ISSN 2395-8618

Dermoscopic Skin Lesion Classification Using Artificial Intelligence

Muhammad Zahid, Tayyaba Anees*, Kinza Sardar, Abdullah

***Abstract*—Skin cancer remains one of the most prevalent malignancies worldwide, with melanoma accounting for the majority of related deaths. Early and accurate detection is critical for improving patient outcomes, yet traditional diagnostic methods are often subjective and resource-intensive. This study investigates the efficacy of deep learning-based binary classification of dermoscopic skin lesions into benign and malignant categories using the publicly available HAM10000 dataset. Malignant cases were defined to include melanoma and basal cell carcinoma, while all other lesion types were grouped as benign. We systematically compared two convolutional neural network (CNN) architectures, EfficientNetB7 and VGG19, under identical preprocessing conditions, including resizing, normalization, noise reduction, and artifact removal (e.g., hair inpainting). Model performance was evaluated using accuracy, loss, and validation metrics. Experimental results demonstrated that while EfficientNetB7 achieved marginally higher training accuracy (87.04%), VGG19 exhibited superior generalization to unseen data, attaining consistent training and validation accuracy (86.44%) with significantly lower and balanced loss values (0.3972), indicating reduced overfitting. These findings suggest that VGG19 offers greater robustness and reliability for real-world clinical applications. Future work will focus on optimizing transfer learning, advanced data augmentation, and regularization techniques to further enhance model generalizability and interpretability.**

***Index Terms*—Skin cancer detection, melanoma classification, dermoscopic images, convolutional neural networks (CNNs), HAM10000 dataset, EfficientNetB7, VGG19, binary classification, deep learning, medical image analysis.**

I. Introduction

The detection of melanoma, which is the most aggressive type of skin cancer, remains an enduring challenge in the medical field. While early detection can significantly improve patient outcomes and survival rates, most traditional diagnostic methods rely on visual examinations and invasive biopsies, which are often subjective, tedious, and resource-intensive [1-5].

This project aimed at harnessing advances in deep learning to create another reliable and non-invasive method for diagnosing melanoma via dermoscopic images. The idea is to incorporate state-of-the-art convolutional neural networks (CNNs) like VGG19 and EfficientNetB7 to enhance the body of knowledge in AI-powered medical diagnostics. It is an application of artificial intelligence in health care, the most important application, potentially to detect melanoma. Importance leans towards application in many domains, such as AI-powered diagnostic tools that assist dermatologists in identifying melanomas, reducing diagnostic errors, and improving efficiency by providing high-confidence predictions as a second opinion. AI-enabled melanoma detection in telemedicine allows remote consultations, overcoming barriers to healthcare access, particularly for patients in underserved or isolated areas [6-10].

Automated melanoma detection systems can be integrated into mass skin cancer screening programs, identifying high-risk individuals where dermatologists are unable to assess each patient individually. We trained models on large annotated datasets that can serve as educational tools for medical students and novice dermatologists, simulating real-world diagnostic scenarios [11].

This study explores the following key research questions:

- How effective are CNN architectures, specifically VGG19 and EfficientNetB7, in distinguishing melanoma from benign lesions in dermoscopic images?
- What preprocessing and augmentation techniques are most effective in enhancing model performance for this dataset?
- How do different evaluation metrics, such as accuracy, F1-score, and ROC-AUC, inform the reliability of the models?
- What strategies can be implemented to improve the interpretability and explainability of AI-based melanoma detection systems?
- How can the trained models be adapted for integration into practical healthcare applications?

Skin cancer is the most common cancer on earth, and melanoma accounts for most deaths due to this type of cancer. Though it has curative potential if detected in time, making a diagnosis is tough because of visual similarity to benign lesions. Such problems are exacerbated by poor access to dermatology in less resourceful areas. In this project, the challenges above can be addressed using a deep learning-based system for automated melanoma detection that will be accurate, scalable, and interpretable; this can save lives being lost to late detection.

Manuscript received on 22/12/2024, accepted for publication on 17/05/2025. Corresponding author is Tayyaba Anees.

Muhammad Zahid is with SolutionInn, 1800 Pembrook Dr Suite 300, Orlando, FL 32810, United States (e-mail: zahidmetha@solutioninn.com).

Tayyaba Anees and Kinza Sardar are with the Department of Software Engineering, School of Systems and Technology (SST), University of Management and Technology, Lahore 54770, Pakistan (e-mail: tayyaba.anees@umt.edu.pk; kinza.sardar@umt.edu.pk).

Abdullah is with the Department of Computer Science, Bahria University Lahore Campus, Lahore 54600, Pakistan (e-mail: Abdullah.bulc@bahria.edu.pk).

ISSN 2395-8618

TABLE I
SUMMARY OF RESEARCH PAPERS.

Paper	Work Done	Data Used	Methods Used	Results and Conclusions
[19]	Investigated ensemble techniques for melanoma detection.	ISIC 2017 Challenge dataset	Ensemble of CNNs and feature-based machine learning models, with data augmentation.	Achieved ROC-AUC of 0.874, demonstrating the potential of ensemble methods in improving diagnostic accuracy.
[20]	Compared models with dermatologists' diagnostic accuracy.	HAM10000 dataset	Transfer learning with ResNet50 and InceptionV3 architectures.	Models achieved 91% accuracy, outperforming dermatologists in melanoma detection.
[21]	Developed an interpretable deep learning model with visual explanations for melanoma detection.	The HAM10000 Classification dataset	Modified EfficientNetB0 with Grad-CAM for interpretability.	Achieved an F1-score of 0.85 and emphasized the role of model interpretability for clinical adoption.
[22]	Explored self-supervised learning techniques for melanoma detection.	Synthetic datasets generated with variations.	Rotation prediction and corruption removal-based pretext tasks.	Self-supervised models showed accuracy improvements, highlighting potential for scalability with extended training.
[23]	Proposed a lesion-focused convolutional technique to enhance model generalizability.	ISIC 2020 dataset	Focused convolution layers emphasizing lesion-specific features.	Increased mean AUC from 0.9 to 0.922, demonstrating improved diagnostic accuracy.
[24]	Introduced a unified lightweight melanoma classification model (Mela-D).	Multiple public dermoscopic datasets	Lightweight CNN architecture suitable for real-time applications.	Achieved 88.8% accuracy and significantly faster runtime, enhancing feasibility for consumer-level hardware.
[25]	Compared the AWE ensemble algorithm performance against dermatologists in a multicenter study.	Clinical dermoscopic datasets	Ensemble model combining predictions from multiple CNNs.	AWE showed higher balanced accuracy and sensitivity, indicating its support potential for clinicians in challenging cases.
[26]	Developed a deep learning model for melanoma detection.	Dermoscopic image dataset	CNN architecture optimized for binary classification tasks.	Achieved sensitivity of 87.4% and specificity of 95.1%, demonstrating strong diagnostic performance.

This research holds significant implications for the medical field, particularly in reducing the burden on dermatologists by providing an AI-driven diagnostic tool. Expanding access to skin cancer detection for underserved populations, particularly in rural or remote regions.

Supporting early detection efforts, which are critical for reducing melanoma mortality rates. Advancing the state of research in medical AWE by providing insights into the application of deep learning to skin cancer detection.

Binary classification (melanoma versus non-melanoma) using dermoscopic images from the HAM10000 dataset [12] is the focus of this project.

However, while the dataset offers a wide variety of annotated images, its real-world applicability might need further validation across different demographics and types of image acquisition.

Additionally, the models are designed to serve as diagnostic aids and not to replace medical judgment from professionals.

II. LITERATURE REVIEW

Detection of melanoma using artificial intelligence is of interest in recent studies because of its potential to change cancer diagnosis. The increased availability of abundantly annotated data sets and advances in deep learning architecture mark this site, which has witnessed significant progress in recent years. This work reviews existing studies that use technologies to detect skin cancer, specifically melanoma, and critically examines their contributions. The goal is to explore the strengths and weaknesses of previous research, identify the gap, and justify the methodology adopted for the present. Literature on this WORK is selected keeping in mind the applicability to melamine detection through AI. The selection is also limited to peer-reviewed journals and conference papers of the past 10 years to obtain the newer advancements.

