IJCI V9-02(2022) 74–87
International Journal of Computers and Information (IJCI)
Available online at *https://ijci.journals.ekb.eg/*

An Enhanced Technique for Skin Lesion Diagnosis using Dermoscopic Images

Aya Mostafa Mosa [a], Ahmed Afifi [a,b], Khaled Mohammed Amin [a]

[a] Information Technology dept., Faculty of Computers and Information, Menoufia University, Menoufia 32511, Egypt
[b] Department of Computer Science, College of Computer Science and Information Technology, King Faisal University, P.O. Box 400, Al-Ahsa 31982, Saudi Arabia
aya.moosa@gmail.com, afifi@ci.menofia.edu.eg, k.amin@ci.menofia.edu.eg

Abstract

There are many types of skin cancer, the most harmful among them is melanoma. Skin cancer occurs through the abnormal growth of the body cells, which may be caused by continuous exposure to ultraviolet rays resulting from the sun. Early diagnosis of skin cancer is essential as it can reduce the burden, make the treatment more effective and save the patient life. In this work, therefore, we develop an enhanced ensemble approach to improve the classification accuracy of eight types of skin cancer. Three pretrained Convolutional Neural Network (CNN) models, namely ResNet18, DensNet121, and Inception v4, are used as a base for this ensemble. Firstly, we fine-tune each pre-trained CNN model separately on an augmented dataset. Afterwards, the prediction probabilities obtained from all the base learners are combined as a new feature vector for each case. These features are used to build another classifier that employs the outputs of all the networks used in the previous stage. Thus, this ensemble learns how to weigh the output of each base learner without any user interaction. Support vector machine (SVM), and random forest (RF) classifiers are utilized in this stage and compared to the average ensemble which gives the same importance to all base learners. The pre-trained CNN models were fine-tuned and evaluated using 17731 and 3800 images from different types of skin cancer, respectively. The ResNet18, DensNet121, and InceptionV4 achieved individual accuracy of 79.5%, 81.2%, and 82.6%, respectively. The proposed ensemble method using SVM, and RF classifiers gives the best accuracy result with 85% for the SVM classifier and 86.2% for the Random Forest classifier. These results show that the proposed ensemble method using SVM, and RF classifiers is superior to the single classifiers of ResNet18, DensNet121 and InceptionV4 as well as the commonly used average ensemble which achieved an accuracy of 83.2%.
Keywords: Skin Cancer; Classification; Dermoscopic Images; Deep learning; Conventional neural network; Ensemble.

1. Introduction

A person's life is threatened when he/she is infected with cancer because it can lead to the death. Many cancers affect humans, including skin cancer, which must be treated early so that it does not spread quickly and cause death [1]. The most common factor that causes skin cancer is exposure to ultraviolet rays as it affects the DNA of skin cells [2][3][4]. Skin cancer is classified into two types, benign tumor, and malignant tumor, and there are 8 types of malignant tumors, and they are Actinic Keratosis, Squamous Cell Carcinoma, Basal Cell Carcinoma, Seborrheic Keratosis, Solar Lentigo, Dermatofibroma, Nevi, and Melanoma. However, the most dangerous one is the Melanoma [5]. Skin cancer is more common in the United States, Australia, and Canada [6] where the skin is sensitive to the sun. Therefore, early detection and treatment of skin cancer can reduce the consequences. Accordingly, researchers are seeking to develop computer systems to accurately diagnose skin cancer. In this research, therefore, we propose an improved ensemble approach that learns the importance of each base learner and utilizes it to obtain more accurate results. All experiments in this were conducted using ISBI 2019 benchmark dataset [7][8].

This paper is organized as follows: after this introduction, the related work is discussed in Section 2. In Section 3, the proposed methodology is presented. The results are explained in Section 4 and the paper is concluded in Section 5.

