Research Article

Enhanced Multi-Class Skin Lesion Classification of Dermoscopic Images Using an Ensemble of Deep Learning Models

Kyi Pyar Zaw * and Atar Mon

Faculty of Electronic Engineering, University of Technology, (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar; e-mail : kyipyarzaw89@gmail.com; atarmon@gmail.com
* Corresponding Author : Kyi Pyar Zaw

Abstract: This study presents an advanced approach to multi-class skin lesion classification by leveraging an ensemble model comprising the Inception-V3, ResNet-50, and VGG16 architectures. The classification task focuses on categorizing skin lesions into distinct classes, including Melanoma, basal cell carcinoma (BCC), and squamous cell carcinoma (SCC), using the ISIC dataset, a comprehensive collection of dermoscopic images. In order to properly balance the dataset, the oversampling strategy is utilized, as some lesion types are underrepresented due to inherent imbalances in the dataset. By ensuring that the model is trained on a more representative dataset, this balancing improves the algorithm's capacity to categorize all lesion types properly and impartially. By combining the complementary features of ResNet-50, Inception-V3, and VGG16, the ensemble technique improves the overall classification performance. ResNet-50 is chosen for its deep feature extraction capabilities, which help capture fine details in lesion patterns. Inception-V3 is selected for its multi-scale processing, allowing it to effectively analyze lesions at varying resolutions and sizes. VGG16 is included due to its simple yet highly effective architecture for image classification tasks. The ensemble model with data augmentation significantly outperforms individual models in skin lesion classification for both the original and balanced ISIC datasets regarding accuracy, precision, recall, and F1-score. This method offers a robust solution for skin lesion classification, contributing to more accurate and reliable diagnostic tools in dermatology.

Keywords: Inception-V3; Oversampling; ResNet-50; Skin lesion classification; VGG16.

Received: September, 26th 2024
Revised: November, 10th 2024
Accepted: November, 11th 2024
Published: November, 13th 2024

1. Introduction

The three most prevalent forms of skin cancer, which is among the most common cancers worldwide, are Melanoma, BCC, and SCC. Early and precise diagnosis is essential to improve patient outcomes and enable effective treatment. The identification of skin lesions has historically depended mostly on dermatologists' knowledge, which can be laborious and subjective[1]–[3]. The increasing availability of dermoscopic images, combined with advancements in deep learning, has opened new avenues enabling automated skin lesion classification, which could help medical professionals identify patients more quickly and accurately.

This work focuses on classifying skin lesions into many classes using an ensemble of three cutting-edge: Inception-V3, VGG16, and ResNet-50. Each model has demonstrated significant success in various image classification tasks due to their distinct architectural advantages[4], [5]. VGG16 is known for its simplicity and depth, ResNet-50 for its ability to train very deep networks through residual learning, and Inception-V3 for its efficient multi-scale processing. By combining these models in an ensemble, we aim to use their complementary abilities to increase classification accuracy for various skin lesion classes.

A significant challenge in this domain is the imbalance in the available datasets, particularly the ISIC dataset, which contains a disproportionate number of images for each lesion type. The model is more likely to accurately classify the majority classes while underperforming the minority classes due to this imbalance, which might result in biased model predictions.

DOI : 10.62411/jcta.11530

To address this issue, an oversampling technique is applied to balance the dataset, ensuring that each class is equally represented during training. This approach helps mitigate bias and enhances the model's generalization ability across different lesion types.

Integrating the ensemble model with a balanced dataset through oversampling represents a novel approach to improving the accuracy and reliability of automated skin lesion classification. By providing a more equitable representation of lesion types and harnessing the power of multiple deep-learning models, this study aims to develop more effective diagnostic tools that can assist dermatologists in clinical practice. The contribution of this paper is:

1. Introduced an ensemble method combining VGG16, ResNet-50, and Inception-V3 models, leveraging their complementary strengths in feature extraction, depth, and multi-scale processing for enhanced skin lesion classification accuracy.
2. Tackled the challenge of class imbalance in the ISIC dataset by implementing an oversampling technique, ensuring equitable representation of all skin lesion types during training.
3. Demonstrated that the ensemble model, trained on the balanced dataset, outperforms individual CNN models and traditional methods with evaluation criteria.
4. Contributed to developing more accurate and reliable automated diagnostic tools, potentially assisting dermatologists in making faster and more precise diagnoses, thereby improving patient outcomes.

The remaining sections of this document are organized as follows: the relevant literature is explored in Section 2. An explanation of the suggested system architecture and its constituent parts is provided in Section 3. The performance evaluation methodology is then described in Section 4, which also details the experimental design and the metrics used. Section 5 summarizes the study's main conclusions and closing thoughts on the investigation as a whole.

2. Literature Review

2.1. Related Works

Previous studies highlight various deep-learning approaches to improve skin lesion classification. Study [6] introduced a knowledge distillation framework where a simpler student model learned from a more complex teacher model, significantly enhancing melanoma classification, especially for the minority class. Study [7] presented a CNN integrated with soft-attention mechanisms, allowing the model to focus on critical dermatoscopic image segments, improving accuracy despite lesion variability. Study [8] combined an Extreme Learning Machine (ELM) with Teaching-Learning-Based Optimization (TLBO) to optimize parameter selection, resulting in improved classification accuracy, precision, recall, and F1-score for skin cancer detection, outperforming traditional machine learning models.