The study should necessarily be related to melanoma detection; only those studies have been included that specifically talk of melanoma diagnosis or classification. Application of Deep Learning Architectures. Only actual applications like CNNs, transfer learning, or ensemble techniques would be studied in most cases from papers. Application of Relevant Dataset Preference would be given to studies in this regard that would use dermoscopic image data sets that are more representative of the HAM10000 Classification data. Evaluation Metrics include studies reporting a lot of evaluation metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC. Applicability to Clinical Settings: Preferably, studies examining interpretability or real-world implementation of models.

Advanced developments in deep learning have provided many improvements in the detection of melanoma through sophisticated models and methodologies. Such as a study initiated by [13] using an automatic melanoma detection based on deep learning through the ISIC 2017 Challenge dataset. Their method combined traditional machine learning algorithms with convolutional neural networks (CNNs), thereby achieving a high ROC-AUC score of 0.874 that

ISSN 2395-8618

revealed the potential of ensemble techniques in improving diagnosis accuracy. Similar to this, [14] evaluated the specificity and sensitivity of human experts as compared to models such as ResNet50 and InceptionV3 in the detection of melanoma with the use of the HAM10000 dataset. The models outperformed dermatologists, gaining an accuracy of 91% in detecting melanoma, which is indeed a testament to the power of transfer learning for skin cancer classification.

The authors propose an interpretable deep learning model for melanoma detection that emphasizes its explainability alongside accuracy [15]. This very same dataset that provides the HAM10000 Classification was used to implement a modified EfficientNetB0 with Grad-CAM for visual explanation, which indeed came out with an F1-score of 0.85. Again, the work strongly advocates for interpretability to facilitate clinical adoption.

The most recent study to date is with [16], wherein self-supervised neural networks were studied for melanoma diagnosis. He had rotation prediction and corruption removal as the different techniques he employed in evaluating self-supervision for encouraging model accuracy, with suggestions for improving it via longer training and expansion of the dataset.

How deep learning models can be generalized [17], they present a lesion-oriented convolutional procedure, improving the diagnostic accuracy by means of the ISIC 2020 dataset from a mean AUC score of about 0.9 up to 0.922. Focusing on lesion-specific features can improve the performance of a model by increasing diagnostic confidence.

Next, Kim et al. are coming forward with an integrated classification scheme of melanoma whose support comprises a variety of databases and architectures. Their very compact model, Mela-D, produced similar predictive performance as ResNet50 (with an accuracy of 88.8%) at speeds 33 times faster, enabling accurate yet facile melanoma detection on consumer-class desktops. The use of dermoscopic images in a deep learning model for melanoma detection is reported in a study by [18], which shows a sensitivity of 87.4% and specificity of 95.1% in the high-performing diagnostic potential and further applicability in the clinical field, as shown in Table 1.

These studies collectively underscore the significant progress in AI-driven melanoma detection, highlighting the importance of model accuracy, interpretability, and real-world applicability in developing effective diagnostic tools [27].

III. Methodology

The technical and practical activities carried out in the development and implementation of this melanoma detection system have been described in this. Every coding, experimentation, and optimization activity was performed by me. We concentrated on the design and evaluation of deep learning models to achieve a meaningful identity for melanoma diagnosis. Below is a breakdown of the dataset, data preparation, models, metrics, techniques, and optimizations employed in the project. Data Preparation Another major step of this project was to prepare the actual dataset for efficient and effective training and testing purposes of the deep learning models. The dataset, which is the HAM10000 Classification

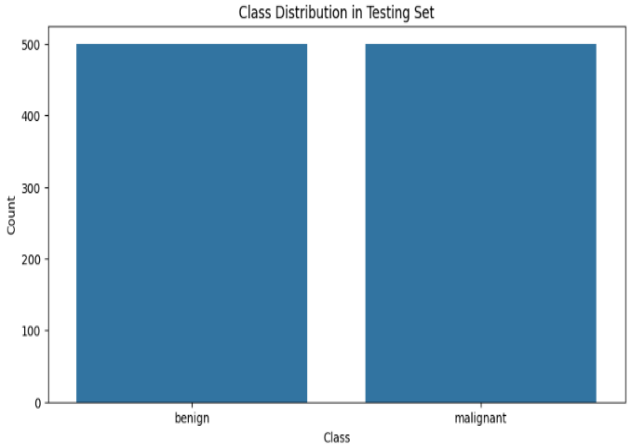

Fig. 1. Class distribution of the training and testing set.

Dataset, consists, in general, of high-resolution dermoscopic images that are heterogeneous among themselves in size and format. WE perform an image resizing technique for standardizing the input sizes for such neural networks. Image Resizing: All images in the dataset were resized to 224x224 pixels, aligning with the input requirements of models like VGG19 and EfficientNetB7. This standardization, achieved using OpenCV, ensured uniformity in the feature extraction process while reducing computational load. The images were also converted to the BGR color channel format, optimizing them for input. Bilinear interpolation was used during resizing to minimize distortion and maintain quality.

Apart from other things, the exploratory data analysis (EDA) has a major part in the project, as it gives an open door to understand the dataset, to discover likely problems of quality, and to decide the preprocessing and selection of the model steps. The dataset is named after the HAM10000 Classification dataset, as it includes labeled dermoscopy images or images about dermoscopic lesions, along with other types of metadata, such as patient age and gender, location of the lesions.

This study employs the HAM10000 dataset (Human Agains t Machine with 10,000 training images), a commonly utilized benchmark dataset for skin lesion analysis. This study reformulates the original dataset into a binary classification problem to simplify the prediction task and improve clinical interpretability.

Specifically, the dataset is divided into two classes: Class 0, representing non-cancerous lesions (including melanocytic nevi, benign keratosis-like lesions, vascular lesions, and dermatofibroma), and Class 1, representing cancerous lesions (melanoma and basal cell carcinoma). This binary grouping aligns with real-world clinical screening practices, where the primary objective is to distinguish between malignant and benign lesions.

As illustrated in Figure 1, the dataset comprises 9,605 images in the training set and 1,000 images in the testing set. Within the training set, there are 5,000 benign and 4,605 malignant images, while the test set is balanced, containing 500 images for each class. Understanding this class distribution is essential for ensuring reliable model training and unbiased evaluation. To facilitate this analysis, the distribution of images across classes in both training and testing sets is visualized using bar plots, providing a clear representation of dataset balance.

ISSN 2395-8618

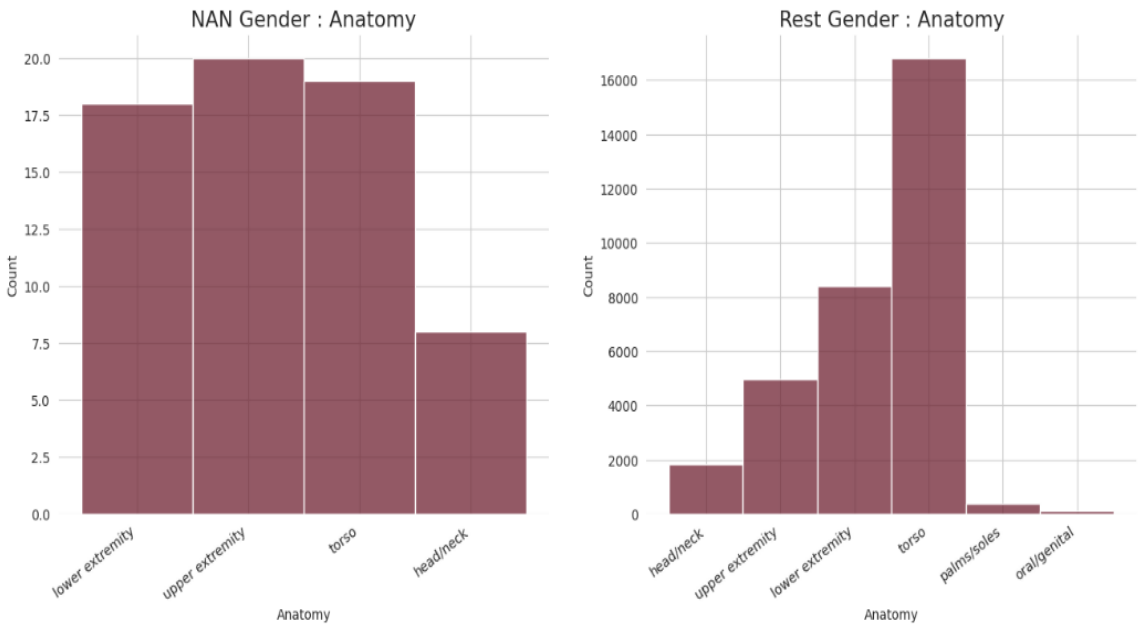

Fig. 2. Analysis of missing gender values.

