Classification of Skin Cancer Using CNN with Transformer Layer

K. P. R. Surya, N. Sunil Kumar, N. Sai Rama Krishna, K. Avinash, N. Prakash, and Abdul Rahaman Shaik(✉)

Department of Electronics and Communication Engineering, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India
Abdulrahman.s@vishnu.edu.in

Abstract. Skin cancer, a major global health concern, demands early detection for optimal patient outcomes. Traditional methods relying on subjective dermatological examinations often prove time-consuming and susceptible to human error. Fortunately, recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have showcased positive outcomes in improving skin cancer classification accuracy. Building upon this success, this research delves into the potential of integrating transformer layers within a Convolutional Neural Network (CNN) framework for further advancements. By explicitly capturing intricate spatial dependencies and enhancing feature extraction capabilities, our proposed hybrid model seeks to surpass the limitations of individual architectures and offer a robust and comprehensive tool for skin disease classification. This study aims to demonstrate the model's efficacy on a large-scale skin cancer dataset, evaluating its performance against established approaches and offering valuable perspectives on the capacity of hybrid architectures for improved skin cancer diagnosis.

Keywords: Skin Cancer · HAM10000 Dataset · Transformer · CNN

1 Introduction

Skin cancer ranks among the top frequently diagnosed types of cancer globally, accounting for a substantial portion of all cancer cases worldwide [1]. It tends to be more prevalent among individuals with fair skin tones. Various types of skin cancer exist, with melanoma being widely recognized. Notably, within the United States, skin cancer incidence rates are notable, particularly among the elderly population. Between 2009 and 2016, there was a noteworthy rise in incidence rates among individuals aged 70 and above, highlighting concerns amidst an aging demographic [2].

The issue of categorizing skin lesions has garnered attention from the machine learning community, with automated lesion classification aiming to assist healthcare professionals in their daily practices and facilitate swift and cost-effective access to critical diagnoses, even beyond clinical settings, through mobile app installations [3].

P. Pareek et al. (Eds.): IC4S 2024, LNICST 597, pp. 151–162, 2025.
https://doi.org/10.1007/978-3-031-77075-3_13

Pre-2016, research predominantly adhered to the conventional process of machine learning, encompassing preprocessing, segmentation, feature extraction, and classification [4–6]. Nonetheless, this approach demands significant domain-specific expertise, especially for feature extraction, which is inherently time-intensive. After 2016, a notable shift emerged in the realm of lesion classification techniques, evident in the methodologies presented at the 2016 International Symposium on Biomedical Imaging (ISBI) [7]. Among the 25 engaged sides, there was a conspicuous absence of traditional standard machine learning approaches; rather, all teams opted for the utilization of deep learning techniques, specifically CNNs [8]. CNNs are a class of neural networks distinguished by their specialized architecture, renowned for their remarkable efficacy in tasks like image recognition and classification [9]. While CNN has demonstrated proficiency in image classification tasks, it does harbor limitations. One such drawback is its propensity to diminish the size of the feature space, leading to the deprivation of significant findings. To address this constraint, transformers, extensively employed in various natural language processing (NLP) endeavors, have garnered substantial attention in the realm of computer vision. Leveraging the self-attention mechanism, transformers-based models offer a wider perspective. This enables them to glean long-distance spatial relationships and concentrate on pertinent regions within the image [10].

This paper introduces a classification of skin cancer using CNN with transformer layer model's potential efficacy on a large-scale dataset, comparing its performance with established approaches. This model achieved the accuracy of 96%.

2 Literature Survey

In recent years, there has been significant research interest in developing accurate and efficient methods for investigating medical images, especially in the context of lesion identification and disease diagnosis. The following studies contribute valuable insights into various techniques and approaches used in medical imaging and machine learning for classification tasks:

Prabhakara Rao, et al., conducted a study comparing different classifying models for recognizing breast anomalies in ultrasound images [11]. The research focused on evaluating machine learning and deep learning techniques applied to medical imaging for accurate lesion detection and classification. This study is relevant to our research as it provides insights into methodologies and performance metrics essential for skin lesion classification tasks.