Prior studies explore advanced deep-learning techniques for skin lesion classification. A study [9] conducted a detailed analysis using dermoscopy images, comparing CNN architectures to identify the most effective model for distinguishing Melanoma from benign lesions. The deep CNN approach outperformed traditional methods, achieving high accuracy, precision, recall, and F1-score, demonstrating robustness across various scenarios. Study [10] employed the Inception-ResNet architecture, which excelled in feature extraction and learning efficiency, outperforming conventional and deep learning methods for melanoma detection with high classification accuracy. Study [11] utilized an adversarial framework with attention mechanisms to merge clinical and dermoscopic images, significantly improving classification performance by preserving key features from both modalities, resulting in superior accuracy, precision, recall, and F1-score.

Earlier research presents various advanced methods for skin lesion classification. Study [3] developed a deep learning model to assess lesion symmetry, which is crucial for melanoma diagnosis, achieving significant improvements in accuracy, precision, recall, and F1-score over traditional methods by effectively distinguishing between symmetrical and asymmetrical lesions. Study [12] introduced the CS-AF framework, combining multiple classifiers with a cost-sensitive approach to minimize misclassification errors, significantly improving accuracy, precision, recall, and F1-score while reducing false negatives. Study [13] systematically reviewed AI methodologies in skin cancer detection, highlighting the effectiveness of CNNs, SVMs, and ensemble models in improving diagnostic precision through advanced deep learning techniques and integration with imaging technologies. Study [14] explored various deep learning models, including CNNs, for analyzing dermoscopic images, achieving substantial performance improvements and high accuracy in classifying skin lesions, particularly excelling in recall and early detection. The author utilized [15] the K-nearest neighbor (KNN) classifier combined with the Gray Level Co-occurrence Matrix (GLCM) for classifying two types of skin cancer, with an average filter applied for pre-processing. A comprehensive analysis was performed on the ISIC dataset through 480 experiments, testing various dataset sizes using random sampling techniques with 3297, 1649, 825, and 210 images. Different KNN parameters, including the number of neighbors (k=1) and distance metrics (d=1 to 3), were evaluated at angles of 0, 45, 90, and 135 degrees. The maximum accuracy achieved was 79.24%, 79.39%, 83.63%, and 100% for the respective dataset sizes. An advanced automated system using Deep Neural Networks [16], specifically MobileNetV2, was presented to detect acute lymphoblastic leukemia (ALL) blast cells in microscopic blood smear images, achieving an impressive 97% accuracy. The system shows high sensitivity and specificity in identifying multiple ALL sub-types. Additionally, the study introduces innovative telediagnosis software that provides real-time support for clinicians, enabling prompt and accurate diagnosis of ALL subtypes from blood smear images.

Despite advances in deep learning for skin lesion classification, several critical research gaps remain. Current models often focus on binary classification, neglecting the complexity of multi-class problems involving Melanoma, BCC, and SCC, leading to suboptimal generalization across lesion types. Additionally, the imbalance in medical datasets skews model performance, particularly for underrepresented classes, with inadequate balancing strategies applied. While effective in specific cases, single-model approaches fail to capture the full variability of dermoscopic images, necessitating an ensemble approach to leverage diverse model strengths. Limited application of advanced data augmentation techniques restricts generalizability across diverse patient populations, and the lack of focus on model efficiency hampers real-time clinical deployment. This study addresses these gaps by proposing an ensemble model with enhanced multi-class performance, balanced training data, robust augmentation, and optimizations for clinical relevance.

2.2 Oversampling

One method for addressing the class imbalance in datasets—where some classes are underrepresented compared to others—is oversampling [17]. This is particularly important in machine learning and statistical modeling, where imbalanced data can lead to biased models that perform poorly on the minority class. Oversampling involves generating additional synthetic samples or duplicating existing samples of the minority class to create a more balanced dataset. One common approach is the Synthetic Minority Over-sampling Technique (SMOTE), which creates synthetic samples by interpolating between existing minority class instances. Another method, called Random Oversampling, involves duplicating minority class instances until the desired balance is achieved. This system applies random oversampling. By employing these techniques, the model can learn more effectively from the minority class, thereby improving its performance and generalization. By preventing the model from becoming biased in favor of the majority class and improving its ability to identify and categorize underrepresented instances, oversampling contributes to producing more accurate and dependable predictions. Suppose the original number of minority class samples is n_{minority}, and the total number of samples required for balance is n_{target}. In that case, the number of samples to generate is used in Equation (1).

$$n_{\text{generate}} = n_{\text{target}} - n_{\text{minority}}, \tag{1}$$

In random oversampling, the new dataset S_{new} is constructed as Equation (2).

$$S_{\text{new}} = S_{\text{minority}} \cup \left(\text{Randomly duplicated samples from } S_{\text{minority}}\right) \tag{2}$$

2.3. VGG16

The Visual Geometry Group at the University of Oxford developed the well-known deep learning model VGG16 [18] for picture categorization applications. Because of its deep convolutional neural network design, it is highly renowned for being user-friendly and efficient. The model is a deep network that can recognize intricate patterns in images since it has 16 layers, 13 convolutional layers, and three fully linked layers. The convolutional layers use small 3x3 filters and are followed by max-pooling layers, which help reduce the spatial dimensions of the input while retaining essential features. VGG16's architecture emphasizes using smaller convolutional kernels and a deep network to enhance feature extraction capabilities. Because of its capacity to learn rich, hierarchical representations of visual input, the model has had a significant impact on computer vision. It has demonstrated outstanding performance on a number of benchmark datasets, including ImageNet. Its design has inspired numerous subsequent models and is a valuable tool for image classification and transfer learning applications. Its design is shown in Figure 1.

