

SKIN CANCER DETECTION: LEVERAGING CONVOLUTIONAL NEURAL NETWORKS FOR BINARY CLASSIFICATION

Amritanjali Swaroop
A. Charan Kumari
K. Srinivas¹

Received 11.02.2024.
Revised : 21.03.2024.
Accepted 17.05.2024.

Keywords:

*Skin Cancer classification,
Convolutional Neural Networks,
Dermoscopic images, Early detection,
Melanoma Classification.*

ABSTRACT

Globally Skin cancer accounts for millions of new cases annually posing a significant threat to public health, emphasizing the importance of timely detection to enhance patient outcomes and decrease mortality rates. In this research article an AI convolutional neural network (CNN) model for the detection and classification of skin lesions into benign and malignant type is presented. A comprehensive dataset comprising 33,126 dermoscopic images from the ISIC 2020 challenge, a tailored CNN architecture was developed and validated for accurate lesion classification. The good performance of the developed CNN model on various benchmark metrics highlight its capacity to precisely distinguish between benign and malignant skin lesions. The outcomes of this investigation indicate that the developed CNN model has the potential to significantly enhance diagnostic protocols by minimizing manual errors and streamlining patient care pathways through automation. Through the amalgamation of sophisticated machine learning methodologies with clinical protocols, the model not only supports immediate clinical implementations but also lays the groundwork for future advancements by incorporating larger datasets and real-world clinical information. This study provides a scalable CNN model for building an automated tool that can integrate into healthcare systems to facilitate early detection and treatment of skin cancer, resulting in cost reductions and improved health outcomes. This endeavor makes a remarkable impact in the realm of dermatology by emphasizing the utilization of artificial intelligence to transform the diagnosis of skin cancer and enhance patient care.



© 2024 Journal of Trends and Challenges in Artificial Intelligence

1. INTRODUCTION

Skin cancer is a serious health concern, with millions of cases worldwide each year. It is caused by abnormal cell proliferation in the skin tissue as a consequence of extended sun exposure. Squamous cell carcinoma, Melanoma and basal cell carcinoma are the three types of skin cancer that are most common. While melanoma in particular can be aggressive and potentially fatal if not found and treated early, most skin tumors are generally

benign and easily treatable. Improving patient outcomes and lowering the mortality rate from skin cancer require early detection and action.

Late diagnosis is still a major problem that raises morbidity and death rates, even with advances in medical technology and awareness campaigns (Malvia et al. 2017). This research is to improve patient outcomes and the efficacy of skin cancer screening programmes by creating a reliable and efficient approach for binary classification of benign and malignant skin lesions.

¹ Corresponding author: K. Srinivas
Email: ksri12@gmail.com

There are still gaps and limitations in the present approaches for skin cancer identification, particularly in the binary classification of lesions, despite major advancements in diagnostic imaging technology and clinical protocols. Traditional diagnostic methods, such as ocular inspection and histological investigation, can be time-consuming, arbitrary, and prone to mistakes. Classifying skin lesions as benign or malignant requires the use of objective, standardized criteria.

The contributions of the paper include:

1. This paper introduces a CNN model which is specially designed to handle the dermoscopic image data, leveraging over 7.6 million trainable parameters to learn and differentiate between benign and malignant lesions effectively.
2. By employing the ISIC 2020 dataset, which consists of 33,126 images, the research leverages one of the most extensive and diverse datasets available. This dataset ensures good robustness and generalizability of the CNN model.
3. The good performance of this model in classification accuracy (93.09% on the test dataset) and other performance metrics indicate the model efficacy in not only classifying skin lesions accurately but also in reducing the number of false positives and negatives—critical factors in medical diagnostics.
4. The research outlines significant potential benefits for clinical dermatology, including reducing manual diagnostic errors, streamlining patient care pathways, and facilitating early and accurate skin cancer detection.
5. The automated classification system can help in early intervention, thereby improving patient outcomes and reducing healthcare costs.
6. The study lays a foundational framework for future advancements by proposing integration with larger datasets and real-world clinical data.
7. This approach opens avenues for continuous improvement of the CNN model and adaptation to evolving medical imaging technologies and needs.
8. The research meticulously details the selection and optimization of hyperparameters such as the Adam optimizer, batch size, activation functions, and a structured training regime over 100 epochs. These elements are crucial for the training stability and performance enhancement of the CNN model.
9. The methodical division of the dataset into training, validation, and testing sets, and the subsequent detailed analysis of model performance on these sets, provide a transparent and replicable model evaluation strategy.
10. The inclusion of confusion matrices and performance metrics across different datasets highlights the model's consistency and reliability. The study indicates the ability of the model to generalize across different data subsets (training, validation, and testing), indicating that the model can be effectively used in real-world settings without overfitting to the training data.