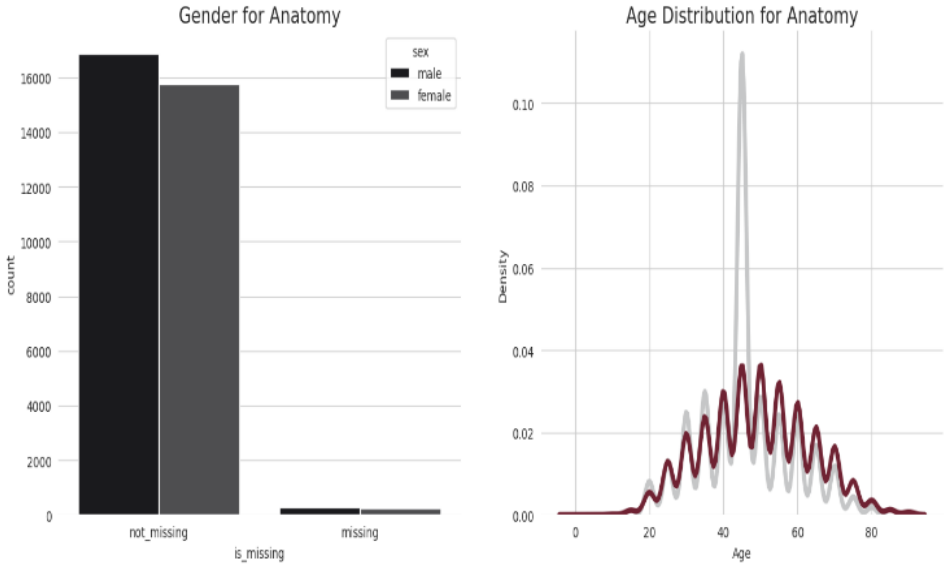

Fig. 3. Missing data for anatomy and age.

A. Missing Gender Values and Distribution Analysis

The feature concerning the sexually described aspect of this data set has missing values, and so has therefore produced two data sets, one with a consuming gender value (nan_sex) and one with a present value, as shown in Figure 2.

Lesional distribution indicated more lesions with absent gender within the torso and limbs, while the available gender showed more generalized, albeit varying distributions. Also, a majority of the identified cases lacking gender properties were benign, with a smaller fraction realized as malignant.

B. Handling Missing Data for Anatomy and Age Features

For missing age values, an age-specific imputation using the median age of male patients with benign lesions from specific sites, such as the lower extremity, upper limb, and torso areas, was used, as shown in Figure 3. This median value, relevant to the context, ensured that the imputed ages were less biased by the data distribution. For missing anatomy events, just one binary indicator column, "is_missing," is created; values are either "missing" or "not_missing." The comparative analysis, therefore, has gender-specific patterns and significant differences in the age distributions between the groups.

C. Analysis of Missing Anatomy Data in Test Dataset

The anatomy feature in the test dataset was analyzed for missing values by introducing a binary indicator column (is_missing). This enabled a clear distinction between records with and without missing anatomical data. Gender Distribution. A count plot revealed that missing anatomy data showed a distinct pattern concerning gender, as shown in Figure 4.

This highlights potential demographic factors influencing the completeness of anatomical data, which may introduce biases if not accounted for. Age Distribution Kernel density estimation

ISSN 2395-8618

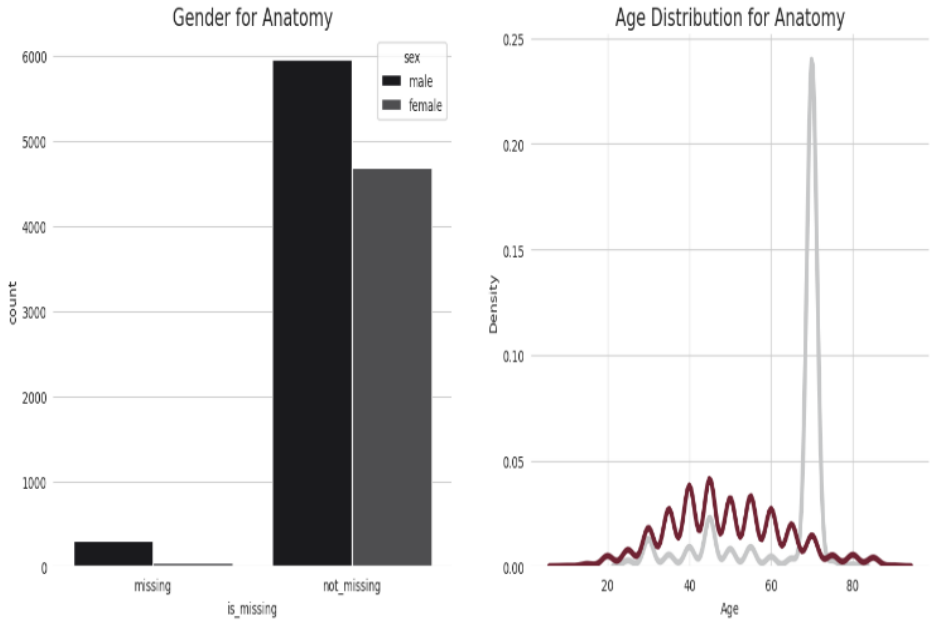

Fig. 4. Gender and age distribution for anatomy.

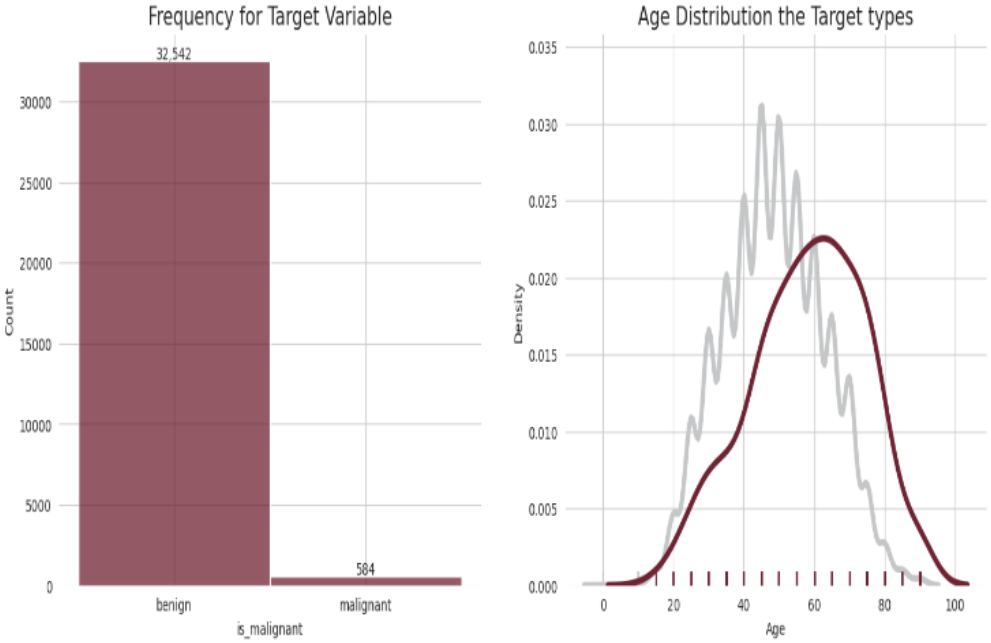

Fig. 5. Frequency and age distribution of the target variables.

(KDE) plots compared the age distributions for records with and without missing anatomy data. The age distribution for records with missing anatomy appeared narrower and skewed compared to those with complete anatomical data, reflecting underlying differences that could impact model generalization.