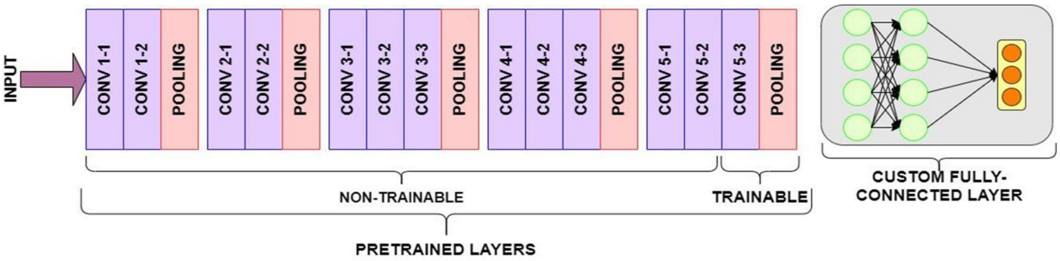

Figure 1. VGG16 Architecture [19]

2.4. ResNet-50

ResNet-50 [20] is a deep convolutional neural network architecture designed to address the challenges of training very deep networks. Developed as part of the Residual Networks (ResNet) family by Microsoft Research, ResNet-50 stands out for its creative use of skip connections, also known as residual connections, which lessen the effects of the vanishing gradient issue and increase training effectiveness. The "50" in ResNet-50 indicates that the network consists of 50 layers, making it a relatively deep network that balances complexity and performance. The core idea behind ResNet-50 is to introduce residual blocks that allow the network to learn residual mappings instead of the original unreferenced map-pings. These blocks consist of convolutional layers with shortcut connections that bypass one or more layers, facilitating the flow of gradients during backpropagation. Because of its design, ResNet-50 can learn intricate features and patterns from massive amounts of data, leading to impressive accuracy on benchmark datasets like ImageNet. The architecture's ability to maintain high performance with increasing depth has made ResNet-50 a well-liked option for many computer vision applications, such as segmentation, object detection, and picture classification. It is described in Figure 2.

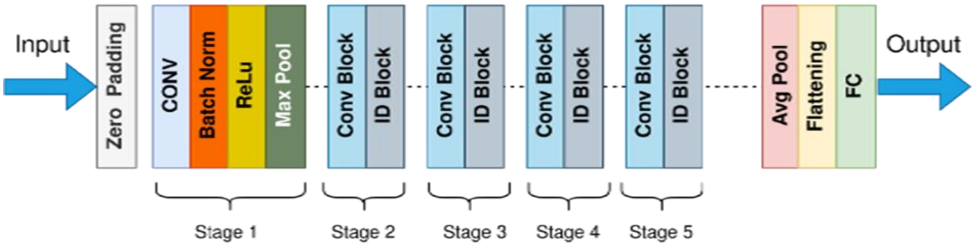

Figure 2. ResNet-50 Architecture[21]

2.5. Inception-V3

Inception-V3[22] is a sophisticated convolutional neural network architecture developed by Google, known for its efficiency and high performance in image classification tasks. It is part of the Inception series, which focuses on optimizing the depth and width of neural networks to achieve better computational efficiency. Inception-V3 introduces several key innovations, including inception modules that employ multiple convolutional filter sizes and pooling operations within the same layer. This allows the network to capture various features at different scales. Additionally, Inception-V3 incorporates batch normalization and factorized convolutions to improve training speed and model performance.

Moreover, the network gains from dimensionality reduction by employing 1x1 convolutions, which lower computational costs and parameter counts. These enhancements enable Inception-V3 to achieve state-of-the-art accuracy on benchmark datasets like ImageNet while maintaining a relatively low computational footprint. Its effectiveness has made it a well-liked option for several computer vision applications, including transfer learning, object identification, and picture categorization[23]. Figure 3 presents its design.

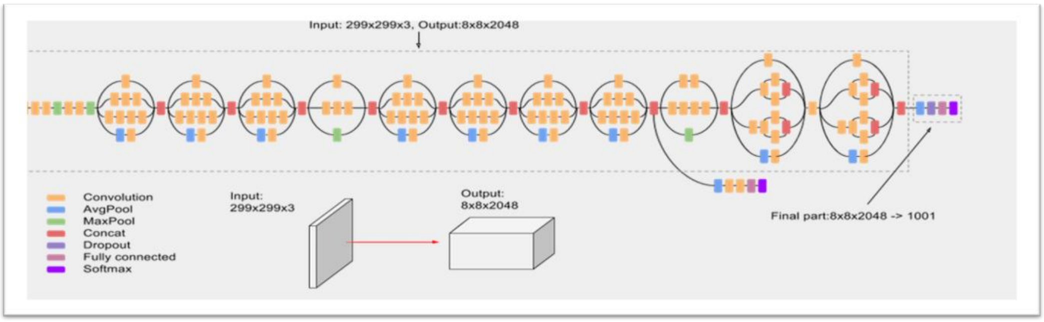

Figure 3. Inception-V3 Architecture[24]

3. Proposed Method

The proposed method introduces a novel ensemble approach that combines the strengths of multiple deep learning models—ResNet-50, Inception-V3, and VGG16—to address the limitations of single-model architectures in multi-class skin lesion classification. This ensemble technique improves robustness and accuracy, particularly for underrepresented lesion types like squamous cell carcinoma (SCC). Unlike conventional methods, this approach integrates advanced oversampling to mitigate class imbalance and emphasizes clinically relevant metrics such as recall and specificity over accuracy alone.