These contributions underscore the importance of integrating advanced machine learning techniques with clinical applications to enhance diagnostic protocols, reduce healthcare burdens, and improve patient outcomes in dermatology.

The organization of the remaining sections of the paper is as follows: Section 2 presents a Literature review, the methodology is discussed in Section 3, Section 4 presents Results and Discussion. The last Section 5 presents Conclusion and Future Work.

2. LITERATURE REVIEW

This section presents a brief overview of the research in this field. Convolutional neural networks were used by (Subramanian et al., 2021) to classify skin cancer types with the aim of obtaining good accuracy, a low false negative rate, and high precision. They employed HAM10000 dataset in their research. Two methods for the diagnosis of skin cancers were proposed by (Tanna & Sharma 2021), specifically using imaging data from melanoma malignant cells. In the first method, three-layer convolutional neural networks were used, while in the second, a basic Support Vector Machine (SVM) model with the default Radial Basis Function kernel was employed. Convolution Neural Networks and the SVM classifier produced reasonably accurate classification results.

A CNN model was proposed by (Shetty et al., 2022) and trained on an enlarged dataset of HAM10000. (Junayed et al. 2021) presented a unique method for the detection and classification of skin cancer using a deep Convolutional Neural Network (CNN). On test data, the suggested CNN model outperformed both GoogleNet and MobileNet pre-trained models with an impressive accuracy. Moreover, the suggested model proved to be better than other modern models while retaining computing efficiency.

Victor and Ghalib (2017), focused on the automatic diagnosis and classification of skin cancer utilizing segmentation, pre-processing, feature extraction, and classifiers with different accuracies, such as SVM, KNN, Decision Tree, and Boosted Tree. In their work, the median filter was used to identify and categorize benign and normal images, and classification methods were compared.

Ananda et al., (2022) focused on skin cancer classification using CNN models. They reported that InceptionV3 achieved the highest accuracy, outperforming other models like MobileNetV2, VGG19, and EfficientNetB7. (Sigurdur et al. 2004) detected skin cancer through Raman spectra classification using a neural network, with high accuracy rates for malignant melanoma and basal cell carcinoma, showcasing the method effectiveness.

Chaugule et al., (2017), achieved skin melanoma cancer detection and classification through dermoscopy image database, preprocessing, segmentation, feature extraction, and MSVM classification for early and

accurate diagnosis. (Torti et al., 2020) presented a parallel pipeline utilizing hyperspectral imaging for skin cancer detection, reducing computational times and improving accuracy in classifying pigmented skin lesions.

Paliwal (2016) achieved Skin cancer detection and classification through image segmentation, feature extraction and pattern analysis using hybrid image processing techniques for early diagnosis and treatment. (Muniteja, Bee & Suresh 2022) in enhanced detection and classification of skin cancer using Convolutional Neural Network, surpassing Coactive Neuro Fuzzy Inference System in melanoma image analysis.

Blackledge and Dubovitskiy (2008). presented a system for skin cancer screening using object detection and classification based on fractal parameters, texture analysis, and fuzzy logic decision- making, aiding in distinguishing normal from abnormal cases.

3. METHODOLOGY

The methodology encompasses several key aspects, including the dataset used, the architecture of the CNN model, the hyperparameters chosen for training and the assessment metrics utilized to evaluate the performance of the model. Each of these components is essential for comprehensive understanding of the approach taken in this study.

