International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2781- 2795
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0385

Exploratory Analysis of Skin Cancer Dermatoscopic Image Datasets and Classification Methods

Prof. Dipali Ghatge[1], Prathmesh Sankpal[2], Vedantika Gharge[3], Rutuja Patil[4], Shivam Zanjurne[5]
[1,2,3,4,5]Department of Computer Science and Engineering, Karmaveer Bhaurao College of Engineering Satara, Maharashtra, India
Emails: *dipali.ghatge@kbpcoes.edu.in[1], prathmeshsankpal483@gmail.com[2], vedantikagharge24@gmail.com[3], patilrutuja294@gmail.com[4], z.shivam093@gmail.com[5]*

Abstract
Skin cancer is a critical global health issue, where early detection significantly improves treatment outcomes. In this review paper, titled Exploratory Analysis of Skin Cancer Dermatoscopic Image Datasets and Classification Methods, we systematically explore the intersection of artificial intelligence (AI) and skin cancer diagnosis. Our approach began with a detailed literature survey of 40 research studies, providing insights into advancements and challenges in AI-based classification methods. This survey emphasizes the role of machine learning algorithms, particularly convolutional neural networks (CNNs), and datasets such as HAM10000, ISIC 2019, and ISIC 2024 in improving diagnostic performance. Next, we conducted exploratory data analysis (EDA) on the HAM10000, ISIC 2019, and ISIC 2024 datasets, uncovering critical patterns related to lesion distribution, anatomical sites, and demographic factors. These analyses highlight biases and imbalances in the datasets, which are crucial to address for robust model training. Finally, we discuss the creation and evaluation of a machine learning model trained on a separate dataset. Initial experiments revealed challenges such as overfitting and class imbalance. Through advanced data augmentation techniques and the integration of an Augmentor pipeline, we mitigated these issues, achieving improved accuracy and generalizability. This paper provides a comprehensive framework for integrating literature insights, dataset analysis, and iterative model improvement to develop effective AI-based solutions for skin cancer detection. It underscores the importance of addressing dataset biases, adopting diverse datasets, and refining methodologies to advance AI applications in dermatology.

Keywords: *AI for dermatology; Convolutional neural networks; Data augmentation; Dermatoscopic image analysis; Exploratory data analysis.*

1. Introduction

Artificial intelligence (AI) has garnered significant attention in recent years for its potential to enhance skin cancer detection from dermoscopic images. Early detection is crucial for effective skin cancer treatment, and AI-driven models, especially those utilizing deep learning techniques such as convolutional neural networks (CNNs), have demonstrated accuracy rates exceeding 90%. These models provide dermatologists with a valuable tool to distinguish between benign and malignant tumors with high precision. While various methods, such as Support Vector Machines (SVMs) and K-Nearest Neighbors (KNNs), have been explored for skin cancer classification, CNNs remain the most commonly employed due to their ability to identify complex patterns in image data. These models have been trained and evaluated on several publicly available datasets, including the ISIC Archive, HAM10000, and specialized datasets, which allow for the comparison of model performance across different image types. Recent research has expanded AI models to include more diverse datasets, such as clinical images from smartphones and hyperspectral imaging, aiming to increase their applicability to real-world scenarios. Papers have utilized datasets like the ISIC, HAM10000, PH2, and hyperspectral skin cancer images, enabling the training of advanced models such as the AICO self-feature selected ECNN and 34-layer ResNet. These datasets are critical in ensuring the robustness of the models across different skin cancer categories Despite these advancements, challenges such as the need for more varied datasets, improving AI decision-making interpretability, and addressing biases remain. There are also concerns about the generalizability of AI models in real-world clinical settings, as highlighted in studies using datasets like MIDAS. Overcoming these challenges is essential for integrating AI techniques into clinical practice. Overall, AI holds significant promise for improving the accuracy and efficiency of skin cancer diagnostics, though further research is needed to address the remaining issues and ensure its successful application in medical practice. [1]

2. Literature Survey

Datasets play a crucial role in the development and validation of machine learning models for skin cancer detection. The availability of diverse and well-curated datasets enables researchers to train models that can accurately distinguish between benign and malignant lesions. In recent years, several high-quality datasets, such as the ISIC, HAM10000, and other specialized collections, have been utilized to train advanced models, including convolutional neural networks (CNNs) and decision trees. These datasets, often accompanied by metadata like biopsy-proven diagnoses and dermoscopic images, provide the foundation for robust image classification systems. This section provides a literature survey summarizing recent studies that have leveraged different datasets for skin cancer detection and highlights the methodologies and outcomes of each. The paper [1] focuses on leveraging two prominent datasets, the ISIC dataset and the MNIST dataset, for skin cancer detection. The ISIC dataset contains 10,015 dermoscopic images encompassing a wide array of diagnostic categories crucial for identifying skin cancer. On the other hand, the MNIST dataset consists of 2,357 images, including both malignant and benign oncological cases. The datasets were instrumental in training the AICO self-feature selected ECNN model, which was mathematically formulated in the research. The images from the ISIC dataset were classified based on the standard ISIC classification, ensuring equal representation across subsets, except for a slightly higher proportion of images related to melanomas and moles. Prior to being processed by the model, the images underwent pre-processing techniques such as morphological and blur filters, which effectively minimized noise and removed artifacts from the input data. To evaluate the model's performance, several key metrics were used, including accuracy, critical success index (CSI), false positive rate (FPR), and false negative rate (FNR). The comprehensive use of these datasets, along with robust pre-processing and evaluation metrics, highlights the significance of dataset selection and processing in achieving reliable skin cancer detection results. The paper [2] makes use of an extensive dataset comprising 129,450 clinical images, representing 2,032 different diseases. The dataset, significantly larger than those used in previous research, includes images from a variety of open-access sources, such as the ISIC Dermoscopic Archive, the Edinburgh Dermofit Library, and data from Stanford Hospital. Importantly, these images are biopsy-proven, ensuring a high level of reliability for the classification tasks undertaken by the CNN. During the training process, blurry or distant images were excluded from the test and validation sets but were still utilized for training. Additionally, careful attention was given to avoid splitting images of the same lesion (captured from multiple angles) between the training and validation sets, ensuring the integrity of the test data. The test sets were derived from independent, high-quality repositories of biopsy-proven images, with no overlap between the test and training/validation data. The model was trained using transfer learning, building on features learned from the ImageNet dataset, which contains 1.28 million images. This approach enabled the CNN to leverage pre-trained natural image features, improving its capability to classify dermatological conditions. The paper [3] leverages a hyperspectral skin cancer dataset consisting of 76 images of skin lesions from

