Research Article GU J Sci 38(4): 1846-1865 (2025) DOI: 10.35378/gujs.1549629

Gazi University

Journal of Science

http://dergipark.org.tr/gujs

Ensemble Model with Color Spaces Transformations for Improving Skin Cancer Classification

Feyza YILMAZ[1,*], Rifat EDIZKAN[1], Omer Nezih GEREK[2]

[1]*Eskisehir Osmangazi University, Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, 26480 Eskisehir, Türkiye*
[2]*Eskişehir Technical University, Faculty of Engineering, Department of Electrical and Electronics Engineering, 26555 Eskisehir, Türkiye*

Highlights

- Improvement of skin cancer diagnosis using a color-space-based ensemble model.
- Reducing the misclassification rate by combining the outputs of three color-based classifiers.
- Improvement on the F1 score is around 1% higher than the RGB color space score.

Article Info

Received: 13 Sep 2024
Accepted: 25 July 2025

Keywords

Color space
Deep neural networks
Ensemble model
Majority voting
Skin cancer classification

Abstract

This study introduces a novel ensemble model leveraging color space transformations for enhancing skin cancer classification accuracy. The proposed model enhances the accuracy of distinguishing between benign and malignant skin lesions by using three baseline classifiers, each specialized in a different color representation (RGB, HSI, and YCbCr), and employing majority voting decision rule. The experimental study was conducted on ISIC database using four CNN architectures; InceptionV3, ResNet101V2, InceptionResNetV2, and MobileNetV2, for three color spaces. The results reveals that the proposed model consistently outperformed three classifiers, demonstrating reduction in misclassification rates and an enhancement in the F1 score. In this study, the improvement in F1 score is approximately about 1% on the ISIC database. This achievement is obtained without applying any preprocessing. The F1 scores obtained from of the baseline classifiers and the proposed ensemble model are analyzed by the Friedman test. The generalizability of the proposed model is evaluated by conducting the same experiments on the PH2 dataset. Our findings indicate that incorporating multiple color spaces into an ensemble model can enhance classification performance, providing a promising approach for early and accurate skin cancer diagnosis.

1. INTRODUCTION

Skin cancer ranks as one of the most lethal cancers, experiencing the highest surge in cases worldwide [1]. Originating from the skin—the body's outermost layer—this disease is significantly influenced by environmental elements, with sunlight identified as a primary catalyst due to its UV radiation. Research indicates that a substantial 86% of individuals with skin cancer have been excessively exposed to UV rays [2].

The disease stems from the unchecked and erratic multiplication of skin cells, leading to the development of lesions categorized as either melanoma or non-melanoma. Melanomas, which can be either benign or malignant, typically manifest on the upper bodies of men and, less frequently, on women's legs [3]. Notoriously aggressive, malignant melanoma is distinguished by its high fatality risk and ability to spread to nearby tissues [4]. Early detection and treatment can lead to a 95% or greater survival rate in patients with malignant melanoma [5], whereas late diagnosis can lead to a 95% or greater survival rate [6]. Therefore, the most crucial factor influencing survival is the early diagnosis of malignant melanoma. To diagnose this condition, dermatologists employ a variety of methods, such as the ABCD rule, the 7-point checklist, and the Menzies method, each varying in effectiveness based on the lesion type and the clinician's expertise [7].

*Corresponding author, e-mail: f.yilmaz@ogu.edu.tr

Depending on the lesion type and clinician's expertise, the effectiveness of each dermoscopic technique varies [8]. AI-based decision support systems will assist clinicians in improving the diagnosis of malignant lesions since the precise identification of malignant lesions is essential. The aim of this study is to improve the accuracy of skin cancer classification. To do this, a color space-based majority voting-based ensemble model approach has been developed. According to related research, it has been proven that combining the outputs of base classifiers in the ensemble increases classification performance.

Using a single classifier alone may not be sufficient for accurately detecting skin cancer lesions. In such cases, combining the outputs of two or more classifiers in an ensemble model can lead to improved outcomes. Majority voting is one of the most commonly used decision rule for combining the outputs of base classifiers into a single definitive outcome [9]. In this rule, each classifier's decision is treated as a vote, and the final decision is made based on the accumulated votes from different classifications. If all classifiers carry the same weight, the decision of each classifier for an input example is equally considered, and the decision with the highest number of votes is selected as the winning class for the input sample [10].

In the literature, ensemble model seeks to enhance the effectiveness and generalization of models by integrating diverse deep learning models that handle different features or subsets of data. In our research, datasets represented in RGB, HSI, and YCbCr color spaces were classified using a homogeneous ensemble model. The proposed model evaluates four convolutional neural network (CNN) architectures: InceptionV3, ResNet101V2, InceptionResNetV2, and MobileNetV2. The novelty of our study is in the way of using three color spaces in a homogenous ensemble model. This approach differs from the studies on skin cancer diagnosis the literature that typical focus using heterogeneous ensemble model utilizing only one color space.

The paper is organized as follows: Section 2 gives the literature reviews. The materials and the proposed majority voting-based ensemble model are described in Section 3. The experimental results and discussions are presented in Section 4. Section 5 contains the conclusion remarks and discussion on future work.

2. LITERATURE REVIEWS

In recent years, skin cancer detection has emerged as a crucial research area due to its substantial impact on public health. To effectively evaluate and treat skin lesions, dermatologists follow a sequence of visual examination, dermoscopy, and biopsy. Studies emphasize that image pre-processing methods, deep learning techniques, and ensemble models enhance accuracy and efficiency in the interpretation, segmentation, and classification of skin cancer. This section reviews the literature on image pre-processing methods, deep learning techniques, and ensemble models employed by researchers in the field of skin cancer detection.

The exploration of color spaces in machine learning algorithms for the detection and classification of melanoma has been a longstanding area of research. Santos et al. have examined the impact of imaging devices and lighting variations on deep learning-based skin cancer diagnosis systems. The researchers utilized deep learning techniques such as LCDPNet, LLNeRF, and DSN to correct the lighting and color properties of images. The study results showed that these preprocessing steps improved classification accuracy by 3% to 4%. Additionally, it was observed that different deep neural network architectures exhibited varying levels of sensitivity to these corrections, with some models being more resilient to such changes. These findings highlight the importance of correcting the lighting and color properties of images used in skin cancer diagnosis [11]. Abbas et al. developed a melanoma recognition system by applying ABCD rules to diagnose and classify malignant and nevus lesions [12]. In this study, the CIE L*a*b* color space was used. During image preprocessing, the homomorphic transform filter technique was used to remove specular reflection and adjust contrast. A linear search-based scanning method was employed to eliminate the black frame observed in the dermoscopic images. Then, hair pixels in the image were removed using the derivatives of Gaussian, morphological function and fast progression techniques. The lesion segmentation was performed using skin tumor area extraction, and feature extraction was performed with the ABCD rule. The sequential forward selection method was also applied, and in the experiments, 88% area under the curve (AUC) was obtained using the SVM classification algorithm [12]. Another study,

Barata et al. proposed a color model learning-based diagnosis method for classifying common nevi, atypical nevi, and melanoma. A model based on color recognition has been developed for images since black, blue, gray, white, dark, and light brown colors are more common in malignant lesions than in benign melanoma lesions. In the preprocessing, derivatives of Gaussian morphological function and rapid advance techniques were taken to remove hair pixels. A Gaussian mixing model was used to construct black, blue–gray, white, dark, and light brown models. These color models were applied to segment the lesions in each image. Then, the CIE L*a*b* and HSV color space conversions were applied, and dermatologist-provided labels were used to evaluate diagnostic performance. The Spearman rank correlation coefficient was calculated for the concordance between the dermatologist and the system. As a result, a 0.7981 Spearman rank correlation ratio was obtained [13]. These studies demonstrate the application of advanced image processing techniques, color space transformations, and machine learning algorithms to enhance the accuracy of melanoma detection and classification. The use of CIE Lab* color space in both studies highlights its effectiveness in capturing color information critical for distinguishing between malignant and benign lesions. Shalu et al. developed a system for automatic detection of melanoma skin cancer. In this study, features were extracted in HSV and YCbCr color spaces using digital images from MED-NODE dataset and it was found that the combination of these features performed better compared to features extracted from other color spaces. In particular, the system achieved 84.11% accuracy with features extracted from HSV and YCbCr color spaces [14]. Oukil et al. proposed an automatic segmentation and classification algorithm for melanoma detection. Features were extracted from skin lesions in five different color spaces such as RGB, HSV, Lab, XYZ and YCbCr and these features were given as input to classifiers such as K-nearest neighbor, support vector machines and artificial neural networks. The results showed that the combination of color and texture features are powerful and relevant features in melanoma detection [15]. Mamadou et al. analyzed the effect of different color spaces in skin cancer diagnosis. The color features of skin lesions were studied in detail using various color spaces such as RGB, Lab, HSV and YUV. It was shown that the YUV-RGB combination exhibited superior performance in the detection and classification of different skin cancer types with an accuracy rate of 98.51% [16].