3.1 Dataset

The 2020 SIIM-ISIC Melanoma Classification Challenge dataset (Sutradhar et al. 2022) was used in this study. There are two formats for the dataset. The Digital Imaging and Communication in Medicine (DICOM) standard is the first format. The metadata for the second format JPEG is contained in a comma-separated values (CSV) file, which is utilized in this study.

Splitting the available data into distinct sets for training, testing, and validation is essential to accurately assess and maximize the performance of the model. The machine learning model is trained using the training dataset. Throughout the training phase, the model learns to recognize the patterns and relationships in the data by adjusting its parameters (weights and biases) in response to the input characteristics and matching target labels. Batches of data from the training set are processed iteratively by the model, which updates its parameters in order to minimize a specified loss function that measures the discrepancy between the actual labels and the anticipated outputs. By use of this iterative optimization procedure (often called backpropagation), the model progressively enhances its performance on the training data with the goal of achieving good generalization to unknown data.

The validation set consists of a small subset of all the accessible data. The validation dataset is used to assess the performance of the model throughout training and make necessary hyperparameter adjustments. While

training the model, the performance on the validation set is monitored after each epoch. This allows for early stopping or adjusting hyperparameters based on the validation performance to prevent overfitting. The validation set provides an unbiased evaluation of the performance of the model and helps determine how well the model will generalize to new data.

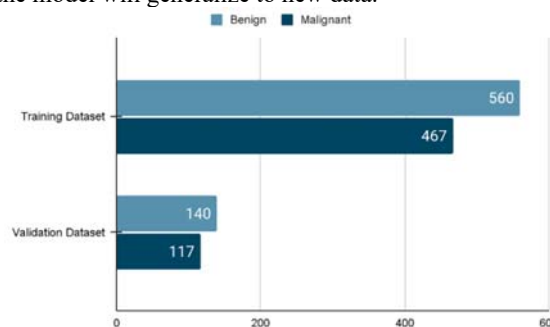


Figure 1. Number of images in Training and Validation Datasets

After the training and validation stages are finished, the testing dataset is used to assess the ultimate performance of the trained model. It simulates real-world scenarios where the model encounters unseen data.



Figure 2. An image from a Benign class



Figure 3. An image from a Malignant class

For evaluating the performance of the model on the new data and estimating its capacity for generalization, the testing dataset is essential. Performance measures like accuracy, precision, recall, and F1-score are calculated based on the model’s predictions on the testing data. The number of images in training and validation datasets is shown in figure 1 and a sample image from each class is shown in figures 2 and 3.

3.2 CNN architecture of the proposed model

The architecture of the proposed model is shown in Table 1. The network architecture consists of several layers: a convolutional layer with 62 filters, followed by a MaxPooling2D layer with a pool size of 2x2.

Table 1. CNN Architecture

Layer (type)	Output Shape	Param #
squential_2 (Sequential)	(64, 512, 512, 3)	0
conv2d_3 (Conv2D)	(64, 510, 510, 64)	1,792
max_pooling_3 (MaxPooling2D)	(64, 255, 255, 64)	0
conv2d_4 (Conv2D)	(64, 253, 253, 32)	18,464
max_pooling_4 (MaxPooling2D)	(64, 126, 126, 32)	0
conv2d_5 (Conv2D)	(64, 122, 122, 16)	12,816
max_pooling_5 (MaxPooling2D)	(64, 61, 61, 16)	0
flatten_1 (Flatten)	(64, 59536)	0
dense_5 (Dense)	(64, 128)	7,620,736
dense_6 (Dense)	(64, 128)	16,512
dropout_1 (Dropout)	(64, 128)	0
dense_7 (Dense)	(64, 64)	8,256
dense_8 (Dense)	(64, 32)	2,080
dense_9 (Dense)	(64,1)	33

This is succeeded by another convolutional layer with 32 filters, followed by another MaxPooling2D layer with a pool size of 2x2. Subsequently, there's a convolutional layer with a filter size of 16, followed by another MaxPooling2D layer with a pool size of 2x2. The output is then flattened and passed through two consecutive dense layers, each with 128 units. A dropout layer with a dropout rate of 0.3 is then applied, followed by three consecutive dense layers with 64, 32, and 1 unit, respectively. The total number of trainable parameters in the model is 7,680,689.