61 subjects, with 40 benign and 36 malignant lesions. The images were captured using a snapshot camera (Cubert UHD, Cubert GmbH), covering the 450-950 nm range across 125 spectral channels. The dataset was collected in collaboration with two hospitals in the Canary Islands, Spain: Hospital Universitario de Gran Canaria Doctor Negrin and Complejo Hospitalario Universitario Insular-Materno Infantil.
Expert dermatologists and pathologists labeled the images according to a defined taxonomy, ensuring accurate classification. The study utilized this dataset to test Support Vector Machine (SVM), Random Forest (RF), and eXtreme Gradient Boosting (XGB) algorithms, both in serial and parallel modes, to classify the hyperspectral images. In addition to classification performance, the study measured the average classification time for each algorithm, demonstrating significant improvements in diagnostic speed through the application of hyperspectral imaging and machine learning techniques. This highlights the potential for accelerated medical diagnoses using advanced image processing methods. The paper [4] utilizes the HAM10000 dataset, which comprises images of skin lesions classified into seven categories: actinic keratosis, basal cell carcinoma, benign keratosis, dermatofibroma, melanocytic nevi, melanoma, and vascular lesions. The dataset was primarily used to evaluate the performance of the DUNEScan application, focusing on the most common malignant and benign lesion types: melanoma and melanocytic nevi. The training dataset was derived from the International Skin Imaging Collaboration (ISIC) archive, containing a total of 23,900 skin lesion images, including 2,287 malignant and 21,613 benign lesions. To address class imbalance, the paper randomly sampled 10,000 benign lesion images and combined them with all available malignant lesion images, forming a meta dataset for the analysis. The meta dataset was then split into an 80-20 train-test ratio, ensuring balanced representation of benign and malignant cases in the test set. A fivefold cross-validation approach was adopted during training, dividing the training set into five groups to comprehensively assess the loss and accuracy of the CNN model. This approach ensured robust evaluation, providing insights into the model's performance across multiple subsets of the dataset. The paper [5] utilizes a combination of 22 microarray datasets and 5 RNA-seq datasets, containing samples related to multiple skin pathological states (SPSs) associated with skin cancer. The datasets are integrated to enhance the analysis of gene expression and to identify differentially expressed genes (DEGs) across various SPSs. Each dataset undergoes a preprocessing phase that includes quality analysis to remove potentially erroneous samples, ensuring the integrity of the data used for analysis. The integration process involves summarizing expression values of genes with the same identifier and correcting for batch effects to achieve effective data integration. The analysis employs a cross-validation approach, splitting the integrated dataset into training and testing sets to ensure representativeness of each SPS. The paper emphasizes the importance of selecting informative DEGs through feature selection algorithms, which reduces the search space and improves the reliability of the results. The paper [6] uses the International Skin Imaging Collaboration (ISIC) database, which contains over 20,000 labeled dermoscopic images. The ResNet model was trained on more than 12,000 images from this database to classify dermoscopic images of melanocytic lesions as benign or malignant. The performance of dermatologists was assessed using 200 test images, where they were asked to make biopsy/treat or reassure decisions similar to previous studies. The dataset's strict quality standards ensure the reliability of the training and testing processes. This paper emphasizes the importance of a well-curated dataset in training machine learning algorithms, ultimately improving diagnostic accuracy in dermatology. The paper [7] utilizes the HAM10000 dataset, which contains images of pigmented lesions categorized into seven diagnostic categories: malignant (melanomas, basal cell carcinomas, actinic keratoses, and intraepithelial carcinomas) and benign (melanocytic nevi, benign keratinocytic lesions, dermatofibromas, and vascular lesions). The dataset was used to train a convolutional neural network (CNN), specifically a 34-layer residual network (ResNet34), for lesion classification. The training of

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2781- 2795
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0385

the CNN was performed on NVIDIA graphics processing units (GPUs) using the Pytorch framework. The mean recall of the CNN across all disease categories was reported to be 77.7, with an accuracy of 80.3 when tested on a publicly available benchmark test set. Additionally, the study analyzed images collected from a randomized controlled trial on self-examinations in high-risk patients, where participants submitted self-made photographs of suspicious lesions for telediagnosis. In total, 1,521 self-made photographs of 596 suspicious lesions were analyzed, although the CNN was primarily trained on curated images of pigmented lesions. The paper [8] discusses various datasets used in previous studies for skin cancer diagnosis, specifically highlighting the International Skin Imaging Collaboration (ISIC) dataset from 2020. This dataset includes images of different skin cancer categories such as actinic keratosis, basal cell carcinoma, dermatofibroma, melanoma, nevus, seborrheic keratosis, squamous cell carcinoma, and vascular lesions. The dataset comprises a total of 2,357 dermoscopic images, which contain both malignant and benign oncological disease images, with a slight dominance of melanoma and mole images. The paper emphasizes the importance of dataset availability for training machine learning and deep learning algorithms, noting that a lack of diverse datasets can lead to subpar model performance. It also mentions that some datasets do not include benign lesions, which are common in dermatological practice, potentially leading to missed skin cancer diagnoses. To address dataset imbalances, researchers are employing data augmentation techniques such as cropping, rotation, and filtering to increase the number of training images. In this paper [9] it utilizes the 2012 TTQS (Taiwan TrainQuali System) central Taiwan review database to explore human training quality and identify critical assessment indicators for the TTQS. The dataset is analyzed using back-propagation neural networks to evaluate classification accuracy and performance, achieving prediction accuracies greater than 95% for both training and testing samples. Additionally, the study employs K-Means clustering analysis to identify critical indicators selected by the decision tree algorithms. The decision tree algorithms analyzed include C5.0, CART, and CHAID, with C5.0 demonstrating the highest accuracy rate of 89.41%. The dataset is processed to facilitate analysis, including scaling of variable value fields to ensure efficient data processing. The paper [10] utilizes the Melanoma Research Alliance Multimodal Image Dataset for AI-based Skin Cancer (MIDAS), the largest publicly available dataset containing paired dermoscopic and clinical images of biopsy-proven, dermatopathology-labeled skin lesions. The dataset includes 3830 images representing 1290 unique lesions from 796 patients, collected under an IRB-approved protocol with informed consent. The images cover a wide diagnostic range, including malignant, benign, and inflammatory lesions, such as melanocytic nevi, invasive cutaneous melanomas, and melanoma in situ. One of the unique aspects of MIDAS is its ability to assess the performance of AI algorithms on real-world clinical images, which is crucial for determining their effectiveness in practical settings. In this paper, four state-of-the-art AI models, previously published and widely recognized for their high performance in skin cancer detection, were evaluated using the MIDAS dataset. The results demonstrated a notable decrease in model performance when applied to this dataset, highlighting the ongoing challenge of generalizability in AI-based diagnostic systems. The difficulty AI models faced in maintaining their accuracy with the MIDAS dataset suggests that real-world variations in clinical and dermoscopic images can significantly impact model efficacy. In the paper [11], the PH2 dataset is utilized to evaluate the performance of the proposed melanoma detection system. The dataset contains 100 dermoscopic images, with 80 images used for training and 20 images for testing. The dataset is employed to train and test a Support Vector Machine (SVM) classifier, where features such as color, shape, and texture are analyzed to classify the skin lesions as either normal or melanoma. The performance of the system is evaluated using metrics such as sensitivity, specificity, and accuracy, calculated based on the PH2 dataset. [2-5]