Shaik Abdul Rahaman, et al., presented a study on identifying glaucoma using segmentation and fusion techniques in medical image analysis [12]. While their focus was on detecting eye diseases, the methods and techniques discussed in this paper offer potential insights into similar approaches for analyzing skin lesion images. Understanding segmentation and fusion techniques is crucial for extracting meaningful features from medical images for classification purposes.

Although primarily focused on crack length estimation using image processing methods, Budumuru Prudhvi Raj, et al., study provides valuable knowledge about image analysis techniques [13]. These techniques, such as preprocessing, feature extraction, and segmentation, are applicable and can be adapted for analyzing skin lesion images, contributing to more accurate and effective classification models.

Gayathri, et al., introduced a breast cancer prediction model based on an optimized Convolutional Neural Network (CNN) [14]. This study demonstrates the application of deep learning techniques in medical image analysis and offers strategies for optimizing CNN architectures. Such strategies are relevant for developing robust and accurate skin cancer classification models.

Anand, et al., discussed optimizing machine learning and deep learning algorithms for cancer diagnosis [15]. Their research gives important insights into improving classification model accuracy and efficiency, which is essential for developing effective skin lesion classification systems.

Arshed MA, et al., presented a study on skin cancer classification using Vision Transformer Networks and CNN based pre-trained models [16]. This research explores advanced techniques for skin cancer classification, leveraging ViT architectures and pre-trained CNN models to achieve improved accuracy and performance in multi-class classification tasks.

3 Methodology

3.1 Dataset

In this paper, we expand upon the widely used HAM10000 dataset [17] by incorporating a diverse range of skin conditions to create a more comprehensive dataset for our research on skin cancer classification. While the original dataset encompasses 10,015 thermoscopic images across seven categories, we introduce healthy skin images and two additional disease types: squamous cell carcinoma and pigmented benign keratosis. This enriched dataset, totaling (12786), provides a more realistic representation of the complexity of skin conditions encountered in clinical practice.

We'll resize each image to a consistent size of 32 × 32 pixels, convert it into a NumPy array, and reshape it to represent the image's dimensions and channels (RGB). Subsequently, we'll save these NumPy arrays. These saved arrays are crucial for training a deep learning model, facilitating efficient handling and processing of image data. Sample images are shown in Fig. 1.

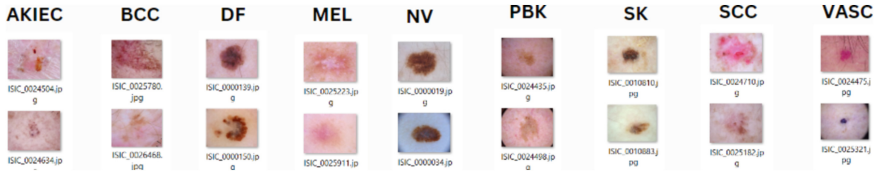

Fig. 1. Sample Images from the dataset used.

3.2 CNN Model

A CNN is a type of neural network designed for processing and analyzing visual data, making it well-suited for tasks like image classification and object detection. Key components include convolutional layers that capture localized patterns, pooling layers that

diminish spatial dimensions, fully connected layers for global patterns, and an output layer for final predictions. The architecture of a CNN can vary based on the specific task and requirements. The selection of architecture is contingent upon the complexity of the problem and the available computational resources.

The CNN [18] model employed in this study comprises several key architectural components tailored for image classification tasks. The model begins with an input layer, Conv2D_2_Input, designed to accept images of dimensions 32 × 32 pixels with three color channels (RGB). Following the input layer, a Conv2D layer named Conv2D_2 is employed, applying 32 filters of size 3 × 3 to the input images, resulting in feature maps of size 30 × 30. Subsequently, a MaxPooling2D layer [19], MaxPooling2D_2, is utilized to decrease the size of the feature maps by half, enhancing computational efficiency.