D. Analysis of Target Variable and Age Distribution

The relationship between the target variable (is_malignant) and the age distribution was explored using visualizations. Frequency Distribution of Target Variable: A histogram depicted the frequency of benign (target=0) and malignant (target=1) cases within the dataset. Annotation of bar heights provided a clear numerical representation of the class imbalance, offering insights into the dataset composition.

E. Age Distribution for Benign and Malignant Cases

A kernel density estimation (KDE) plot compared the age distribution for benign and malignant cases. Benign cases exhibited a broader age range, while malignant cases showed a narrower, skewed distribution.

The overlapping regions in the KDE curves highlighted the challenges of age-based discrimination between the classes, as shown in Figure 5.

These analyses underscore the importance of the Age feature and its potential influence on predictive modeling. They also emphasize the need to address class imbalance to ensure fair and unbiased model performance.

ISSN 2395-8618

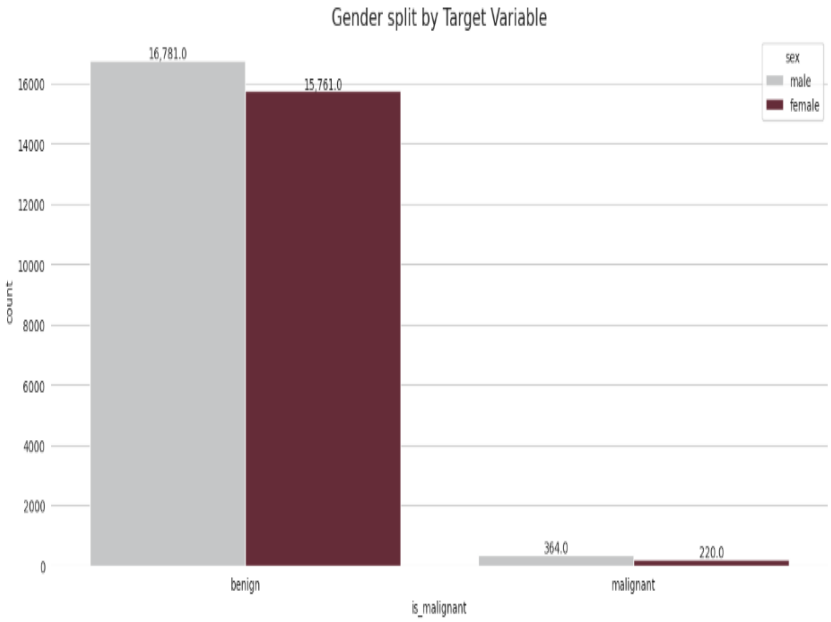

Fig. 1. Gender split by target variable.

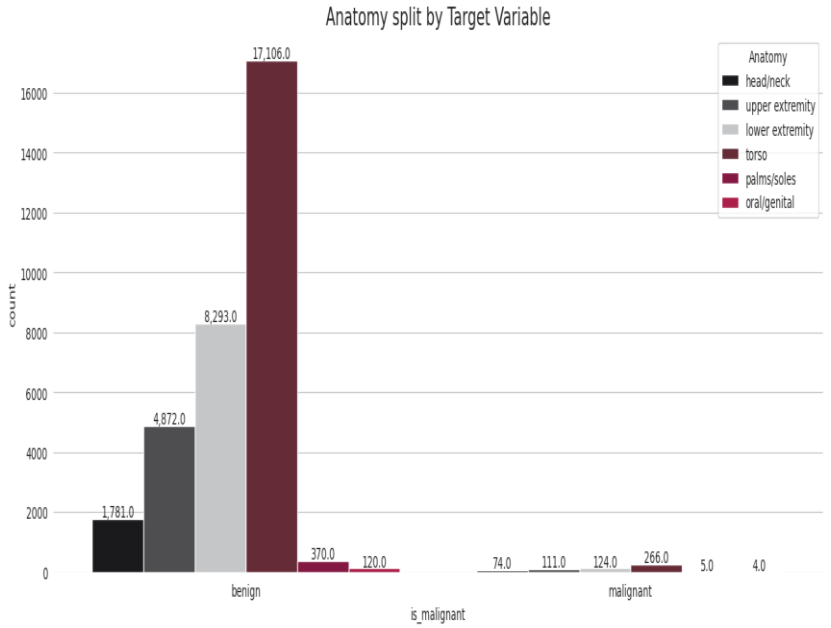

Fig. 2. Anatomy split by target variable.

F. Gender Distribution by Target Variable

It is essential to assess with respect to gender distribution across the target variable (is_malignant) for any possible emerging pattern of benign versus malignant cases. In this attempt, a count plot was used to give a visual breakdown of male and female cases across the target classes.

The annotations on the bars highlighted on the bars showed the number of cases and hence, the existence of pronounced differences in gender distribution for benign (target=0) and malignant (target=1) cases.

They, of course, give important insights into how gender could affect the prevalence of melanoma in the dataset and further influence how the analyst may consider the predictive value of the attribute as a result in Figure 6.

Anatomy Distribution by Target Variable: The anatomical sites' distribution related to the target variable (malignant or benign) was the focus of this visualization. A count plot was used to show the contributions of the anatomical sites for melanoma cases and non-melanoma cases. It indicates that some sites, for example, the torso or lower extremities, occurred more often among benign cases, and others may show a significant presence in malignant cases.

Marks on the bars also give the exact count number, thereby helping in understanding the distribution patterns. Such analysis would greatly help in pointing out the potential biases embedded in the dataset, as it identifies over-representation of certain anatomical sites within specific classes. Identifying these imbalances brings a complete understanding of the dataset and informs how performance evaluation and usage of the model have to be conducted, as Figure 7.

ISSN 2395-8618

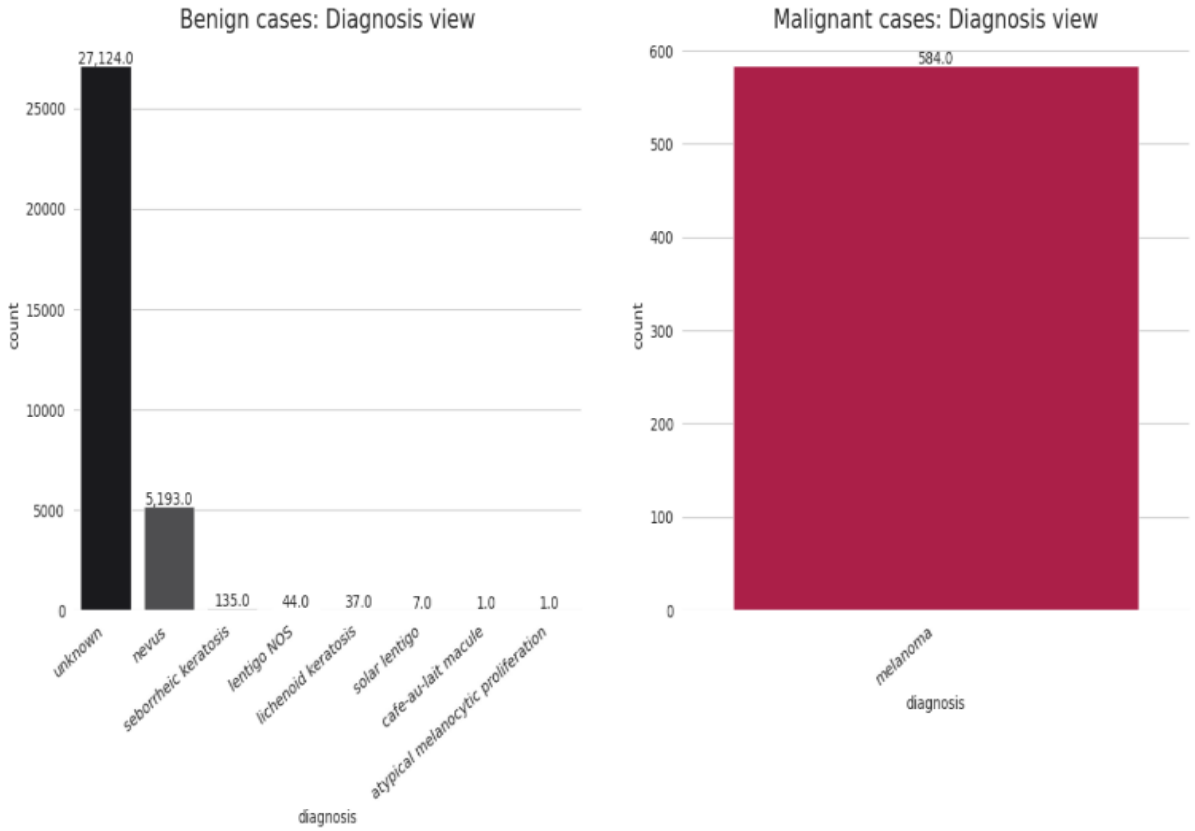

Fig. 3. Benign and malignant cases diagnosis view.