Table 1 presents recent studies from the past five years focusing on classifying skin cancer using CNN models. The classification results achieved from binary and multiclass datasets are presented on the table. These studies evaluate the model performance using either accuracy or area under curve (AUC) [17-30].

In the literature, various deep learning techniques and ensemble models built upon them are utilized for skin cancer classification. Nersison et al. extract features from images with CNN using the ISIC 2016 skin cancer dataset. These features are combined with texture and color features extracted from the lesion area of the images. The combined features are then fed into the fully connected network and classified with 94% accuracy [31]. In another study, Gu et al. proposed an ensemble model based on the maximum correntropy criterion (MCC) and CNN to classify benign seborrheic keratosis accurately and malignant basal cell carcinoma skin cancers. The MCC method was used to learn to weigh more than one single model and reduce the effects of noise in seborrheic keratosis and basal cell carcinoma data. The DenseNet, ResNet, and Xception models were used in the ensemble model, and the proposed ensemble model achieved the highest classification with an accuracy (ACC) of 97.07% [32]. Oliveira et al. developed three models based on different input features to increase the accuracy of diagnosing benign and malignant skin lesions on dermoscopic images. The optimum-path forest classifier model was used for each ensemble model. In the first subset selection model, specific feature groups were used. The second one is a correlation-based feature selection algorithm, and feature selection is performed by applying this model to feature subsets used in the first model. Finally, feature selection was performed using feature selection algorithms on all the extracted features. The first subset selection model gave promising results for the ensemble model. Skin lesion classification achieved 94.3% accuracy [33]. In another study, Schafer et al. presented an ensemble classifier model for skin cancer diagnosis from dermoscopic images of skin lesions. Automatic border detection was applied to split the lesion from the background skin. Three hundred fifty-four color features, 72 texture features, and 11 shape features were extracted for each lesion split from the skin, and these features were used in the training phase of the ensemble classifier. First, single-classification models were used in the ensemble classifier model, and redundant classifier models were excluded from the ensemble classifier model based on the fuzzy diversity measure. The selected classifier models are combined with a neural network fuser. Classification performance was achieved with a specificity of 93.84% and a

sensitivity of 93.76% [34]. The study by Cao et al. [35] proposed a deep learning-based ensemble model for the classification of malignant and benign melanocytic nevus. The mixed skin lesion picture generation method based on the mask R-CNN was proposed to solve the problem of imbalance in the number of images used in the model. Inception V4, ResNet152, DenseNet121, and DenseNet145 deep learning models were used in the ensemble model. The highest performance was achieved with mask-DenseNet, for which the area under the curve (AUC) increased from 91.57% to 95.02%. Harangi used the GoogleNet, AlexNet, ResNet, and VggNet deep learning models to classify skin lesion images of nevus, melanoma, and seborrheic keratosis and developed an ensemble model with these deep learning models. The predicted class was obtained using a voting-based decision. The proposed ensemble model achieved better classification performance than did the individual CNN models [36]. Le et al., proposed a ResNet50 ensemble model utilizing focal loss and class weighting to address the challenges posed by the data imbalance in the HAM10000 dataset. The proposed model achieved an accuracy of 93% [37].

Table 1. *A summary of recent studies on the classification of skin cancer*

References	Datasets	Number of Skin Classes	Methods	Metrics	Results (%)
[17]	ISIC (2019)	7	ResNet-18	Accuracy	94.47
[18]	ISIC (2017)	3	MB-DCNN	AUC	95.70
[19]	ISBI (2016)	2	Deep residual network (DRN)+ Local texton XOR pattern (LTxXORP)/SVM	Accuracy	88.92
	ISIC (2017)	3			85.30
[20]	ISIC (2018)	7	CNN +scratch with balance	Balanced-accuracy	64.00
			CNN+scratch without balance		57.00
			VGG19		79.00
			ResNet50		74.00
			Inception V3		76.00
			DenseNet 121		76.00
			Xception		76.00
			VGG16		77.00
[21]	ISIC (2017)	3	FCRN-88	Accuracy	85.70
[22]	ISIC (2018)	2	RF	Accuracy	85.70
			SVM		91.20
			DenseNet-161		96.30
[23]	Dermquest/DermIS	2	CNN	Accuracy	96.00
[24]	ISIC	8	CNN+Gray wolf optimization	Accuracy	93.00
[25]	ISIC (2017)	3	DE-Net	Pixelwise accuracy	94.50
[26]	HAM10000	7	EfficientNets B4	AUC	97.53
			EfficientNets B5		97.54
[27]	ISBI (2016)	2	DCNN	Accuracy	95.10
	ISBI (2017)	2			94.80
[28]	ISIC (2017)	2	ResNet152	Accuracy	90.40
[29]	ISIC	2	InceptionV3/LSTM	F1 Score	99.40
	DermsIS				98.00
[30]	ISIC (2022)	2	DenseNet121	Accuracy	99.60
			Xception		99.10
			InceptionResNetV2		98.45
			ResNet50		98.10
			MobileNet		99.40

3. MATERIALS AND METHODS

This section describes the dataset and the color representations used in the study. Also, the framework of the proposed homogeneous ensemble model and the properties of the four CNN architectures are introduced.

3.1. Dataset

Our research leverages a Kaggle dataset, accessible to the public, for both the training and evaluation phases of our proposed approach [38]. This dataset comprises RGB dermoscopic images compiled from the Skin Imaging Collaboration (ISIC) archive, encompassing both benign and malignant skin lesions. Notably, these images from the ISIC collection contain a variety of noises and artifacts, and were incorporated into our Deep Neural Network (DNN) models without undergoing any preprocessing treatments like hair removal, contrast adjustments, or resizing [39]. Organized into three distinct sets-training, validation, and testing—the images, resized to 300×300 pixels, include a specified distribution of benign and malignant lesions, detailed further in Table 2.

Table 2. *Categorization of benign and malignant samples*

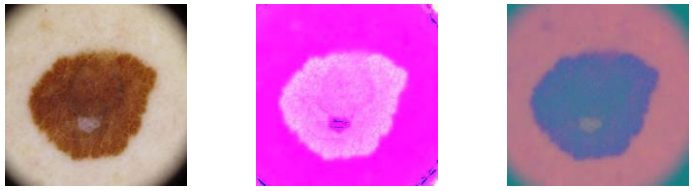
Diagnostic Category	Number of Images		
	Train	Validation	Test
Malignant	4105	500	500
Benign	4500	500	500
Total	8605	1000	1000

3.2. Color Space Transformations

In this research, the critical role of color in distinguishing between benign and malignant skin lesions was explored by converting RGB dermoscopic images into HSI and YCbCr color spaces. These transformations were chosen for their ability to represent colors differently from RGB while maintaining image brightness, an aspect crucial to human visual perception which can discern brightness, hue, and saturation [40]. Within the HSI space, 'H' stands for hue, identifying colors like red, yellow, green, or blue, while 'S' represents saturation, the degree to which color mixes with white. 'I' signifies intensity, reflecting the color's brightness. YCbCr, often used in digital image and video encoding, efficiently compresses RGB colors, splitting them into luminance (Y) and two chrominance channels (Cb for chroma blue and Cr for chroma red), aligning with the human eye's heightened sensitivity to luminance over chrominance [41]. The study utilized these color spaces to classify skin lesions, as illustrated in Figure 1, demonstrating their efficacy in distinguishing between lesion types.



a) Benign



b) Malignant

Figure 1. *Three color space representations of some (left to right: RGB, HSI, and YCbCr)*

3.3. CNN Architectures

CNNs are pivotal in the realm of computer vision, significantly enhancing the analysis, segmentation, and classification of visual content. Illustrated in Figure 2 is the structured layout of a CNN, showcasing a sequence of interconnected layers that include convolutional layers, pooling layers, and fully connected

layers, integral to the model's architecture [42]. Employing a feedforward network design, CNNs utilize filters across the image in the convolutional layer, computing dot products at each filter position to generate a variety of feature maps. The application of nonlinear activation functions, such as rectified linear units (ReLU), sigmoid, and tanh, to these feature maps introduces nonlinearity, a critical aspect for learning complex patterns. Following the convolutional stages, pooling layers serve to compress the feature maps by summarizing their contents, thereby diminishing the overall computational load while preserving essential information. This process bridges the gap to the fully connected layer, where the network's classification decisions are formulated, often leveraging the softmax classifier to predict the probability distribution across various classes. In our research, we utilized prominent CNN architectures like InceptionV3, ResNet101V2, InceptionResNetV2, and MobileNetV2, known for their effective image classification capabilities within the academic and practical fields.