Another Conv2D layer, Conv2D_3, continues the feature extraction process by applying additional filters, yielding feature maps of size 13 × 13. Max pooling is again performed via MaxPooling2D_3 to further down sample the feature maps. The Flatten layer, flatten_1, serves to flatten the 3D feature maps into a 1D vector, facilitating seamless integration with fully connected layers.

Following this is Dense_2 layer followed by ReLU function [20] which is employed for feature transformation and extraction. Finally, the output layer, Dense_3, employs the soft-max to generate class likelihoods for the 10 target classes in the classification task.

This comprehensive framework, which includes convolutional layers, max-pooling layers, and fully connected layers, enables effective feature extraction and classification, thus rendering it suitable for image classification applications. The CNN model is shown in Fig. 2.

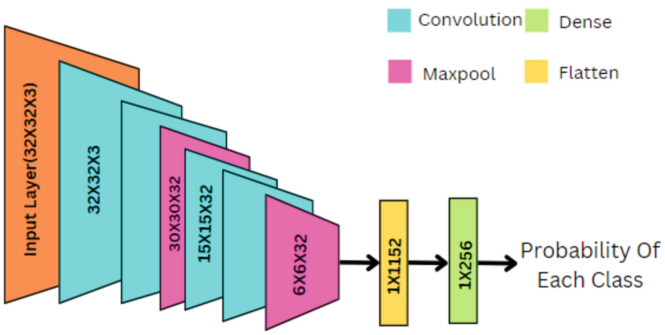

Fig. 2. CNN block diagram without Transformer Layer

3.3 CNN with a Transformer Layer

The CNN model, augmented with a transformer layer, represents a fusion of two powerful deep learning architectures, poised to further elevate effectiveness in tasks related to image classification. The introduction of a transformer layer, strategically positioned

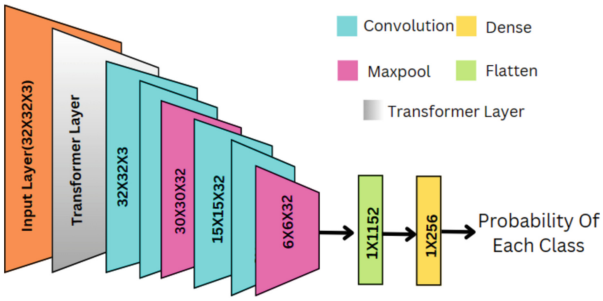

Fig. 3. CNN block diagram with a Transformer Layer

after the input layer, injects a dose of attention mechanism into the traditional CNN framework. Figure 3 displays the CNN architecture with a Transformer Layer.

This transformer layer enables the model to grab intricate spatial parameters across various regions of the input images, enhancing its ability to discern complex patterns and features. By leveraging self-attention mechanisms, the transformer layer allows the network to focus on relevant image regions while suppressing noise and irrelevant information, thereby improving feature extraction efficiency. Consequently, this hybrid CNN with a transformer layer amalgamates the strengths of both architectures, harnessing the hierarchical feature learning capabilities of CNNs with the self-attention mechanisms of transformers.

The synergistic integration of these components not only enhances the model's interpretability but also elevates its performance in image classification tasks, empowering it to achieve state-of-the-art results across diverse datasets and benchmarks.

4 Performance Indicators

We use loss, accuracy, C-matrix (Confusion Matrix), precision, recall, and F1-Score as valuation metrics [21]. These metrics provide more in-depth awareness into the model's performance for each individual skin lesion category.

4.1 Loss and Accuracy

The chosen loss function, cross-entropy, continuously monitors the gap between predicted and actual values, updating weights based on any discrepancies. Formally, it can be expressed as:

$$L = \frac{1}{R}\sum_i L_i = -\frac{1}{R}\sum_i \sum_{c=1}^{M} q_{ic}\log(p_{ic}) \tag{1}$$

- R represents the total no of types,
- q_{ic}represents the binary ground truth label.
- $p_{ic}$is the predicted likelihood that category i corresponds to category s

$$Accuracy = \frac{Tpos + Tneg}{Tpos + Tneg + Fpos + Fneg} \tag{2}$$

- Tpos –True Positive
- Tneg – True Negative
- Fpos – False Positive
- Fneg – False Negative

4.2 Estimation of Various Indicators

Precision estimates the percentage of positively predicted cases that are truly positive.

$$Precision = \frac{Tpos}{Tpos + Fpos} \tag{3}$$

Recall measures the fraction of actual positive cases that are rightly detected.

$$Recall = \frac{Tpos}{Tpos + Fneg} \tag{4}$$

The F1-Score provides a balanced assessment by offering the harmonic mean of precision and recall.