The flattening layer, which turns the 2D feature maps into a 1D vector, comes after the convolutional and max-pooling layers. The flattened output is then passed through six dense layers, starting with 128 neurons, with one dropout layer (dropout rate = 0.3) inserted between the dense layers to regularize the network and improve generalization. An Adam optimizer is employed along with a learning rate of 0.001 in the model for optimization purposes, coupled with a categorical cross-entropy loss function, rendering it suitable for applications involving multi-class classification. Subsequently, this enables the

model to continuously learn from the training data and optimize its parameters to minimize the loss function.

3.3 Hyperparameters

The tunable hyperparameters in the proposed model are listed in Table 2.

Table 2. Hyperparameters

Parameter	Value
Batch size	64
Image size	512 x 512
Activation function	Rectified Linear Unit (ReLU)
Optimizer	Adam
Loss Function	Binary cross entropy
Epochs	100

3.4 Evaluation Metrics

The performance evaluation metrics to assess the efficacy of the proposed model are presented below.

3.4.1 Accuracy

Accuracy describes how often the model's prediction comes true out of the total predictions it makes.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \quad (1)$$

3.4.2 Precision

Precision of the model focuses on the positive predictions that the model makes. It shows how often the model predict positive instances without mislabeling negative as positive.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2)$$

3.4.3 Recall

Recall measures how well the model is at finding every positive instance present in the dataset. It indicates whether the model can catch all the positives without missing any.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (3)$$

3.4.4 F1-score

It is a harmonic mean of precision and recall which provides a balance between precision and recall when the classes in the dataset are imbalanced. It ranges between 0 to 1. A higher F1 – score indicates that model performed well in terms of both precision and recall.

$$\text{F1 score} = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

3.4.5 Confusion Matrix

A table that shows how well a classification system performs by contrasting expected and actual labels. It constitutes of four cells for a binary classification: True

Table 3. Structure of Confusion matrix for a binary classification problem

		Actual Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), as can be seen in Table 3.

1. A true positive (TP) occurs when a positive data point is accurately anticipated by the model.
2. A true negative (TN) occurs when the model properly predicts a negative data point.
3. A false positive (FP) occurs when the model predicts a positive data point incorrectly.
4. False negatives (FN) occur when the model misclassifies a negative data item.

4. RESULTS AND DISCUSSION

The results obtained by the proposed model are presented in this section.

4.1 Confusion matrix

Figure 4 shows the Confusion matrix generated based on the predictions made by the model on the Training dataset. The cell on the top-left corner (TP) indicates instances correctly classified as positive, totaling 528. Conversely, the top-right cell (FP) signifies instances inaccurately classified as positive, with a count of 33. T

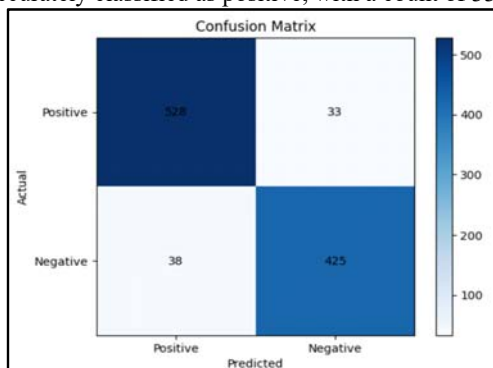


Figure 4. Confusion Matrix of Training dataset

he bottom-left cell (FN) denotes instances mistakenly classified as negative, recorded at 38. Lastly, the bottom-right cell (TN) accurately classifies instances as negative, standing at 425. The model demonstrates balanced sensitivity and specificity as evidenced by the comparable counts of TP and TN. The high values along

the main diagonal (true positives and true negatives) indicate strong overall performance, with relatively few misclassifications.