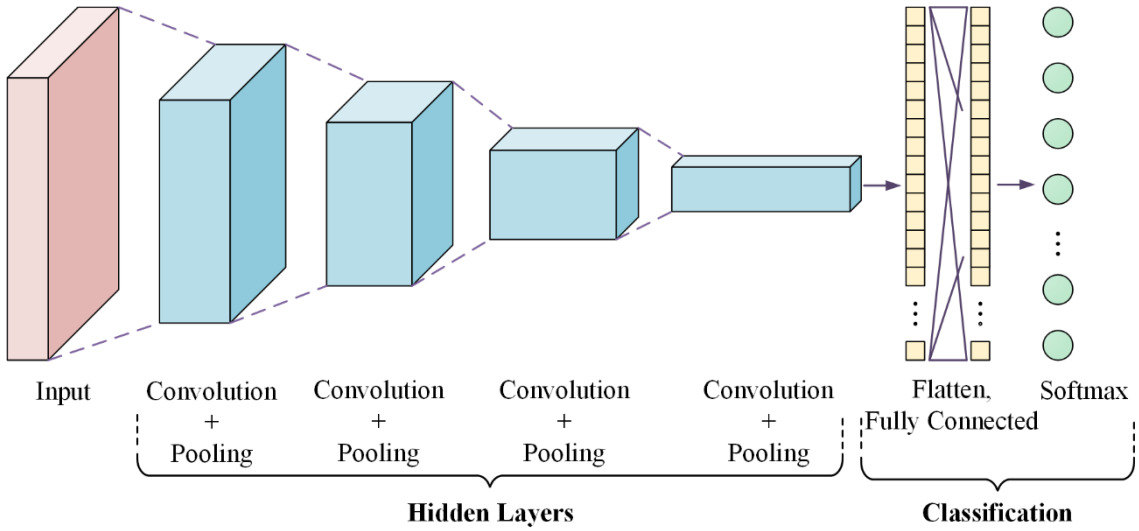


Figure 2. *The basic architecture of the CNN*

3.3.1. Inception V3

InceptionV3 introduces enhancements to the original GoogleNet design, comprising three principal modules along with additional layers aimed at deepening the model's structure while maintaining lower computational demands compared to its predecessors [43]. This version employs a unique approach by splitting the convolutional core into two segments with dimensions 1×n and n×1, facilitating efficient feature extraction and combination through filter concatenate layers, incorporating layers for pooling, dropout, softmax functions, and connectivity.

3.3.2. ResNet101V2

ResNet is a deep convolutional neural network introduced by a group of Microsoft researchers. ResNet contains multiple residual blocks. Each block consists of 2-3 layers and a connection from the input to the output. The operation principle of ResNet is based on optimizing a residual function. The vanishing gradient phenomena observed in CNN architectures with large layer numbers are overcome in the ResNet. ResNet101V2 is an enhanced version of ResNet. ResNet101V2 has a deep residual network architecture and consists of large numbers of layers and large numbers of parameters. Despite its drawback of high computational cost, ResNet is widely recognized as one of the leading CNN architectures in the field [44].

3.3.3. InceptionResNetV2

InceptionResNetV2 combines an Inception structure and residual connections. In this architecture, the Inception structure provides high accuracy and performance at a relatively low computational cost. The residual connection overcomes the degradation problem caused by large layer numbers (deep structure) and reduces the training time. InceptionResNetV2 consists of 164 images [45].

3.3.4. MobileNetV2

MobileNet, introduced by Google, is an optimized neural network architecture for mobile devices. In the architecture, the number of parameters and mathematical operations is kept as low as possible so that deep networks can be run on mobile devices. Nineteen residual bottleneck layers follow an initial fully connected convolution layer with 32 filters. The idea of filtering images in this architecture is different from that of conventional convolution. MobileNetV2 uses depth convolution and point convolution to filter images. First, depth convolution is applied to each image layer and channel to obtain a filtered image from separate image layers. After that, point convolution is applied to these filtered images [46].

3.4. Evaluation Metrics

For assessing the binary classification efficacy of the CNN architectures, we utilized five key metrics: specificity, precision, recall, accuracy, and F1 score, derived from four fundamental values: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Accuracy measures the proportion of correct predictions out of all samples. Precision quantifies the accuracy of positive predictions, while recall assesses the rate of correctly identified positive cases. Specificity evaluates the classifier's ability to identify negative cases accurately. The F1 score, commonly employed in binary classification, reflects the test's accuracy by calculating the harmonic mean between precision and recall [45]

$$Accuracy = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) \quad (1)$$

$$Precision = \left(\frac{TP}{TP+FP}\right) \quad (2)$$

$$Recall = \left(\frac{TP}{TP+FN}\right) \quad (3)$$

$$Specificity = \left(\frac{TN}{TN+FP}\right) \quad (4)$$

$$F1\ score = \left(\frac{2\times Precision\times Recall}{Precision+Recall}\right). \quad (5)$$

3.5. Proposed Ensemble Model

Ensemble model refers to an approach that aims to achieve better results by combining the outputs of many classifiers that use either machine learning or deep learning models. Ensemble model provides the ability for making a stronger and more stable prediction. There are two ensemble strategies to select baseline classifiers: homogeneous ensemble and heterogeneous ensemble. Homogeneous ensemble is constructed from baseline classifier with same type. Each classifier is trained with different data. In heterogeneous ensemble, classifiers are different type, but all of them use the same data [47]. Bagging, boosting, and stacking are some of well-known ensemble learning methods. In bagging, different datasets are created using repeated sampling method and the same model is trained with different parameters on each dataset and then the results are combined. Boosting trains models sequentially, where each subsequent model focuses on correcting the mistakes made by previous models. Stacking method generates a new model by combining information from multiple predictive models. In an ensemble model, outputs of baseline classifiers are integrated using fusion methods. Voting or meta-learning methods can be used as a fusion method [47]. In our study, the output of classifiers in the homogeneous ensemble model combined using majority voting.

One effective decision-making strategy is majority voting [48, 49]. Consider a classifier ensemble model comprising of L classifiers in the set $\mathcal{D} = \{D_1, \ldots, D_L\}$. Let $\Omega = \{\omega_1, \ldots, \omega_C\}$ denote the set of class labels. The majority voting method operates as follows: each classifier D_i assigns a class label $s_i \in \Omega = \{\omega_1, \ldots, \omega_C\}$ for an object. Then, using the class label outputs $\mathrm{s} = [s_1, \ldots s_L]^T \in \Omega^L$, the number of votes for each class label ω_k, $k = 1, \ldots, C$ in majority voting method is determined by Equation (6) [50]:

$$P(k) = \sum_{i=1}^{L} I(s_i, \omega_k) \quad (6)$$

where $I(a, b) = 1$ if $a = b$ and 0 otherwise.

Finally, the ensemble classifier assigns the class label k^* to the object, where

$$k^* = arg \max_{1 \le k \le C} P(k). \quad (7)$$

In the proposed homogeneous ensemble model, three color spaces are applied to the baseline classifier and a majority voting based decision rule is used as fusion method. However, in the studies of DNN-based skin cancer, heterogeneous ensemble models have been generally used for combining the classification results of different deep learning models on the same color space [17-23].

Color is an important criterion in skin cancer diagnosis, and our approach reduces the misclassification rate by combining the outputs of three color-based classifiers that learn common features of skin lesions, shape, and color. The proposed ensemble model is illustrated in Figure 3. In this model, three color space representations of the input image are applied to the classifiers. The same CNN model is used in all classifiers. Each classifier predicts the class label of the input image as malignant or benign. The final decision is obtained using the majority voting method. In our model, the number of classifier (L) is equal three and the number of class (C) is equal to two.

