applied sciences

MDPI

Article

Skin Lesion Classification Using a Deep Ensemble Model

Su Myat Thwin and Hyun-Seok Park *

Department of Computer Science and Engineering, Ewha Womans University, Seoul 03760, Republic of Korea; sumyatthwin@ewhain.net
* Correspondence: neo@ewha.ac.kr

Abstract: Skin cancer, particularly melanoma, is a severe health threat that necessitates early detection for effective treatment. This research introduces a skin lesion classification system that harnesses the capabilities of three advanced deep learning models: VGG16, Inception-V3, and ResNet-50. By integrating these models into an ensemble, the system leverages their individual strengths to improve classification accuracy and robustness. Every model in the ensemble brings its distinctive contribution, having undergone pre-training on ImageNet and subsequent fine-tuning using dermoscopic images. We evaluated our ensemble approach using the ISIC 2018 dataset, a standard benchmark for skin lesion classification. By employing a weighted averaging method to merge predictions from the three models, our ensemble model demonstrated superior performance. The results show an improvement in classification accuracy, achieving an accuracy of 91% on the original dataset and 97% on the dataset balanced by oversampling compared to the individual models. This system was then evaluated using the HAM10000 dataset. The results using the HAM10000 dataset show an improvement in classification accuracy, achieving an accuracy of 90% on the original dataset and 96% on the dataset balanced by oversampling compared to the individual models. This enhanced performance highlights the effectiveness of our ensemble method in capturing diverse features of skin lesions, leading to more accurate diagnoses. Our findings suggest that this approach can significantly assist dermatologists in early and precise skin cancer detection, potentially improving patient outcomes.

Keywords: skin lesion classification; VGG16; ResNet-50; Inception-V3; ensemble

check for updates

Citation: Thwin, S.M.; Park, H.-S. Skin Lesion Classification Using a Deep Ensemble Model. *Appl. Sci.* **2024**, *14*, 5599. https://doi.org/10.3390/app14135599

Academic Editor: Thomas Lindner

Received: 27 May 2024
Revised: 22 June 2024
Accepted: 25 June 2024
Published: 27 June 2024

1. Introduction

Skin cancer poses a significant threat, marked by abnormal and malignant proliferation of cells within the skin. The skin is composed of three primary layers: the epidermis, dermis, and hypodermis. The epidermis contains melanocytes, which produce melanin, a pigment that is particularly sensitive to ultraviolet radiation [1,2]. Abnormal development of melanocytes can lead to melanoma, a highly malignant tumor. Dermoscopy is a valuable imaging technique that reduces skin reflection, allowing for detailed examination of structures within the dermis and epidermis. The heightened magnification and minimized reflection artifacts characteristic of this approach enhance the acquisition of detailed visual data from deeper layers of the skin, thereby enhancing systems used for computer-aided diagnosis.

Skin abnormalities are typically classified into three main categories: basal cell carcinoma (BCC), melanoma, and squamous cell carcinoma (SCC). BCC originates from basal cells in the epidermis, while SCC arises from squamous cells in the upper epidermis. Melanoma is often regarded as the most hazardous form of skin cancer due to its propensity for rapid dissemination and its potential to pose life-threatening risks. However, early detection and treatment significantly increase the likelihood of successful outcomes. Early and accurate diagnosis is critical for effective treatment and improved patient survival rates. Traditional diagnostic methods rely heavily on dermatologists' expertise and are subject to variability in accuracy [3,4]. Therefore, there is a pressing need for automated, reliable, and precise diagnostic tools to assist in skin lesion classification.

Recent advancements in deep learning have shown great promise in the automated categorization of skin abnormalities [5–7]. Convolutional neural networks (CNNs), in particular, have achieved significant success in image recognition tasks. This study leverages the strengths of an ensemble of three state-of-the-art CNN architectures—VGG16, Inception-V3, and ResNet50—to classify skin lesions into BCC, SCC, and melanoma using the ISIC dataset [8], as shown in Figure 1. The ISIC dataset, a benchmark in dermatological research, presents a diverse collection of skin lesion images. However, the dataset is imbalanced, with certain lesion types underrepresented. To mitigate the issue of dataset imbalance, we utilize an oversampling method [9], which ensures that our deep learning models are trained with a balanced distribution of each lesion category. This approach aims to improve the models' performance and generalizability.

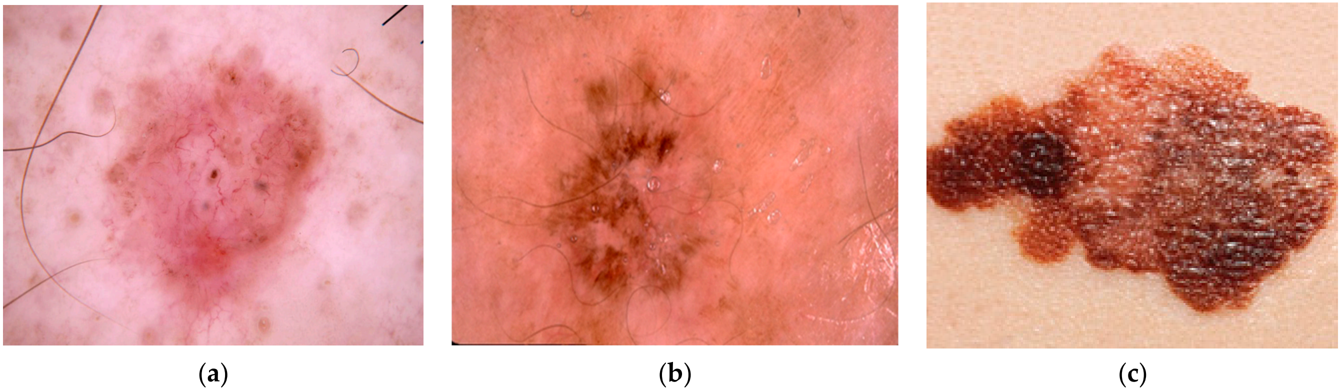

Figure 1. ISIC dataset: (**a**) BCC; (**b**) SCC; and (**c**) melanoma.

In this paper, we detail the methodology for classifying skin lesions using the ensemble of VGG16, Inception-V3, and ResNet50 models. We describe the preprocessing steps, including the oversampling technique used to balance the ISIC dataset, and the training procedures for each model. We also evaluate the ensemble model's performance compared to the individual models and discuss the implications of our findings for clinical practice and future research. This work contributes to the expanding field of deep learning in medical diagnostics, highlighting the potential of ensemble methods to enhance classification accuracy and reliability in skin lesion analysis. By combining multiple CNN architectures and addressing dataset imbalance, this study aims to develop a robust solution for the automated classification of skin lesions.

Our main contributions to this endeavor are as follows:

1. An ensemble approach combining VGG16, Inception-V3, and ResNet50 is developed to classify skin lesions into basal cell carcinoma, squamous cell carcinoma, and melanoma.
2. This approach achieves superior classification performance compared to individual deep learning models, as demonstrated by comprehensive experiments using the ISIC dataset.
3. An effective oversampling technique is applied to address class imbalance in the ISIC dataset, enhancing model robustness and generalization.
4. This technique leverages transfer learning by fine-tuning pre-trained models, optimizing the training efficiency and achieving high classification accuracy with reduced computational resources.
5. The proposed model offers a reliable and accurate tool for dermatologists, aiding in the early detection and accurate classification of skin lesions, potentially improving patient outcomes.
6. Advanced data augmentation techniques are employed to broaden the range of training samples, consequently enhancing the model's efficacy while mitigating overfitting.

The structure of this paper is as follows: Section 2 provides an overview of related work in the field of skin lesion classification using deep learning models. Section 3 describes the methodology, including data collection, preprocessing, model training, and the ensemble approach. Section 4 presents the experimental results and evaluation metrics. Finally, Section 5 concludes the paper and suggests directions for future research.

2. Literature Review

Researchers had undertaken an exhaustive exploration of various algorithms employed for feature selection and melanoma prediction, as elaborated in [10]. These predictive frameworks were intricately devised for diagnosing melanoma, incorporating real-world data-gathering approaches. The development of the model hinged on a series of guidelines formulated in collaboration with domain experts. Noteworthy is the substantial investment necessitated by the need for swift image processing and the acquisition of nuanced insights to fine-tune the optimal rule set. Interestingly, many researchers chose to incorporate deep learning methodologies into their system. In their study [11], the authors presented a novel concept focused on employing a one-class deep neural network learning framework. This innovative approach was contrasted with alternative classifiers such as Isolation Forest, Gaussian Mixtures, and OC-SVM techniques. Following a comprehensive assessment of these methodologies, the authors determined that the Isolation Forest technique was the most suitable for the dataset under investigation. This method operates on the premise that features linked to anomalies diverge notably from those of normal samples. Central to this approach is the creation of an ensemble consisting of isolation trees, where anomalies exhibit significantly shorter average path lengths. In their study, the researchers devised an automated system designed to analyze skin lesions, with a particular emphasis on identifying melanoma and performing semantic segmentation [12]. Utilizing advanced deep learning techniques and publicly available dermoscopic images, the system significantly improved melanoma detection. Initially, the image classification utilized a deep convolutional neural network, integrating feature extraction methods. Subsequently, these extracted features were inputted into a random forest classifier. By combining the outputs from both classifiers, the system achieved notable performance metrics with a precision of 0.81, an accuracy of 80.3%, and an AUC score of 0.69. Furthermore, segmentation utilized a convolutional–deconvolutional architecture, resulting in a dice coefficient of 73.5%, illustrating precise delineation of segmented regions. In the study described in [13], a five-layered CNN with the PH2 dataset was utilized. The CNN proposed in the study demonstrated remarkable performance, achieving approximately 95% in accuracy, 94% in sensitivity, 97% in specificity, and a perfect AUC score of 100% on the test set, as evaluated by four essential performance metrics.