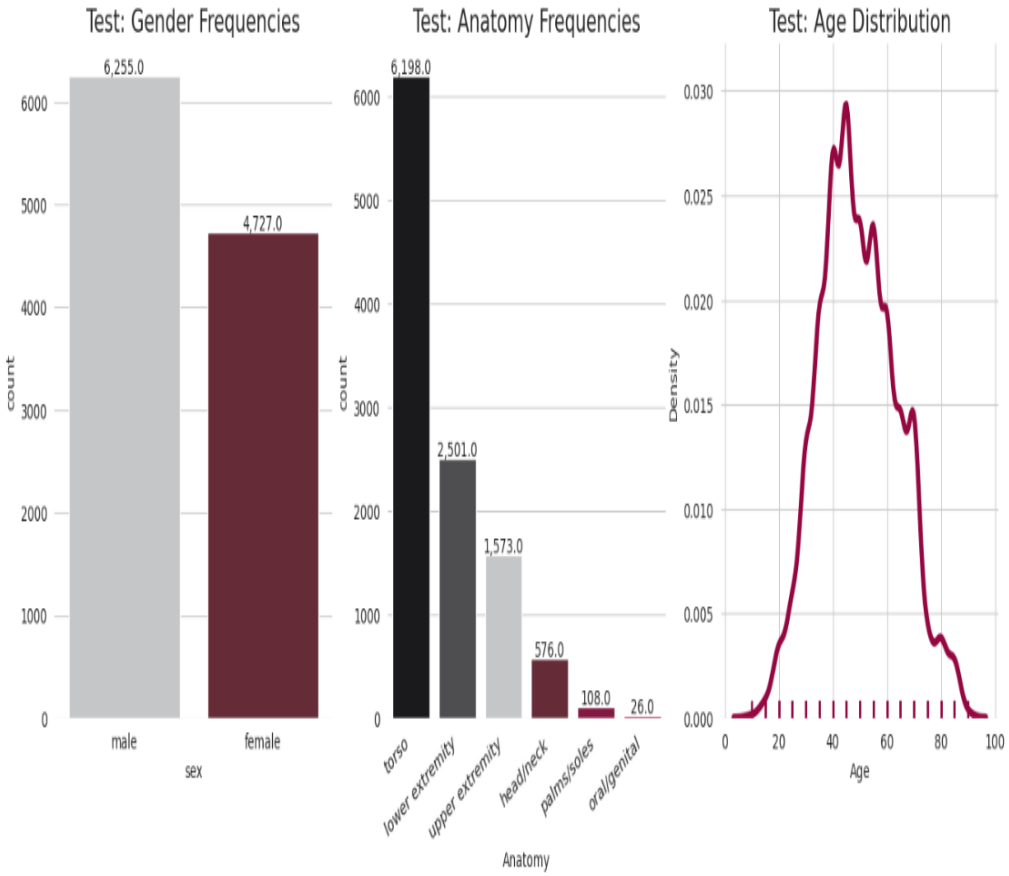

Fig. 4. Test dataset distribution.

ISSN 2395-8618

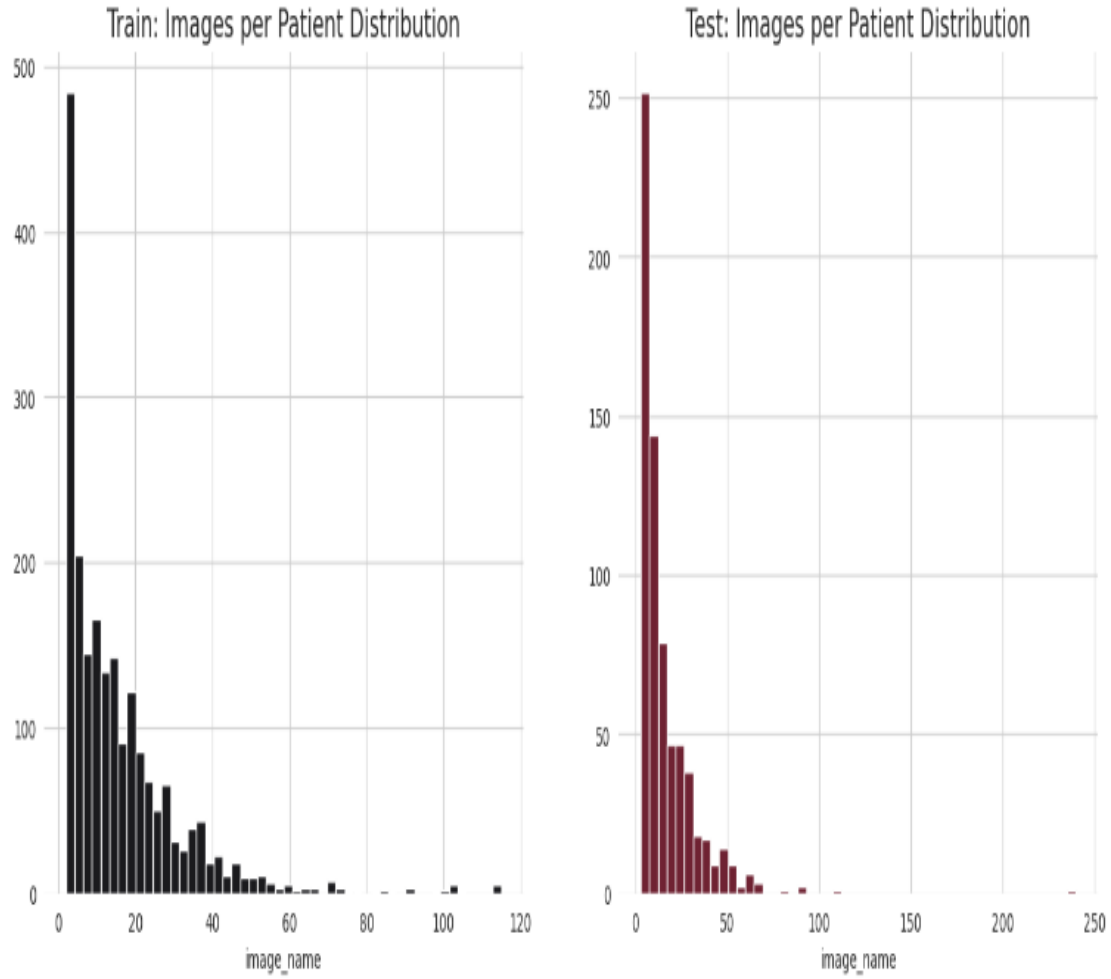

Fig. 5. Train and test images per patient distribution.

G. Diagnosis Distribution by Target Class

A comparison of the distributions of particular diagnoses across benign and malignant cases. The first subplot showed benign cases with a prevalence of specific benign diagnoses, along with patterns or the most dominant categories among them. The second subplot executed a similar exercise with malignant cases, showing the diagnosis types associated with melanoma, as shown in Figure 8.

H. Test Dataset Distribution

The present figure shows the overview of test datasets through three different types of plots. Here, the first plot shows the distribution of genders, meaning the proportion between males and females in the test data. The second plot is all about the distribution of anatomical regions where different frequencies of skin lesions are displayed over several body parts. And the third plot is age distribution across the test dataset, which differentiates age variations as smoothed for an idea of how age is covered through test cases. The count annotations on the first two plots represent each category in an accurate framework.