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

4.3 C-Matrix

The C-matrix aids in visualizing the categorization results by displaying the frequency of distinct skin lesions categorized by the model within each classification. This matrix provides a concise and organized representation of the model's efficacy in categorizing various types of skin lesions.

5 Results

This portion examines the efficacy of both the CNN model and the CNN with transformer layer model for classifying skin lesions. The comparison between the models primarily focuses on accuracy. Additionally, we assess the models' performance by examining the Confusion Matrix, which offers a comprehensive insight into their classification accuracy and effectiveness.

5.1 Comparison of Confusion Matrices

A confusion matrix acts as a structured table that allows visualization of the performance of an algorithm by displaying the number of correct and incorrect classifications made by the model on a dataset.

Confusion matrices are particularly useful for evaluating the performance of classification models, especially in scenarios where there might be class imbalance or where different types of errors have varying levels of importance. They provide a more detailed

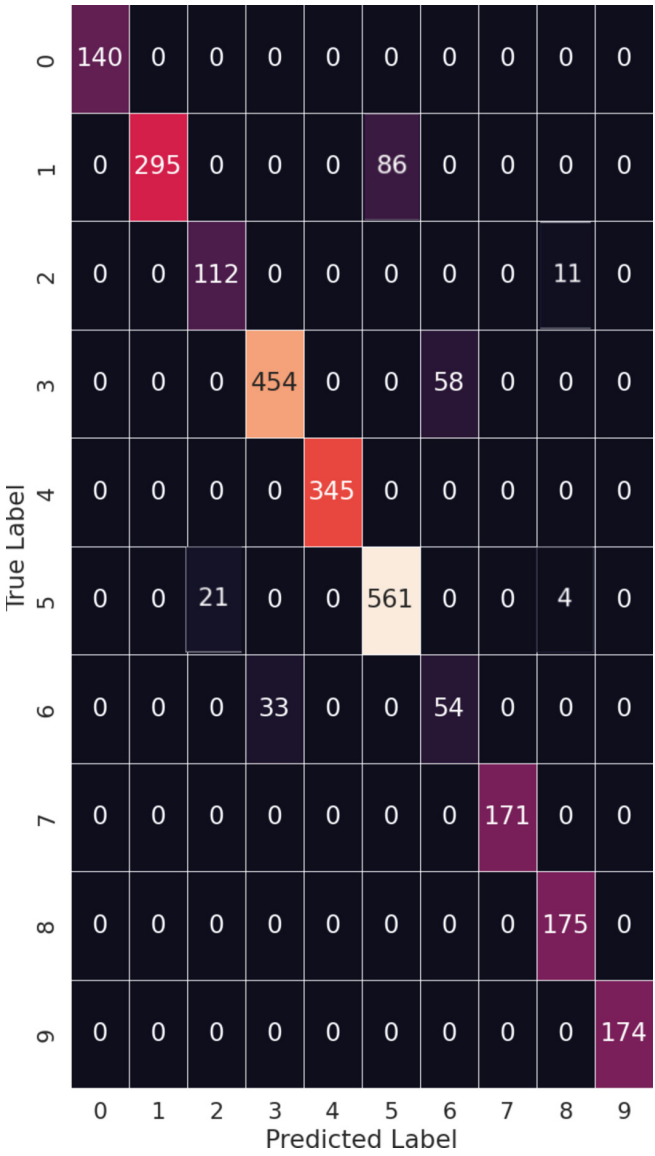

Fig. 4. Confusion Matrix for CNN Model without Transformer Layer

understanding of the model's behavior beyond simple accuracy metrics. Figures 4, 5 show the confusion matrices for the CNN model without and with the Transformer Layer respectively.