3. Overview of Datasets and Methodologies

Table 1 **Overview of Datasets**

Paper	Dataset	Number of Images	Methods/Algorithms	Additional Insights/Implications
1	ISIC dataset, MNIST dataset	12,372	AICO self-feature selected ECNN	Highlights the importance of dataset diversity in training.
2	ISIC Dermoscopic Archive, Edinburgh Dermofit Library, Stanford Hospital Data	129,450	CNN with transfer learning from ImageNet	Emphasizes robustness through large, biopsy-proven datasets.
3	Hyperspectral skin cancer dataset	76	SVM, RF, XGB	Demonstrates improved diagnostic speed through advanced imaging.
4	HAM10000 dataset	23,900	CNN (DUNEScan application)	Focuses on uncertainty estimation in skin cancer detection.
5	22 microarray datasets, 5 RNA-seq datasets	N/A	Feature selection algorithms	Integrates diverse data types to enhance gene expression analysis.
6	International Skin Imaging Collaboration (ISIC) database	20,000+	ResNet	Stresses the importance of high-quality, labeled datasets.
9	TTQS central Taiwan review database	N/A	Back-propagation neural networks, K-Means, C5.0, CART, CHAID	Identifies critical assessment indicators for training quality.
10	MIDAS dataset	3,830	Four state-of-the-art AI models	Assesses generalizability of AI algorithms in real-world scenarios
11	PH2 dataset	100	SVM	Evaluates feature importance in melanoma detection.

4. Dataset Insights with Visualization

4.1. HAM10000 Dataset

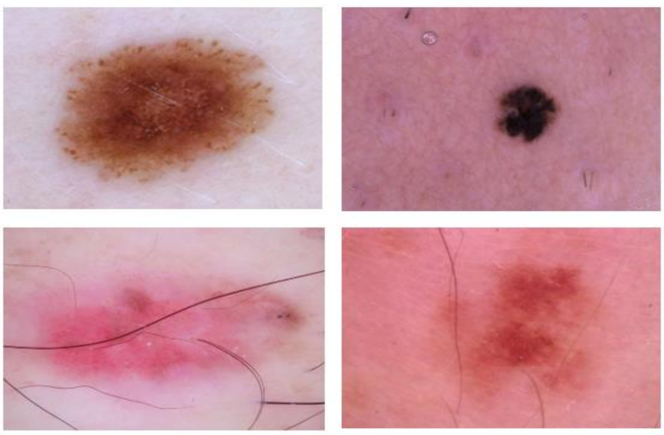

Figure 1 **Sample Images (HAM10000)**

- **Full Name:** Human Against Machine with 10,000 training images (HAM10000)
- **Description:** The HAM10000 dataset contains 10,015 dermatoscopic images, including images of various common pigmented skin lesions.
- **Lesion Types:** Seven types of skin lesions including melanoma (mel), benign keratosis-like lesions (bkl), and dermatofibroma (df).
- **Sample** Size: 10,015 images Figure 1 shows Sample Images (HAM10000)

4.2. Visual Insights from HAM10000 Dataset

4.2.1. Lesion Type Distribution

- Nevus (NV) is the most common lesion type, accounting for 66.9% of cases, followed by Melanoma (Mel) and Benign Keratosis-like Lesions (BKL), each representing around

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2781- 2795
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0385

11% of cases. The presence of malignant lesions such as melanoma highlights the critical need for accurate classification models.

- Less frequent lesion types, such as Basal Cell Carcinoma (BCC) (5.1%), Actinic Keratosis (AKIEC) (3.3%), Vascular Lesions (VASC) (1.4%), and Dermatofibroma (DF) (1.1%), emphasize the dataset's variety and the importance of training models to detect rare lesions. Figure 2 shows Lesion Type Distribution (HAM10000) [6-10]

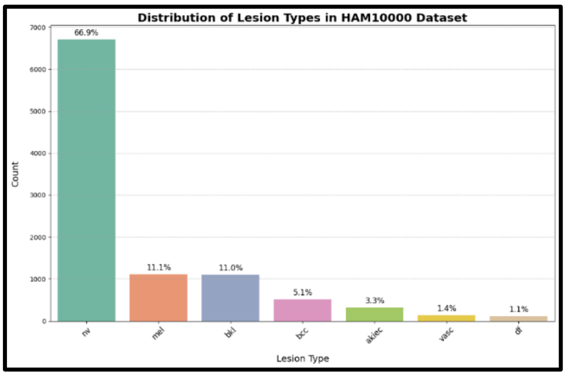

Figure 2 Lesion Type Distribution (HAM10000)

4.2.2. Lesion Localization Distribution

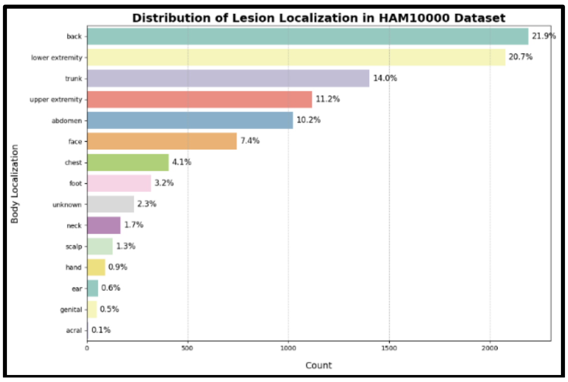

Figure 3 Lesion Location Distribution (HAM10000)

- Lesions are most frequently located on the **lower extremity** (**21.9%**) and the **back**
- **20.7%**). These locations are common sites for both benign and malignant lesions, especially melanoma. Figure 3 shows Lesion Location Distribution (HAM10000)
- The **trunk** (**14%**), **upper extremity** (**11.2%**), and **abdomen** (**10.2%**) are also significant sites, providing a diverse set of lesion localizations, crucial for improving model accuracy in various body areas. Figure 4 shows Lesion Type by Gender (Ham10000)
- Less common sites like the **foot**, **face**, and **chest** demonstrate that the dataset covers a wide range of lesion locations, although with lower prevalence. [11]