The authors of [14] presented a novel deep learning-based method designed to enhance the accuracy of classification systems. The method employs a multi-scale decomposition model in the preprocessing phase, utilizing texture as input for the CNN to eliminate separate texture extraction and feature selection. The experimental findings demonstrate that the proposed approach attains superior accuracy when contrasted with other established algorithms and methodologies documented in the existing literature. In another research study [15], CNN was employed to detect malignant and benign lesions utilizing the ISIC2018 dataset, comprising 3533 images on a variety of skin lesions, ranging from benign and malignant tumors to nonmelanocytic and melanocytic growths. Initially, the images underwent enhancement using ESRGAN. During preprocessing, augmentation, normalization, and resizing techniques were applied to the images. The CNN method was utilized to classify the skin lesion images, based on aggregated results from multiple iterations. Furthermore, various transfer learning models, including ResNet50, InceptionV3, and Inception ResNet, underwent fine-tuning. The custom CNN model achieved an accuracy of 83.2%, which was comparable to the pre-trained models: Resnet50 (83.7%), InceptionV3 (85.8%), and Inception Resnet (84%). In the research cited in [16], a deep learning model was developed to detect skin cancer using the HAM10000 dermoscopic image database that includes 513 BCC, 790 benign, 327 actinic and intraepithelial carcinoma (AKIEC), and 115 dermatofibroma events. In this research, a CNN was created to detect benign and malignant groups. AlexNet was utilized as the pre-trained model. This model directly processes raw images, learning crucial features for classification without lesion segmentation or manual feature extraction. It achieved an accuracy of 84%, specificity of 88%, an area under the receiver operating characteristic (ROC) curve of 0.91, and sensitivity of 81%, with a confidence score threshold of 0.5. In this paper [17], deep learning was utilized to identify malignant and benign tumors using skin images of the RGB channel. A combination of lesion segmentation and classification was employed to detect malignant and benign tumors. For segmentation, U-Net was utilized, and an algorithm was used for classification.

Other researchers have developed a CNN model for accurately identifying skin cancer [18]. AlexNet was employed with the HAM10K dataset that is highly imbalanced across classes, which typically results in lower training accuracy. Moreover, a novel activation function was developed to mitigate the vanishing gradient problem and tested across multiple benchmark architectures. It provided superior accuracy compared to existing activation functions with an accuracy of 98.20%, and precision, recall, and F-score of 98.20%, outperforming current models. Another study [19] introduced a deep learning model for detecting skin lesions at the benign and malignant stages utilizing transfer learning. The method built upon a pre-existing VGG16 architecture, enhancing it by incorporating additional layers. Their integration aimed to enhance accuracy in the classification task. The evaluation utilized a Kaggle dataset, implementing data augmentation methods to diversify the input data and bolster model robustness. Numerous hyperparameters were explored, encompassing batch sizes ranging from 8 to 128, varied epochs, and different optimizers. The optimized model demonstrated exceptional performance, attaining an accuracy of 89.09%. This achievement was realized with a batch size of 128, utilizing the Adam optimizer and undergoing training for 10 epochs. Notably, these results surpassed existing methodologies. The system proposed in [20] detected skin cancer from skin images by first applying a median filter to reduce noise. Following this, the images were segmented using the Mean Shift segmentation technique. Following segmentation, feature extraction was conducted on the segmented images, with emphasis on extracting features such as GLCM (Gray-Level Co-occurrence Matrix), Moment Invariants, and GLRLM (Gray-Level Run Length Matrix). These extracted features were subsequently classified using multiple techniques, including Support Vector Machine (SVM), Probabilistic Neural Networks (PNNs), and random forest (RF). Notably, the combined SVM+RF classifier emerged as the most successful approach, yielding the highest performance. The paper [21] introduced a combination of faster region-based CNN and fuzzy k-means clustering (FKM) to segment melanoma. The method was tested on several clinical images to aid dermatologists in the early diagnosis of this serious and potentially life-threatening condition. Initially, the images in the dataset underwent preprocessing to address issues such as noise and uneven illumination, which aims to improve the clarity of visual data. Following this, the faster R-CNN algorithm was utilized to produce a fixed-length feature vector from the images. Following this step, the fuzzy k-means (FKM) algorithm was utilized to delineate the boundaries of melanoma-affected skin regions, which may vary in size and shape. The efficacy of this methodology was evaluated using three well-established datasets: ISBI-2016, PH2, and ISIC-2017. The results demonstrate superior performance compared to existing methods, with average accuracies of 95.40%, 95.6%, and 93.1% for the ISIC-2016, PH2, and ISIC-2017 datasets.

3. Materials and Methods

This section outlines the proposed deep ensemble methodology proposed for classifying three distinct lesion types—basal, squamous, and melanoma—using a dermoscopic image (ISIC) dataset.

3.1. Oversampling

Oversampling is used to address imbalances in datasets with uneven class distributions. It replicates the minority class's samples to establish a more balanced illustration. Oversampling employs a procedure called simple random sampling with replacement (SRSWR). Unlike simple random sampling without replacement (SRSWOR), SRSWR allows each selected instance to be returned to the dataset, enabling multiple selections of the same instance. This process results in an augmented dataset where the minority class is more substantially represented. Oversampling with SRSWR is particularly effective in situations where one class (the majority) is much more prevalent than the other (the minority). By expanding the number of instances within the minority class, over-sampling aids in mitigating dataset imbalance.

3.2. Transfer Learning

Transfer learning is a machine learning approach in which a pre-existing model is applied for a specific task, which is reused and fine-tuned for a distinct yet related objective [22]. Instead of training a model from scratch, transfer learning leverages the knowledge gained from a large dataset and applies it to a smaller, specific dataset. This approach is particularly beneficial when data for the new task are limited, as it enhances performance and reduces training time by utilizing the pre-trained model's established features and patterns. Transfer learning is widely used in applications such as image classification, natural language processing, and other areas where acquiring large amounts of labeled data is challenging.

3.3. VGG16

VGG16 is a convolutional neural network (CNN) architecture known for its simplicity and uniform design [23]. It applies 3 × 3 filters by one stride and the same padding convolutional layers, and 2 × 2 filters and two-stride max pooling layers. This consistent pattern is maintained throughout the network. The architecture concludes with two fully connected layers and a Softmax activation function for classification. VGG16 is notable for its depth and capacity, containing approximately 138 million parameters, where its capacity to effectively capture intricate patterns in image data is attributed to its architecture. Figure 2 illustrates the VGG16 design.

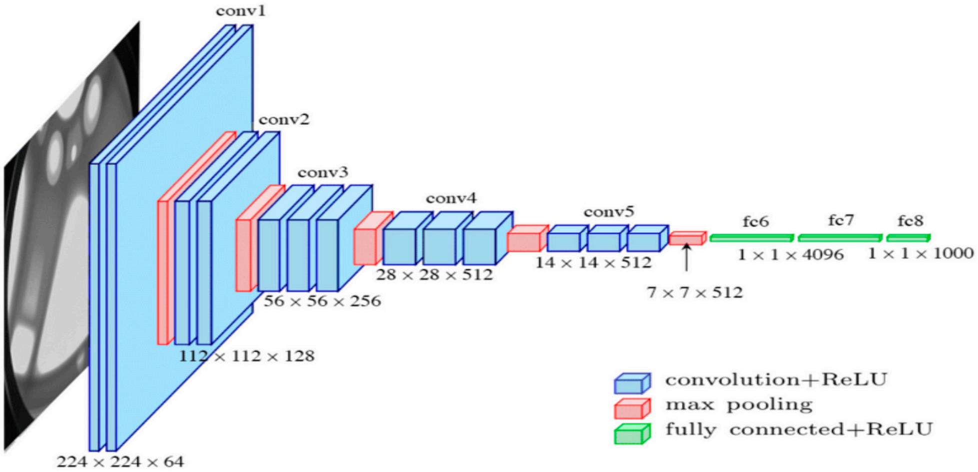

Figure 2. VGG16 design.

3.4. ResNet-50

ResNet-50 is a deep convolutional neural network (CNN) architecture distinguished by its use of residual learning [24]. Comprising 50 layers, it integrates residual blocks that include shortcut connections. Each residual block typically consists of a few convolutional layers. By enabling direct paths for gradients to flow, ResNet-50 significantly enhances training efficiency and performance, making it highly effective for image classification tasks. The architecture of ResNet-50 is shown in Figure 3.