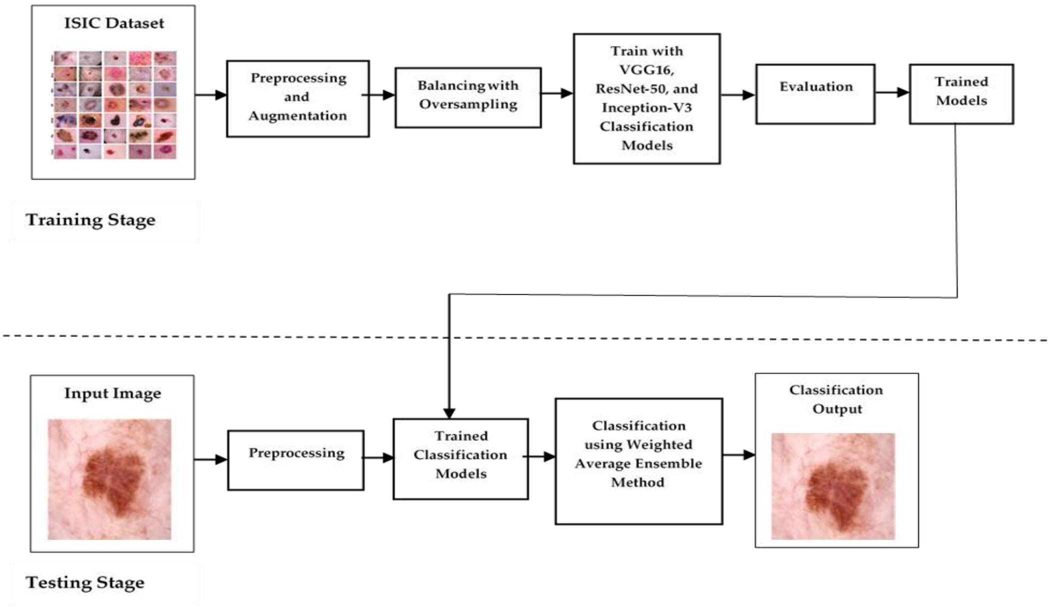
ISIC Dataset
Preprocessing and Augmentation
Balancing with Oversampling
Train with VGG16, ResNet-50, and Inception-V3 Classification Models
Evaluation
Trained Models
Training Stage
Input Image
Preprocessing
Trained Classification Models
Classification using Weighted Average Ensemble Method
Classification Output
Testing Stage

Figure 4. System Design Overview

Additionally, enhanced data augmentation techniques improve generalization across diverse patient populations. It aims to achieve high accuracy and robustness in identifying various skin lesions. The system employs a sophisticated ensemble approach, integrating three well-established convolutional neural networks— VGG16, ResNet-50, and Inception-V3—to take advantage of each one's advantages and enhance classification performance. VGG16's deep and consistent structure allows it to effectively identify subtle patterns and features crucial for differentiating between various skin lesions. By creatively utilizing residual connections to lessen the vanishing gradient issue in deep networks, ResNet-50 enhances the system.

ResNet-50 allowed it to maintain performance and learn complex features from deep layers, enhancing its ability to accurately classify skin lesions by preserving important information and gradients during training. The multi-scale feature extraction enables Inception-V3 to capture diverse patterns and textures in the skin images, improving its ability to recognize different lesion types. Its efficient use of factorized convolutions and dimensionality reduction balances computational efficiency and model performance.

The proposed system incorporates an oversampling technique to address the challenge of class imbalance in the ISIC dataset. This approach generates synthetic samples for underrepresented classes, thereby balancing the dataset and ensuring that all classes are equally represented. By doing so, the system mitigates the risk of biased predictions and enhances the reliability of the classification results. The VGG16, ResNet-50, and Inception-V3 outputs are combined with the ensemble model's predictions to yield a more dependable and accurate categorization. By utilizing each network's distinct capabilities, this integration improves diagnostic performance for classifying many skin lesions. The system design of this research is shown in Figure 4.

3.1. Data Collection

The process begins by collecting images from the International Skin Imaging Collaboration (ISIC) dataset available on Kaggle[25], focusing on three types of skin lesions: melanoma (438 images), basal cell carcinoma (376 images), and squamous cell carcinoma (181 images). ISIC Dataset is shown in Figure 5. The International Skin Imaging Collaboration (ISIC) gathered 2357 photographs of benign and malignant oncological illnesses to create this dataset. Except for melanomas and moles, whose photos are somewhat predominant, all photographs were sorted following the ISIC categorization, and each subgroup was comprised of the same number of images. The following illnesses are included in this ISIC dataset: basal cell cancer (376 images), actinic keratosis (114 images), dermatofibroma (95 images), Melanoma (438 images), nevus (357 images), benign keratosis with pigmentation (462 images), seborrheic keratosis (77 images), squamous cell carcinoma (181 images), and lesion in the vein (139 images). Melanoma (438 images), basal (376 images), and squamous (181 images) skin lesions were selected from this dataset to be included in the system evaluation.