4.2.3. Gender Distribution of Lesion Types

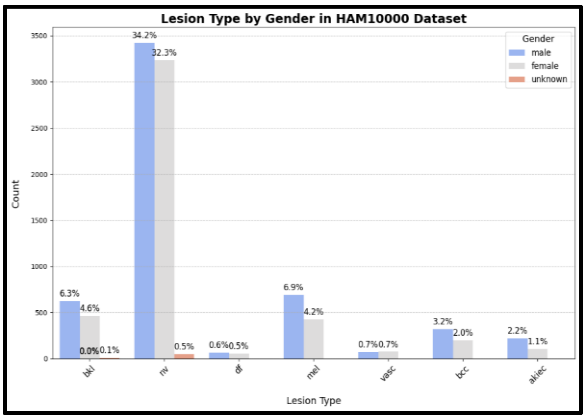

Figure 4 Lesion Type by Gender (Ham10000)

While **Nevus (NV)** is the most common lesion type across both genders, it is slightly more, prevalent in females (34.2%) compared to males.

1. **AgeDistribution:**

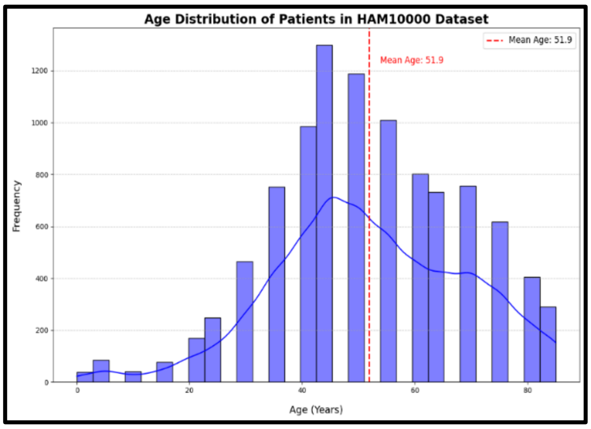

Figure 5 Age Distribution (Ham10000)

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2781- 2795
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0385

- The **mean age** of patients is **51.9 years**, with a majority of lesions occurring in patients between **30 and 70 years old**. The dataset is skewed toward older individuals, who are at higher risk for developing skin cancers like melanoma.
- The right-skewed distribution reflects the increased likelihood of skin lesions, especially malignant ones, in older adults, reinforcing the importance of early detection strategies for this age group. [13]

4.3. ISIC 2019 Dataset

Full Name: International Skin Imaging Collaboration (ISIC) 2019 Challenge Dataset

- Description: The ISIC 2019 dataset includes over 25,000 dermatoscopic images for the detection of melanoma and other skin lesions. It was used as part of a global challenge for melanoma detection. Figure 6 Sample Images shows (ISIC 2019)
- Lesion Types: Eight categories of skin lesions, including melanoma and various benign and malignant lesions.
- Sample Size: 25,331 images

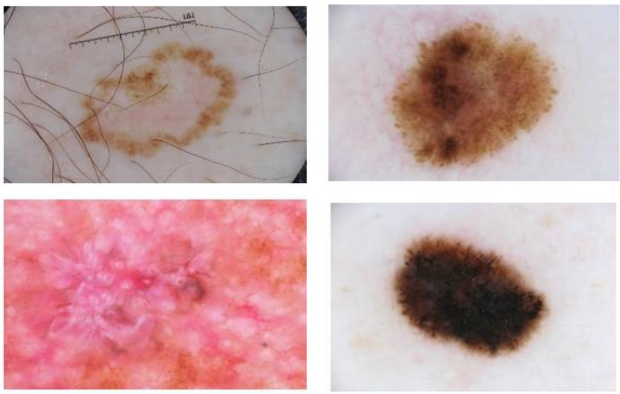

Figure 6 Sample Images (ISIC 2019)

4.4. Visual Insights from ISIC 2019 Dataset

4.4.1. Age Approximation Distribution of Patients

The ISIC 2019 dataset provides a comprehensive age distribution ranging from 0 to 90 years, making it a robust resource for studying the correlation between age and skin cancer occurrence.

- Peak Age Range: The majority of cases fall within the 45 to 50-year range, highlighting that middle-aged adults are the most affected demographic. The Kernel Density Estimation (KDE) curve peaks in this age group, emphasizing a high concentration of lesions during this phase of life. Figure 7 shows Age Distribution (ISIC 2019) [12]
- Fewer Younger and Elderly Patients: The dataset shows a relative scarcity of cases in younger and older patients, underlining the need for better screening efforts in these age groups. This skew suggests that while skin cancer primarily impacts middle-aged adults, it's important to extend research and prevention efforts to younger and elderly populations, who may face unique risk factors. Figure 8 shows Anatomical Sites Distribution (ISIC 2019)

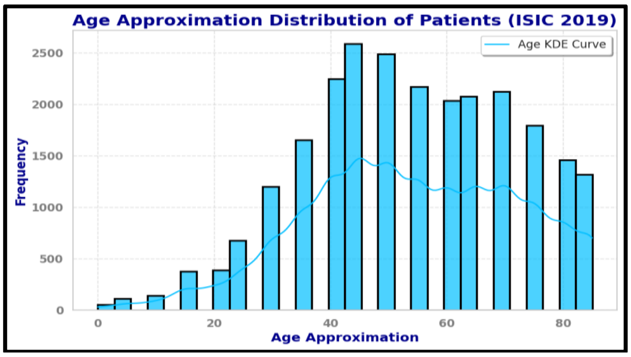

Figure 7 Age Distribution (ISIC 2019)

This age distribution insight underscores the critical need for **targeted screening programs** aimed at middle-aged adults, with additional emphasis on encouraging early detection practices across all age groups.

4.4.2. Distribution of Anatomical Sites

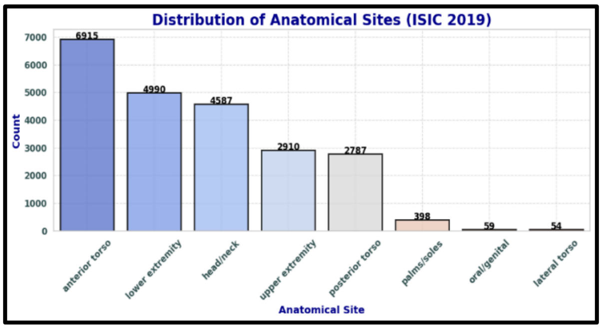

Figure 8 Anatomical Sites Distribution (ISIC 2019)

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2781- 2795
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0385

An analysis of lesion localization reveals key anatomical sites where skin lesions most commonly appear, which is vital for targeted diagnostic approaches

- Anterior Torso: The most frequently affected area, with 6,915 instances, indicates that the torso, a typically sun-exposed region, is prone to developing skin lesions, especially melanoma. Clinicians should prioritize this area during examinations.
- Lower and Upper Extremities: Together, the lower extremity (4,990 instances) and upper extremity (4,587 instances) account for a significant portion of lesions. These areas, often exposed to sunlight, highlight the importance of limb monitoring in routine skin examinations. Figure 9 shows Correlation of Anatomical Sites and Gender (ISIC 2019)
- Less Common Sites: Areas such as the oral/genital regions (398 cases) and lateral torso (59 cases) have fewer occurrences, but their presence in the dataset reflects the need to consider atypical anatomical sites in clinical diagnostics. [14]

This distribution insight demonstrates that focusing on high-frequency areas like the torso and limbs, while remaining vigilant about less common locations, is key for effective lesion detection.