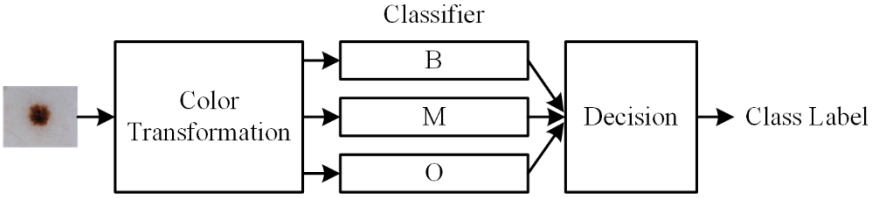


Figure 3. *Framework of the proposed ensemble model*

In this model, three color spaces were used in the binary classifiers. To implement an ensemble model utilizing three color spaces, the RGB images in the dataset are transformed into the HSI and YCbCr color spaces. Three color spaces are then assigned to three classifiers, according to the highest classification score for benign and malignant. The color space with the highest benign correct classification is assigned to the classifier B (Benign). The color space with the highest malignant correct classification is assigned to the classifier named M (Malignant). The classifier named O (Other) uses the color space that is not assigned to any classifiers.

4. RESULTS AND DISCUSSION

The proposed ensemble model was evaluated by testing four CNN-based deep neural network architectures to assess their classification performance on the ISIC dataset. These models adjust the size of dermoscopic images. The CNN parameters (batch size, learning rate, number of epochs, and patience) listed in Table 3 were used in four CNN architectures and they were not tuned. The batch size denotes how many samples are processed before the network updates, while the learning rate dictates the adjustment speed of the model's parameters. The epoch count represents the total number of complete passes through the training dataset. Patience, set uniformly at 30 for all models, is the threshold for halting training upon no further decrease in validation loss, ensuring efficiency and preventing overfitting.

Table 3. *Parameters of the CNN architectures*

Parameter	InceptionV3	Others
Image size	229×229	224×224
Batch size	16	16
Learning rate	0.0001	0.0001
Optimizer	SGD	SGD
Epoch	50	50

In this study, the code is developed in the Python language, and the training and testing of CNN architectures are implemented using the TensorFlow library functions. An Intel(R) Xeon(R) E-2136 CPU-based computer running at 3.30 GHz with 16 GB of RAM is used in the experiments. In the experiments,

the InceptionV3, ResNet101V2, InceptionResNetV2, and MobileNetV2 CNN architectures were evaluated in the proposed ensemble model. To evaluate the performance of both the baseline classifiers and the ensemble model, five-fold cross-validation was performed. The use of each set in a CNN will be denoted as Run-1, Run-2, etc. The results are evaluated using four metrics: specificity, precision, recall, accuracy, and F1 score. In the evaluation metrics, the malignant and benign classes correspond to the positive and negative classes, respectively.

The metrics provide a general overview of the model's performance, but they do not offer any insights into the interpretability of the model. Therefore, the gradient class activation mapping (Grad-CAM) method was used to visualize where the deep learning models were looking [51]. The color space representations of some labeled ground truth test images and their Grad-CAM representations are shown in Table 4. The output of Grad-CAM gives a heatmap, where the hot parts correspond to a place with high activation, indicating a greater focus on the region. When the Grad-CAMs in Table 4 are examined, the places with high heat in three color spaces show the lesion area. This indicates that each CNN was trained properly to capture the most important features from the lesion not from other regions.

Table 5 presents the assignment of the color spaces to the baseline classifiers in the proposed model for four CNN models. In this table, true positives (TP) and true negatives (TN) values, and the average classification performances (average of Run 1-5) of the classifiers and the ensemble model are also given. The results indicate that the InceptionV3 and InceptionResNetV2 architectures achieve the highest F1 scores at 0.9425, improving the RGB color space results by 0.36% and 0.93%, respectively. Figure 4 shows the graphical representation of the classification performances of the classifiers and the ensemble model.

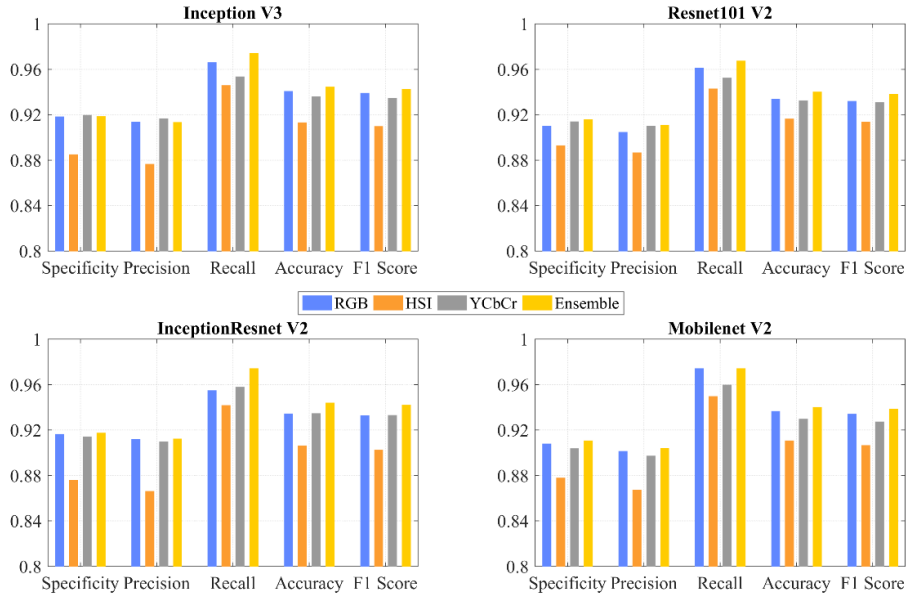


Figure 4. *The graphical representation of the highest classification performances obtained for the RGB, HSI, and YCbCr color spaces, as well as the ensemble model, for four CNN architectures*

Table 4. *Grad-CAM visualization of test samples obtained from the CNN architectures*

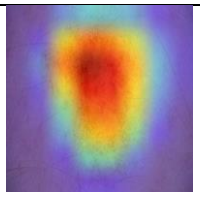
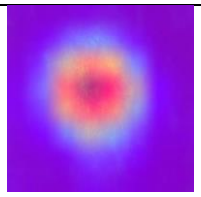
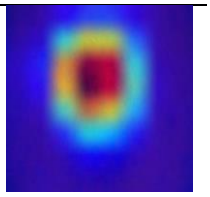
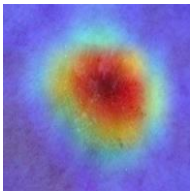
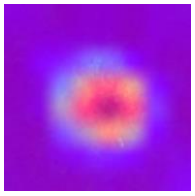
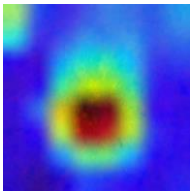
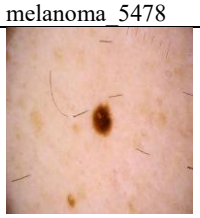
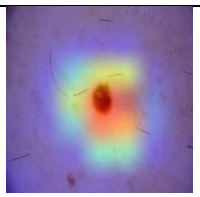
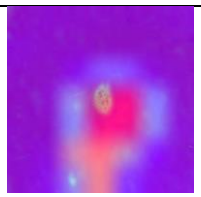
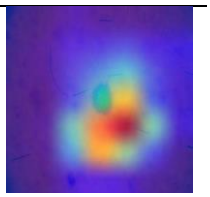
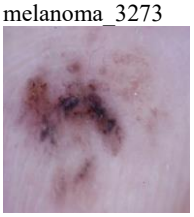
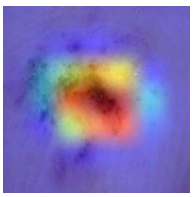
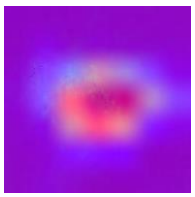
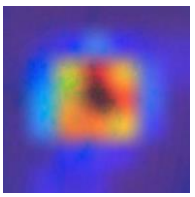
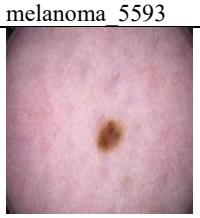
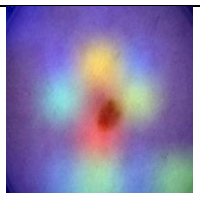
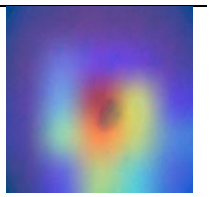
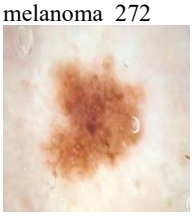
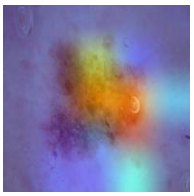
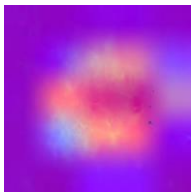
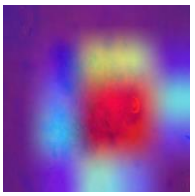
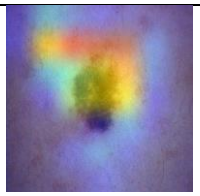
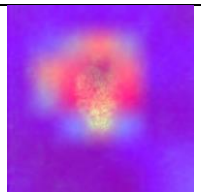
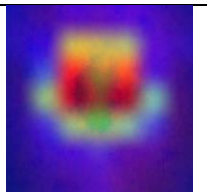
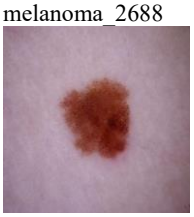
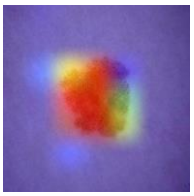
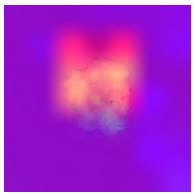
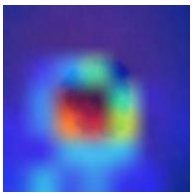
DNN Model	Class	Original Image	RGB	HSI	YCbCr
InceptionV3	Benign	 melanoma_1278			
	Malignant	 melanoma_5478			
ResNet101V2	Benign	 melanoma_3273			
	Malignant	 melanoma_5593			
InceptionResNetV2	Benign	 melanoma_272			
	Malignant	 melanoma_8478			
MobileNetV2	Benign	 melanoma_2688			
	Malignant	 melanoma_5327			