The results are compatible. Overall, the incorporation of a transformer layer into a CNN architecture enhances the model's capabilities in capturing contextual information, handling long-range dependencies, improving feature representations, adapting to

varied data distributions, and promoting regularization and generalization. These factors collectively contribute to the hybrid model achieving more correct classifications compared to the standalone CNN model.

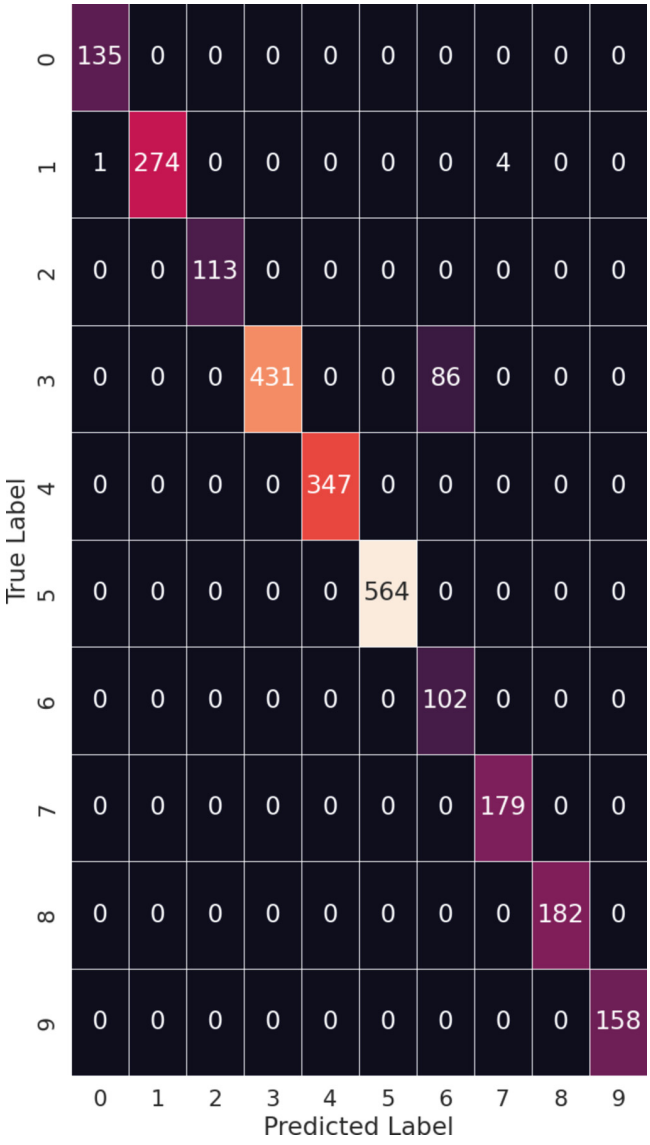

Fig. 5. Confusion Matrix for CNN Model with Transformer Layer

5.2 Comparison of Classification Reports

A classification provides a detailed breakdown of metrics. Precision measures the correctness of positive forecasts, whereas recall gauges the model's ability to include all positive instances and the F1 score amalgamates these two parameters. Support indicates the actual count of occurrences for each type, offering perceptions into class dispensations. While not explicitly part of the classification report, accuracy is also crucial, representing the effectiveness of the model's predictions. By examining these metrics collectively, a classification report enables analysts and data scientists to assess a model's robustness and areas for betterment, aiding in the iterative process of model refinement and optimization in supervised learning scenarios.

