convolutional layer learns weights that are specific to its respective receptive fields, and the smaller and larger convolution kernels cooperate. This approach enhances the model's overall accuracy by exploring a broader pathological region. To perform non-linear and channel-integrated processing, the feature maps from all the convolutions with different receptive fields are combined and sent into the dropout + ReLU layers. The feature fusion includes a GeM pooling technique [13], [14] to assess the problematic regions derived from the features. The input for the pooling procedure is a 1×1×1536 feature vector F, and the output is a vector V. This vector V_n^{mp} is provided when max pooling is used.

$$V^{mp} = [V_1^{mp} \dots V_n^{mp} \dots V_C^{mp}]^T,\ V_n^{mp} = \max_{f \in F_C} f \tag{3}$$

In which C represents the feature map's channel count. F_C is the set of feature maps where n is a member of (1, C). The network produces a total of C feature maps. All of the features F_C are represented by f, and mp is the max pooling operation. By utilizing average pooling, the vector V_n^a is provided by (4).

$$V^a = [V_1^a \dots V_n^a \dots V_C^a]^T, \; V_n^a = \frac{1}{F_C}\sum_{f \in F_C} f \tag{4}$$

Alternatively, the vector V_n^g is provided by taking advantage of GeM pooling.

$$V^g = [V_1^g \dots V_n^g \dots V_C^g]^T, \; V_n^g = \left(\frac{1}{F_C}\sum_{f \in F_C} f^{hp}\right)^{1/hp} \tag{5}$$

A hyperparameter called hp controls how much each of the two pooling processes weighs. Pooling with maximum and average values is a subset of GeM pooling. Maximum and average pooling are used when hp approaches infinity, and if hp equals 1, it's 0. In the end, V^g is composed of the attributes of every feature map following pooling of GeM, and its dimensionality is C. Various hp values were taught and evaluated, and suitable parameters were chosen, in order to achieve superior classification results in the subsequent comparative tests. To accomplish multiclassification of skin lesions, we employ the SoftMax classifier in the multiclassification module. Algorithm 1 outlines the steps involved in the proposed method.

Algorithm 1. MLFF-Net

```
Input: skin disease images S = {S1, S2, ……Sn}, the initialized network MLFFNet consists of
an improved DenseNet and ViT, maximum epochs E = 60 with a batch size of BS = 16 for the
network, and a learning rate of 0.001.
Output: The optimized MLFF-Net network
While E1 ≤ En × BS do
         Sample a batch of skin disease images.
         Apply data preprocessing (image resizing and standardization) and feed it to the
network.
          Improved DenseNet model by equation (1).
          Transform each layer by equation (2).
          Adding the feature fusion block by equations (3-5).
         Jointly optimize MLFFNet.
End
```