Figure 5 displays the Confusion matrix, which represents the predictions made by the model on the Validation dataset. The matrix is presented in a 2x2 format. The number of cases that were correctly identified as positive, 68 in total, is displayed in the cell located in the top-left corner (TP). Adjacently, the top-right cell (FP), signifies instances erroneously classified as positive, recorded at 8. Conversely, the bottom-left cell (FN) indicates instances mistakenly classified as negative, amounting to 6. Lastly, the bottom-right cell (TN) accurately classifies instances as negative, standing at 46.

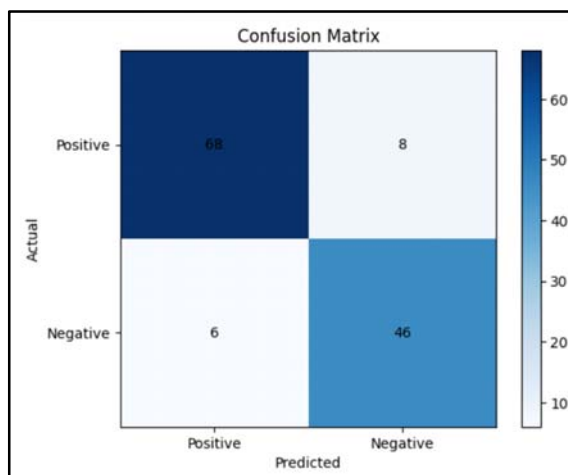


Figure 5. Confusion Matrix of Validation dataset

Figure 6 shows the Confusion matrix generated based on the predictions made by the model on the Testing dataset.

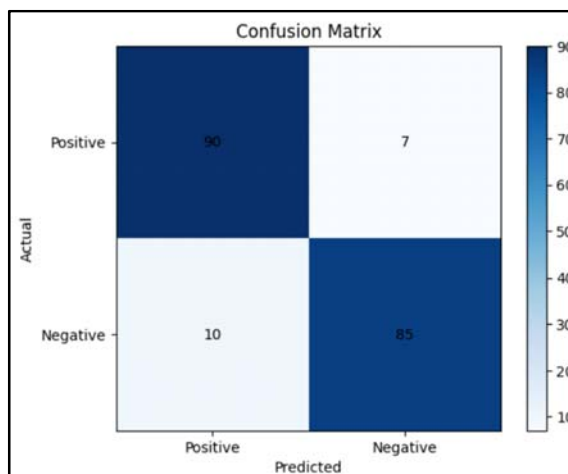


Figure 6. Confusion Matrix of Testing dataset

The count of TP predictions stands at 90, reflecting instances correctly classified as positive. Conversely, the count of FP predictions is observed to be 7, indicating instances inaccurately classified as positive. The FN predictions, totaling 10, denote instances mistakenly

classified as negative. Correspondingly, the count of TN predictions is recorded at 85, representing instances accurately classified as negative. These findings collectively underscore the model's robustness in accurately classifying instances within the given classification task.

4.2 Performance Metrics

The results of the evaluation metrics of the CNN model developed to classify benign and malignant cases of skin cancer are presented in Table 4.

Table 4. Performance evaluation metrics obtained by the proposed model

	Accuracy	Precision	Recall	F1-Score
Training	0.99	0.93	0.93	0.93
Validation	0.91	0.89	0.89	0.89
Testing	0.93	0.91	0.91	0.91

Accuracy: The model demonstrates high accuracy across all datasets, with training accuracy at 0.99, validation accuracy at 0.91, and testing accuracy at 0.93. This indicates that the model performs exceptionally well in correctly classifying instances of both benign and malignant cases.

Precision: Precision measures the ratio of true positive predictions to the total positive predictions made by the model. The training, validation, and testing precisions are all relatively high, with values of 0.93, 0.89, and 0.91, respectively. This suggests that the model has a low false positive rate and effectively identifies malignant cases without misclassifying benign cases as malignant.

Recall: Recall, also known as sensitivity, evaluates the model's ability to correctly identify true positive instances from all actual positive instances. The recall scores for training, validation, and testing are consistent at 0.93, indicating that the model effectively captures the majority of malignant cases in the dataset.