This figure may come in handy for understanding class imbalances or trends. Smoothed age distribution over kernels indicates a further and continued spectrum of age profiles that might lead to insight into population demographics under consideration for the test group. These representations are very critical for investigating learning and testing data and detecting latent biases that would otherwise influence the model performance, as shown in Figure 9.

I. Image Distribution per Patient

The figure shows the distribution of images per patient in both the training and test datasets. The histograms illustrate the number of images assigned to each patient, with bins set to 50. The training set shows the spread of image counts, revealing any patterns such as variation in the number of images per patient. The test set is compared to assess consistency with the training data. These distributions help guide data processing and model training decisions, as imbalanced distributions may impact model performance, as shown in Figure 10.

J. Gaussian Blur

The images were then turned into grayscale before applying the Gaussian blur technique. This was done to reduce the complexity of the images while defining the grayscale values rather than how many colors were used, diversely affecting their classifications. After applying a grayscale to the image, a Gaussian blur was applied to smooth high-frequency noise

ISSN 2395-8618

while retaining edges and textures important for feature extraction in melanoma detection.

This preprocessing step turned out to be quite effective in improving the signal-to-noise ratio in the images, thus improving the accuracy during feature extraction and classification during training. To further improve the quality of dermoscopic images and increase model accuracy, the image preprocessing technique applied removed unwanted artifacts, such as hair, which tend to interfere with melanoma classification. Hair in dermoscopic images can obscure or confuse important features pertaining to the skin lesions, which could give rise to the possibility of misclassifying them. To solve this problem, a technique for hair removal was developed through morphological operations and inpainting.

K. Hair Removal Technique

The hair removal process involves several steps. Grayscale Conversion: The images are first converted to grayscale, simplifying the data and focusing on intensity variations necessary for skin lesion detection. This preprocessing technique improves image quality by removing irrelevant details, helping the model focus on essential features for melanoma detection and boosting accuracy. Model Selection In this WORK, I'll explain the two models that have been deep-learned for melanoma detection purposes, VGG19 and EfficientNetB7. The reason both of these models were selected was that they were believed to be very efficient at image classification tasks and were also very suitable for the fine-tuning process for the detection analysis of medical images.

L. Used Neural Networks

VGG19 is a deep convolutional neural network that is well-known for its high performance in image classification tasks. It basically has 19 different layers, 16 of which are convolutional layers, and the other three are fully connected layers. VGG19 uses small 3x3 kernels and reduces dimensions of the feature maps by applying max-pooling layers. The architecture of VGG19 is a high-level abstraction and has hierarchical learning features to identify objects in the image, including medical image analysis. Justifications for VGG19: Its simplicity and fantastic performance in ImageNet and other datasets make it a good candidate for the current project, since the pre-trained weights on ImageNet have already allowed me to work with features that were learned and apply them to find melanoma with very little specific tuning. Optimizations. We replaced the last fully connected layers of the model with custom dense layers to adapt the network to melanoma detection into two classes- malignant and benign.

We also fine-tuned the pre-trained layers on the SIIM-ISIC dataset to boost performance on the specific characteristics of dermoscopic images. VGG19 Architecture and Configuration. The main establishment in convolutional neural networks is the VGG19 model. This model has basic architecture properties, but it is too deep. It consists of 16 convolution layers and 3 fully connected networks. This very simple architecture of VGG19 using small 3x3 filters makes it so easy to understand and even fine-tune for specific tasks such as melanoma detection. Here are the parameters we set in configuring this VGG19 model for this project:

- input_shape= (128, 128, 3) The input images were resized to 128x128 pixels to balance computational efficiency with the need to capture fine-grained features in the images.
- classes=2 Since the task at hand is binary classification (malignant or benign), the final layer was configured with two output units.
- classifier_activation="sigmoid" However, a sigmoid activation function is used in the output layer of the classifier. This signifies that the classifier can be used for binary classification, providing a probability score between 0 and 1 for either class.

M. Model Compilation and Optimization

The model was compiled with the following parameters:

- Optimizer 'adam' Adam optimizer was utilized, which has an adaptive learning rate, which helps in quicker convergence of the model.
- Loss function 'sparse_categorical_crossentropy'. It has chosen sparse categorical crossentropy as the dataset consists of categorical labels (malignant and benign) in sparse form (not one-hot encoded).
- Metrics ['accuracy'] Accuracy is used as a metric for evaluation of the model performance, which is assessed by the ratio of correct classification during training and testing.
- All the above information would be included in the first description. Now, components for compiling the model are as follows: 1) Loss function-sparse_categorical_crossentropy,2) Metrics-accuracy.

Post-compilation of the model, a summary was generated concerning the VGG19 architecture. Highlights of the summary included, but were by no means limited to, the number of layers, parameters, and the output shape for each. This summary has helped in understanding the framework of the model and has put lights on the complexity of the network.

N. EfficientNetB7

Thus, for this project, EfficientNetB7 has been preferred since it is one of the best models for efficiency and accuracy in image classification among all other models available. The highest variant of the EfficientNet family, EfficientNetB7, uses a compound scaling method for model depth, width, and resolution to provide an even distribution. This scaling method permits even higher accuracy with fewer parameters than regular deep neural networks. EfficientNetB7 employs a depth-separable convolution scheme that decreases computational cost compared with filter and mapping operations in convolution.

Further, squeeze and excitation blocks calibrate responses to features per channel in the model; thus, the efficiency. EfficientNetB7 gave results with great efficiency and performance upon usage on different benchmarks of image recognition. In terms of object detection and medical imaging, this model provides the best state-of-the-art results in its use for melanoma detection, where great details at high resolution need very fine differences in the indexation. Implementation of the Keras application for the EfficientNetB7 model included configuration as per the following parameters:

ISSN 2395-8618

- input_shape=(128, 128, 3) The input images were resized to 128x128 pixels to optimize computational efficiency while maintaining enough detail for classification. classes=2 Since the task involves binary classification (malignant vs. benign), the output layer was set to have two classes. classifier_activation="sigmoid" A sigmoid activation function was used in the final layer to produce a probability score for each class, as it is well-suited for binary classification tasks.
- optimizer='adam' The Adam optimizer was selected for its adaptive learning rate, which aids in faster convergence and prevents overfitting. loss='sparse_categorical_crossentropy' Sparse categorical crossentropy was chosen as the loss function because the classification task is categorical, and sparse labels (without one-hot encoding) are used. metrics=['accuracy']. Accuracy was selected as the primary evaluation metric to monitor how well the model is performing during training and testing.

Then, the model was compiled and summarized, giving an overview of its architecture, number of parameters, and details on each layer. This configuration thus has a very effective and scalable prototype for melanoma detection. In training the EfficientNetB7 model, we used learning rate dynamic adjustment by learning rate scheduling, early stopping to prevent overfitting, as well as data augmentation to increase the variance of the training set and to improve generalization of the model.

Combining VGG19 with EfficientNetB7 allowed me to take advantage of the good things from both models and do an exhaustive evaluation to find out the best-performing architecture for melanoma classification. Model Training and Evaluation. In this section, we will explain how WE trained both the models, VGG19 and EfficientNetB7, on a dataset aimed at classifying images of melanoma lesions and assessing performance using validation data.

Training of the VGG19 model was done using the fit () method in Keras. It trained the model for a total of 15 epochs using the training set (train_generator) and validated it using the validation set (x_val, y_val). The training scheme comprised

The training data includes train_generator, created through data augmentation to diversify and enlarge the volume of training images. It has improved the generalizability of the model with respect to new and previously unseen data for its better performance. The model was trained for more than 15 epochs based on initial testing and convergence behavior of the training loss. Validation was to use x_val, y_val as a validation set to observe the model's progress through the epochs. Early Stopping: An EarlyStopping callback halts training early if validation performance does not improve for a defined number of epochs. It is done to avoid overfitting and save computational resources. Stored all training history for VGG19 into the history variable so we can track metrics like loss or accuracy for each epoch, to see more clearly how the model learns and changes. Training the EfficientNetB7 Model

Like the other models, this model, EfficientNetB7, was also trained using the same methodology. Consequently, it underwent training for 15 epochs, similar to VGG19, with the training and validation data. The only difference between the configurations for training EfficientNetB7 and VGG19 is the architecture of the model. Training data train_generator, augmented by providing more varied examples to the model. Epochs Trained for 15 epochs, the same number as for VGG19.Validation Data The same validation set (x_val, y_val) was used for both models to ensure consistency in evaluation. The EarlyStopping callback was also applied to EfficientNetB7 to monitor the model's performance and prevent overfitting.