2. Related Work

Dermatologists diagnose skin cancer by analyzing dermoscopic images and determining whether they have cancerous cells or not. Cancerous cells are diagnosed as benign or malignant by the dermatologist who analyzes skin cancer by using a biopsy. In this method, samples of skin lesion are extracted for examination to determine whether there are cancerous cells or not. This medical process is costly, painful, and time-consuming. Therefore, the development of an automated system for skin lesions diagnosis is very important. In literature, there are different systems that have been developed to detect skin lesions for years as we will briefly describe in this section.

In [9], Skin lesion was detected by transfer learning of the AlexNet model using 10-classes dataset. The original fully connected layers for the AlexNet were converted to convolution layers with the same weights. This system achieved an accuracy of 81.8%. However, the authors used a small and non-standard dataset. In [10], the Author developed a weighted ensemble network using Inception, AlexNet, and VGG models for binary classification of skin lesions. This model was trained and evaluated using ISBI 2017 dataset and achieved an accuracy of 83.8%. For final prediction, the author inserted a joint SoftMax activation with a fully connected layer after removing classification and final fully connected layers of the base learners. However, they used ISIC 2017 dataset which is less challenging than ISIC 2019 dataset which contains various skin cancer types.

In [11], Seven skin diseases were classified by using some ensemble transfer learning models such as SeResNeXt, ResNeXt, SeNet, DenseNet, and others. The class imbalance problem was solved by using different techniques such as normalized class frequency and simple loss weighting which achieved the best result. This system used 54 CNN models for the ensemble. These models were selected by analyzing the performance of 5-fold cross-validation. This ensemble achieved the best-weighted accuracy (WACC) of 85%. However, this research takes more time to ensemble 54 CNN models using ISIC 2018 and HAM10000 datasets. In [12], the Authors classified seven skin diseases by training all layers of DenseNet and MobileNet pre-trained models. They also applied class balancing of the seven classes in the used dataset. The two models were trained and evaluated using the processed HAM10000 dataset. The misbalancing problem was resolved by down-sampling then splitting and augmenting the dataset. MobileNet achieved a high performance of 92.7% accuracy after solving the class imbalance problem trough augmentation and down-sampling. However, the authors used a non-standard and small dataset.

In [13], the Authors detect seven skin diseases using CNN. The features are extracted from the ISIC 2018 dataset after the segmentation process. The features were extracted and classified by the convolutional neural network. This model achieved a validation accuracy of 89.5% and a accuracy of 93.7%. They achieved good accuracy but there was information loss by shrinking the size of the image during the convolution operation when applying the filters on images. In [14], the Authors segment the ISIC 2018 dataset using a full Resolution Convolutional Network (FRCN) and classifies seven skin diseases by using DenseNet-201, InceptionV3, Inception-ResNet-V2, and ResNet-50 pre-trained models. Dataset size was increased by using flipping and rotation data augmentation. The class Imbalance problem was solved by using the weighted cross-entropy method. This experiment achieved the highest accuracy of 89.28% for ResNet50. This technique is evaluated on the ISIC 2018 dataset which contains different types of skin lesions.

In [15], Seven skin diseases were classified by transfer learning of ResNet50, VGG16, MobileNet, and EfficientNet B1 models. ResNet50 model outperformed other models with an accuracy of 90%. Author modifying and reusing pre-trained model architecture with class-weighted and focal loss techniques to classify skin lesion images in the HAM10000 dataset. They achieved average accuracy of 93% and precision

in the range [0.7, 0.94]. However, HAM10000 is a small dataset and easy dataset. In [16], Steppan et al. classify nine skin diseases by developing an ensemble model using transfer learning. Data augmentation techniques were used after combining 6 datasets. Moreover, the author used 3 methods to handle the class imbalance problem after pre-processing all images to eliminate the black area that surrounds the dataset images. EfficientNet has achieved the best accuracy of 63%. However, their approach is not accurate despite using a large database.

Although a lot of research has been proposed for diagnosing skin cancer, there are imbalance problems in the dataset which usually affects accuracy. Also, many researchers used nonstandard datasets and used CNN models individually which does not give the best accuracy as ensemble models. In this work, we used a standard benchmark dataset with more challenges, handle class imbalance problems, and proposed an improved ensemble methodology to enhance the diagnosis accuracy.