F1-Score: The F1-Score represents the harmonic mean of precision and recall, providing a balanced measure of a model's performance. Across all datasets, the F1-Score is consistent at 0.93, 0.89, and 0.91 for training, validation, and testing, respectively. This indicates that the model achieves a good balance between precision and recall, reflecting its overall effectiveness in classifying both benign and malignant cases.

Overall, the results suggest that the CNN model performs well in distinguishing between benign and malignant cases of skin cancer, exhibiting high accuracy, precision, recall, and F1-Score across training, validation, and testing datasets.

4.3 Training Versus Validation accuracy plot

The training versus validation curves plotted over 100 epochs is depicted in figure 7 showcase the model's

progression in accuracy throughout the training process. Initially, both training and validation accuracies show significant fluctuations, indicating the model's learning process. The training accuracy starts at 0.53 and steadily increases, eventually reaching a high of 0.99 by the end of the 100 epochs, indicating that the model effectively learns from the training data and achieves near-perfect accuracy on the training set. Conversely, the validation accuracy starts at a lower point of 0.1 but also demonstrates a steady upward trend, reaching 0.91 by the end of the epochs. This suggests that the model generalizes well to unseen data, as evidenced by the consistent improvement in validation accuracy. The narrowing gap between training and validation accuracies indicates minimal overfitting, signifying that the model effectively learns the underlying patterns in the data without memorizing noise. Overall, the results suggest that the model successfully learns from the training data and generalizes well to unseen data, showcasing its robustness and effectiveness in classifying benign and malignant cases of skin cancer.

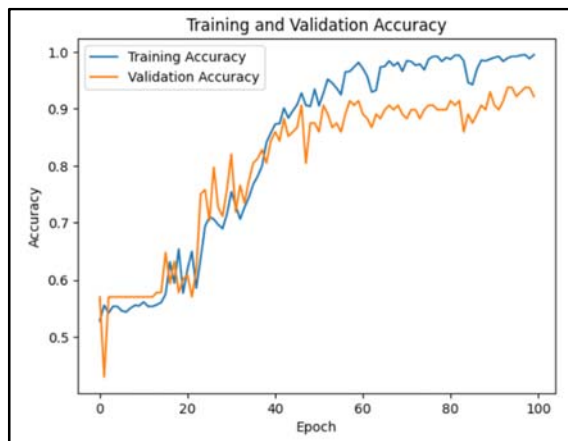


Figure 7. Training Vs Validation accuracy curves

The overall findings demonstrate the effectiveness of the proposed model in accurately classifying pigmented skin lesions and differentiating between benign and malignant cases. High test accuracy, combined with strong performance measures like recall, precision, and F1-score, highlights the potential of machine learning methods for automated skin cancer diagnosis.

5. CONCLUSION AND FUTURE WORK

In this study using a large dataset from the ISIC 2020 challenge, our tailored CNN architecture has proven effective in distinguishing between benign and malignant skin lesions with high accuracy across various datasets. The model's reliability and consistency in performance highlight its potential to revolutionize the early detection and treatment of skin cancer, ultimately enhancing patient outcomes and reducing the burden on healthcare systems.

The integration of advanced machine learning techniques with clinical protocols has not only facilitated the automation of diagnostic processes but also minimized manual errors, thereby expediting patient care pathways. The successful integration of this model can lead to significant improvements in clinical practices, reducing unnecessary biopsies and surgeries, and potentially lowering healthcare costs through early and accurate diagnosis.

Future Directions

To further enhance the efficacy and applicability of the CNN model, the following future directions are proposed:

- Expanding the training dataset to include a wider range of skin lesion types from diverse demographic and geographic backgrounds can help improve the model's generalizability and accuracy. Future studies can consider the inclusion of multi-racial and multi-ethnic data to ensure the model's effectiveness across different skin types and conditions.
- Creating a user-friendly diagnostic platform that integrates this CNN model could facilitate its adoption in clinical settings. Such a platform could assist dermatologists by providing preliminary diagnosis suggestions, thus improving the efficiency of the diagnostic process.