Table 5. *The average classification performances of the classifiers and the ensemble model for four CNN architectures*

CNN	Classifier	Color Space	TP	TN	Specificity	Precision	Recall	Accuracy	F1 score
InceptionV3	B	RGB	457	484	0.9182	0.9136	0.9659	0.9406	0.9389
	O	HSI	438	475	0.8849	0.8764	0.9457	0.9130	0.9097
	M	YCbCr	458	478	0.9196	0.9164	0.9534	0.9358	0.9345
	Ensemble		457	488	0.9184	0.9132	0.9740	0.9444	0.9425
ResNet101V2	B	RGB	452	482	0.9099	0.9044	0.9609	0.9338	0.9317
	O	HSI	443	473	0.8929	0.8864	0.9427	0.9162	0.9136
	M	YCbCr	455	477	0.9139	0.9100	0.9522	0.9322	0.9306
	Ensemble		455	485	0.9158	0.9108	0.9673	0.9400	0.9381
InceptionResNetV2	M	RGB	456	478	0.9161	0.9120	0.9548	0.9342	0.9326
	O	HSI	433	473	0.8759	0.8660	0.9417	0.9062	0.9022
	B	YCbCr	455	480	0.9140	0.9096	0.9576	0.9346	0.9329
	Ensemble		456	488	0.9174	0.9121	0.9740	0.9438	0.9419
MobileNetV2	B	RGB	451	486	0.9077	0.9012	0.9742	0.9364	0.9340
	O	HSI	434	477	0.8778	0.8672	0.9494	0.9104	0.9063
	M	YCbCr	449	481	0.9036	0.8972	0.9595	0.9296	0.9272
	Ensemble		452	488	0.9105	0.9040	0.9742	0.9400	0.9383

Table 6. *Unanimous prediction counts of four CNN architectures*

Model	Run	Malignant Unanimous Count	Benign Unanimous Count	Model	Run	Malignant Unanimous Count	Benign Unanimous Count
InceptionV3	1	419	452	InceptionResNetV2	1	414	456
	2	428	454		2	406	440
	3	428	437		3	416	432
	4	421	462		4	408	458
	5	414	452		5	414	443
ResNet101V2	1	425	446	MobileNetV2	1	411	463
	2	418	456		2	411	458
	3	425	443		3	418	447
	4	425	455		4	412	461
	5	416	449		5	405	458

The proposed ensemble model, which uses a majority voting method, consistently achieves higher performance than the individual classifiers used in the model. This is partly due to the diverse capture conditions of the ISIC dataset images, including varying cameras and environmental settings, which can affect the consistency of classifier predictions. In cases where all classifiers agreed on the classification, the ensemble method did not enhance performance, as the benefit of ensemble learning lies in its ability to leverage differing opinions. The enhancement in accuracy primarily comes from the instances where classifiers disagreed, and majority voting was applied to determine the outcome. Table 6 illustrates the unanimous prediction counts obtained from the three individual classifiers. For instance, in Run-1, InceptionV3 achieved a unified count of 419 for malignant case, indicating an agreement among the three classifiers for the same malignant samples. Since all single classifiers agree on class labels up to the

unanimous count, the remaining samples in the test set are classified using our proposed majority voting-based ensemble model.

Table 7. *The numbers of correct classifications obtained from the classifiers and the proposed ensemble model*

Model	Run	Classifier	Color Space	Single-Model				Ensemble			
				Malignant TP		Benign TN		Malignant TP		Benign TN	
				Count	%	Count	%	Count	%	Count	%
InceptionV3	1	B	RGB	450	90.0	484	96.8	451	90.2	485	97.0
		O	HSI	440	88.0	477	95.4				
		M	YCbCr	456	91.2	472	94.4				
	2	B	RGB	467	93.4	493	98.6	465	93.0	495	99.0
		O	HSI	441	88.2	471	94.2				
		M	YCbCr	463	92.6	484	96.8				
	3	B	RGB	462	92.4	477	95.4	464	92.8	485	97.0
		O	HSI	444	88.8	469	93.8				
		M	YCbCr	466	93.2	473	94.6				
	4	B	RGB	445	89.0	487	97.4	449	89.8	490	98.0
		O	HSI	437	87.4	482	96.4				
		M	YCbCr	455	91.0	480	96.0				
	5	M	RGB	460	92.0	478	95.6	454	90.8	484	96.8
		O	HSI	429	85.8	475	95.0				
		B	YCbCr	451	90.2	479	95.8				
ResNet101V2	1	B	RGB	466	93.2	483	90.2	461	92.2	489	97.8
		O	HSI	440	88.0	472	94.4				
		M	YCbCr	462	92.4	478	95.6				
	2	M	RGB	453	90.6	479	95.8	450	90.0	484	96.8
		O	HSI	436	87.2	479	95.8				
		B	YCbCr	450	90.0	479	95.8				
	3	M	RGB	457	91.4	477	95.4	460	92.0	480	96.0
		O	HSI	446	89.2	465	93.0				
		B	YCbCr	460	92.0	479	95.8				
	4	B	RGB	444	88.8	486	97.2	449	89.8	488	97.6
		O	HSI	442	88.4	476	95.2				
		M	YCbCr	452	90.4	478	95.6				
	5	B	RGB	441	88.2	483	96.6	457	91.4	482	96.4
		O	HSI	452	90.4	473	94.6				
		M	YCbCr	451	90.2	472	94.4				
InceptionResNetV2	1	M	RGB	466	93.2	483	96.6	465	93.0	493	98.6
		O	HSI	433	86.6	478	95.6				
		B	YCbCr	462	92.4	486	97.2				
	2	B	RGB	443	88.6	480	96.0	445	89.0	485	97.0
		O	HSI	431	86.2	470	94.0				
		M	YCbCr	449	89.8	472	94.4				
	3	M	RGB	461	92.2	466	93.2	464	92.8	485	97.0
		O	HSI	435	87.0	474	94.8				
		B	YCbCr	465	93.0	475	95.0				
	4	B	RGB	447	89.4	490	98.0	449	89.8	492	98.4
		O	HSI	433	86.6	470	94.0				
		M	YCbCr	449	89.8	490	98.0				
	5	M	RGB	463	92.6	472	94.4	457	91.4	484	96.8
		O	HSI	433	86.6	474	94.8				
		B	YCbCr	449	89.8	476	95.2				
MobileNetV2	1	B	RGB	458	91.6	490	98.0	462	92.4	493	98.6
		O	HSI	432	86.4	486	97.2				
		M	YCbCr	458	91.6	477	95.4				
	2	M	RGB	446	89.2	481	96.2	447	89.4	489	97.8
		O	HSI	434	86.8	480	96.0				
		B	YCbCr	444	88.8	485	97.0				
	3	B	RGB	453	90.6	484	96.8	454	90.8	483	96.6
		O	HSI	439	87.8	467	93.4				
		M	YCbCr	453	90.6	478	95.6				
	4	M	RGB	447	89.4	486	97.2	443	88.6	492	98.4

	O	HSI	429	85.8	478	95.6				
	B	YCbCr	443	88.6	488	97.6				
	B	RGB	449	89.8	488	97.6				
5	O	HSI	434	86.8	473	94.6	454	90.8	483	96.6
	M	YCbCr	445	89.0	477	95.4				

Table 7 presents the correct prediction counts (TP and TN) and their percentage values for the baseline classifiers and the ensemble model. In our study, the number of malignant and benign samples in test set is 500. The data underline the enhanced accuracy typically seen in single classifier working with RGB images over those processing images in alternative color spaces. For instance, in Run-3 for the InceptionV3 model, the RGB-based classifier achieved 462 true positives (TP) and 477 true negatives (TN). In contrast, the proposed ensemble model improved these results to 464 TP and 485 TN, demonstrating a notable enhancement in the classification of true samples. This pattern of superior performance by the ensemble model is consistent across all runs, illustrating its effectiveness over single classifier.