Table 1. Classification Report without Transformer Layer

Classification Report

0-	1.00	1.00	1.00
1-	1.00	0.77	0.99
2-	0.80	0.91	1.00
3-	0.93	0.88	0.91
4-	1.00	1.00	1.00
5-	0.86	0.95	1.00
6-	0.48	0.62	0.70
7-	1.00	1.00	0.99
8-	0.92	1.00	1.00
9-	1.00	1.00	1.00
Accuracy-	0.93	0.93	0.93
Macro avg-	0.91	0.93	0.92
Weighted avg-	0.93	0.91	0.92
	Precision	Recall	F1-Score

It is a standard instrument employed in machine learning to assess the effectiveness of a classification model. It furnishes a comprehensive synopsis of different evaluation metrics. These metrics offer insights into how well the model is performing for different classes and can help in understanding its strengths and weaknesses. Tables 1 and 2 show the classification reports for the CNN without and with Transformer Layer respectively. Analyzing precision, recall, and F1-score is crucial for evaluating the effectiveness of classification models, especially in scenarios with multiple classes. These metrics offer

deeper insights than overall accuracy, particularly in cases of imbalanced data or varying class importance. By examining these metrics for different classes, one can identify biases, prioritize improvements, and ensure the model performs well across all categories, leading to more informed decision-making and better model performance overall. A comprehensive classification report encompasses a range of metrics and visualizations that collectively offer a deep understanding of a classification model performance, facilitating decisions and optimizations in machine learning workflows.

Figure 6 shows the plot of Accuracy and loss versus number of Epochs. Furthermore, the plot in Fig. 6 supports the performance superiority of the CNN model with a transformer layer. A higher accuracy curve and a lower loss curve in Fig. 6 for the CNN combined with a transformer layer demonstrates improved accuracy and reduced loss throughout the training phase, indicating enhanced model performance. This suggests that the model converges faster and more effectively to optimal solutions, leading to improved classification performance.

When analyzing the classification reports presented in Tables 1 and 2, it becomes apparent that the CNN model augmented with a transformer layer achieves an impressive accuracy rate of 96%, which surpasses the accuracy achieved by the standalone CNN model.

Table 2. Classification Report with Transformer Layer

Classification Report

	Precision	Recall	F1-Score
0-	0.99	1.00	1.00
1-	1.00	0.98	0.99
2-	1.00	1.00	1.00
3-	1.00	0.83	0.91
4-	1.00	1.00	1.00
5-	1.00	1.00	1.00
6-	0.54	1.00	0.70
7-	0.98	1.00	0.99
8-	1.00	1.00	1.00
9-	1.00	1.00	1.00
Accuracy-	0.96	0.96	0.96
Macro avg-	0.95	0.98	0.96
Weighted avg-	0.98	0.96	0.97

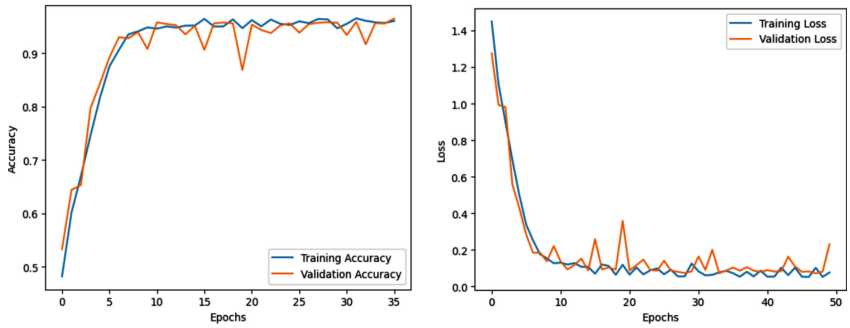

Fig. 6. Training Vs Testing Accuracy and Loss for CNN Model with Transformer Layer

6 Conclusion

In this paper, we delved into a comprehensive investigation aimed at improving the precision of skin cancer classification. We achieved this by leveraging a synergistic combination of Convolutional Neural Networks (CNNs) and Transformer layers. The outcomes, as evidenced by the metrics of accuracy and loss, underscored the effectiveness of our novel approach.

In conclusion, our project not only makes a significant contribution to the ongoing advancements in the field of skin cancer classification but also highlights the tremendous potential of integrating transformer layers with CNN architectures. This fusion of methodologies opens exciting avenues for future research and innovation in medical image analysis, paving the way for more accurate and reliable diagnostic systems in healthcare.