3. The Proposed Approach

The proposed approach tends to classify eight different categories of skin cancer using deep learning-based techniques. The idea of this strategy is to improve the CNN model's performance by combining multiple CNN models with different characteristics such as Resnet18, Densnet121, and Inception v4. The proposed ensemble approach will learn the importance of each model directly from the data.

3.1 Methodology

We built an accurate automated model by utilizing three deep CNN models namely ResNet18 [17], Densnet121 [18], and Inception v4 [19]. A color constancy algorithm is applied to resized images to alleviate the effect of capturing condition and the computational complexity of the proposed approach. Data augmentation is also applied to increase the data variability and we handle class imbalance problem using a weighted loss function. Then we improved the accuracy of these models by developing several ensemble methods. We applied the average ensemble method for ResNet18 with DensNet121 and Inception v4, which gave a higher accuracy than each model. Finally, we used the trainable ensemble method to improve the accuracy of the ensemble process by using SVM [20] and Random Forest [21] classifiers based on the prediction probabilities of all models as shown in Figure 1. The details of the proposed approach will be explained in the following sections.

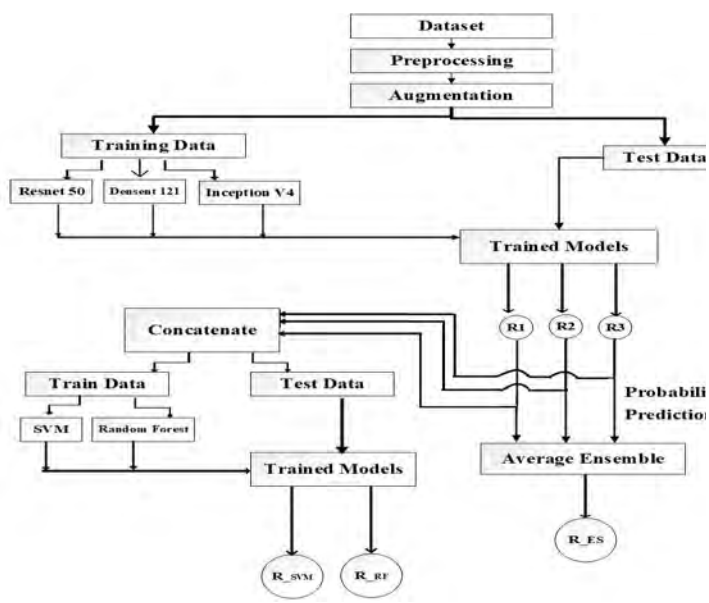

Fig. 1. Proposed CNNs models and ens

1) Data preprocessing

In this section, we resize the dataset images to 224×224 to reduce the computational complexity. As an additional step of preprocessing, we applied a color constancy algorithm on all dataset images. Color constancy is a major issue for any visual system doing a task that requires a consistent sense of color. Color constancy was performed to transform the colors of an image acquired under an unknown light source so that under a source of canonical light, they appear identical. This transformation is accomplished by estimating the light source color in RGB color space and then transforming the image using the estimated illumination as shown in Figure 2. There are common algorithms to estimate illuminant color such as the Shades of Gray, Gray World, and Max RGB methods which correct the image color by estimating the R, G, B illumination vector components after calculating the gain factors for each color channel. Max-RGB and Gray World achieve satisfying results if average color of the scene is the maximum is white or approximating grey. Shades of Grey [22] algorithm achieve satisfying result between Max-RGB and Gray World algorithm. The Shades of Gray algorithm uses the Minkowski norm to estimate the illumination color. Finlayson and Trezzi computed the Shades of Grey as in Equation (1).

$$Ke = \left(\frac{\int (f(X))^p dX}{\int dX}\right)^{1/p} \quad (1)$$

Where p is the Minkowski norm, Ke is the color correction factor for component C of the image, C ∈ {R, G, B} and f(x) is the selected pixel value of the component C. The estimated illumination vector is formed via the normalized which calculates the weighted average of the pixels by assigning high weights to pixels of higher intensities. The proposed weight function is based on the Minkowski-norm p which when p equal to ∞, 1, 1< p < ∞ the color constancy is performed by Max-RGB, Gray World and Shades of Grey algorithm respectively. From many experiments, the Shades of Grey algorithm achieves the best result when p is equal to 6.