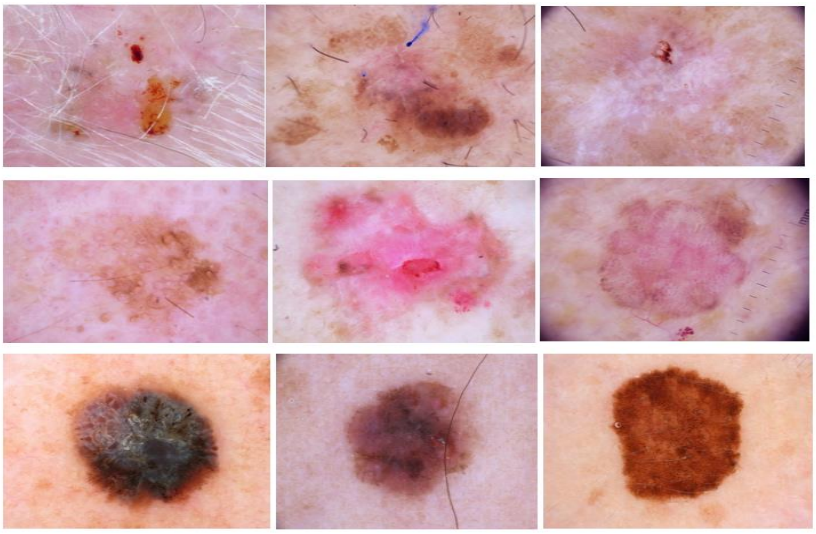

Figure 5. Sample of ISIC Dataset

3.2. Pre-processing

The dataset is then split into a training set, comprising 75% of the images, and a validation set, comprising the remaining 25%. To prepare the images for input into the neural networks, they are resized to 224x224 pixels for models like VGG16 and ResNet-50 and 299x299 pixels for Inception V3. After resizing, the images are normalized, and the pixel values are transformed into a tensor form suitable for model training. Extra pre-processing techniques, like data augmentation (e.g., rotation, flipping, and zooming), significantly boosts model performance by diversifying the dataset. These techniques involve center cropping to emphasize central features, such as random rotation, grid distortion, horizontal and vertical flipping, optical distortion, and affine transformations. Each is applied with a probability of 0.1 to introduce controlled variations and distortions, enhancing the model's robustness to varied input patterns. Table 1 depicts the image augmentation approaches.

Table 1. Augmentation Approaches.

Augmentation Approaches	Parameter	Performance
Center Crop	True	Crop to a 0.1 height-to-width ratio.
Random Rotate	90	Rotate from - 90 to 90° with a probability of 0.1.
Grid Distortion	True	Grid distorts with a probability of 0.1.
Horizontal Flip	True	Flip horizontally with a probability of 0.1.
Vertical Flip	True	Flip vertically with a probability of 0.1.
Optical Distortion	True	Optical distort with a probability of 0.1.
Affine	True	Affine with a probability of 0.1.
Piecewise Affine	True	Piecewise Affine with a probability of 0.1.
Transpose	True	Transpose with a probability of 0.1.

3.3. Balancing the Dataset by Random Oversampling

Given the imbalance in the dataset, where the number of images for each lesion type varies significantly, a random oversampling method is employed to balance the dataset. This stage makes sure that the imbalance in the number of photos does not cause the models to be skewed toward any certain class. This technique is crucial because the dataset contains various images for different classes, such as Melanoma, basal cell carcinoma (BCC), and squamous cell carcinoma (SCC). In multi-class skin lesion classification tasks, applying random oversampling directly on images can be inefficient and potentially problematic due to increased redundancy and risk of overfitting. Instead, oversampling is more effective when applied to vector or tabular data representations of images, such as feature vectors generated during the pre-processing pipeline. The typical workflow begins with reading and normalizing images to ensure consistent input quality. Following this, data augmentation techniques can be applied to increase diversity within each class, providing the model with varied representations of each lesion type. Once this augmented set of images is ready, a feature extraction process, often utilizing the intermediate layers of pre-trained CNN models, transforms the images into vectorized representations. These feature vectors capture essential characteristics of the lesions in a lower-dimensional format, which is more suitable for oversampling.

Random oversampling is applied to the feature vectors rather than the raw image data at this stage. This approach avoids the redundancy and resource intensity of duplicating large image files, focusing instead on balancing classes by duplicating these compact feature representations. By oversampling at the feature level, the model receives a balanced dataset without excessive memory use or computational load, improving training stability. As depicted in Figure 4, this process follows a clear structure from image normalization and augmentation through feature extraction, allowing for balanced oversampling that enhances model performance in classifying diverse skin lesion types while reducing overfitting risk. This approach mitigates the risk of biased predictions towards the majority class and enhances the model's ability to classify skin lesions across all classes accurately. Moreover, it fosters a more robust and reliable evaluation of the model's performance by reducing the impact of class imbalance on training outcomes. This balanced dataset enables our models, such as VGG16, ResNet50, and Inception V3, to learn effectively from all classes, improving overall classification accuracy and generalization capabilities in skin lesion diagnosis.

3.4. Classification

The training process involves using the training dataset to train the models for 150 epochs to achieve high accuracy. Once the models reach the desired level of accuracy, they are saved for further use. Finally, the models—VGG16, ResNet-50, and Inception V3—are combined using an Ensemble Weighted Average method to leverage the strengths of each model and improve the overall classification performance. Ensemble Weighted Average is mathematically depicted as Equation (3).

$$p' = \frac{1}{n}\sum_{i=1}^{n} \sigma_i(\vec{y}) \tag{3}$$

where σ is the weight values that multiply with the weight vector $\vec{y}$ and n is the number of ensemble deep-learning models.