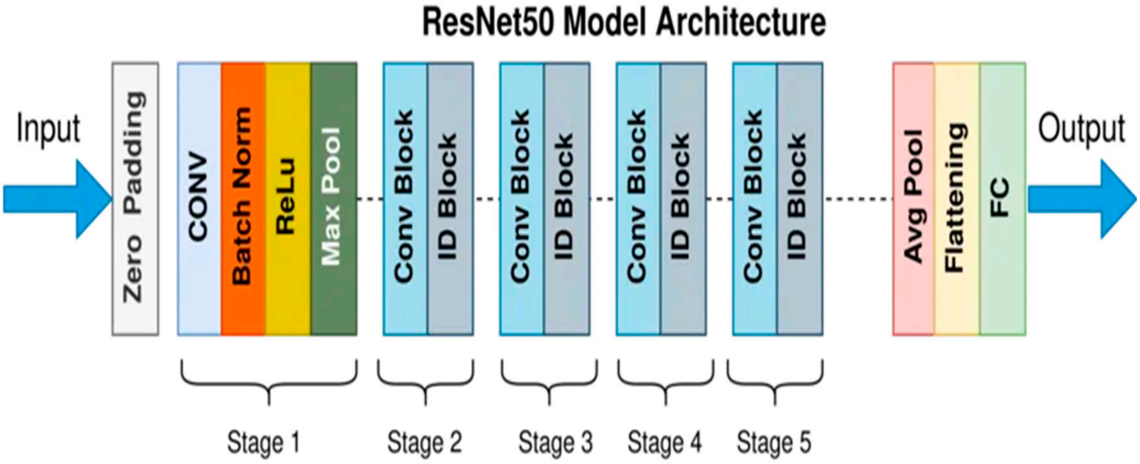

Figure 3. ResNet-50 design.

3.5. Inception-V3

Inception-V3 is a convolutional neural network (CNN) known for its efficient and effective design, featuring an advanced Inception module that processes multiple filter sizes (1 × 1, 3 × 3, 5 × 5) simultaneously to capture various feature scales [25]. It employs techniques like factorized convolutions, breaking down larger convolutions into smaller ones (e.g., 7 × 7 into two 3 × 3), and asymmetric convolutions (e.g., 3 × 3 into 1 × 3 followed by 3 × 1) to enhance computational efficiency without compromising accuracy. Additionally, Inception-V3 includes auxiliary classifiers that provide intermediate training signals, improving the overall learning process. These innovations make Inception-V3 highly effective for image classification tasks. Its design is described in Figure 4.

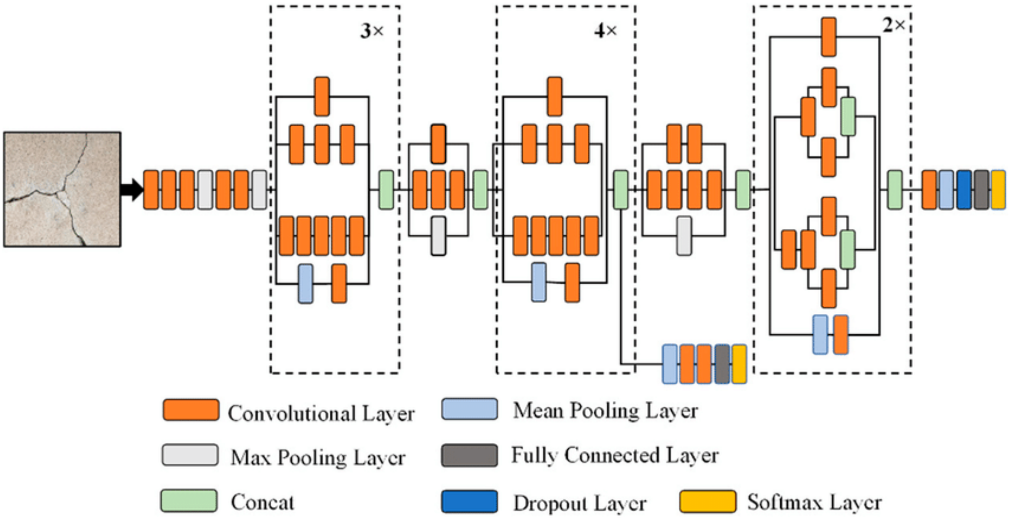

Figure 4. Inception-V3 design.

3.6. Data Collection

The ISIC dataset on Kaggle is an extensive compilation of dermoscopic images aimed at improving skin lesion analysis and melanoma detection. It includes thousands of high-resolution images, each annotated by dermatology experts with detailed labels for various skin lesion types such as melanoma, nevi, and keratoses. Accompanying the images is a metadata set that provides patient demographics and the anatomical site of each lesion. The dataset is organized into training, validation, and test sets to support the development and evaluation of machine learning models. This dataset is pivotal in enhancing the precision and dependability of automated systems for detecting skin cancer. This dataset consists of 2357 images of malignant and benign oncological diseases, which were collated by the International Skin Imaging Collaboration (ISIC). All images were sorted according to the classification proposed by the ISIC, and all subsets were divided into the same number of images, with the exception of melanomas and moles, whose images are slightly predominant. This ISIC dataset contains the following diseases:

- Actinic keratosis (114 images);
- Basal cell carcinoma (376 images);
- Dermatofibroma (95 images);
- Melanoma (438 images);
- Nevus (357 images);
- Pigmented benign keratosis (462 images);
- Seborrheic keratosis (77 images);
- Squamous cell carcinoma (181 images);
- Vascular lesion (139 images).

From this dataset, melanoma (438 images), basal (376 images) and squamous (181 images) skin lesions were used for system evaluation.

3.7. Proposed Deep Ensemble Model for Skin Lesion Classification

Our study proposes a skin lesion classification system to classify skin lesions into BCC, SCC, and melanoma by applying a deep ensemble model. The system design integrates three widely recognized deep learning architectures: VGG16, Inception-V3, and ResNet-50. These models are selected for their proven effectiveness in image classification tasks and their widespread adoption in medical image analysis. The ISIC dataset, a well-established and widely used repository of dermoscopic images, serves as the primary dataset for training and evaluation. This dataset comprises a varied assortment of skin lesion images, encompassing basal cell carcinoma, squamous cell carcinoma, and melanoma, thereby offering a comprehensive and diverse resource for training and validating models. One of the primary challenges in skin lesion classification revolves around the inherent class imbalance within the dataset, where certain classes may be underrepresented in comparison to others. To mitigate this imbalance, we employ oversampling techniques. Specifically, we augment the size of the minority classes (e.g., melanoma) by generating synthetic instances, thereby ensuring a more balanced distribution across all classes. Preprocessing enhances the learning capability and precise classification is achieved regarding the minority class.

The core of our proposed system lies in the integration of multiple deep learning models through ensemble learning. By combining the strengths of VGG16, ResNet-50, and Inception-V3, we aim to leverage their complementary features and improve the overall classification performance. Each model contributes unique insights and representations, which are aggregated to make the final predictions. Ensemble learning has been shown to enhance robustness and generalization capabilities, making it an ideal approach for complex classification tasks such as skin lesion categorization. By leveraging deep ensemble modeling techniques and balancing the ISIC dataset through oversampling, our proposed system represents a significant advancement in automated skin lesion classification, offering potential applications in early detection and diagnosis of skin cancer, thereby contributing to improved patient outcomes and clinical decision-making in dermatology. The proposed skin lesion classification system design is illustrated in Figure 5.

The system aims to develop a reliable assistance system for medical doctors for the classification of skin lesions using an original dataset and a balanced dataset by applying the oversampling technique. Initially, the original dataset was split for training (75%) and testing (25%). The images were resized to 224 × 224 for VGG16 and ResNet50 inputs, and 299 × 299 for Inception V3 inputs. They were then normalized and transformed into tensor form. To balance the dataset, a random oversampling method was used. Pre-trained models were loaded using the deep learning framework TensorFlow, with the feature extraction layers frozen to prevent updates during training. A new fully connected layer was added, replacing the old one. The models were compiled with a specified loss function and the Adam optimizer. Training was conducted on the training dataset for 150 epochs. Models that achieved the desired accuracy were saved. Finally, the models were combined using the ensemble weighted average method.