References

1. Silpa, S.R., Chidvila, V.: A review on skin cancer. Int. Res. J. Pharm. **4**(8), 83–88 (2013)
2. Le, H., Van., Le, C.H.I.H.U.U.H., Le, P.H.U.U.U., Truong, C.H.I.T.H.I.L.E.: Incidence and trends of skin cancer in the United States, 1999–2016. J. Clin. Oncol. **38**(15), 10077 (2020)
3. Foraker, R.E., et al.: EHR-based visualization tool: adoption rates, satisfaction, and patient outcomes. eGEMs **3**(2), 1159 (2015)
4. Lynn, N.C., Kyu, Z.M.: Segmentation and classification of skin cancer melanoma from skin lesion images. In:18th international conference on parallel and distributed computing, applications and technologies (PDCAT), pp. 117–122. IEEE, Taiwan (2017)
5. Alom, M.Z., Aspiras, T., Taha, T.M., Asari, V.K.: Skin cancer segmentation and classification with improved deep convolutional neural network. In: Medical Imaging 2020: Imaging informatics for healthcare. research, and applications, pp. 291–301. SPIE, Houston (2020)
6. Oliveira, R.B., Papa, J.P., Pereira, A.S., Tavares, J.M.R.S.: Computational methods for pigmented skin lesion classification in images: review and future trends. Neural Comput. Appl. **29**, 613–636 (2018)
7. Barata, C., Ruela, M., Francisco, M., Mendonça, T., Marques, J.S.: Two systems for the detection of melanomas in dermoscopy images using texture and color features. IEEE Syst. J. **8**(3), 965–979 (2013)

8. Marchetti, M.A., et al.: Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J. Am. Acad. Dermatol. **78**(2), 270–277 (2018)
9. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
10. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv 2010, 11929 (2020)
11. Prabhakara Rao, A., Prasanna Kumar, G., Ranjan, R.: Performance comparison of classification models for identification of breast lesions in ultrasound images. In: Pattern Recognition and Data Analysis with Applications, pp. 689–699. Springer Nature, Singapore (2022)
12. Shaik, A.R., Chandra, K.R., Raju, B.E., Budumuru, P.R.: Glaucoma identification based on segmentation and fusion techniques. In: International Conference on Advances in Computing, Communication, and Control (ICAC3), pp. 1–4. IEEE (2021)
13. Budumuru, P.R., Shaik, A.R., Satyanarayana, B.V.V., Manikanta, S.P., Sharmila, K.S., Prasad, D.D.: Normalized algorithm with image processing methods for estimation of crack length. In: 6th International Conference on Electronics, Communication and Aerospace Technology, pp. 1436–1439. IEEE (2022)
14. Gayathri, T., Madhavi, T., Kumari, K.R.: A prediction of breast cancer based on mayfly optimized CNN. In: International Conference on Computing, Communication and Power Technology (IC3P), pp. 176–180. IEEE (2022)
15. Anand, M., Saravanan, D., Pushpalatha, K.: Others: optimization of machine learning and deep learning algorithms for diagnosis of cancer. ECS Trans. **107**(1), 9389 (2022)
16. Arshed, M.A., Mumtaz, S., Ibrahim, M., Ahmed, S., Tahir, M., Shafi, M.: Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pre-trained models. Information **14**(7), 415 (2023)
17. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**(1), 1–9 (2018)
18. Malo, D.C., Rahman, Md.M., Mahbub, J., Khan, M.M.: Skin cancer detection using convolutional neural network. In: 12th Annual Computing and Communication Workshop and Conference (CCWC), pp. 169–176. IEEE (2022)
19. Zafar, A., et al.: A comparison of pooling methods for convolutional neural networks. Appl. Sci. **12**(17), 8643 (2022)
20. Ide, H., Kurita, T.: Improvement of learning for CNN with ReLU activation by sparse regularization. In: International joint conference on neural networks (IJCNN), pp. 2684–2691. IEEE (2017)
21. Qian, Y., Zeng, G., Pan, Y., Liu, Y., Zhang, L., Li, K.: A prediction model for high risk of positive RT-PCR test results in COVID-19 patients discharged from Wuhan Leishenshan hospital, China. Front. Public Health **9**, 778539 (2021)