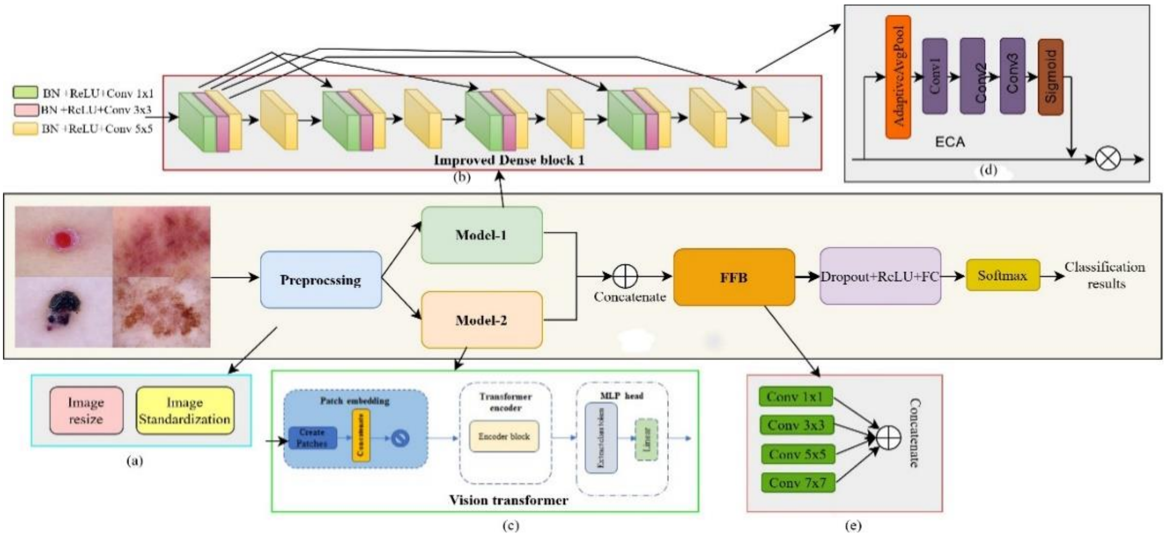

Figure 1. The proposed architecture of the MLFF-Net: (a) the data preprocessing stages, (b) the enhanced DenseNet block's structure, (c) the architecture of the ViT, (d) the architecture of the ECA block, and (e) the FFB

2.6. Performance evaluations

The effectiveness of algorithms for automatically classifying images of skin lesions is often measured by their F1-score, recall, precision, and accuracy. True positives (T_P), false negatives (F_N), and false positives (F_P) proportions provide the basis for these measures' calculation:

$$accuracy = \frac{T\_P+T\_N}{T\_P+T\_N+F\_P+F\_N} \tag{6}$$

$$precision = \frac{T\_P}{T\_P+F\_P} \tag{7}$$

$$recall = \frac{T\_P}{T\_P+F\_N} \tag{8}$$

$$f1 = \frac{2 \times precesion \times recall}{precision+recall} \tag{9}$$

2.7. Dataset

The proposed model was trained and validated using the ISIC2016 [15], ISIC2017 [9], and HAM10000 [16] datasets. The HAM10000 dataset is particularly noteworthy, as over fifty percent of its skin lesions have been pathologically confirmed, establishing it as a gold standard in the field. The dataset was split into a 70:30 ratio for training and validation, as shown in Table 1. The dataset's significant intraclass similarities pose a challenge for categorization.

Table 1. A breakdown of the dataset's distribution

Dataset	Total	Training	Test
ISIC2016 [15]	1,279	900	379
ISIC2017 [9]	2,750	2,000	600
HAM10000 [16]	10,015	7,011	3,004

2.8. Implementation details

The proposed model was implemented using the open-source Keras library, and all experiments were conducted on Google Colab with K80, P100, and T4 GPUs. A training set and a test set were randomly selected from a dataset containing instances of various skin diseases for this study. During the training process for the CNN, the image sizes were randomly reduced to 256×224 and then horizontally flipped. To achieve optimal performance, the following strategies were employed: the finalization vector control fully connected layer and the final output layer of every pre-trained network were replaced with an initialization vector control fully connected layer. Extra fully connected layers were added and filled with weights initialized using Gaussian random weights with a standard deviation of 0.001 and a mean of zero.

3. RESULTS

This section provides a thorough analysis in addition to explaining the research findings. Figures, graphs, tables, and other reader-friendly formats can be used to present results [17], [18]. There are multiple subsections that can cover the same ground.

3.1. Results of individual models using three datasets

Models trained with DenseNet-121 and ViTb16 have an average F1-score of 80 and 79, recall of 79 and 75, precision of 81 and 75, and accuracy of 83% and 79%, respectively, on ISIC2016 dataset. DenseNet-121 and ViTb16 models, respectively, achieve an average of 84% and 80% accuracy, 78 and 77 precision, 83 and 81 recall, and 82 and 79 F1-score on ISIC2017. Models trained with DenseNet-121 and ViTb16 have an average F1-score of 81 and 80, recall of 81 and 78, precision of 83 and 79, and accuracy of 85% and 81%, respectively, on the HAM10000 dataset. Therefore, the best accuracy is 85% for the DenseNet-121 model on the HAM10000 dataset. The results are broken down and summarized in Table 2.

Table 2. Summarize classification results of single network on three datasets

	Model	Precision	Recall	F1-score	Testing accuracy (%)
Single network - ISIC -2016 dataset	DenseNet-121	81	79	80	83
	ViTb16	75	75	79	79
single network - ISIC -2017 dataset	DenseNet-121	78	83	82	84
	ViTb16	77	81	79	80
Single network - HAM10000 dataset	DenseNet-121	83	81	81	85
	ViTb16	79	78	80	81

3.2. Results of late fused models using three datasets

Experiments on late fused models with DenseNet-121+ViTb16 have an average F1-score of 78, recall of 79, precision of 80, and accuracy of 81% on ISIC2016 dataset. Experiments on late fused models with DenseNet-121+ViTb16 have an average F1-score of 82, recall of 80, precision of 81, and accuracy of 82% on ISIC2017 dataset. Experiments on late fused models with DenseNet-121+ViTb16 have an average F1-score of 83, recall of 80, precision of 78, and accuracy of 84% on the HAM10000 dataset. The best accuracy is 84% of the fused model on the HAM10000 dataset. Three datasets were used to display the classification results of the late fusion network, as shown in Table 3.