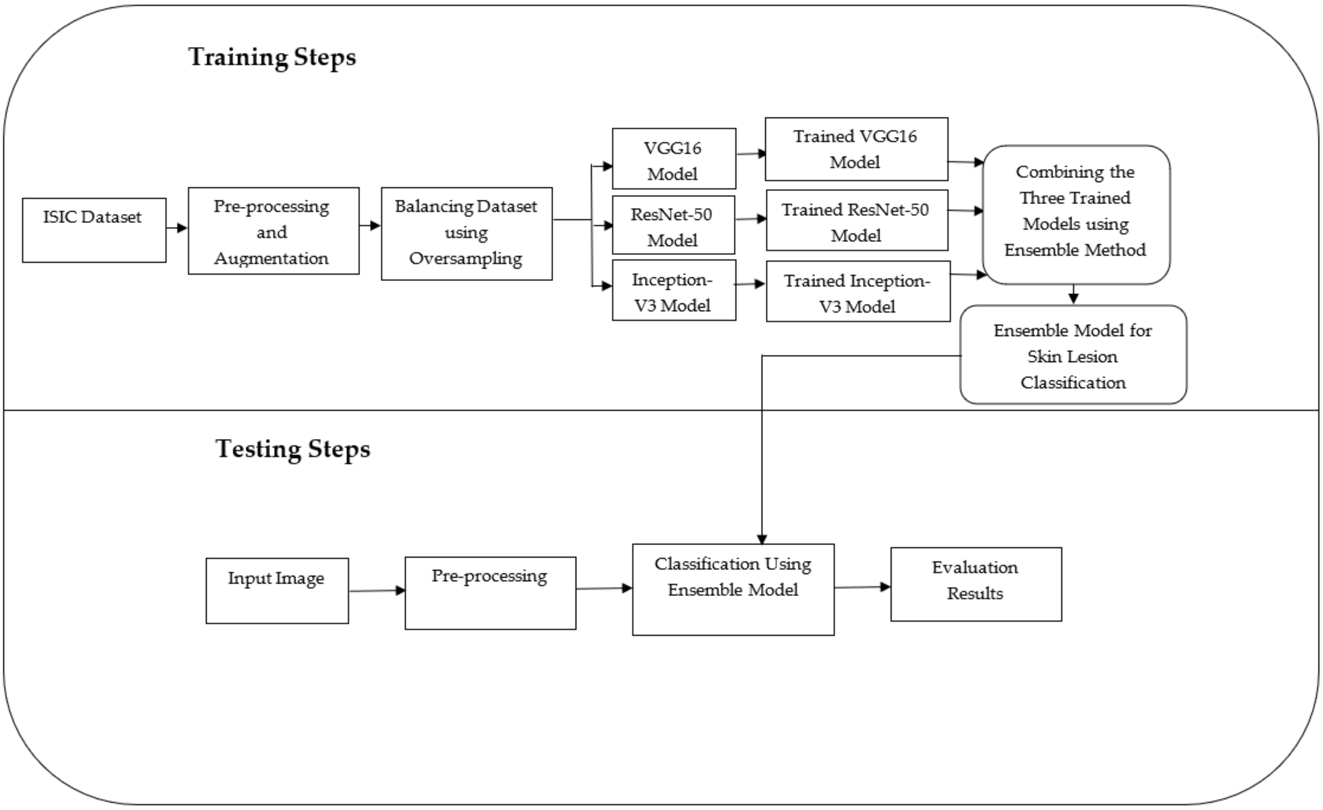
Training Steps
ISIC Dataset
Pre-processing and Augmentation
Balancing Dataset using Oversampling
VGG16 Model
ResNet-50 Model
Inception-V3 Model
Trained VGG16 Model
Trained ResNet-50 Model
Trained Inception-V3 Model
Combining the Three Trained Models using Ensemble Method
Ensemble Model for Skin Lesion Classification
Testing Steps
Input Image
Pre-processing
Classification Using Ensemble Model
Evaluation Results

Figure 5. Proposed skin lesion classification system design.

During testing, the images were resized to 224 × 224 for VGG16 and ResNet-50 inputs and to 299 × 299 for Inception-V3 inputs. They were then normalized and transformed into tensor form for further processing. Classification was performed using an ensemble model. Initially, the system was tested with the original, unbalanced dataset. Subsequently, the dataset was balanced using oversampling and augmentation methods, and the system was tested again. Finally, model evaluation was conducted to assess the performance of the classification system using evaluation metrics including precision, recall, F1-score, and accuracy under both the original and balanced dataset conditions.

3.8. Performance Evaluation Metrics

To assess the effectiveness of our proposed system, we employ standard evaluation criteria including precision, F1-score, accuracy, and recall. These metrics offer comprehensive assessments of a model's classification performance across various classes and aid in quantifying its ability to accurately differentiate between basal, squamous, and melanoma lesions. Each metric is computed using specific equations, referenced as Equation (1), Equation (2), Equation (3), and Equation (4), respectively. These equations provide quantitative metrics for assessing the performance of the model in terms of its ability to correctly classify instances across different categories.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \tag{1}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{2}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{3}$$

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

4. Results

This section evaluates the performance of the proposed deep ensemble model against the individual models, VGG16, ResNet-50, and Inception-V3, focusing on accuracy metrics. In this system, the ISIC dataset from Kaggle is employed, specifically targeting images depicting the classification of BCC, SCC, and melanoma. At the outset, the system's performance is evaluated using the original dataset, which is inherently unbalanced. Subsequently, the evaluation process is repeated using a balanced dataset, which is achieved through the application of oversampling techniques. These models are created and trained using Keras, which runs on the TensorFlow framework. For thorough evaluation, the dataset is split into training and testing sets with a 75–25% ratio. Specifically, 75% of the data are randomly chosen for training, while the remaining 25% are set aside for testing. Table 1 describes the parameters of VGG16, ResNet-50, and Inception-V3.

Table 1. Parameter settings of three models.

Parameter	VGG16	ResNet-50	Inception-V3
Input Image size	224 × 224	224 × 224	299 × 299
Batch Size	32	32	32
Learning Rate	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam
Epochs	150	150	150
Loss Function	Categorical Crossentropy	Categorical Crossentropy	Categorical Crossentropy
Dropout Rate	0.5	0.5	0.5
Data Augmentation	Yes	Yes	Yes
Training/Validation Split	75%/25%	75%/25%	75%/25%

Figure 6 presents the comparative performance results of the VGG16 model on both the original unbalanced dataset and the dataset balanced through oversampling. Figure 7 describes the comparative performance results of the ResNet-50 model on both the original unbalanced dataset and the dataset balanced through oversampling. Figure 8 depicts the comparative performance outcomes of the Inception-V3 model on both the initial unbalanced dataset and the dataset balanced through oversampling. Similarly, Figure 9 showcases the comparative performance outcomes of the deep ensemble model on the original dataset and the dataset balanced through oversampling.

Table 2 describes the confusion matrix for each class. Table 3 presents the sensitivity and specificity of each class.

Table 2. Confusion matrix for each class.

	Predicted BCC	Predicted SCC	Predicted Melanoma
Actual BCC	320	30	26
Actual SCC	15	150	16
Actual Melanoma	20	25	393

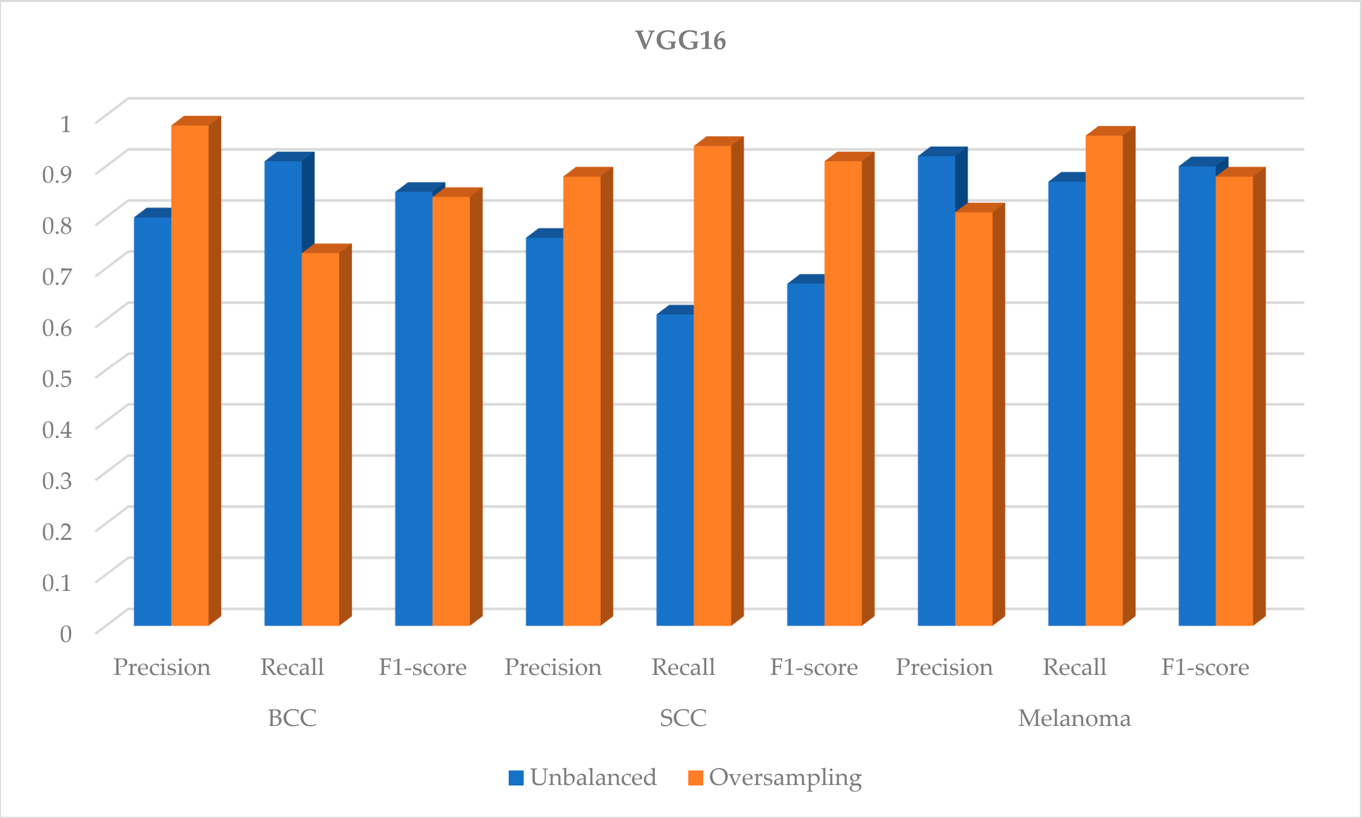

Figure 6. Results of VGG16 model on unbalanced and balanced datasets.