4.5. Correlation Between Anatomical Site and Gender

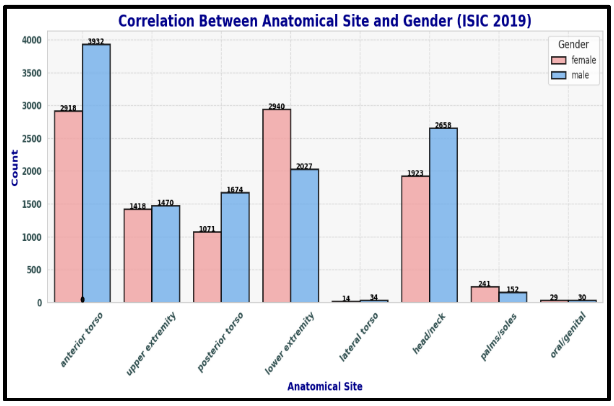

Figure 9 Correlation of Anatomical Sites and Gender (ISIC 2019)

The ISIC 2019 dataset reveals important **gender-based differences** in lesion localization, which can guide the development of more personalized diagnostic models

- Anterior Torso: Exhibits a gender disparity, with 3,932 cases in females and 2,918 cases in males. This suggests that females may be more prone to developing lesions in this region, possibly due to lifestyle or biological factors, necessitating targeted screening.
- Upper Extremities: With a relatively even distribution between females (2,027 cases) and males (1,923 cases), the upper extremities emerge as a critical site for monitoring lesions in both genders.
- Head/Neck Region: Notably, head/neck lesions are more common in females (241 cases) than in males (152 cases). This could be linked to gender-specific factors such as cosmetic use or sun protection practices, highlighting the need for gender-sensitive awareness campaigns.
- Palms/Soles and Oral/Genital Areas: These regions, while less frequent across both genders, underscore the necessity for thorough and inclusive skin examinations that cover even rare anatomical sites.

The gender-based insights emphasize the need for personalized diagnostic and prevention strategies that cater to the specific lesion distribution patterns observed in males and females. [15]

4.6. ISIC 2024 Dataset

- **Full Name:** ISIC 2024: Skin Lesion Segmentation Challenge (SLICE-3D)
- **Description:** The ISIC 2024 dataset, also known as SLICE-3D, is a novel dataset that focuses on 3D dermatoscopic imaging for skin lesion analysis. It provides dermoscopic images that are used for advanced lesion segmentation and classification tasks. This dataset represents the cutting edge of skin lesion analysis, emphasizing volumetric imaging techniques.
- **Lesion Types:** The dataset includes both benign and malignant lesions, primarily focusing on

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2781- 2795
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0385

melanoma and keratosis, but with greater emphasis on lesion structure through 3D imaging.

- **Sample Size:** As the ISIC 2024 dataset is relatively new and focused on 3D segmentation, the number of samples is still growing, but early reports suggest over 10,000 3D dermoscopic images. Figure 10 shows Sample Images (ISIC 2024)

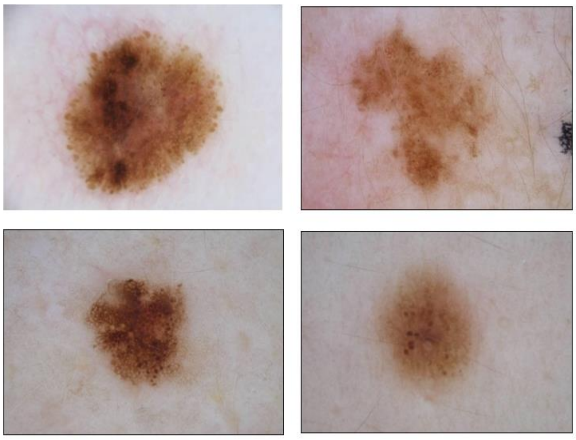

Figure 10 **Sample Images (ISIC 2024)**

4.7. Visual Insights from ISIC 2024 Dataset

4.7.1. Age Distribution

- The age distribution in ISIC 2024 follows a bimodal pattern, with notable peaks at age 20 and age 50, suggesting that both younger adults and middle-aged individuals are particularly affected.
- The mean age is 47.8 years, while the median age is 50.0 years, indicating a slight skew toward middle age.
- This pattern highlights the necessity of age-targeted screening and awareness programs, particularly focusing on both younger adults and individuals approaching middle age, who may otherwise overlook skin health concerns.

4.7.2. Lesion Location Distribution

- The lower extremities (19.8%) and torso (anterior torso at 17.1% and posterior torso at 16.7%) dominate as the most common lesion sites. This concentration in the trunk and extremities underscores the importance of comprehensive skin examinations, especially in these regions
- Upper extremities account for 12.3% of lesions, while head/neck regions contribute to 10.8%, indicating their relatively high exposure to lesions. Figure 11 shows Age Distribution (ISIC 2024)
- Less common areas include palms/soles, oral/genital regions, and lateral torso, all contributing under 1% to the total, suggesting a lesser need for routine examinations in these areas, though they should not be entirely neglected. Figure 12 shows Lesion Location Distribution (ISIC 2024)
- These findings stress the importance of prioritizing frequent examination of the lower extremities and torso in clinical practice, particularly in high-risk patients.