The testing process begins by resizing the collected images to match the input requirements of the chosen deep learning models: 224x224 pixels for VGG16 and ResNet-50, and 299x299 pixels for Inception-V3. This resizing ensures that the images are compatible with the architectures of these models. After resizing, the images are normalized to standardize the pixel values, which helps improve the training stability and performance of the models. The pixel values are then transformed into tensor form, making them suitable for neural networks to process. To improve the accuracy and resilience of the skin lesion classification, an ensemble model that integrates the capabilities of ResNet-50, Inception-V3, and VGG16 is used to carry out the classification task. Initially, the system is tested on the unbalanced nature of the original dataset, which reflects the real-world distribution of skin lesion types. After oversampling and augmenting the dataset to achieve balance, the system is tested on a balanced dataset to solve the issues raised by this imbalance. These methods increase the representation of underrepresented classes, thereby mitigating the risk of biased predictions. Lastly, model evaluation is carried out to evaluate the ensemble model's performance in both balanced and unbalanced settings, offering information about the overall classification accuracy and the efficacy of the balancing procedures.

4. Results and Discussion

This method uses the ISIC dataset and the skin lesion dataset from Kaggle. Images of SCC, Melanoma, and BCC are taken from these datasets in order to classify them. The initial dataset's imbalance is used to test this system. Next, oversampling is used to test the system's balanced nature. For this suggested system study, the performance metrics' accuracy, recall, f-measure, and precision are assessed. Table 2 describes the system configurations for three models.

Table 2. Model Configurations.

Configuration	VGG16	ResNet-50	Inception-V3
Input Size	224x224 pixels	224x224 pixels	299x299 pixels
Epochs	150	150	150
Batch Size	32	32	32
Loss function	categorical_crossentropy	categorical_crossentropy	categorical_crossentropy
Optimizer	SGD	SGD	SGD
Learning rate	0.001	0.001	0.001
Data Augmentation	Center Crop Random Rotate Grid Distortion Horizontal Flip Vertical Flip Optical Distortion Affine Piecewise Affine Transpose	Center Crop Random Rotate Grid Distortion Horizontal Flip Vertical Flip Optical Distortion Affine Piecewise Affine Transpose	Center Crop Random Rotate Grid Distortion Horizontal Flip Vertical Flip Optical Distortion Affine Piecewise Affine Transpose

Table 3 presents the results of VGG16 on the original and balanced dataset. The VGG16 model's performance in skin lesion classification shows notable differences between unbalanced and balanced datasets, particularly after data augmentation. On the unbalanced dataset, VGG16 achieves moderate results with higher accuracy in Melanoma detection compared to Basal and Squamous cell carcinomas. After balancing the dataset through oversampling and augmentation, the model's performance improves significantly across all classes, especially in Basal and Squamous cell carcinoma, where it demonstrates enhanced accuracy, precision, recall, and F1-scores. The balanced dataset allows VGG16 to achieve nearly perfect results in Melanoma classification, highlighting the importance of addressing class imbalance for more accurate predictions.

Table 3. VGG16's Results in Unbalanced and Balanced Dataset.

Dataset	Class	Precision	Recall	F1-Score	Accuracy
Unbalanced	basal	0.85	0.78	0.81	0.80
	melanoma	0.80	0.91	0.85	
	squamous	0.67	0.57	0.61	
Balanced	basal	0.94	0.94	0.94	0.94
	melanoma	0.98	0.93	0.96	
	squamous	0.88	0.98	0.93	

The results of ResNet-50 on the original and balanced dataset are shown in Table 4. In classifying skin lesions using the ResNet-50 model, the performance metrics reveal significant differences between the unbalanced and balanced datasets, particularly after data augmentation. On the unbalanced dataset, ResNet-50 demonstrates strong precision but moderate recall for Basal cell carcinoma, with an F1-score of 0.84. The model's performance for Squamous cell carcinoma is more varied, with challenges due to class underrepresentation, reflected in a lower F1-score of 0.75. However, ResNet-50 performs exceptionally well for Melanoma, achieving an F1-score of 0.96. When evaluated on the unbalanced dataset, the system achieved an accuracy of 87%. The model's performance significantly improves across all classes when the dataset is balanced through oversampling and augmentation. With F1-scores ranging from 0.98 to 0.99, ResNet-50 almost reaches ideal metrics for SCC, Melanoma, and BCC, demonstrating the value of dataset balance in improving prediction accuracy. The system showed significant improvements on the balanced dataset, achieving an overall accuracy of 98%.

Table 4. ResNet-50's Results in Unbalanced and Balanced Dataset.

Dataset	Class	Precision	Recall	F1-Score	Accuracy
Unbalanced	basal	0.93	0.77	0.84	0.87
	melanoma	0.95	0.97	0.96	
	squamous	0.66	0.87	0.75	
Balanced	basal	0.99	0.97	0.98	0.98
	melanoma	0.99	0.99	0.99	
	squamous	0.96	1.00	0.98	

The findings of Inception-V3 on the original and balanced dataset are shown in Table 5. The Inception-V3 model demonstrates significant differences in performance when classifying skin lesions on unbalanced versus balanced datasets, both with augmentation. The system achieved an overall accuracy of 85% on the unbalanced dataset. Also, it shows high sensitivity for Basal cell carcinoma but lower precision, achieving a precision of 0.83, recall of 0.92, and F1-score of 0.87. The model struggles with Squamous cell carcinoma, with lower scores across the board, while performing well on Melanoma with a precision of 0.94, recall of 0.91, and F1-score of 0.93. The balanced dataset significantly improved the system's performance, reaching an accuracy of 98%. After balancing the dataset through oversampling, the model's performance improves dramatically across all classes, particularly achieving near-perfect results for Mela-noma and significant enhancements for Basal and Squamous cell carcinomas.