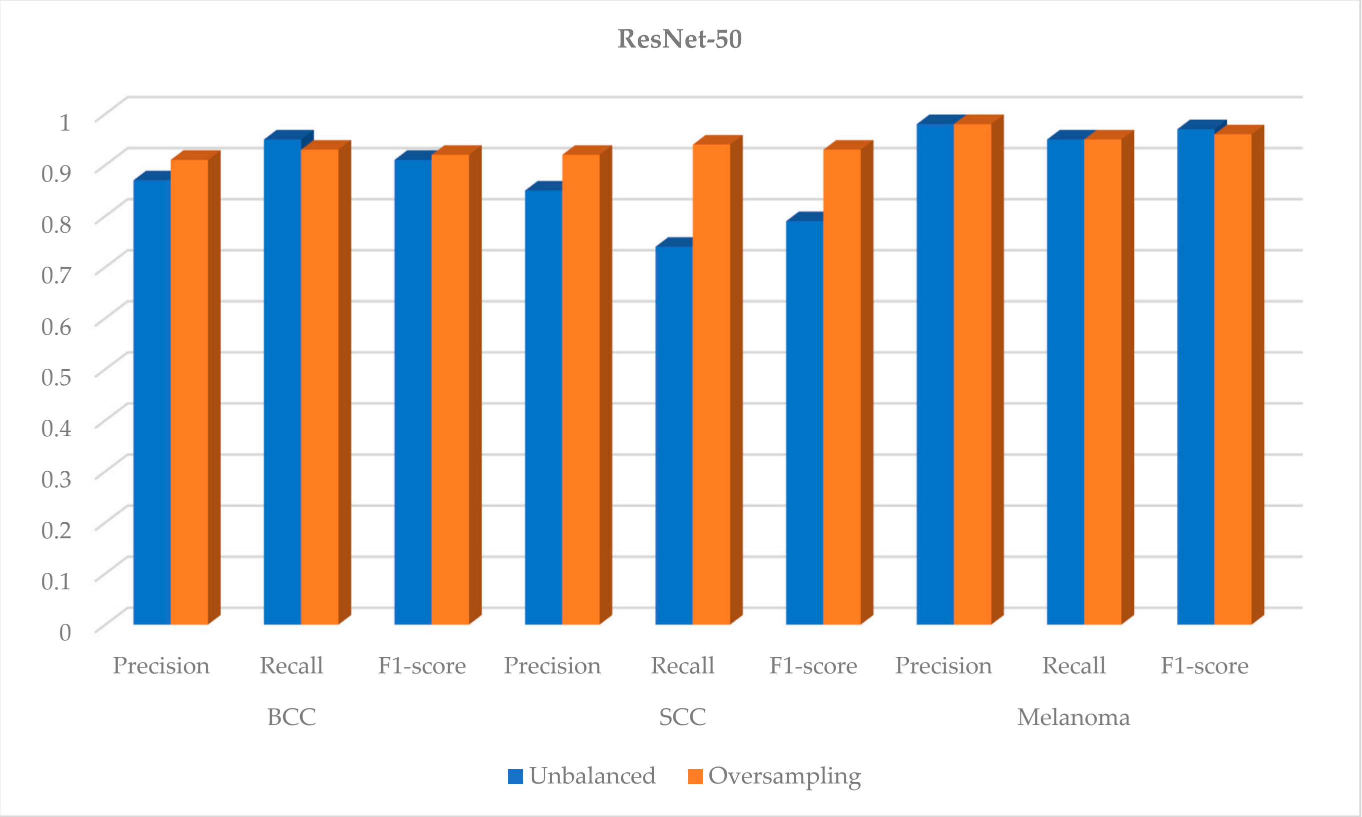

Figure 7. Results of ResNet-50 model on unbalanced and balanced datasets.

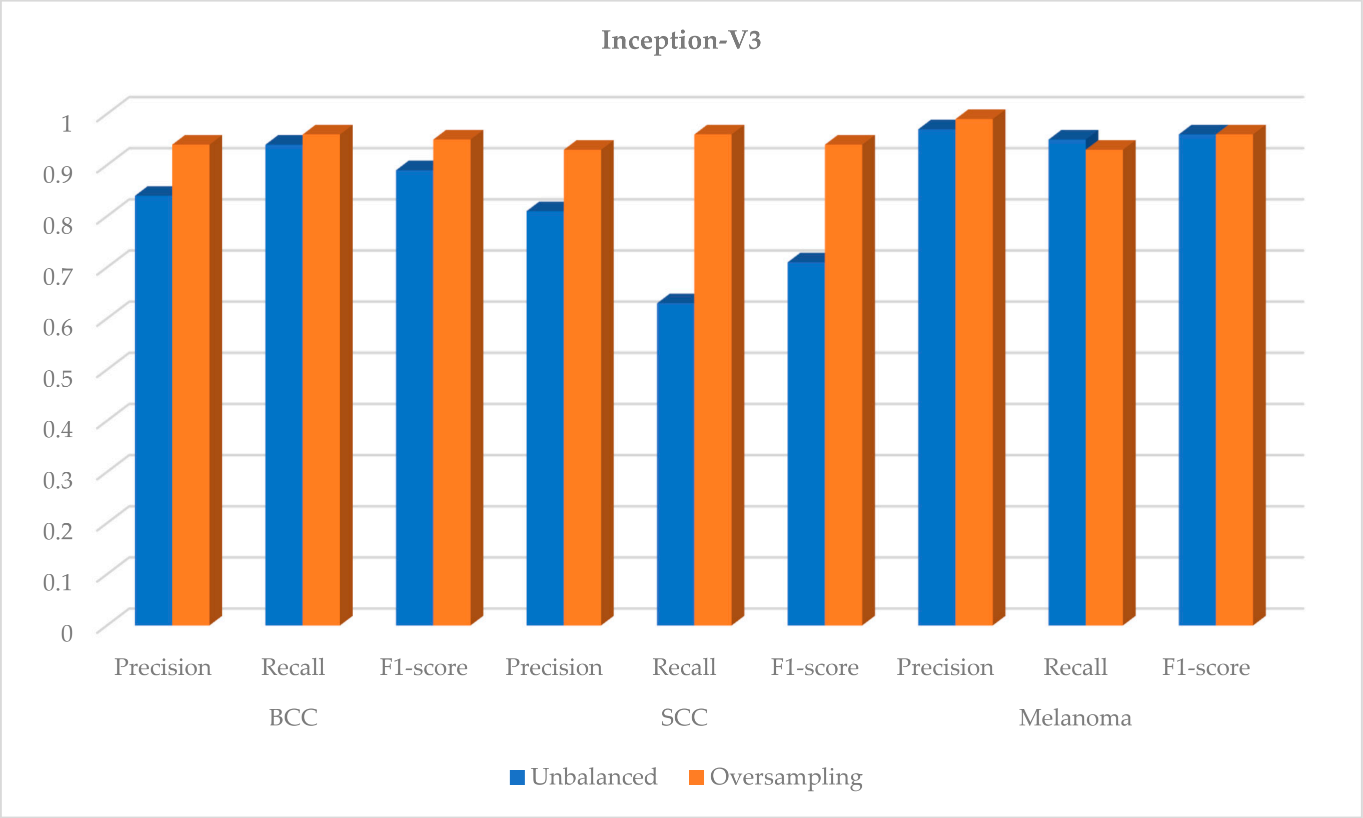

Figure 8. Results of Inception-V3 model on unbalanced and balanced datasets.

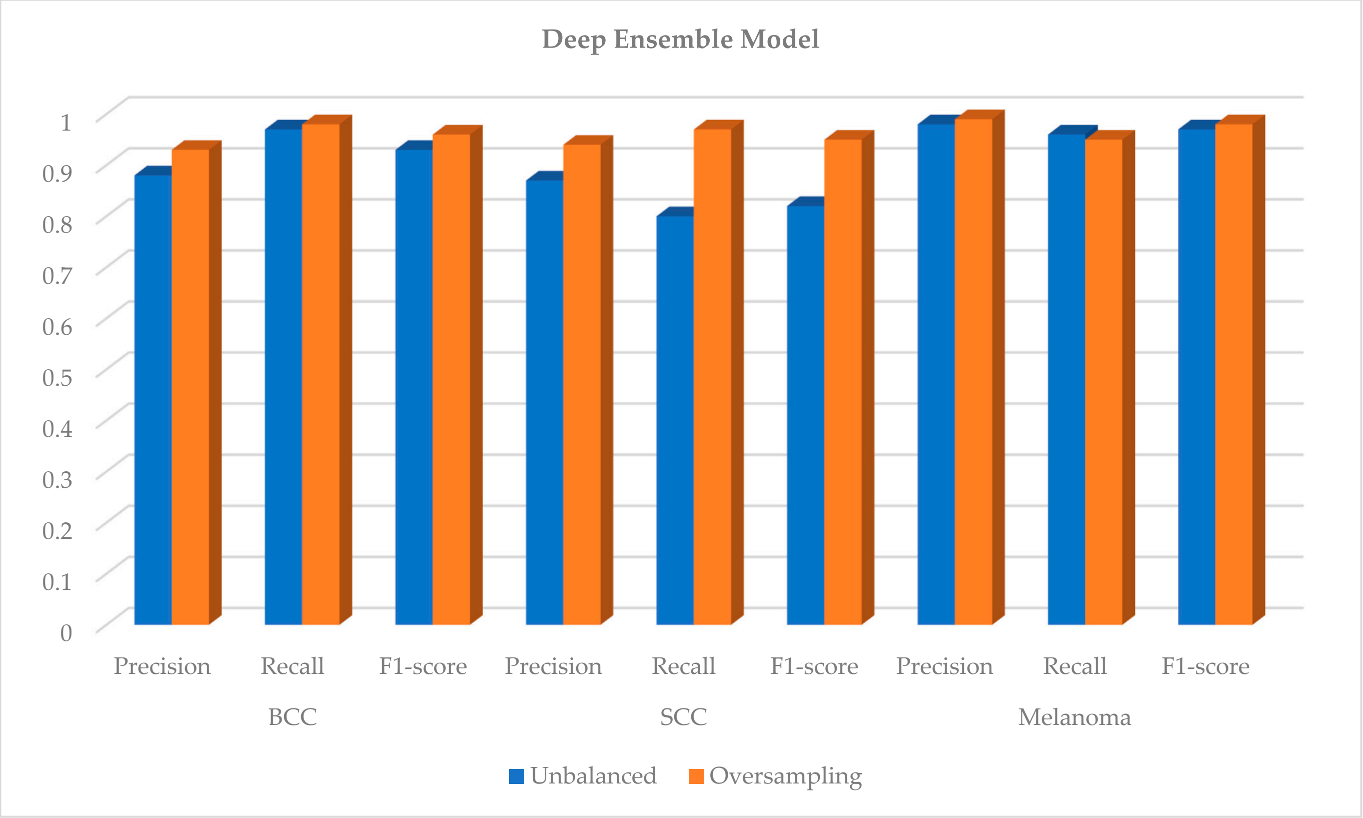

Figure 9. Results of deep ensemble model on unbalanced and balanced datasets.