Table 3. Summarize classification results of late fusion network on three datasets

	Model	Precision	Recall	F1-score	Testing accuracy (%)
Late fusion - ISIC -2016 dataset	DenseNet-121+ViTb16	80	79	78	81
Late fusion - ISIC -2017 dataset	DenseNet-121+ViTb16	81	80	82	82
Late fusion - HAM10000 dataset	DenseNet-121+ViTb16	78	80	83	84

3.3. Results of proposed approach using three datasets

The proposed model, respectively, achieves an average accuracy of 82%, 79 precision, 80 recall, and an 80 F1-score on ISIC2016 dataset. The proposed model, respectively, achieves an average accuracy of 84%, 82 precision, 80 recall, and 78 F1-score on ISIC2017 dataset. The proposed model, respectively, achieves an average accuracy of 86%, 77 precision, 82 recall, and an 84 F1-score on the HAM10000 dataset shown in Table 4.

Table 4. Summarizes classification results of proposed method on three datasets

	Model	Precision	Recall	F1-score	Testing accuracy (%)
ISIC -2016 dataset	Proposed method	79	80	80	82
ISIC -2017 dataset	Proposed method	82	80	78	84
HAM10000 dataset	Proposed method	77	82	84	86

3.4. Comparative analysis of DenseNet-121 and ViTb16

Analyzing the parallels and dissimilarities between the DenseNet-121 and ViTb16 models presented in Figure 2. On the HAM10000 dataset, DenseNet-121 achieves an accuracy of 85% in classification, surpassing the second-best method by more than 1%. The second-best accuracy is 84% with DenseNet-121 on the ISIC2017.

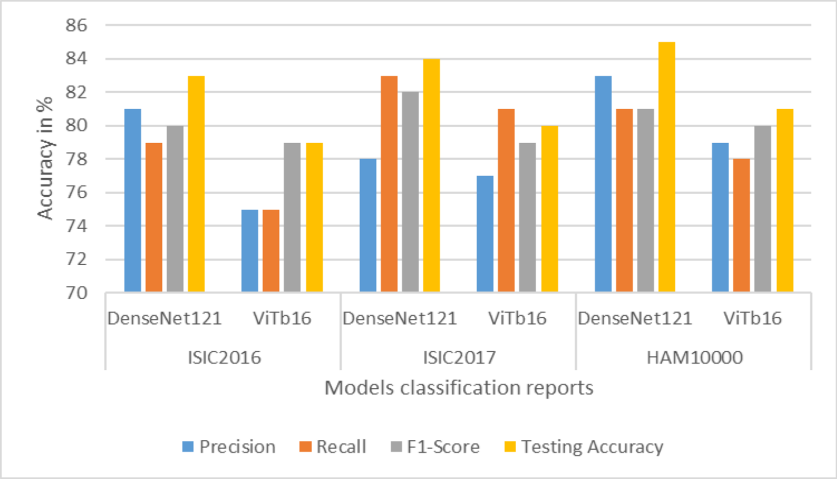

Figure 2. Graphical representation of the classification report on various parameters: accuracy, precision, recall, and F1-score on single models

3.5. Comparative analysis of late-fused models

In this analysis, we compared and contrasted the late fusion shown in Figure 3. The fused model outperforms the gold standard by more than 2% on the HAM10000 dataset, with a classification accuracy of 84%. Hence, the most accurate fused model using HAM100000 is the one that combines DenseNet-121 and ViTb16.