Table 6 displays the ensemble model's output on the original, balanced dataset. The ensemble model performs much better in skin lesion classification with data augmentation when comparing balanced and unbalanced datasets. On the unbalanced dataset, the model achieves an accuracy by 0.90, a balanced performance for Basal cell carcinoma with a precision of 0.88, recall of 0.90, and an F1-score of 0.89. Due to class imbalance, it performs poorly in the classification of squamous cell carcinoma (accuracy of 0.79, recall of 0.74, F1-score of 0.76), but does exceptionally well in the classification of Melanoma (precision of 0.95, recall of 0.96, F1-score of 0.96). With the dataset balanced through oversampling, the model's performance greatly improves across all categories, achieving near-perfect scores: precision, recall, and F1-score of 0.99 for Basal cell carcinoma, 0.98, 1.00, and 0.99 for Squamous cell carcinoma, and 0.99, 1.00, and 1.00 for Melanoma, showcasing enhanced accuracy by 0.99 and consistency.

Table 5. Inception-V3's Results in Unbalanced and Balanced Dataset.

Dataset	Class	Precision	Recall	F1-Score	Accuracy
Unbalanced	basal	0.83	0.92	0.87	0.85
	melanoma	0.94	0.91	0.93	
	squamous	0.68	0.57	0.62	
Balanced	basal	1.00	0.97	0.98	0.98
	melanoma	0.98	0.99	0.99	
	squamous	0.96	1.00	0.98	

Table 6. Ensemble Model's Results in Unbalanced and Balanced Dataset.

Dataset	Class	Precision	Recall	F1-Score	Accuracy
Unbalanced	basal	0.88	0.90	0.89	0.90
	melanoma	0.95	0.96	0.96	
	squamous	0.79	0.74	0.76	
Balanced	basal	1.00	0.98	0.99	0.99
	melanoma	0.99	1.00	1.00	
	squamous	0.98	1.00	0.99	

As demonstrated by the skin lesion classification system's performance evaluation, the ensemble model with augmentation outperforms all individual models for both the original and balanced ISIC dataset produced by oversampling. Accuracy for ResNet-50, Inception-V3, and VGG16 are 87%, 85%, and 80%, respectively; accuracy for the ensemble model that uses augmentation with the original dataset is 90%. With 99% accuracy on the balanced dataset, the ensemble model surpasses 94% by VGG16, 98% by ResNet-50, and 98% by Inception-V3. Table 7 compares the results of the research. Compared with existing literature, the proposed ensemble model for classification outperforms previous methods.

Table 7. Model Configurations.

Method	Dataset	Method	Accuracy
Ref [9]	PH2	CNN	98.3
	ISIC 2016		80.47
	ISIC 2017		81.16
	HAM10000		81
Ref [10]	ISIC 2018	Inception-ResNet	96.21
Proposed	ISIC2018	Ensemble Model	99

5. Conclusions

This paper's results highlight the effectiveness of combining multiple deep-learning models to enhance classification performance across various skin lesion types. When used on a balanced dataset, the ensemble technique that integrates VGG16, ResNet-50, and Inception-V3 shows notable increases in model performance. Balancing the dataset through oversampling effectively mitigates the challenges associated with class imbalance, which is evident in the improved performance metrics across all skin lesion categories. The performance evaluation of the skin lesion classification system reveals that the ensemble model, enhanced with data augmentation, consistently outperforms individual models on both the original and oversampled balanced ISIC datasets. Specifically, the accuracies of ResNet-50, Inception-V3, and VGG16 are 87%, 85%, and 80%, respectively. The ensemble model achieves an accuracy of 90% on the original dataset and an impressive 99% on the balanced dataset, surpassing the individual models' accuracies of 94% (VGG16), 98% (ResNet-50), and 98% (Inception-V3). The ensemble model achieves near-perfect classification results, particularly for Melanoma, SCC, and BCC, showcasing the benefits of this approach in providing more reliable and consistent diagnostic support. The results affirm that combining different architectures enhances the robustness and generalization of the model, leading to superior performance compared to individual models. This comprehensive approach advances the classi-fication and underscores the importance of data balancing and augmentation in developing effective machine learning solutions for medical image analysis. Despite these advancements, limitations remain, including the need for large computational resources and potential over-fitting due to the complexity of the ensemble model. Future work will focus on reducing the model's computational overhead through further optimizations, such as more aggressive pruning and quantization while exploring its applicability to real-time diagnosis on larger, more diverse datasets. Additionally, integrating attention mechanisms and exploring self-supervised learning techniques could improve the model's robustness and clinical usability.

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

Funding: This research received no external funding.

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

References

[1] A. Murugan, S. A. H. Nair, A. A. P. Preethi, and K. P. S. Kumar, "Diagnosis of skin cancer using machine learning techniques," *Microprocess. Microsyst.*, vol. 81, p. 103727, Mar. 2021, doi: 10.1016/j.micpro.2020.103727.

[2] G. Akilandasowmya, G. Nirmaladevi, S. Suganthi, and A. Aishwariya, "Skin cancer diagnosis: Leveraging deep hidden features and ensemble classifiers for early detection and classification," *Biomed. Signal Process. Control*, vol. 88, p. 105306, Feb. 2024, doi: 10.1016/j.bspc.2023.105306.