The accuracy and loss metrics for the different color spaces used in four CNN architectures are shown in Figure 5. The plots demonstrate that each CNN architecture learns the images represented in one of the three color spaces over 50 epochs. The validation accuracies achieved range from approximately 92.5% to 94.1%, and the validation losses decrease to a range of 20-31%.

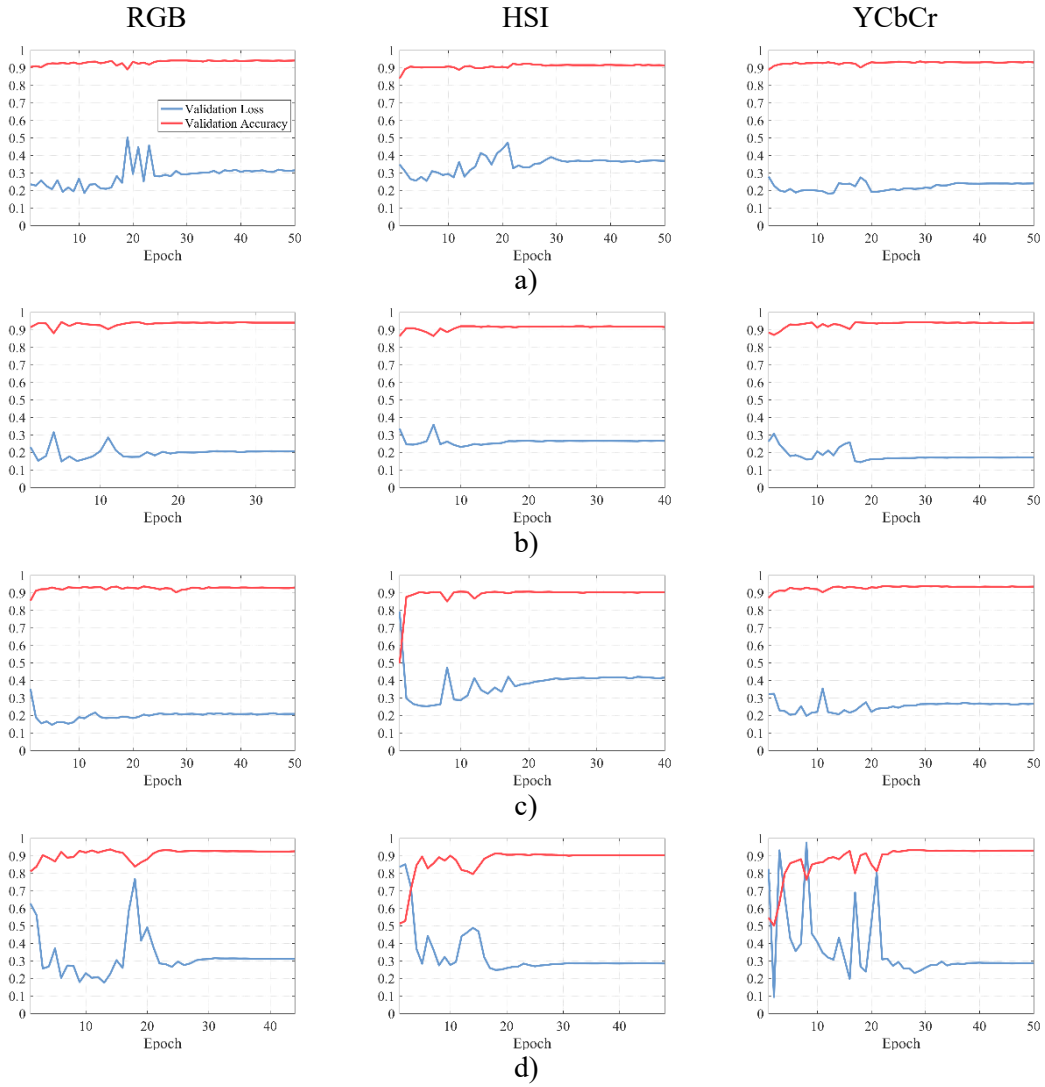
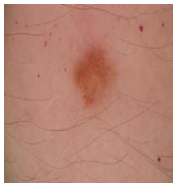
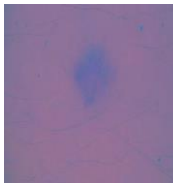
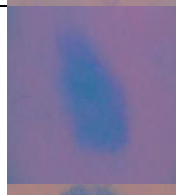
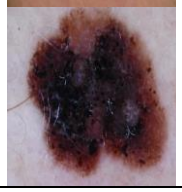
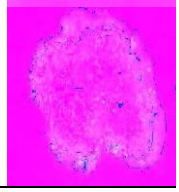
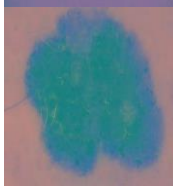
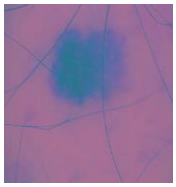
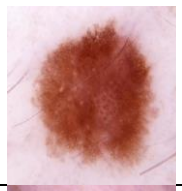
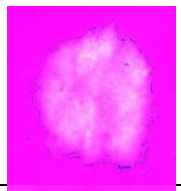
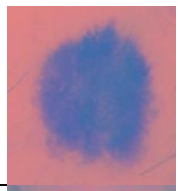
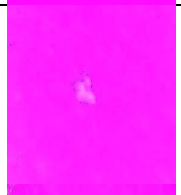
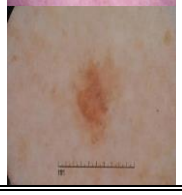


Figure 5. *Accuracy and loss plots of four CNN architectures for three-color spaces:a) InceptionV3, b) ResNet101V2, c) InceptionResNetV2, and d) MobileNetV2*

In the ISIC dataset, images were captured with a variety of cameras, under different lighting conditions, and at varying scales. Some images in the dataset are particularly small or featured a circular mask, challenging the Deep Neural Network (DNN) models' ability to consistently predict accurately in line with the ground truth. Some misclassified images by the proposed model are presented in Table 8.

Table 8. *Some examples of misclassified images*

CNN Model	Label	Ground Truth	Classification Result	RGB	HSI	YCbCr
InceptionV3	melanoma_3084	Benign	Malignant			
	melanoma_9486	Malignant	Benign			
ResNet101V2	melanoma_2463	Benign	Malignant			
	melanoma_5151	Malignant	Benign			
InceptionResNetV2	melanoma_9653	Benign	Malignant			
	melanoma_9256	Malignant	Benign			
MobileNetV2	melanoma_1602	Benign	Malignant			
	melanoma_8951	Malignant	Benign			

In addition, the performance of the InceptionV3, Resnet101V2, InceptionResNetV2 and MobileNetV2 models in each color space dataset and the performance of the proposed ensemble model were analyzed using the Friedman test [44], which is a statistical method. The Friedman test can be applied to two or more

models. It is used to determine whether there is a difference between the performances obtained from group members. With this non-parametric test, the performances of group members are analyzed for the significance level 0.05 (5%). If the results are less than 0.05, there is a statistical difference between group members. The results are ranked, and the importance of model is statistically emphasized. The results of the Friedman test for our study are given in Table 9. When the p-value in the table is examined, it is observed that this value was less than 0.05 for four CNN architectures. Therefore, there are significant differences between the models. The rankings of RGB, HIS, YCbCr and the proposed ensemble model, and also total rank are given in Table 10. The total rank is obtained by sorting the mean rank in descending order. The model with the smallest mean rank value indicates that the selected color space gives the best performance. It is observed that the proposed ensemble model is the best model respect to both mean rank and total rank values.