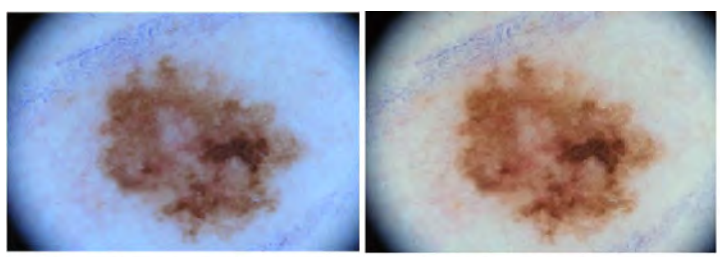

Fig. 2. The effect of color consistency operation

2) Data Augmentation

The data augmentation process creates new synthetic data from the existing training data to increase the variability of the training data and make it rich. It usually improves the performance of deep learning models by reducing overfitting, increasing model generalization ability. Accordingly, deep learning models can achieve better results by applying an appropriate amount of data augmentation [23]. In this work, data augmentation using horizontal and vertical flipping, rotation with a range from 0 to 180 degrees, and color normalization.

3) Imbalance Handling

The number of samples in each class of skin lesion types is significantly different and this usually led to biased classifiers. To reduce the effect of the class imbalance problem; we used a weighted loss function to finetune the base classifiers. The class weight is determined by using the inverse of the class's frequency in training data as in Equation (2).

$$\text{class\_weigh}_i = \frac{1}{N_i}, \quad (2)$$

Where, N_i is the number of samples in i^{th}class.

4) Base learners

a) ResNet18

Residual Network (ResNet) [17] is a CNN architecture which uses skip connection to enhance the gradient flow and the whole training process, as shown in Figure 3. It has several blocks of convolution and pooling layers that are fully connected and stacked one over the other. There are several variants of ResNet known as ResNet18, ResNet34, ResNet50, ResNet101 and ReseNet152 with different number of blocks and training parameters. In this work, we use ReseNet18 which has the lowest number of parameters as the dataset is not very large.

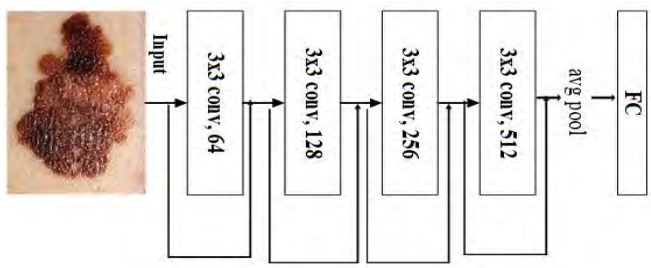

Fig.3. ResNet-18-Arc

b) Densnet121

DenseNet [18] is a CNN architecture which uses dense connections to preserve the natural of feed-forward. It concatenates the previous layer output with the future layer so, it is different from ResNet because ResNet merges the previous layer with the future layer. DenseNet is extremely powerful because it connects every layer to every other layer. Its input layer is feature maps concatenation from previous layers. DenseNet consists of 4 Dense Blocks with different layers number as DenseNet-121 has [6, 12, 24, 16] layers in its 4 dense blocks. The advantages of densnet121 are that strengthens the propagation of features, reduces the number of the parameter, encourages the reusing of features, and alleviates the vanishing gradient problem. The following Figure 4 illustrates DenseNet-121 architecture [24].