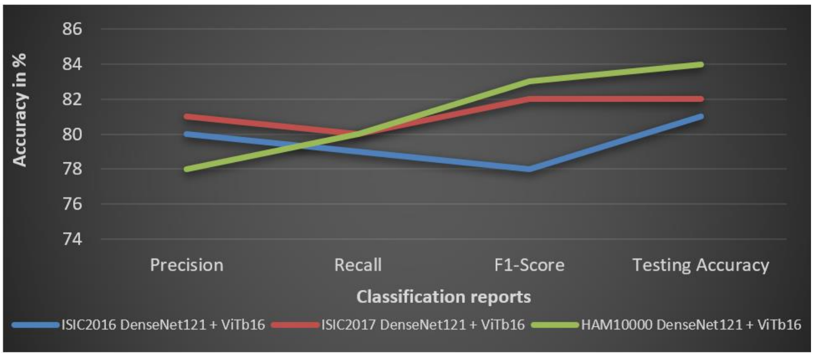

Figure 3. Line graph representation of the classification report on various parameters (F1-score, recall, precision, and testing accuracy) on fused models

3.6. Comparative analysis of the proposed model

Here, we looked at the proposed approach in Figure 4 and compared it to others. A classification accuracy of 86% shows that the proposed approach surpasses the gold standard by over 2% on the HAM10000 dataset. These results show that the proposed technique successfully and efficiently categorizes skin diseases.

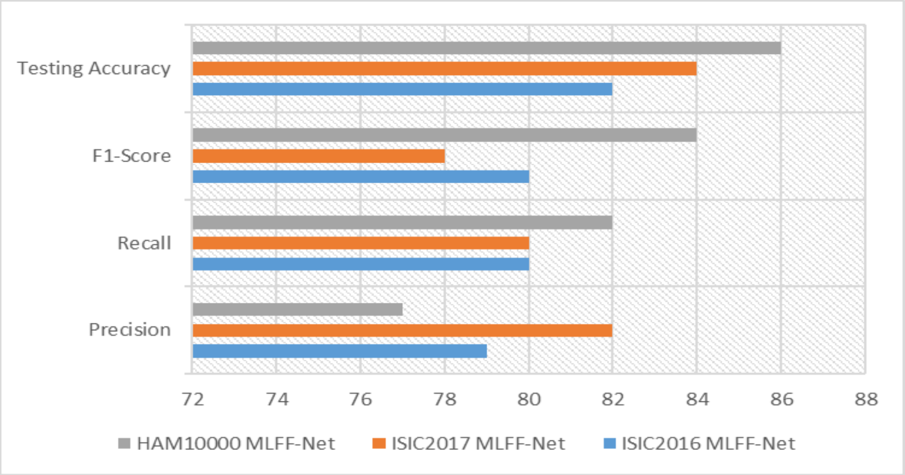

Figure 4. Graphical form of classification report on various parameters (F1-score, recall, precision, and accuracy) of the proposed approach

3.7. Comparative analysis of the proposed model with the state of the arts

Analysis of all the test data showed that our proposed model correctly classified 86% of the data. Using the necessary data extraction techniques, we are able to determine that our DenseNet-121 and ViTb16 do the best job of identifying images in the ISIC 2016, ISIC2017, and HAM10000 datasets. According to Satheesh *et al.* [19], He *et al.* [20], and Hsu and Tseng [21], our proposed method achieves an impressive accuracy of 86%, which is 6.1% higher than the next highest accuracy. Several alternatives to traditional deep learning models and segmentation techniques have been proposed [22]–[25]. Table 5 displays a comparison with the current level of technology.

Table 5. Comparative analysis of the proposed approach

Method	Accuracy (%)
Satheesha *et al.* [19]	77.04
He *et al.* [20]	76.8
Hsu and Tseng [21]	79.90
MLFF-Net (ours)	86.00

4. DISCUSSION

The main contribution is a two-stream network that classifies skin diseases using multimodal late feature fusion. The network performs well on three distorted datasets. DenseNet-121 improves the deep neural network's classification rate and lessens gradient dispersion brought on by the overly deep network model due to its lower parameter count and better feature propagation through densely connected dense blocks. An alternative ViTb16 network model uses residual structures. Before and after layers merge, leftover structures are retained. Identity mapping and residual mapping can transfer layer properties. By training and merging more models using various data preprocessing methods, we hope to improve medical diagnosis classification performance. Our strategy requires us to create the experiment using a lightweight network, but the model isn't. Only three ISIC datasets have been examined, but we will select more and use real-world clinic data. There were still several issues and limitations with our proposed model, even though it demonstrated descent classification performance on datasets with imbalanced or few samples. For example, our proposed model was found to consume high computing resources and have a comparatively slow training pace. More benchmark datasets are needed for training to enhance our model's performance, as it only recognizes a subset of skin diseases.