Table 3. Sensitivity and Specificity of each class.

	BCC	**SCC**	**Melanoma**
Sensitivity	0.85	0.83	0.91
Specificity	0.93	0.96	0.97

Figure 10 presents the comparative accuracy results using the original unbalanced dataset. Figure 11 describes the comparative accuracy results using the dataset balanced through oversampling.

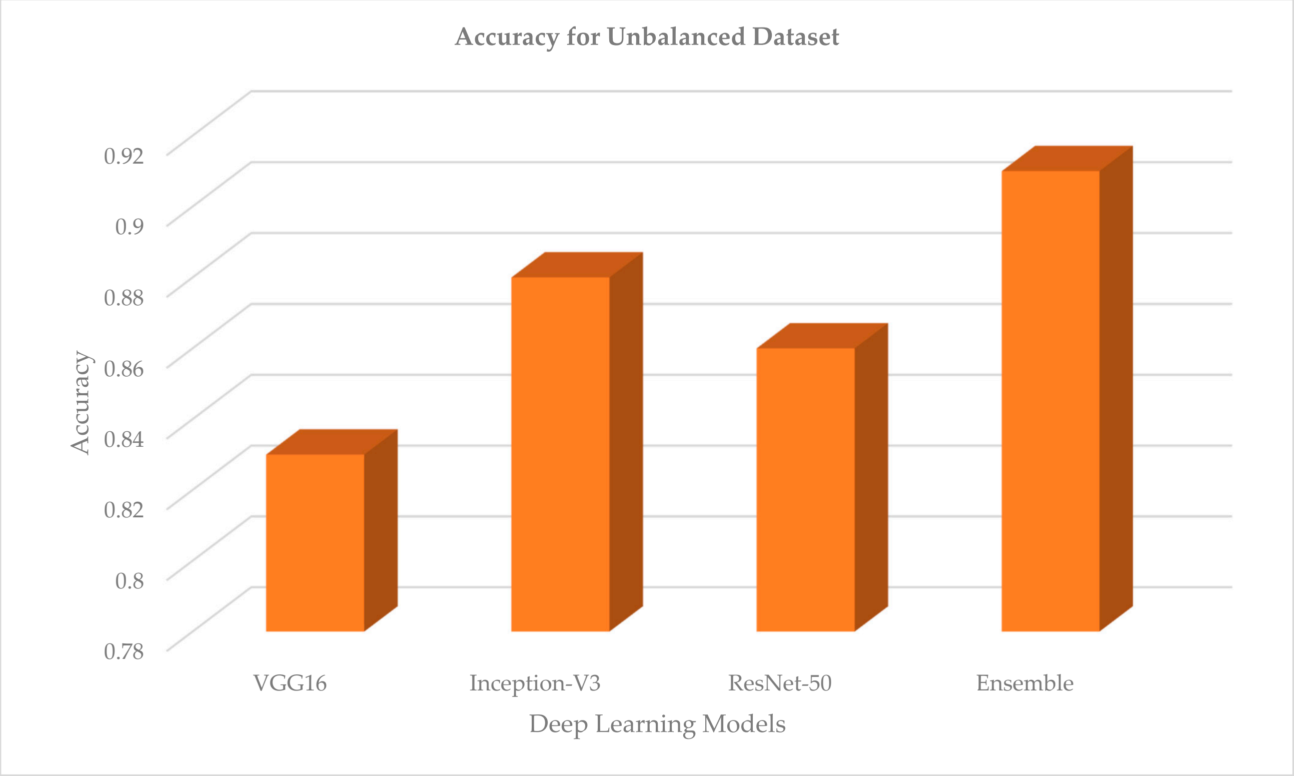

Figure 10. Accuracy results on original unbalanced dataset.

The performance evaluation of the skin lesion classification system reveals that the deep ensemble model comprising VGG16, ResNet-50, and Inception-V3 significantly outperforms each individual model on both the original ISIC dataset and the balanced ISIC dataset achieved through oversampling. For the original dataset, the ensemble model achieves an accuracy of 91%, compared to 83% for VGG16, 86% for ResNet-50, and 88% for Inception-V3. On the balanced dataset, the ensemble model further improves with an accuracy of 97%, demonstrating its robustness in handling class imbalance.

Then, the system is evaluated using the HAM10000 dataset containing seven categories: melanocytic nevi (nv): benign melanocytic nevi or moles; melanoma (mel): malignant melanoma; benign keratosis-like lesions (bkl): solar lentigines/seborrheic keratoses and lichenoid keratoses; basal cell carcinoma (bcc): a common form of skin cancer; actinic keratoses (akiec): a precancerous skin lesion that can develop into squamous cell carcinoma; vascular lesions (vasc): angiomas, angiokeratomas, and pyogenic granulomas; and dermatofibroma (df): a benign fibrous nodule. From this HAM10000 dataset, images of 860 melanoma, 327 actinic keratosis squamous cell carcinoma, and 513 basal cell carcinoma (BCC) cases were used for system evaluation. Figure 12 presents the comparative accuracy results using the original unbalanced dataset. Figure 13 describes the comparative accuracy results using the dataset balanced through oversampling.

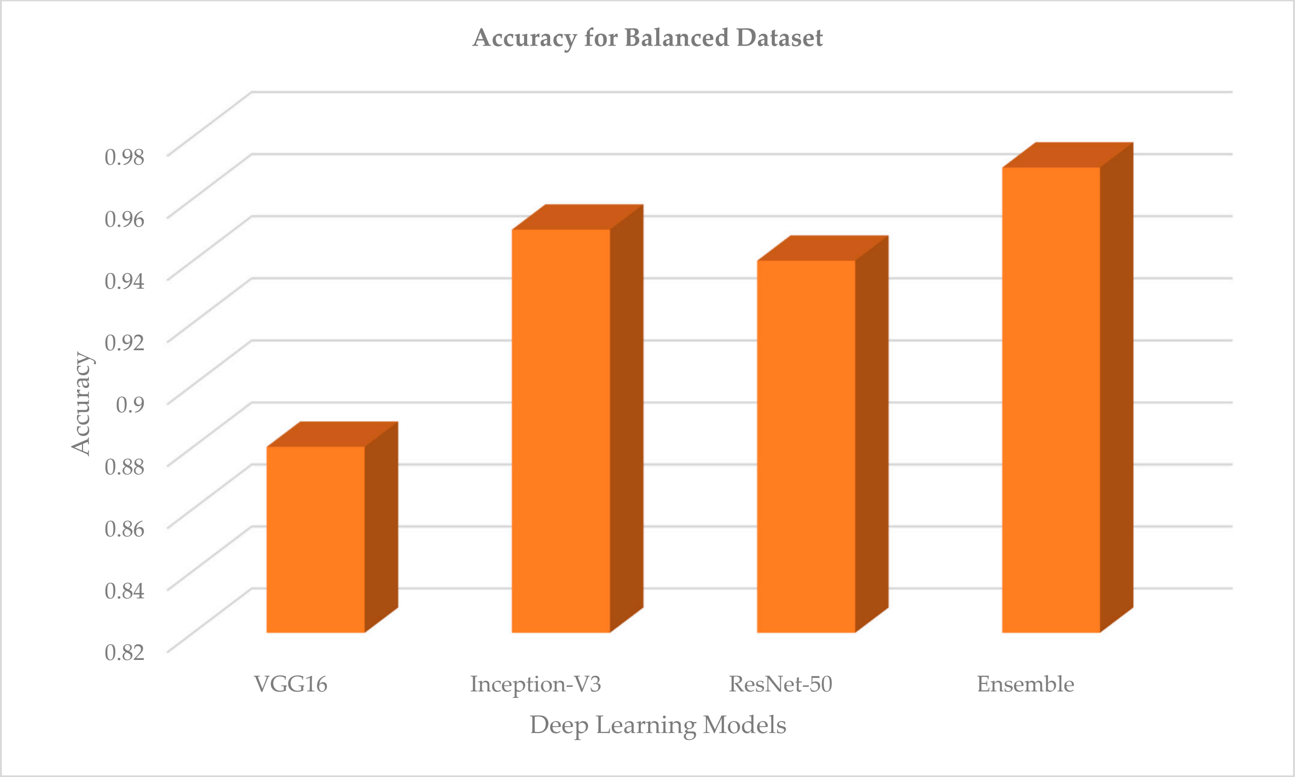

Figure 11. Accuracy results on balanced dataset.