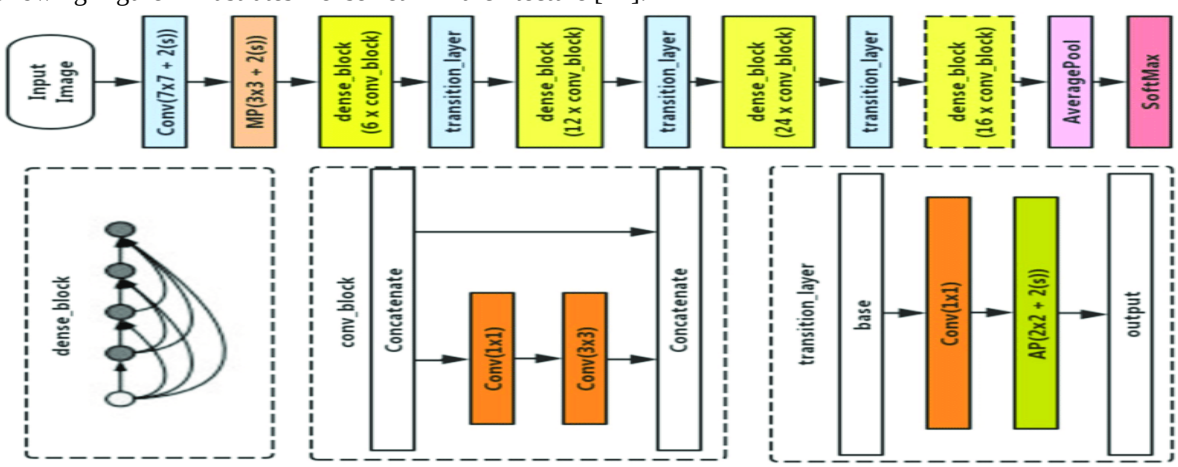

Fig. 4. Left-DenseNet121-architecture-Right-Dense-block-conv-block-and-transition-layer [24]

c) Inception v4

Inception-v4 [19] is a CNN architecture that achieves good performance with low computational cost. It is an updated version of Inception v3 with a more adding simplified architecture. Inception-v4 was extended by adding a residual connection to merge high-level features with low-level features that improved model performance. The stem layer of the Inception-v4 architecture preprocesses data. Then comes the feature reuse stage, where low-level features (general features) that were extracted from the earlier CNN layers, such as edges, shapes, and colors of lesions, concatenate with high-level features extracted by the Inception-v4 model. The advantages of inception v4 are that optimizes memory to backpropagation, train without replica splitting and it is a pure variant of inception variants with no remaining connections. The following Figure 5 illustrates Inception-v4 architecture [25].

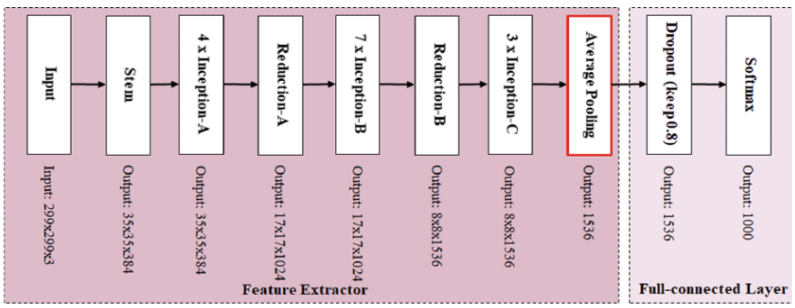

Fig.5. The-overall-schema-of-the-Inception-V4-network [25]

5) Model Ensemble

The proposed system diagnoses eight types of skin cancer by ensemble of three CNN models; ResNet18, DenseNet121, and Inception V4 which are fine-tuned using the ISIC 2019 dataset. It is also known that increasing the number of classes led to a more complex model and it may lead to low performance. Therefore, we develop an ensemble approach to deal with the multiclass complexity problem by combining the predictions of individual base learners to enhance the result accuracy. There are many techniques to develop ensemble learning to generate accurate and robust combined models such as stacking, voting, and averaging strategies. Average ensemble learning is the most popular ensemble method and is performed by averaging the output of final SoftMax layer of all base learners as in Equation (3).