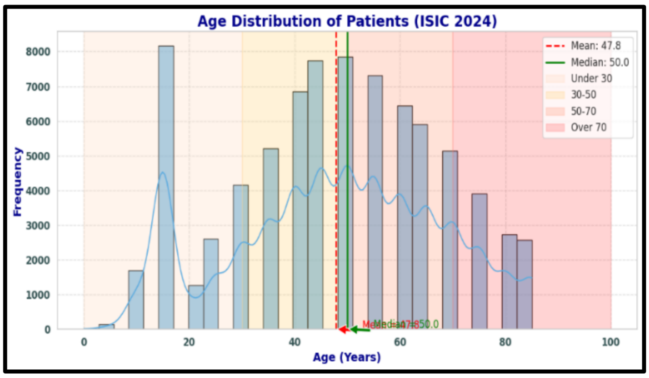

Figure 11 **Age Distribution (ISIC 2024)**

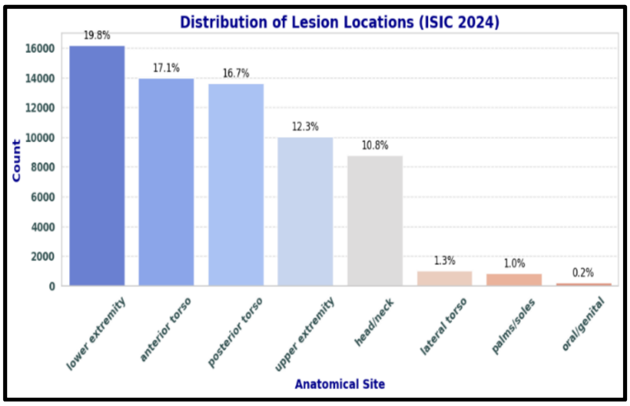

Figure 12 **Lesion Location Distribution (ISIC 2024)**

- Upper extremities account for 12.3% of lesions, while head/neck regions contribute to 10.8%,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2781- 2795
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0385

indicating their relatively high exposure to lesions.

- Less common areas include palms/soles, oral/genital regions, and lateral torso, all contributing under 1% to the total, suggesting a lesser need for routine examinations in these areas, though they should not be entirely neglected. Figure 13 shows Anatomical Sites by Gender
- (ISIC 2024)
- These findings stress the importance of prioritizing frequent examination of the lower extremities and torso in clinical practice, particularly in high-risk patients.

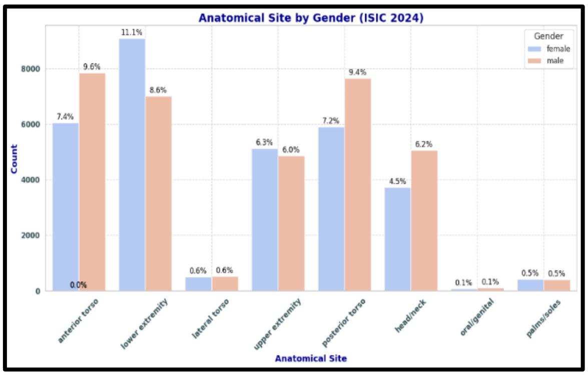

Figure 13 Anatomical Sites by Gender (ISIC 2024)

- Lower extremities emerge as the most common lesion site across genders, with females (11.1%) showing a slightly higher prevalence compared to males (8.6%), indicating the potential need for gender-specific awareness campaigns focusing on areas where women are more affected.
- For males, anterior torso (9.6%) and head/neck (9.4%) areas stand out, suggesting a more frequent occurrence of lesions in these regions.
- The distribution emphasizes the necessity for gender-specific diagnostic protocols, where screenings for men should focus more on the torso and head/neck, and for women, on the lower extremities.
- These anatomical differences reflect potential lifestyle or biological factors that may contribute to the gender disparity in lesion occurrences, pointing toward a more customized approach in skin cancer diagnostics.

4.7.3. Lesion Type by Gender

- Nevus is the most prevalent lesion type across genders, affecting 21.1% of males and 17.8% of females, marking it as the most common type of skin lesion. The higher rate in males suggests the need for more vigilant screening practices for men. Figure 14 shows Lesion Type by Gender (ISIC 2024)
- Melanoma, the most dangerous form of skin cancer, shows a higher prevalence in males (5.0%) than in females (3.6%), underlining the need for more aggressive early detection efforts and preventative strategies for men.

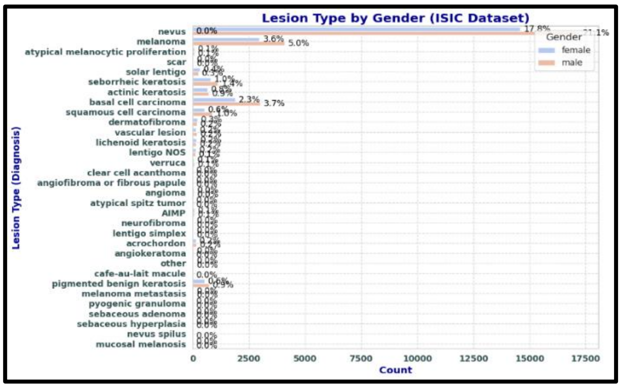

Figure 14 Lesion Type by Gender (ISIC 2024)

- Other types like basal cell carcinoma and seborrheic keratosis also exhibit higher rates in males, further emphasizing the gender-based disparity in lesion types. Figure 14 shows Lesion Type by Gender (ISIC 2024)
- These findings are crucial for designing gender-focused preventive measures and public health campaigns, especially given the higher incidence of aggressive skin cancer types in men.

5. Implementation

The implementation of the skin cancer classification project is divided into multiple phases, each addressing specific challenges associated with dataset processing, model training, and evaluation.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2781- 2795
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0385

Below, we provide a detailed account of each phase, highlighting the methodologies employed and their underlying motivations.

5.1. Dataset Preparation

The dataset used for this study is the ISIC Dataset, comprising dermoscopic images of various skin lesions. Given the inherent challenges of medical imaging, such as class imbalance and variability in image characteristics, several steps were undertaken to prepare the dataset for training:

- Raw Dataset Analysis: The dataset was analyzed to understand the distribution of classes, revealing significant imbalance among categories.
- Data Augmentation: To enhance model generalization, augmentation techniques like random flipping, rotation, and zooming were applied. These transformations were implemented using TensorFlow's RandomFlip, RandomRotation, and RandomZoom layers within a sequential augmentation pipeline.
- Class Balancing: To mitigate the effects of class imbalance, oversampling techniques were applied using Augmentor. Synthetic images were generated for minority classes, ensuring a more balanced representation in the training dataset.

5.2. Model Development

Three different Convolutional Neural Network (CNN) models were developed to evaluate the impact of augmentation and class balancing on classification performance. The models were trained using TensorFlow/Keras, and each variation was designed to build upon the insights from its predecessor:

- **Baseline CNN:**The baseline model serves as a foundational approach, with a straightforward CNN architecture. It includes a series of convolutional layers followed by max-pooling layers to extract features, and dense layers for classification. The network concludes with a softmax activation function for multi-class classification. The Adam optimizer (learning rate = 0.001) and Sparse Categorical Crossentropy loss function were used, with accuracy as the primary evaluation metric.
- **CNN with Data Augmentation:** The second model builds upon the baseline by incorporating data augmentation techniques. These augmentations enhance generalization by introducing variability in the training data, likely through transformations such as horizontal flipping, random rotations, or zooming.
- **CNN with Augmentation And Class Balancing:** The third model addresses class imbalance by employing advanced augmentation strategies. Techniques such as oversampling minority classes or applying class-weighted loss functions were utilized to balance the dataset, improving the network's sensitivity to underrepresented classes.