References:

- Anand, V., Gupta, S., Nayak, S. R., Koundal, D., Prakash, D., & Verma, K. D. (2022). An automated deep learning models for classification of skin disease using Dermoscopy images: A comprehensive study. *Multimedia Tools and Applications*, 81(26), 37379-37401.
- Blackledge, J., & Dubovitskiy, D. (2008). Object Detection and Classification with Applications to Skin Cancer Screening. *ISAST Transactions on Intelligent Systems*, 1(1), 34-45, doi:10.21427/D7M32K
- Chaugule, B., Bomble, K., Jundare, S., Maske, N., & Gagare, V. (2023). Skin Melanoma Cancer Detection and Classification using Machine Learning. *International Journal of Scientific Research in Science and Technology*, 10(3), 519-524. DOI: 10.32628/IJSRST523103110
- Junayed, M. S., Anjum, N., Noman, A., & Islam, B. (2021). A deep CNN model for skin cancer detection and classification. *Computer Science Research Notes*, CSRN 3101 871-80. DOI:10.24132/CSRN.2021.3101.
- Malvia, S., Bagadi, S. A., Dubey, U. S., & Saxena, S. (2017). Epidemiology of breast cancer in Indian women. *Asia-Pacific Journal of Clinical Oncology*, 13(4), 289-295.
- Muniteja, M., Bee, M. M., & Suresh, V. (2022, October). Detection and classification of Melanoma image of skin cancer based on Convolutional Neural Network and comparison with Coactive Neuro Fuzzy Inference System. In *2022 International Conference on Cyber Resilience (ICCR)* (pp. 1-5). IEEE.
- Paliwal, N. (2016). Skin cancer segmentation, detection and classification using hybrid image processing technique. *International Journal of Engineering and Applied Sciences*, 3(4), 257678.
- Shetty, B., Fernandes, R., Rodrigues, A. P., Chengoden, R., Bhattacharya, S., & Lakshmana, K. (2022). Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Scientific Reports*, 12(1), 18134.
- Sigurdsson, S., Philipsen, P. A., Hansen, L. K., Larsen, J., Gniadecka, M., & Wulf, H. C. (2004). Detection of skin cancer by classification of Raman spectra. *IEEE transactions on biomedical engineering*, 51(10), 1784-1793.
- Subramanian, R. R., Achuth, D., Kumar, P. S., Kumar Reddy, K. N., Amara, S., & Chowdary, A. S. (2021, January). Skin cancer classification using Convolutional neural networks. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 13-19). IEEE.
- Sutradhar, A., Tajmen, S., Dhaly, A. A., Shamrat, F. J. M., Talukder, M. S. R., & Khater, A. (2022, October). Skin cancer classification and early detection on cell images using multiple convolution neural network architectures. In *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 1089-1094). IEEE.
- Tanna, R., & Sharma, T. (2021, September). Binary classification of melanoma skin cancer using svm and cnn. In *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)* (pp. 1-4). IEEE.
- Torti, E., Leon, R., La Salvia, M., Florimbi, G., Martinez-Vega, B., Fabelo, H., ... & Leporati, F. (2020). Parallel classification pipelines for skin cancer detection exploiting hyperspectral imaging on hybrid systems. *Electronics*, 9(9), 1503.
- Victor, A., & Ghalib, M. R. (2017). Automatic Detection and Classification of Skin Cancer. *International Journal of Intelligent Engineering & Systems*, 10(3), 444-451, DOI: 10.22266/ijies2017.0630.50

Amritanjali Swaroop

Dayalbagh Educational Institute
Dayalbagh, Agra
India

amritanjali2002106@dei.ac.in

ORCID: 0009-0004-1228-3025

A. Charan Kumari

Dayalbagh Educational Institute
Dayalbagh, Agra
India

charankumari@dei.ac.in

ORCID: 0000-0002-3160-1912

K. Srinivas

Dayalbagh Educational Institute
Dayalbagh, Agra
India

ksri12@gmail.com

ORCID:0009-0002-3884-6282