[3] L. Talavera-Martínez, P. Bibiloni, A. Giacaman, R. Taberner, L. J. D. P. Hernando, and M. González-Hidalgo, "A novel approach for skin lesion symmetry classification with a deep learning model," *Comput. Biol. Med.*, vol. 145, p. 105450, Jun. 2022, doi: 10.1016/j.compbiomed.2022.105450.

[4] A. Bibi *et al.*, "Skin Lesion Segmentation and Classification Using Conventional and Deep Learning Based Framework," *Comput. Mater. Contin.*, vol. 71, no. 2, pp. 2477–2495, 2022, doi: 10.32604/cmc.2022.018917.

[5] W. Gouda, N. U. Sama, G. Al-Waakid, M. Humayun, and N. Z. Jhanjhi, "Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning," *Healthcare*, vol. 10, no. 7, p. 1183, Jun. 2022, doi: 10.3390/healthcare10071183.

[6] A. K. Adepu, S. Sahayam, U. Jayaraman, and R. Arramraju, "Melanoma classification from dermatoscopy images using knowledge distillation for highly imbalanced data," *Comput. Biol. Med.*, vol. 154, p. 106571, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106571.

[7] A. Alhudhaif, B. Almaslukh, A. O. Aseeri, O. Guler, and K. Polat, "A novel nonlinear automated multi-class skin lesion detection system using soft-attention based convolutional neural networks," *Chaos, Solitons & Fractals*, vol. 170, p. 113409, May 2023, doi: 10.1016/j.chaos.2023.113409.

[8] N. Priyadharshini, S. N., B. Hemalatha, and C. Sureshkumar, "A novel hybrid Extreme Learning Machine and Teaching–Learning-Based Optimization algorithm for skin cancer detection," *Healthc. Anal.*, vol. 3, p. 100161, Nov. 2023, doi: 10.1016/j.health.2023.100161.

[9] H. K. Gajera, D. R. Nayak, and M. A. Zaveri, "A comprehensive analysis of dermoscopy images for melanoma detection via deep CNN features," *Biomed. Signal Process. Control*, vol. 79, p. 104186, Jan. 2023, doi: 10.1016/j.bspc.2022.104186.

[10] S. K. Singh, S. Banerjee, A. Chakraborty, and A. Bandyopadhyay, "Classification of Melanoma Skin Cancer Using Inception-ResNet," in *Frontiers of ICT in Healthcare*, 2023, pp. 65–74. doi: 10.1007/978-981-19-5191-6_6.

[11] Y. Wang *et al.*, "Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images," *Med. Image Anal.*, vol. 81, p. 102535, Oct. 2022, doi: 10.1016/j.media.2022.102535.

[12] D. Zhuang, K. Chen, and J. M. Chang, "CS-AF: A cost-sensitive multi-classifier active fusion framework for skin lesion classification," *Neurocomputing*, vol. 491, pp. 206–216, Jun. 2022, doi: 10.1016/j.neucom.2022.03.042.

[13] U. A. Lyakhova and P. A. Lyakhov, "Systematic review of approaches to detection and classification of skin cancer using artificial intelligence: Development and prospects," *Comput. Biol. Med.*, vol. 178, p. 108742, Aug. 2024, doi: 10.1016/j.compbiomed.2024.108742.

[14] M. Hajiarbabi and R. Chandra, "Skin Lesion Detection Using Deep Learning," *J. Autom. Mob. Robot. Intell. Syst.*, pp. 56–64, Aug. 2023, doi: 10.14313/JAMRIS/3-2022/24.

[15] M. A. Araaf, K. Nugroho, and D. R. I. M. Setiadi, "Comprehensive Analysis and Classification of Skin Diseases based on Image Texture Features using K-Nearest Neighbors Algorithm," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 31–40, Sep. 2023, doi: 10.33633/jcta.v1i1.9185.

[16] M. T. H. Khan Tusar, M. T. Islam, A. H. Sakil, M. N. H. N. Khandaker, and M. M. Hossain, "An Intelligent Telediagnosis of Acute Lymphoblastic Leukemia using Histopathological Deep Learning," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 1–12, May 2024, doi: 10.62411/jcta.10358.

[17] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier Inc., 2012. doi: 10.1016/C2009-0-61819-5.

[18] J. Brownlee, "A Gentle Introduction to Transfer Learning for Deep Learning," *Machine Learning Mastery*, 2019. https://machinelearningmastery.com/transfer-learning-for-deep-learning/

[19] M. ul Hassan, "VGG16 – Convolutional Network for Classification and Detection," *Neurohive.io*, 2018. https://neurohive.io/en/popular-networks/vgg16/

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *arXiv*, Dec. 2015, vol. 2016-Decem, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[21] S. Mukherjee, "The Annotated ResNet-50," *Towards Data Science*, 2022. https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758

[22] C. Szegedy *et al.*, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.

[23] F. M. Firnando, D. R. I. M. Setiadi, A. R. Muslikh, and S. W. Iriananda, "Analyzing InceptionV3 and InceptionResNetV2 with Data Augmentation for Rice Leaf Disease Classification," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 1–11, May 2024, doi: 10.62411/faith.2024-4.

[24] "Inception-v3 Explained," *Papers with Code*. https://paperswithcode.com/method/inception-v3

[25] MNOWAK061, "ISIC2018 and PH2 384x384 JPG," *Kaggle.com*, 2000. https://www.kaggle.com/datasets/mnowak061/isic2018-and-ph2-384x384-jpg (accessed Apr. 10, 2022).