5.3. Training and Evaluation

Each model was trained using the following settings:

- Optimizer: Adam optimizer with a learning rate of 0.001.
- Loss Function: Sparse Categorical Cross-entropy, as the target labels were integer-encoded.
- Batch Size: 32.
- Epochs: 25.

The evaluation metrics included accuracy and confusion matrices, with a focus on identifying improvements in classification for minority classes. Training and validation accuracies were monitored across epochs to identify overfitting or underfitting.

5.4. Results and Findings

The results demonstrated a clear progression in model performance:

- **Baseline CNN:** The baseline CNN achieved moderate training and validation accuracy, serving as a reference point for evaluating the impact of data augmentation and class imbalance techniques. While the training accuracy steadily improved, the validation accuracy plateaued early, indicating potential overfitting. This suggests the need for techniques to enhance the model's generalization capabilities. Figure 15 Accuracy Graph for Baseline Model

IRJAEH
International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2781- 2795
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0385

Incorporating data augmentation led to noticeable improvements in model performance. The transformations introduced variations in the training data, preventing overfitting and enhancing generalization to unseen data. The validation accuracy increased compared to the baseline model, demonstrating the effectiveness of this approach.

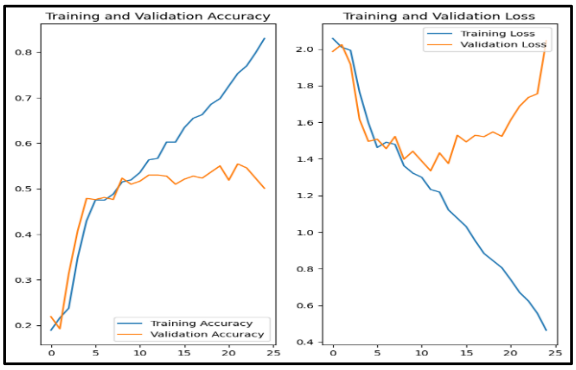

Figure 15 Accuracy Graph for Baseline Model

Table 1 Findings of Model

Model	Training Accuracy	Validation Accuracy	Observation
Baseline CNN	70%	45%	Model is overfitting
CNN with Data Augmentation	60%	50%	Overfitting addressed but at the cost of accuracy
Advanced Augmentation CNN	90%	80%	Class Balanced, Mitigated Overfitting

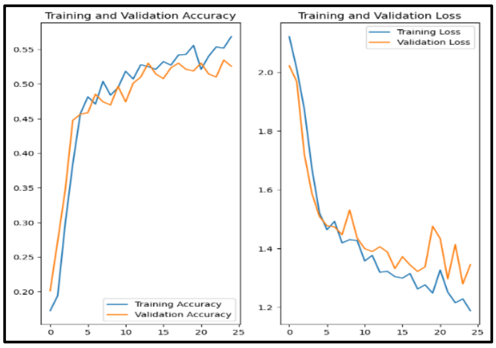

Figure 16 Accuracy Graph for Augmented Model

- **CNN with Advanced Augmentation:** The third model further addressed the class imbalance issue, resulting in enhanced performance on underrepresented classes. This improvement is reflected in the validation accuracy, which not only increased but also stabilized across epochs. Advanced augmentation techniques and class-weighted loss functions contributed significantly to this progress. Figure 16 Accuracy Graph for Augmented Model

5.5. Key Implementation Challenges

- **Class Imbalance:** The ISIC dataset exhibited significant class imbalance, which could lead to biased model predictions. This was addressed through targeted oversampling and evaluation strategies.
- **Computational Overhead:** The generation of augmented and balanced datasets increased processing requirements. Strategies like caching and prefetching were implemented to optimize pipeline efficiency.
- **Evaluation Metrics:** Beyond accuracy, metrics like precision, recall, and F1-score were crucial to assess the model's utility in real-world applications.

5.6. Software and Tools

The following tools were employed during implementation

- **Programming Language:** Python.
- **Libraries:** TensorFlow, Keras, Matplotlib, Seaborn, and Augmentor.
- **Platform:** Google Colab, leveraging GPU acceleration for faster training. Figure 17 Accuracy Graph for Advanced Augmentation