5. CONCLUSION

The mortality rate is significant, and there is considerable similarity and variation between classes of malignant skin lesions. Therefore, a reliable categorization system would greatly benefit doctors in the early diagnosis of skin diseases. To enhance skin disease categorization, we propose a multimodal feature fusion model in this study. Our approach combines the strengths of transformers and CNNs. In our two-stream network, which integrates ViTb16 with an enhanced DenseNet-121 network, we leverage the advantages of both networks. To be more precise, we improve the model's performance by incorporating more parameters by merging the original DenseNet-121 model's residual structure. Subsequently, we capture multi-scale pathological information using feature fusion blocks. On the HAM10000 dataset, our proposed technique attains an 86% classification accuracy. Our model outperforms state-of-the-art models with an astonishing 84% classification accuracy when DenseNet-121 and ViTb16 are combined. Future advancements in the field may be facilitated by EfficientNet and other alternatives to conventional deep learning models and segmentation approaches. Additionally, we plan to explore the potential of training on a large labeled dataset in the near future. To make the proposed model more versatile, we plan to make minor adjustments to it in future work. In addition, we will extensively evaluate the proposed model with extra benchmark datasets, including additional skin diseases.

ACKNOWLEDGEMENTS

We thank Graphic Era (Deemed to be University) Dehradun for the use of their research facilities.

REFERENCES

[1] D. V. Kumar and K. V. Dixit, "Gannet devil optimization-based deep learning for skin lesion segmentation and identification," *Biomedical Signal Processing and Control*, vol. 88, p. 105618, Feb. 2024, doi: 10.1016/j.bspc.2023.105618.

[2] M. Strzelecki, M. Kociołek, M. Strąkowska, M. Kozłowski, A. Grzybowski, and P. M. Szczypiński, "Artificial intelligence in the detection of skin cancer: state of the art," *Clinics in Dermatology*, Jan. 2024, doi: 10.1016/j.clindermatol.2023.12.022.

[3] S. Wang and J. Liu, "Deep learning-assisted automatic classification of skin images," *Chinese Journal of Dermatology*, vol. 53, no. 12, pp. 1037–1040, 2020, doi: 10.35541/cjd.20190660.

[4] C. Y. Zhu *et al.*, "A deep learning based framework for diagnosing multiple skin diseases in a clinical environment," *Frontiers in Medicine*, vol. 8, Apr. 2021, doi: 10.3389/fmed.2021.626369.

[5] S. Khouloud, M. Ahlem, T. Fadel, and S. Amel, "W-net and inception residual network for skin lesion segmentation and classification," *Applied Intelligence*, vol. 52, no. 4, pp. 3976–3994, Mar. 2022, doi: 10.1007/s10489-021-02652-4.

[6] S. Benyahia, B. Meftah, and O. Lézoray, "Multi-features extraction based on deep learning for skin lesion classification," *Tissue and Cell*, vol. 74, p. 101701, Feb. 2022, doi: 10.1016/j.tice.2021.101701.

[7] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017, doi: 10.1109/CVPR.2017.243.

[8] D. Popescu, M. El-Khatib, and L. Ichim, "Skin lesion classification using collective intelligence of multiple neural networks," *Sensors*, vol. 22, no. 12, p. 4399, Jun. 2022, doi: 10.3390/s22124399.

[9] 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," *Informatics in Medicine Unlocked*, vol. 28, p. 100819, 2022, doi: 10.1016/j.imu.2021.100819.

[10] Z.-H. Zhou, "Ensemble learning," in *Machine Learning*, Singapore: Springer Singapore, 2021, pp. 181–210.

[11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[12] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: efficient channel attention for deep convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 11531–11539, doi: 10.1109/CVPR42600.2020.01155.

[13] F. Radenovic, G. Tolias, and O. Chum, “Fine-tuning CNN image retrieval with no human annotation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, Jul. 2019, doi: 10.1109/TPAMI.2018.2846566.

[14] P. Dollár, Z. Tu, P. Perona, and S. Belongie, “Integral channel features,” in *British Machine Vision Conference, BMVC 2009 - Proceedings*, 2009, pp. 91.1-91.11, doi: 10.5244/C.23.91.