Table 9. *The Friedman test results*

Model	Color Space	Mean Rank	Rank	p-value
InceptionV3	RGB	3.00	2	1.09E-02
	HSI	1.00	4	
	YCbCr	2.40	3	
	Ensemble	3.60	1	
Resnet101V2	RGB	2.20	3	1.36E-02
	HSI	1.40	4	
	YCbCr	2.40	2	
	Ensemble	4.00	1	
InceptionResNetV2	RGB	2.60	2	3.60E-03
	HSI	1.00	4	
	YCbCr	2.40	3	
	Ensemble	4.00	1	
MobileNetV2	RGB	2.80	2	2.90E-03
	HSI	1.00	4	
	YCbCr	2.20	3	
	Ensemble	4.00	1	

Table 10. *The ranking of models respect to the mean rank values obtained from Friedman test*

Model	RGB	HSI	YCbCr	Ensemble
InceptionV3	2	4	3	1
Resnet101V2	3	4	2	1
InceptionResNetV2	2	4	3	1
MobileNetV2	2	4	3	1
Mean rank	2.25	4	2.75	1
Total rank	2	4	3	1

The performance of the proposed ensemble model was also evaluated on widely used PH2 dataset [52]. The dataset contains 160 benign and 40 malignant samples. The results for PH2 dataset were obtained from the same experimental setup. The average results are summarized in Table 11. The findings indicate that the proposed ensemble model outperforms individual color space-based classifiers on the PH2 dataset in terms of both F1 score and accuracy.

In this study, it is observed that different colors can enhance the correct classification of malignant and benign lesions. In the proposed model, a diversity of colors is provided using three color representations: RGB, HIS, and YCbCr. The baseline classifiers extracted different features to learn the lesion patterns. The proposed ensemble model reduced the misclassification of TP and TN, improving the F1 score. It is observed that the results obtained on the PH2 dataset are also consistent with those from the ISIC database. Notably, this improvement is obtained without any preprocessing on the ISIC dataset. The output of the ensemble model can give an improvement of more than 1% in the F1 score when the dataset is constructed in a standard way with consistent lighting, size, etc. The ensemble model will be used to support doctors' decisions. Therefore, any improvement in the F1 score is valuable. One can try to enhance the classification

performance by optimizing CNN model parameters. Additionally, CNN architectures can be modified. In health-related applications, the aim will always be high classification performance. Although the proposed ensemble model increases time and memory space complexity, it gives promising results for skin cancer diagnosis.

Table 11. *Average classification metrics obtained using 5-fold cross-validation on the PH2 dataset*

CNN	Classifier	Color Space	TP	TN	Specificity	Precision	Recall	Accuracy	F1 score
InceptionV3	B	RGB	5	32	1.0000	1.0000	0.6250	0.9250	0.9250
	O	HSI	5	26	0.8125	0.4545	0.6250	0.7750	0.5263
	M	YCbCr	6	27	0.8437	0.5455	0.7500	0.8250	0.6316
	Ensemble		8	31	0.9687	0.8889	1.000	0.9750	0.9412
ResNet101V2	B	RGB	7	25	0.7812	0.5000	0.8750	0.8000	0.6364
	O	HSI	5	27	0.8437	0.5000	0.6250	0.8000	0.5556
	M	YCbCr	3	32	1.0000	1.0000	0.3750	0.8750	0.5455
	Ensemble		5	32	1.0000	1.0000	0.6250	0.9250	0.7692
InceptionResNetV2	O	RGB	4	31	0.9687	0.8000	0.5000	0.8750	0.6154
	B	HSI	4	32	1.0000	1.0000	0.5000	0.9000	0.6667
	M	YCbCr	6	27	0.8437	0.5455	0.7500	0.8250	0.6316
	Ensemble		6	31	0.9687	0.8571	0.7500	0.9250	0.8000
MobileNetV2	O	RGB	6	24	0.7500	0.4286	0.7500	0.7500	0.5455
	M	HSI	6	27	0.8437	0.5455	0.7500	0.8250	0.6316
	B	YCbCr	3	32	1.0000	1.000	0.3750	0.8750	0.5455
	Ensemble		8	28	0.8750	0.6667	1.000	0.8500	0.8000

5. CONCLUSION

The correct classification of benign and malignant lesions is crucial for early diagnosis and effective treatment. Color is an important criterion in skin cancer diagnosis, and our method aimed to reduce the misclassification rate by combining the outputs of three different color-based CNN classifiers that learn common features of skin lesions, such as shape and color. Ensemble model studies in the literature are generally carried out by combining more than one deep learning model on the same color space. The novelty of our research is the improvement of skin cancer classification performance using a color-space-based ensemble model that employs majority voting to predict the class label. Our study successfully demonstrates the efficacy of an ensemble model for skin cancer classification, surpassing the performance of single-model classifiers and yielding improved results. The proposed model, which incorporates three color-based CNN classifiers and utilizes the majority voting method, was evaluated using four different CNN architectures (InceptionV3, ResNet101V2, MobileNetV2, and InceptionResNetV2) and three color spaces (RGB, HSI, and YCbCr). The experimental results demonstrate that the ensemble model consistently outperforms the single-model classifiers in terms of classification performance, with higher F1 scores, specificity, and precision values and reduced misclassification of benign and malignant samples. Additionally, the effectiveness of the proposed ensemble model across all four CNN-based deep learning models was validated using the Friedman statistical method, which demonstrated its superiority over the

other models. The results obtained on the ISIC and PH2 datasets demonstrate that the proposed ensemble model produces generalized outcomes.

The choice of color space plays a critical role in the classification performance, and the selected color spaces (RGB, HSI, and YCbCr) in this study were found to preserve the brightness of images and allow for the evaluation of three colors of information in the classifiers. Additionally, the findings highlight the importance of hue in differentiating between benign and malignant lesions.

Moving forward, there are several potential avenues for future research in this area. One potential direction is to further refine and optimize the proposed ensemble model by exploring different combinations of color spaces and CNN models, as well as tuning the parameters of the models. Another possibility is to investigate the impact of image preprocessing techniques, such as applying filters or enhancing contrast, on classification performance. Furthermore, future research could focus on developing a more extensive dataset with standardized conditions for image acquisition, including scale and lighting intensity, which would enable a more accurate and reliable classification of skin lesions. Ultimately, a robust and accurate decision support system for skin cancer classification could have significant clinical implications, enabling early detection and treatment of malignant lesions while reducing unnecessary biopsies for benign lesions.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/. Access date: 26.02.2025

[2] Parkin, D.M., Mesher, D., and Sasieni, P., "13. Cancers attributable to solar (ultraviolet) radiation exposure in the UK in 2010", British Journal of Cancer, 105(2): 66-9, (2011).

[3] Filimon, D.M., and Albu, A., "Skin diseases diagnosis using artificial neural networks", 9th IEEE International Symposium on Applied Computational Intelligence and Informatics, Timisoara, Romania, 189-194, (2014).

[4] Hasan, M.K., Ahamad, M.A., Yap, C.H., and Yang, G., "A survey, review, and future trends of skin lesion segmentation and classification", Computers in Biology and Medicine, 155(106624): 106624, (2023).

[5] Matthews, N.H., Li, W.Q., Qureshi, A.A., Weinstock, M.A., Cho, E., Cutaneous Melanoma: Etiology and Therapy, Codon Publications, Brisbane (AU), 3-22, (2017).

[6] https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2020.html. Access date: 26.02.2025

[7] Rao, B.K., and Ahn, C.S., "Dermatoscopy for melanoma and pigmented lesions", Dermatologic Clinics, 30(3): 413–434, (2012).

[8] Kittler, H., Pehamberger, H., Wolff, K., and Binder, M., "Diagnostic accuracy of dermoscopy", Lancet Oncology, 3(3): 159–165, (2002).

[9] Uhl, F., Græsdal Rasmussen, T., and Oppelt, N., "Classification ensembles for beach cast and drifting vegetation mapping with Sentinel-2 and PlanetScope", Geosciences (Basel), 12(1): 15, (2021).

[10] Kuncheva, L.I., Whitaker, C.J., Shipp, C.A., and Duin, R.P.W., "Limits on the majority vote accuracy in classifier fusion", Pattern Analysis and Applications, 6(1): 22–31, (2003).