Much like VGG19, the training history of EfficientNetB7 is recorded in the history variable, which provides metrics to judge the performance of the model in terms of epochs. Comparison of Model Training: Since both of them were trained using the same number of epochs and similar configurations, it was pretty straightforward to compare the performance of both models. EarlyStopping function proved to be of importance in this case, where the training could not proceed beyond the point where the models were no longer improving on the validation data, thus leading to getting the early stopping point under which the model had achieved a good balance between accuracy on training and validation performance, minimizing overfitting risks. Loss Function: Binary cross-entropy was employed since it is well-suited for binary classification tasks.

O. Hyperparameter Tuning

We experimented with different learning rates (0.001, 0.0001), batch sizes (16, 32), and epochs (10, 20) to find the optimal settings. Validation and Testing During training, we monitored validation loss and accuracy to prevent overfitting. Early stopping was implemented to halt training once validation performance plateaued. The Methodology WORK gives a fair account of how all the technical steps in preparing the model for melanoma detection were accomplished. This WORK details the critical processing activities included in data preparation, selection of the model, training, and optimization.

The data preparation section is where the pre-processing of images is described to maintain consistency and improve the automatic absorption of the images by the model. Image resizing, noise reduction, and hair removal, among other techniques, were used to make the quality of input data sharp. These processes were aimed at standardizing the images by removing artifacts that can mislead the models' focus on the important features for melanoma detection.

The training section elaborates on the training of both models with exactly the same dataset for 15 epochs and applies early stopping to prevent overfitting. Accuracy metrics monitored during the training process disclosed how well the models learned from the training data and generalized performance to previously unseen validation data. Early stopping enforced termination of training once the validation accuracy no longer underwent improvement, which meant prevention of possible overfitting of the models to the training set. In summary, the methodology WORK has elaborated clear, systematic steps the researcher took in preparing data, selecting models, training the model, and applying optimization techniques, paving the way for evaluating the performance of the melanoma detection system. The whole approach was founded to generate robust and reliable results, as discussed in the consecutive WORKs.

ISSN 2395-8618

IV. Results

The results that we have obtained from my analysis of the training of EfficientNetB7 and VGG19 models regarding the melanoma detection task. The core point of this whole discussion involves evaluating the efficiency of the models in terms of their performance measurement criteria, their generalization ability, and the relevance of these results concerning the project. For evaluating both models, the key metrics used were accuracy, loss, validation accuracy, and validation loss. These are the metrics considered in this review because a comprehensive evaluation of performance can be obtained at two stages (training and validation). Accuracy is the number of correctly predicted instances out of all predictions.

This is another measure of the models' effectiveness in the actual application of malignant versus benign lesions. Loss refers to the error or difference between the outputs that the model predicted and the actual outputs. The lower the loss, the better, because it terrifies how far apart a given model's predicted value is from an actual value. Validation accuracy is paramount because it tells how well the model generalizes to data that it has not seen before. For instance, it tells if the model has overfitted on the training data or can make accurate predictions on new unseen data. Validation loss shows how well the model does on the validation set. High validation loss would indicate overfitting, whereas low validation loss indicates that the model is likely to generalize well on unseen data.

The performance of the EfficientNetB7 model after training demonstrates strong classification capability, achieving an accuracy of 0.8704 with a training loss of 0.5475. This indicates that the model correctly classified approximately 87.04% of the training images, reflecting a high ability to capture relevant patterns for melanoma detection. The validation accuracy is slightly lower at 0.8644, suggesting that while the model generalizes reasonably well to unseen data, there is a minor drop in performance. However, the validation loss is extremely high (2823.3994), which may indicate instability during validation or issues such as overfitting, scaling errors, or the need for improved regularization techniques. Despite this anomaly, the overall performance of EfficientNetB7 remains promising for real-world applications.

In comparison, the VGG19 model achieves an accuracy of 0.8644 with a training loss of 0.3977, and maintains the same validation accuracy of 0.8644 with a validation loss of 0.3972. The close alignment between training and validation metrics suggests that VGG19 exhibits more stable and consistent generalization performance, without the large discrepancy observed in EfficientNetB7. While its accuracy is slightly lower than EfficientNetB7, the lower and well-balanced loss values indicate a more reliable and well-regularized model for melanoma classification.

With an accuracy of 0.8644, it gives a clue that the VGG19 model has had an efficient classification rate on the training data. Strangely, the validation accuracy is equal to that of the training accuracy, thus implying that the model has a generalized performance on unseen data. The validation loss of 0.3972 is just a little since it shows that the predicted label by the model is closer to that of the actual labels in the validation dataset, which is a good signal for practical application. Because of very high accuracy and low loss, the model turns out to be extremely dependable for real-world melanoma classification, as it consistently demonstrates the same outcomes in training and validation data.

Comparison of the Two Models When the two models are put side by side, the one that is better in terms of training accuracy is EfficientNetB7 over VGG19. EfficientNetB7 had a value of 0.8704 on training accuracy, while VGG19 had a value of 0.8644. However, both models had the same validation accuracies, suggesting that EfficientNetB7 was slightly better at learning from the training data. On the other hand, it did not generalize to validation data significantly better than VGG19.

The validation loss for EfficientNetB7 appears to be way higher than the validation loss for VGG19. This may signify overfitting of EfficientNetB7 on the training data or ineffectiveness in terms of regularization. On the contrary, validation loss for VGG19 shows lower values, indicating better generalization from the training data into examples that are unseen.

V. Discussion

It is concluded that both the EfficientNetB7 and VGG19 models show quite a promising ability to classify melanoma lesions. None of the evaluation metrics-accuracy, loss, validation accuracy, and validation loss-indicates otherwise. There is a minor difference in performance among the models, advantageously exhibiting the insight into their respective strengths and weaknesses, while there remain further improvements to make them even better suited to clinical use. It seems that these models, with the right tuning, can afford themselves the chance of making an effect on melanoma detection in the real world.

Here, the EfficientNetB7 and VGG19 models are scrutinized on how their results are relevant to the objectives of the project, their relationship to the literature, and their scope for real-life application. These models reveal very necessary knowledge about the strengths and limitations of each model in the crucial task of melanoma detection.

Interpretation of results. The EfficientNetB7 model produced an accuracy of 87.04% in the training set and 86.44% in the validation set, whereas VGG19 produced 86.44% in both training and validation. Thus, both models are performing successfully in terms of acquiring learnings during the training phase since EfficientNetB7 outperformed VGGET, given that EfficientNetB7 scored higher train accuracy, not that different between them in the validation context. This shows that both models can classify melanoma images well, but EfficientNetB7 does so slightly better in learning the training data patterns.

Though EfficientNetB7 has higher training accuracy, the validation loss was significantly larger compared with VGG19. This means that though EfficientNetB7 could fit the training data better, it may overfit, that is, reduce its ability to generalize to unseen data. VGG19 is more consistent, however, since it has a lower validation loss, which generally gives a better picture of its generalized ability. This could be attributed to the simpler architecture of VGG19, from being less prone to overfitting compared to more complex architectures like EfficientNetB7.

Comparison to the Literature: The output produced by both models agrees with that found in existing studies. Melanoma

ISSN 2395-8618

research has several papers that speak of pretrained deep learning models, including but not limited to VGG19 and EfficientNet, which help exceedingly in medical image classifications. However, performance from our side for EfficientNetB7 in this project is not as high as expected due to a lack of pre-trained weights or failure to use advanced regularization methods.