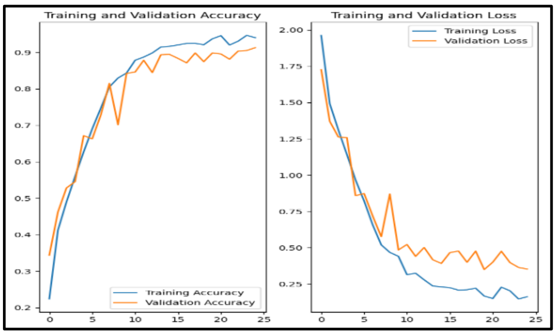

Figure 17 Accuracy Graph for Advanced Augmentation

6. Recent Advances and Future Directions

The field of skin cancer detection has experienced remarkable advancements in recent years, largely driven by the rapid evolution of artificial intelligence (AI) and machine learning (ML) techniques, particularly deep learning models. Early approaches in skin cancer classification relied heavily on handcrafted features and traditional machine learning algorithms, but with the advent of convolutional neural networks (CNNs), the accuracy and reliability of skin cancer diagnosis have seen dramatic improvements. These deep learning models have excelled in handling large and complex datasets such as those provided by the International Skin Imaging Collaboration (ISIC), which contains millions of dermoscopic images annotated with detailed lesion classifications. A significant recent development is the adoption of transfer learning. By using pre-trained models like ResNet, Inception, and VGG, researchers have been able to overcome the challenge of insufficient labeled data in medical image classification. Transfer learning allows models to leverage knowledge gained from large-scale image datasets and fine-tune them for skin cancer detection, thereby improving both accuracy and generalization with fewer resources. This has enabled skin cancer detection systems to achieve near-expert performance, with some models even outperforming dermatologists in specific tasks. Data augmentation has also played a crucial role in recent advancements, particularly when tackling class imbalance in datasets. By artificially increasing the diversity of the training data through transformations such as rotation, scaling, and flipping, models become more robust and less prone to overfitting. Furthermore, techniques like generative adversarial networks (GANs) are being explored to generate synthetic images, addressing both class imbalance and the need for diverse datasets. This approach holds great promise in augmenting existing datasets, particularly in regions where annotated medical images are scarce. Despite these advancements, challenges still persist, particularly in terms of model interpretability. While deep learning models have shown great accuracy, understanding how these models arrive at their predictions remains a hurdle. As AI systems are being integrated into clinical workflows, it is imperative that these models are not only accurate but also transparent and explainable. Recent efforts in model interpretability, such as the use of Grad-CAM(Gradient-weighted Class Activation Mapping), allow clinicians to visualize the areas of an image that influenced a model's decision, fostering trust in AI-driven diagnoses. Looking ahead, several key areas are expected to shape the future of skin cancer detection: Dataset diversity and quality: While current datasets have significantly contributed to model training, expanding datasets to include more diverse populations—considering factors like skin color, age, and geographical location—will be critical for developing globally applicable models. Integration of multimodal data: Incorporating additional clinical data such as patient history, medical imaging (like dermoscopy and confocal microscopy), and even genetic data could enhance model performance, enabling more accurate and personalized predictions. Real-time deployment and clinical adoption: For AI-based skin cancer detection systems to have a real-world impact, they must be deployed in clinical settings with real-time analysis capabilities. This requires models to be efficient, lightweight, and capable of providing results in a short time frame, ideally supporting clinicians during patient consultations. Regulatory approval and ethical concerns: As AI tools are moving closer to real-world deployment, obtaining regulatory approval from medical authorities like the FDA (Food and Drug Administration) will be vital. In parallel, ensuring the ethical use of AI in healthcare, including privacy concerns and ensuring bias-free predictions, will be essential for widespread adoption. Collaboration across disciplines: The future of skin cancer detection will not be shaped by technology alone but by collaborative efforts between AI researchers, dermatologists, healthcare providers, and regulatory bodies. Multi-disciplinary collaboration is necessary to ensure that AI tools are used responsibly, effectively, and safely in clinical practice. In conclusion, while AI and ML have already made significant strides in skin cancer classification, the future holds immense potential. As technological advancements continue to improve

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2024
Page No: 2781- 2795
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0385

model performance, enhance interpretability, and enable clinical integration, AI will play an increasingly central role in early detection and diagnosis, ultimately leading to better patient outcomes and a reduction in the global burden of skin cancer.

Conclusion

In this review, we have explored the significant advancements in skin cancer detection using dermoscopic images and the role of machine learning and artificial intelligence in improving diagnostic accuracy. Over the past few years, the integration of deep learning techniques, particularly convolutional neural networks (CNNs), has revolutionized the field, enabling more precise and automated detection of skin cancer lesions. With advancements in dataset quality, model architecture, and data augmentation, these AI-driven approaches have shown remarkable promise in supporting dermatologists and healthcare providers in diagnosing skin cancer at an early stage. One of the key strengths of AI-based systems in skin cancer detection lies in their ability to handle large and complex datasets, such as the ISIC dataset, which includes thousands of labeled dermoscopic images. With the advent of transfer learning, deep learning models can be trained effectively on smaller datasets by leveraging pre-trained networks, thus reducing the need for extensive labeled data while still achieving high accuracy. Furthermore, the application of data augmentation techniques has been instrumental in addressing class imbalance, which has traditionally been a challenge in skin cancer detection datasets. By artificially increasing the size and diversity of training sets, these methods help to ensure that the models generalize well across different skin types and lesion characteristics. However, despite the substantial progress made, several challenges remain. While deep learning models have proven to be highly accurate, there is still a need for greater model interpretability and transparency. Understanding how these models make decisions is essential for gaining trust from clinicians and patients alike. As AI-based systems move closer to clinical implementation, efforts to improve explainability and ensure ethical considerations in healthcare will be crucial for their widespread adoption.Looking ahead, the future of skin cancer detection lies in multi-modal approaches that integrate various data sources, such as patient medical history, genetic data, and imaging from different modalities. These innovations, along with ongoing advancements in AI model efficiency, will drive the next wave of improvements in diagnostic systems, offering more accurate, personalized, and accessible tools for skin cancer detection. In conclusion, AI and machine learning have the potential to transform the landscape of skin cancer diagnosis, making it more accessible, accurate, and efficient. The collaboration between researchers, clinicians, and regulators will be fundamental in overcoming current challenges and ensuring the ethical deployment of AI tools in healthcare. As technology evolves and datasets become more diverse, AI-driven skin cancer detection systems are poised to become an integral part of clinical practice, improving patient outcomes and helping to reduce the global burden of skin cancer.

References

[1]. Kavita Behara, Ernest Bhero, and John Terhile Agee, Grid-Based Structural and Dimensional Skin Cancer Classification, 2024
[2]. Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, Sebastian Thrun6, Dermatologist–level classification of skin cancer with deep neural networks, 2021
[3]. Bernardo Petracchi, Emanuele Torti, Elisa Marenzi and Francesco Leporati, Acceleration of Hyperspectral Skin Cancer Image Classification through Parallel Machine-Learning Methods, 2023
[4]. Bogdan Mazoure, Alexander Mazoure, Jocelyn Bédard & Vladimir Makarenkov, DUNEScan: a web server for uncertainty estimation in skin cancer detection with deep neural networks, 2022
[5]. Juan M. Gálvez , Daniel Castillo-Secilla , Luis J. Herrera , Olga Valenzuela , Octavio Caba ,José C. Prados , Francisco M. Ortuño , and Ignacio Rojas, Towards Improving Skin Cancer Diagnosis by Integrating Microarray and RNA-Seq Datasets, 2020

[6]. Du-Harpur, F.M. Watt, N.M. Luscombe and M.D.Lynch, What is AI? Applications of artificial intelligence to dermatology, 2020

[7]. Philipp Tschandl, Christoph Rinner, Zoe Apalla Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek and Harald Kittler, Human–computer collaboration for skin cancer recognition, 2020

[8]. Navneet Melarkode, Kathiravan Srinivasan, Saeed Mian Qaisar and Pawel Plawiak, AI-Powered Diagnosis of Skin Cancer: A Contemporary Review, Open Challenges and Future Research Directions, 2023

[9]. Yung-Tsan Jou,Yih-Chuan Wu, Wen-Tsann Lin, Applying Decision Tree and Neural Network to Raise the Performance of Human Training Quality, 2015

[10]. Chiou, A. S., Omiye, J. A., Gui, H., Swetter, S. M., Ko, J. M., Gastman, B., ... & Novoa, R. (2024). Multimodal Image Dataset for AI-based Skin Cancer (MIDAS) Benchmarking. medRxiv, 2024-06.

[11]. Soniya Mane, Dr. Swati Shinde, A Method for Melanoma Skin Cancer Detection Using Dermoscopy Images, 2018