[11] Santos, A.R.F., Aires, K.R.T., and Veras, R.M.S., "Aspects of lighting and color in classifying malignant skin cancer with deep learning", Applied Sciences (Basel), 14(8): 3297, (2024).

[12] Abbas, Q., Celebi, M.E., Garcia, I.F., and Ahmad, W., "Melanoma recognition framework based on expert definition of ABCD for dermoscopic images", Skin Research and Technology, 19(1): e93-102, (2013).

[13] Barata, C., Figueiredo, M.A.T., Celebi, M.E., and Marques, J.S., "Color identification in dermoscopy images using gaussian mixture models", 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 3611-3615, (2014).

[14] Shalu, N.A., Rani, R., and Kamboj, A., "Automated melanoma skin cancer detection from digital images", International Journal of Biomedical Engineering and Technology, 37(3): 275, (2021).

[15] Oukil, S., Kasmi, R., Mokrani, K., and García-Zapirain, B., "Automatic segmentation and melanoma detection based on color and texture features in dermoscopic images", Skin Research and Technology, 28(2): 203–211, (2022).

[16] Mamadou, D., Ayikpa, K.J., Ballo, A.B., and Kouassi, B.M., "Analysis of the impact of color spaces on skin cancer diagnosis using deep learning techniques", Revue d'Intelligence Artificielle, 37(6): 1377–1385, (2023).

[17] Nigar, N., Umar, M., Shahzad, M.K., Islam, S., and Abalo, D., "A deep learning approach based on explainable artificial intelligence for skin lesion classification", IEEE Access, 10: 113715–113725, (2022).

[18] Xie, Y., Zhang, J., Xia, Y., and Shen, C., "A mutual bootstrapping model for automated skin lesion segmentation and classification", IEEE Transactions on Medical Imaging, 39(7): 2482–2493, (2020).

[19] Jayapriya, K., and Jacob, I.J., "Hybrid fully convolutional networks-based skin lesion segmentation and melanoma detection using deep feature", International Journal of Imaging Systems and Technology, 30(2): 348–357, (2020).

[20] Aldwgeri, A., and Abubacker, N.F., "Ensemble of deep convolutional neural network for skin lesion classification in dermoscopy images", 6th International Visual Informatics Conference, Bangi, Malaysia,19-21, (2019).

[21] Li, Y., and Shen, L., "Skin lesion analysis towards melanoma detection using deep learning network", Sensors (Basel), 18(2), (2018).

[22] Saravana Kumar, N.M., Hariprasath, K., Tamilselvi, S., Kavinya, A., and Kaviyavarshini, N., "Detection of stages of melanoma using deep learning", Multimedia Tools and Applications, 80(12): 18677–18692, (2021).

[23] Zhang, N., Cai, Y.X., Wang, Y.Y., Tian, Y.T., Wang X.L., and Badami, B., "Skin cancer diagnosis based on optimized convolutional neural network", Artificial Intelligence in Medicine, 102: 101756, (2020).

[24] Mohakud, R., and Dash, R., "Designing a grey wolf optimization based hyper-parameter optimized convolutional neural network classifier for skin cancer detection", Journal of King Saud University - Computer and Information Sciences, 34(8): 6280–6291, (2022).

[25] Gu, R., Wang, L., and Zhang, L., "DE-Net: a deep edge network with boundary information for automatic skin lesion segmentation", Neurocomputing, 468: 71–84, (2022).

[26] Ali, K., Shaikh, Z.A., Khan, A.A., and Laghari, A.A., "Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer", Neuroscience Informatics, 2(4): 100034, (2022).

[27] Saba, T., Khan, M.A., Rehman, A., and Marie-Sainte, S.L., "Region extraction and classification of skin cancer: a heterogeneous framework of deep CNN features fusion and reduction", Journal of Medical Systems, 43(9): 289, (2019).

[28] Jojoa Acosta, M.F., Caballero Tovar, L.Y., Garcia-Zapirain, M.B., and Percybrooks, W.S., "Melanoma diagnosis using deep learning techniques on dermatoscopic images", BMC Medical Imaging, 21(1): 6, (2021).

[29] Mahum, R., and Aladhadh, S., "Skin lesion detection using hand-crafted and DL-based features fusion and LSTM", Diagnostics (Basel), 12: 2974, (2022).

[30] Al-Saedi, D.K.A., and Savaş, S., "Classification of skin cancer with deep transfer learning method", Computer Science, 202-210, (2022).

[31] Nersisson, R., Iyer, T.J., Joseph Raj, A.N., and Rajangam, V., "A dermoscopic skin lesion classification technique using YOLO-CNN and traditional feature model", Arabian Journal for Science and Engineering, 46(10): 9797–9808, (2021).

[32] Guo, L., "A multimodel ensemble method using CNN and maximum correntropy criterion for basal cell carcinoma and seborrheic keratoses classification", 2019 International Joint Conference on Neural Networks, Budapest, Hungary, 1-6, (2019).

[33] Oliveira, R.B., Pereira, A.S., and Tavares, J.M.R.S., "Skin lesion computational diagnosis of dermoscopic images: Ensemble models based on input feature manipulation", Computer Methods and Programs in Biomedicine, 149: 43–53, (2017).

[34] Schaefer, G., Krawczyk, B., Celebi, M.E., and Iyatomi, H., "An ensemble classification approach for melanoma diagnosis", Memetic Computing, 6(4): 233–240, (2014).

[35] Cao, X., Pan, J.S., Wang, Z., Sun, Z., Haq, A.U., Deng, W.A., and Yang, S., "Application of generated mask method based on mask R-CNN in classification and detection of melanoma", Computer Methods and Programs in Biomedicine, 207: 106174, (2021).

[36] Harangi, B., "Skin lesion classification with ensembles of deep convolutional neural networks", Journal of Biomedical Informatics, 86: 25–32, (2018).

[37] Le, D.N.T., Le, H.X., Ngo, L.T., and Ngo, H.T., "Transfer learning with class-weighted and focal loss function for automatic skin cancer classification", arXiv, (2020).

[38] https://www.kaggle.com/datasets/hasnainjaved/melanoma-skin-cancer-dataset-of-10000-images. Access date: 26.02.2025

[39] Hasan, M.K., Dahal, L., Samarakoon, P.N., Tushar, F.I., and Martí, R., "DSNet: automatic dermoscopic skin lesion segmentation", Computers in Biology and Medicine, 120: 103738, (2020).

[40] MacDonald, L.W., and Ronnier Luo, M., "Colour imaging, vision and technology", Color Research and Application, 27, 6: 455–455, (2002).

[41] Tkalcic, M., and Tasic, J.F., "Color spaces: perceptual, historical and applicational background", The IEEE Region 8 EUROCON 2003, Computer as a Tool, Ljubljana, Slovenia, 304-308, (2003).

[42] Chollet, F., "Xception: Deep learning with depthwise separable convolutions", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 1800-1807, (2017).

[43] Schwenker, F., "Ensemble methods: foundations and algorithms", IEEE Computational Intelligence Magazine, 8(1): 77–79, (2013).

[44] He, K., Zhang, X., Ren, S., and Sun, J., "Identity mappings in deep residual networks", 14th Computer Vision European Conference, Amsterdam, Netherlands, 630–645, (2016).

[45] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A., "Inception-v4, Inception-ResNet and the impact of residual connections on learning", AAAI Conference on Artificial Intelligence, San Francisco, California USA, (2017).

[46] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.C., "MobileNetV2: Inverted residuals and linear bottlenecks", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 4510-4520, (2018).

[47] Mohammed, A., and Kora, R., "A comprehensive review on ensemble deep learning: opportunities and challenges", Journal of King Saud University - Computer and Information Sciences, 35(2): 757–774, (2023).

[48] Day, W., "Consensus methods as tools for data analysis", 1st Conference of the International Federation of Classification Societies, North-Holland, 317–324, (1987).

[49] Grofman, B., Owen, G., and Feld, S.L., "Thirteen theorems in search of the truth," Theory and Decision, 15(3): 261–278, (1983).

[50] Kuncheva, L.I., and Alpaydin, E., "Combining pattern classifiers: methods and algorithms", IEEE Transactions on Neural Networks, 18(3): 964–964, (2007).

[51] Nandi, A., and Pal, A.K., Interpreting Machine Learning Models Berkeley, CA, USA: Apress, 97–115, (2022).

[52] Mendonca, T., Ferreira, P.M., Marques, J.S., Marcal, A.R.S., and Rozeira, J., "PH^2 - a dermoscopic image database for research and benchmarking", Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Osaka, Japan, 5437–5440, (2013).