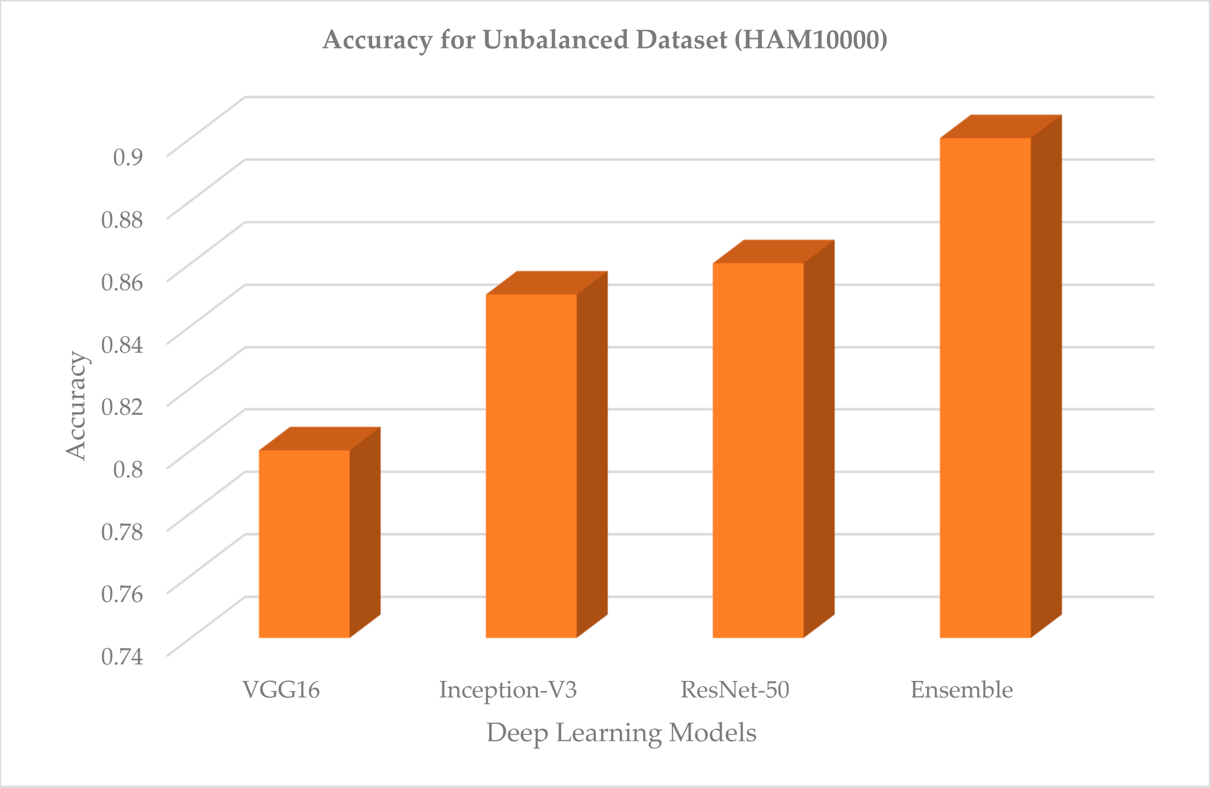

Figure 12. Accuracy results on original unbalanced HAM10000 dataset.

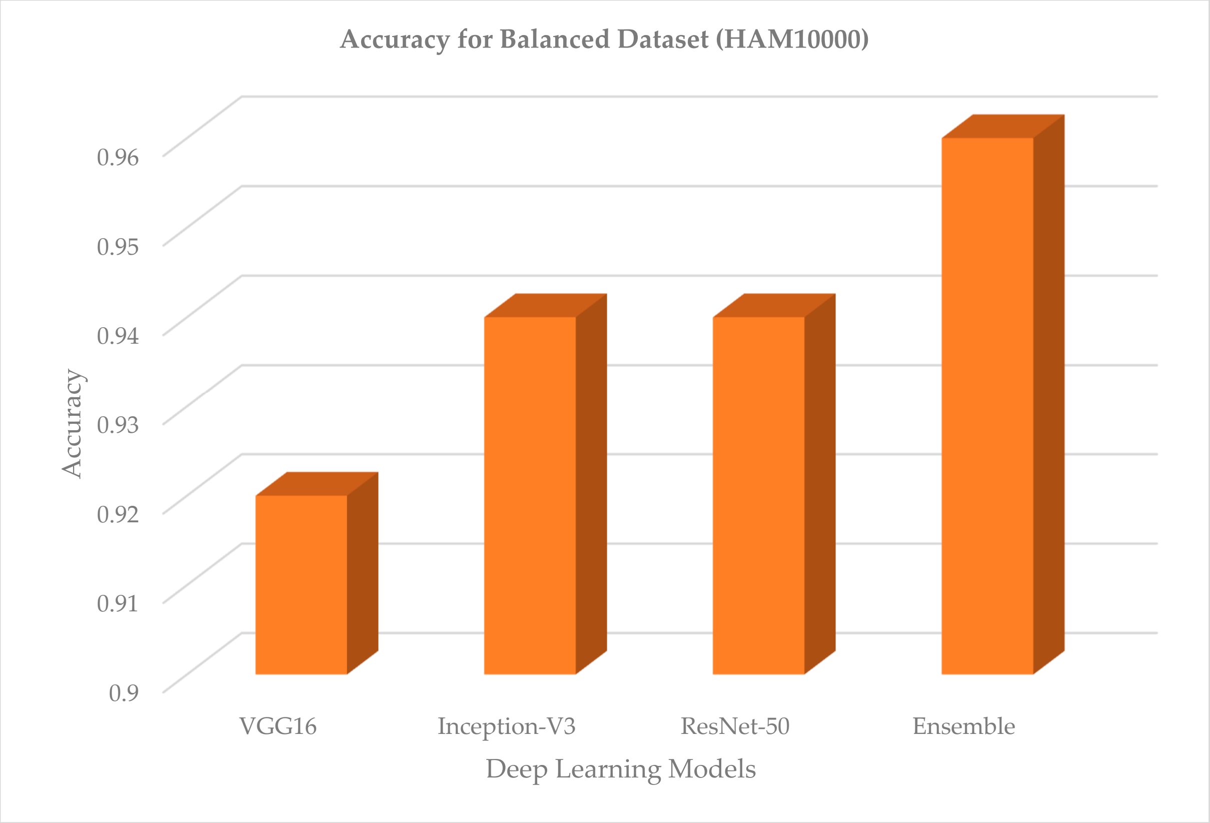

Figure 13. Accuracy results on balanced HAM10000 dataset.

For the original dataset, the ensemble model achieves an accuracy of 90%, compared to 80% for VGG16, 86% for ResNet-50, and 85% for Inception-V3. On the balanced dataset, the ensemble model further improves with an accuracy of 96%, demonstrating its robustness in handling class imbalance. Table 4 shows the comparison with other cutting-edge techniques. These results highlight the ensemble model's superior capability in leveraging diverse model strengths, resulting in more reliable and accurate classification for both imbalanced and balanced datasets.

Table 4. Comparison with other cutting-edge techniques.

Ref.	Diagnosis	Classification Methods	Dataset	Accuracy (%)
[15]	Malignant and benign	CNN ReNet-50 Inception-V3 Inception Resnet	ISIC 2018	CNN (83.2) ResNet-50 (83.7) Inception-V3 (85.8) Inception Resnet (84)
[16]	Melanoma and non-melanoma lesions	AlexNet	ISIC 2018	84
[17]	Malignant melanoma and benign tumors	U-Net	ISIC 2018	80.06
	BCC, SCC, and melanoma	Ensemble model of VGG16, ResNet-50, and Inception-V3	ISIC 2018	97

In comparison with other techniques, our proposed system provides more accurate results than other techniques.

5. Conclusions and Future Research Directions

The skin constitutes a vital aspect of the human body, rendering skin cancer among the most prevalent illnesses. Advancements in computer-based methodologies have notably expedited and streamlined the process of diagnosing skin cancer. A range of noninvasive approaches has been introduced to assess indications of skin cancer. In this study, we developed a robust skin lesion classification system utilizing a deep ensemble model comprising VGG16, ResNet-50, and Inception-V3. Our approach was evaluated using both the original ISIC dataset and a balanced version of the ISIC dataset achieved through oversampling. The ensemble model demonstrated significant improvements in classification performance compared to each individual model, achieving higher accuracy, precision, recall, and F1-scores. On the original ISIC dataset, the ensemble model achieved an accuracy of 91%. When applied to the balanced dataset, the ensemble model further enhanced its performance, attaining an accuracy of 97%. Also, with the original HAM10000 dataset, the ensemble model achieved an accuracy of 90%. When applied to the balanced dataset, the ensemble model further enhanced its performance, attaining an accuracy of 96%. These findings emphasize the efficacy of the ensemble methodology in harnessing the unique strengths of each individual model, thereby yielding a more dependable and precise system for classifying skin lesions. Moreover, the ensemble model's capability to address class imbalance through oversampling underscores its resilience and potential for clinical utilization in the early detection and diagnosis of skin cancer.