[15] N. C. F. Codella *et al.*, “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 *Proceedings - International Symposium on Biomedical Imaging*, Apr. 2018, vol. 2018-April, pp. 168–172, doi: 10.1109/ISBI.2018.8363547.

[16] P. Tschandl, C. Rosendahl, and H. Kittler, “Data descriptor: the HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Scientific Data*, vol. 5, no. 1, Aug. 2018, doi: 10.1038/sdata.2018.161.

[17] L. Yang, Y. Gu, G. Bian, and Y. Liu, “TMF-Net: a transformer-based multiscale fusion network for surgical instrument segmentation from endoscopic images,” *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–15, 2023, doi: 10.1109/TIM.2022.3225922.

[18] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, “DS-TransUNet: dual swin transformer U-Net for medical image segmentation,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, 2022, doi: 10.1109/TIM.2022.3178991.

[19] T. Y. Satheesha, D. Satyanarayana, M. N. G. Prasad, and K. D. Dhruve, “Melanoma is skin deep: a 3D reconstruction technique for computerized dermoscopic skin lesion classification,” *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 5, pp. 1–17, 2017, doi: 10.1109/JTEHM.2017.2648797.

[20] X. He, Y. Wang, S. Zhao, and X. Chen, “Co-attention fusion network for multimodal skin cancer diagnosis,” *Pattern Recognition*, vol. 133, p. 108990, Jan. 2023, doi: 10.1016/j.patcog.2022.108990.

[21] B. W. Y. Hsu and V. S. Tseng, “Hierarchy-aware contrastive learning with late fusion for skin lesion classification,” *Computer Methods and Programs in Biomedicine*, vol. 216, p. 106666, Apr. 2022, doi: 10.1016/j.cmpb.2022.106666.

[22] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, “FAT-Net: feature adaptive transformers for automated skin lesion segmentation,” *Medical Image Analysis*, vol. 76, p. 102327, Feb. 2022, doi: 10.1016/j.media.2021.102327.

[23] Y. Ma *et al.*, “IHA-Net: an automatic segmentation framework for computer-tomography of tiny intracerebral hemorrhage based on improved attention U-Net,” *Biomedical Signal Processing and Control*, vol. 80, Feb. 2023, doi: 10.1016/j.bspc.2022.104320.

[24] J. Wang *et al.*, “XBound-former: toward cross-scale boundary modeling in transformers,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 6, pp. 1735–1745, Jun. 2023, doi: 10.1109/TMI.2023.3236037.

[25] Y. Meng *et al.*, “Graph-based region and boundary aggregation for biomedical image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 3, pp. 690–701, Mar. 2022, doi: 10.1109/TMI.2021.3123567.

BIOGRAPHIES OF AUTHORS

Mr. Ajay Krishan Gairola is a researcher at the Graphic Era Hill University, Dehradun, India. His qualifications are as mentioned Ph.D., MCA, and BCA. His research interests include medical image processing, computer vision, machine learning and deep learning, image retrieval, and medical disease diagnosis. He can be contacted at email: ajaykgairola.cc@geu.ac.in.

Dr. Vidit Kumar is an assistant professor at the Graphic Era Deemed to be University. He has done B.Tech. in Computer Science and Engineering from Uttarakhand Technical University, M.Tech. and Ph.D. in Computer Science and Engineering from Graphic Era Deemed to be University, Dehradun, India. His research of interest is in machine learning, deep learning, video analytics, medical disease diagnosis, and computer vision. He published more than 20 research papers. He can be contacted at email: viditkumar.cse@geu.ac.in.

Dr Ashok Kumar Sahoo is currently working as Professor in Computer Science and Engineering at Graphic Era Hill University, Dehradun. Prior to this, he has worked in Sharda University, Greater Noida, HIMCS, Mathura and Meerut Institute of Engineering and Technology, Meerut. He has more than 26 years of teaching and research experiences. He has obtained his Master of Technology in Computer Science and Engineering from Uttar Pradesh Technical University with II rank in the university. He has obtained his Ph.D. in Computer Science and Engineering from Sharda University. He has organized more than ten international/national conferences. He has participated in more than 50 International and national conferences/seminars. He has published more than 50 research papers in international reputed journals and presented more than 25 research papers in national and international conferences. He also wrote four book chapters published by reputed international publishers. Four Indian patents are also published to his credit. He can be contacted at email: ashok.sahoo@gehu.ac.in.