The VGG19 model performed on par with other studies using similar models for melanoma classification. However, in our case, EfficientNetB7 did not demonstrate a significant improvement over VGG19 regarding generalization (validation accuracy), even though it featured a highly advanced architecture. This can be explained by the fact that the EfficientNetB7 model is highly complex and has not been sufficiently regularized or optimized for this particular dataset. Besides this, the pre-trained weights of EfficientNetB7 were expected to improve its performance, as found in some literature.

Model Performance and Data Characteristics. The same instances can affect the performance of both models; that is, as is common in all cases, these models also include the dataset properties. The actual condition is such that in The HAM10000 Classification Diversity of dermoscopic images introduces variability in data, such as differences in image quality and lighting conditions. Higher training accuracy for EfficientNetB7 could be attributed to its ability to learn more complex patterns within the data. But such diverse datasets do lead to overfitting, as indicated by higher validation loss. It was able to generalize more, but slightly at the expense of lower training accuracy, as would a simpler architecture of VGG19.

EfficientNetB7, being prone to overfitting, could potentially be another limiting factor. Even though early stopping has been applied, its high validation loss shows that the model has learned the training data too much at the expense of its ability to generalize. Such problems can be remedied using advanced approaches like data augmentation, dropout, or regularization to strengthen model robustness.

This research question of the possible usage of deep learning models for lesion detection found an affirmative answer. VGG19 and EfficientNetB7 exhibited good classification ability on melanoma images while being reasonably generalized to unseen data.

VI. Conclusion

This study demonstrates the effectiveness of deep convolutional neural networks for binary classification of dermoscopic skin lesions using the HAM10000 dataset. A comparative analysis of EfficientNetB7 and VGG19 shows that although EfficientNetB7 achieves slightly higher training accuracy, it exhibits signs of overfitting, particularly reflected in its high validation loss. In contrast, VGG19 provides more stable and balanced performance across training and validation, indicating stronger generalization and greater suitability for clinical deployment. These results highlight the potential of AI-assisted dermatology, particularly for telemedicine and large-scale screening in resource-limited settings. Reliable models such as VGG19, especially when combined with explainability techniques, can support clinical decision-making and improve early detection of skin cancer. However, limitations remain, including the lack of pretrained weights and dataset imbalance, which may affect model performance. Future work should focus on transfer learning, improved regularization, ensemble approaches, and the integration of clinical metadata.

References

[1] F. Z. Abdullah, C. G. Sánchez Mejorada, M. A. Ather, J. L. Oropeza Rodríguez, and G. Sidorov, “Fair and explainable multitask deep learning on synthetic endocrine trajectories for real-time prediction of stress, performance, and neuroendocrine states,” *Computers*, vol. 14, no. 12, p. 515, 2025.

[2] Abdullah, F. Ullah, N. Hafeez, I. Latif, G. Sidorov, E. F. Riveron, and A. Gelbukh, “Cyberbullying detection on social media using machine learning techniques,” *Computación y Sistemas*, vol. 29, no. 3, 2025.

[3] M. A. Ather, F. Z. Abdullah, J. L. O. Rodríguez, and G. Sidorov, “An interpretable multi-dataset learning framework for breast cancer prediction using clinical and biomedical tabular data,” *Computers*, vol. 15, no. 2, p. 97, 2026.

[4] F. Z. Abdullah, M. J. T. Ruiz, O. Espinosa-Sosa, C. G. Sánchez-Mejorada, R. Q. Téllez, … and G. Sidorov, “Explainable patient-level cognitive impairment screening via temporal, semantic, and psycholinguistic multimodal AI,” *Journal of Intelligence*, vol. 14, no. 4, p. 66, 2026.

[5] M. A. Ather, R. Q. Tellez, G. Sidorov, C. G. Sánchez-Mejorada, and M. J. T. Ruiz, “Enhancing context-aware SARS disorder management a proposed multi-agent simulation framework with machine learning and bio-sensor data integration,” *Rehabilitation*, vol. 19, p. 20, 2026.

[6] Abdullah, N. Hafeez, M. Shabbir, M. A. Ather, J. L. O. Rodríguez, and G. Sidorov, “Security, privacy, and scalability trade-offs in blockchain-enabled IoT systems a systematic analytical review,” *Applied Sciences*, vol. 16, no. 8, p. 3638, 2026.

[7] M. Shabbir and F. Shabbir, “A machine and deep learning approaches for intrusion detection and attack classification in medical IoT networks,” *Computación y Sistemas*, vol. 30, no. 1, 2026.

[8] Abdullah, M. A. Ather, J. L. O. Rodriguez, C. G. Sánchez-Mejorada, M. J. T. Ruiz, and R. Q. Tellez, “A leakage-aware multimodal machine learning framework for nutrition supply–demand forecasting using temporal and spatial data fusion,” *Computers*, vol. 15, no. 3, p. 156, 2026.

[9] F. Z. Abdullah, H. A. Safder, M. Manzoor, C. G. Sánchez-Mejorada, M. J. Torres Ruiz, and R. Quintero Téllez, “RhythmX™ an interpretable self-supervised contrastive learning framework for heartbeat classification,” *Technologies*, vol. 14, no. 3, p. 148, 2026.

[10] H. Javed, A. Saeed, H. M. ul Haque, Z. Fatima, J. L. O. Rodríguez, and C. G. Sánchez-Mejorada, “Automated detection of breast cancer using machine learning and deep learning models from tabular data,” *Letters in Drug Design & Discovery*, Art. no. 100295, 2026.

[11] M. A. Ather, J. L. O. Rodríguez, and C. G. Sánchez-Mejorada, “A hybrid deep boosting framework with adaptive label stabilization for SEM-based porosity estimation in fly-ash cement mortar,” *Frontiers in Artificial Intelligence*, vol. 9, Art. no. 1766671, 2026.

[12] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset a large collection of multi-sources dermatoscopic images of common pigmented skin lesions,” *Scientific Data*, vol. 5, p. 180161, 2018, doi: 10.1038/sdata.2018.161.

[13] T. J. Brinker *et al.*, “Deep learning outperformed 136 of 157 dermatologists in a head-to-head image classification task,” *European Journal of Cancer*, vol. 113, pp. 47–54, 2019.

ISSN 2395-8618

[14] N. C. F. Codella *et al.*, "Skin lesion analysis toward melanoma detection a challenge at the ISIC 2017," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 501–512, 2018.

[15] Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[16] J. González, J. Pérez, and A. Suárez, "Artificial intelligence for dermatology education the potential of deep learning," *Journal of Medical Imaging*, vol. 6, no. 1, pp. 31–39, 2019.

[17] H. A. Haenssle *et al.*, "Artificial intelligence in dermatology current capabilities, limitations, and future directions," *British Journal of Dermatology*, vol. 179, no. 3, pp. 569–577, 2018.

[18] P. Heinlein *et al.*, "Multicenter evaluation of AWEensemble algorithms in melanoma detection," *Journal of Clinical Dermatology AWEResearch*, vol. 6, no. 1, pp. 10–20, 2024.

[19] N. Codella *et al.*, "Skin lesion analysis toward melanoma detection a challenge at the 2017 International Skin Imaging Collaboration (ISIC)," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 493–503, 2019.

[20] Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[21] Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 20, no. 10, pp. 500–510, 2020.

[22] V. Vusirikala and S. Rajendran, "Self-supervised learning in melanoma diagnosis rotation prediction and corruption removal," *Artificial Intelligence in Dermatology Research*, vol. 12, no. 4, pp. 50–67, 2024.

[23] N. Lama, R. Kumar, and A. Singh, "Lesion-focused convolutional models for improved melanoma classification," *Journal of Biomedical Signal Processing*, vol. 50, pp. 35–47, 2023.

[24] J. Kim, S. Lee, K. Park, and J. Hwang, "Mela-D a lightweight and efficient melanoma detection model," in *Proc. Int. Conf. Dermatological AWEApplications*, 2024, pp. 220–234.

[25] P. Heinlein, M. Fischer, and L. Weinhold, "Multicenter evaluation of AWEensemble algorithms in melanoma detection," *Journal of Clinical Dermatology AWEResearch*, vol. 6, no. 1, pp. 10–20, 2024.

[26] S. Jinnawe *et al.*, "Automated classification of melanoma using deep learning with clinical images and metadata," *Scientific Reports*, vol. 10, p. 14752, 2020.