Deploying a deep ensemble model for skin lesion detection in clinical settings involves several key challenges that must be addressed for effective implementation. Ensuring data privacy and security is paramount, requiring robust encryption, access control, anonymization techniques, and compliance with regulatory standards like HIPAA and GDPR. Seamless integration with existing clinical workflows is essential, necessitating compatibility with electronic health records (EHRs), user-centric design, comprehensive training, and pilot testing to minimize disruptions. Model interpretability is critical for clinician trust, which can be achieved through explainable AI techniques, visualization tools, and clinical validation. Scalability and performance are ensured through model optimization, powerful hardware, cloud computing, and distributed computing solutions. Regulatory compliance requires extensive clinical trials, collaboration with regulatory bodies, and detailed documentation. Continuous learning systems and regular updates are necessary to keep the model current with evolving medical data. Comprehensive training programs, user manuals, and helpdesk support are crucial for user adoption and effectiveness. Ethical considerations must be upheld through ethical guidelines, minimizing biases with diverse training data and maintaining transparency about the model's capabilities and limitations. Addressing these challenges with detailed solutions will facilitate the successful deployment of the skin detection model in clinical environments, enhancing patient care and diagnostic accuracy.

It is crucial to address these instances to gain a comprehensive understanding of the model's limitations and its correlation with assessments made by human experts in challenging cases. Although our study primarily emphasized overall accuracy and performance metrics, we recognize the importance of scrutinizing cases where the model encounters difficulties. These challenges often involve lesions with intricate characteristics, subtle variations, or ambiguous features that pose obstacles for both automated systems and dermatologists. To delve deeper into these cases, our next step involves conducting a thorough analysis to determine if the images flagged by the model as challenging align with those dermatologists would also find difficult to classify. This analysis will entail collaborating closely with dermatologists to glean clinical insights and expert opinions on these specific cases. Furthermore, we intend to strengthen our model by incorporating additional data augmentation techniques, refining feature extraction methods, and exploring ensemble strategies. These enhancements will aim to bolster the model's ability to handle complex scenarios more effectively. By openly discussing and analyzing these findings, we aim to reinforce the reliability and efficacy of our skin lesion classification system. Looking ahead, our commitment remains steadfast in ongoing research and development endeavors aimed at overcoming these challenges. Our ultimate objective is to advance the diagnostic capabilities of AI-driven systems in dermatology, ensuring they align with rigorous clinical standards and practices.

Future research on the skin lesion classification system using the deep ensemble model of VGG16, ResNet-50, and Inception-V3 could explore several promising directions. Firstly, further optimization of the ensemble model can be undertaken by incorporating additional advanced deep learning architectures and exploring different ensemble strategies to enhance performance. Additionally, integrating more diverse datasets and employing cross-dataset validation could improve the model's generalizability and robustness across various clinical settings. Another potential direction is the refinement of the oversampling technique. More sophisticated data augmentation methods, such as synthetic data generation using Generative Adversarial Networks (GANs), could be explored to better address class imbalance and enhance model training. Furthermore, incorporating multimodal data, including patient history and other diagnostic information, could provide a more comprehensive analysis and improve diagnostic accuracy. Lastly, deploying the model in real-world clinical environments and conducting longitudinal studies would be crucial to assess its practical utility and impact on patient outcomes. Feedback from dermatologists and continuous model updates based on new data will be essential for maintaining the system's relevance and efficacy.

Author Contributions: Conceptualization, S.M.T. and H.-S.P.; methodology, S.M.T.; software, S.M.T. and H.-S.P.; validation, formal analysis, and investigation, S.M.T.; resources, H.-S.P.; data curation, H.-S.P.; writing—original draft preparation, S.M.T.; writing—review and editing, H.-S.P.; visualization, S.M.T.; supervision, H.-S.P.; funding acquisition, H.-S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://www.kaggle.com/datasets/mnowak061/isic2018-and-ph2-384x384-jpg (accessed on 10 April 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Han, H.S.; Choi, K.Y. Advances in nanomaterial-mediated photothermal cancer therapies: Toward clinical applications. *Biomedicines* **2021**, *9*, 305. [CrossRef] [PubMed]
2. Fuzzell, L.N.; Perkins, R.B.; Christy, S.M.; Lake, P.W.; Vadaparampil, S.T. Cervical cancer screening in the United States: Challenges and potential solutions for underscreened groups. *Prev. Med.* **2021**, *144*, 106400. [CrossRef] [PubMed]
3. Jinzaki, M.; Yamada, Y.; Nagura, T.; Nakahara, T.; Yokoyama, Y.; Narita, K.; Ogihara, N.; Yamada, M. Development of upright computed tomography with area detector for whole-body scans: Phantom study, efficacy on workflow, effect of gravity on human body, and potential clinical impact. *Investig. Radiol.* **2020**, *55*, 73. [CrossRef] [PubMed]
4. Adegun, A.; Viriri, S. Deep learning techniques for skin lesion analysis and melanoma cancer detection: A survey of state-of-the-art. *Artif. Intell. Rev.* **2021**, *54*, 811–841. [CrossRef]
5. Iqbal, S.; Siddiqui, G.F.; Rehman, A.; Hussain, L.; Saba, T.; Tariq, U.; Abbasi, A.A. Prostate cancer detection using deep learning and traditional techniques. *IEEE Access* **2021**, *9*, 27085–27100. [CrossRef]
6. Vaishnavi, K.; Ramadas, M.A.; Chanalya, N.; Manoj, A.; Nair, J.J. Deep learning approaches for detection of COVID-19 using chest X-ray images. In Proceedings of the 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), Piscataway, NJ, USA, 15–17 September 2021.
7. Duc, N.T.; Lee, Y.M.; Park, J.H.; Lee, B. An ensemble deep learning for automatic prediction of papillary thyroid carcinoma using fine needle aspiration cytology. *Expert Syst. Appl.* **2022**, *188*, 115927. [CrossRef]
8. MNOWAK061. Skin Lesion Dataset. ISIC2018 Kaggle Repository. 2021. Available online: https://www.kaggle.com/datasets/mnowak061/isic2018-and-ph2-384x384-jpg (accessed on 10 April 2022).
9. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*, 3rd ed.; Elsevier Science Ltd.: New York, NY, USA, 2011; pp. 1–703.
10. Rasel, M.A.; Obaidella, H.U.; Kareem, S.A. Convolutional Neural Network-Based Skin Lesion Classification with Variable Nonlinear Activation Functions. *IEEE Access* **2022**, *10*, 83398–83414. [CrossRef]
11. Albahar, M.A. Skin lesion classification using convolutional neural network with novel Regularizer. *IEEE Access* **2019**, *7*, 38306–38313. [CrossRef]

12. Salian, A.C.; Vaze, S.; Singh, P. Skin Lesion Classification using Deep Learning Architectures. In Proceedings of the 2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA), IEEE, Mumbai, India, 3–4 April 2020; pp. 168–173.
13. Alkarakatly, T.; Eidhah, S.; Sarawani, M.A.; Sobhi, A.A.; Bilal, M. Skin Lesions Identification Using Deep Convolutional Neural Network. In Proceedings of the 2019 International Conference on Advances in the Emerging Computing Technologies (AECT), Al Madinah Al Munawwarah, Saudi Arabia, 10 February 2020; pp. 209–213.
14. Filali, Y.; Khoukhi, H.E.; Sabri, M.A. Texture Classification of skin lesion using convolutional neural network. In Proceedings of the 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morocco, 3–5 April 2019.
15. Gouda, W.; Sama, N.U.; Waakid, G.A. Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning. *Healthcare* **2022**, *10*, 1183. [CrossRef] [PubMed]
16. Ameri, A. A deep learning approach to skin cancer detection in dermoscopy images. *J. Biomed. Phys. Eng.* **2020**, *10*, 801. [CrossRef] [PubMed]
17. Kim, C.I.; Hwang, S.M.; Park, E.B.; Won, C.H.; Lee, J.H. Computer-Aided Diagnosis Algorithm for Classification of Malignant Melanoma Using Deep Neural Networks. *Sensors* **2021**, *21*, 5551. [CrossRef] [PubMed]
18. Rajput, G.; Agrawal, S.; Raut, G.; Vishvakarma, S.K. An accurate and noninvasive skin cancer screening based on imaging technique. *Int. J. Imaging Syst. Technol.* **2022**, *32*, 354–368. [CrossRef]
19. Ali, M.S.; Miah, M.S.; Haque, J.; Rahman, M.M.; Islam, M.K. An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Mach. Learn. Appl.* **2021**, *5*, 100036. [CrossRef]
20. Murugan, A.; Nair, S.A.H.; Preethi, A.A.P.; Kumar, K.S. Diagnosis of skin cancer using machine learning techniques. *Microprocess. Microsyst.* **2021**, *81*, 103727. [CrossRef]
21. Nawaz, M.; Mehmood, Z.; Nazir, T.; Naqvi, R.A.; Rehman, A.; Iqbal, M.; Saba, T. Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering. *Microsc. Res. Tech.* **2022**, *85*, 339–351. [CrossRef] [PubMed]
22. Brownlee, J. *A Gentle Introduction to Transfer Learning for Deep Learning*; Machine Learning Mastery: San Juan, PR, USA, 16 September 2019.
23. Ummapure, V.A.; Navya, R.; Desai, K. Skin Disease Detection Using VGG16 and InceptionV3. *Int. J. Comput. Appl.* **2023**, *184*, 27–32.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385. [CrossRef]
25. Harangi, B.; Baran, A.; Hajdu, A. Assisted deep learning framework for multi-class skin lesion classification considering a binary classification support. *Biomed. Signal Process. Control.* **2020**, *62*, 102041. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.