$$\text{Prediction} = \frac{\sum \text{Pi}}{\text{N}} \tag{3}$$

Where *Pi* is the probability obtained by the i^{th} base learner and N is the total number of base learners. As it is can be clearly noticed, this average ensemble technique gives an equal importance for all models. However, it is not suitable when several models with different characteristics are used. Accordingly, weighted ensemble which assigns different weight for each base learner is more suitable here. Traditional weighted average voting ensemble gives importance to the deep learners based on practical values that are calculated by author and it is not easy to determine this value. Therefore, the main contribution of this research is proposing a new ensemble learnable weighted ensemble model which learns weight from all probabilities of all classifiers automatically. So finally, we used this proposed ensemble model to improve the ensemble performance by using the SVM and Random Forest classifier. We concatenated the predictions of Resnet18 with Densnet121 and Inception v4 and used it as an input vector to train and evaluate SVM and Random Forest classifiers. We used a linear kernel to train and evaluate the SVM classifier. The predictions dimension of training data was (17731, 8) and the predictions dimension of testing data was (3800, 8) for each individual base learner model. So, the length of the input feature vector to train SVM and Random Forest classifiers was (17731, 24) and evaluate SVM and Random Forest classifiers was (3800, 24) for developed the proposed learnable weighted ensemble model.

a) SVM

Support Vector Machine (SVM) is a linear model that is used for regression and classification problems. It also solves nonlinear and linear problems. SVM separates data in classes by creating a hyperplane or line. The advantages of SVM are to separate classes well when identifying margin to separate classes, effective in high dimensional spaces, efficient in using memory, and effective when the number of the dimension is greater than the samples number. The following Figure 6 illustrates SVM architecture [26].

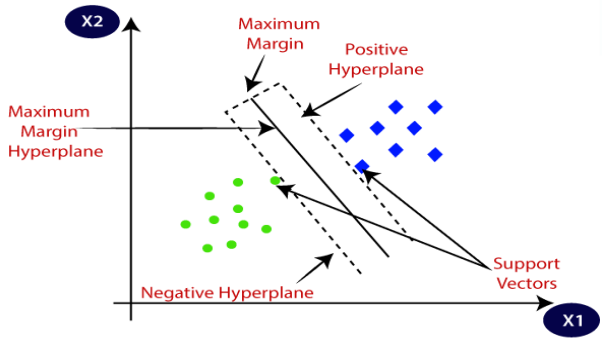

Fig. 6. Support Vector Machine Algorithm [26]

b) Random Forest

A Random Forest classifier is an algorithm of machine learning. It works as a technique of ensemble learning (bagging). It creates many decision trees on a subset of the data and then combines the output of these decision trees. The advantages of Random Forest are that improves accuracy by reducing the variance and overfitting problem, handles missing values, works with continuous and categorical variables, is used in regression and classification, handles outliers, is comparatively less impacted by noise, and is not significantly affected by introducing new data point because it only affects one tree, so it is stable. The following Figure 7 illustrates Random Forest architecture [27].

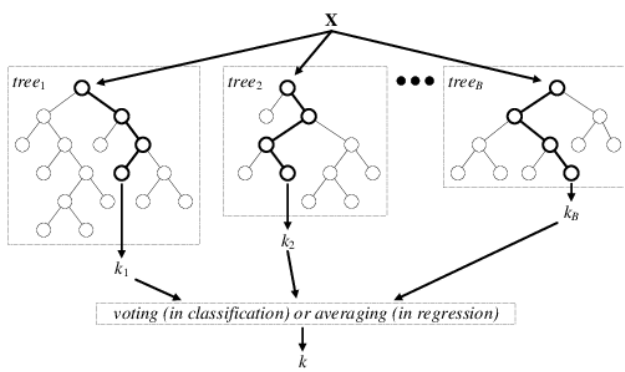

Fig.7. Architecture-of-the-random-forest-model [27]

3.2 Performance Measures

Several metrics recommend by ISIC 2019 dataset to evaluate the performance including accuracy, balanced accuracy, precision, recall, and f1 score.

1) Accuracy

Accuracy (ACC) is a metric that measures the model's ability to recognize correctly classified lesions in the dataset. It calculates many correctly classified negative and positive samples. Let, N_tp, N_tn, N_fp and N_fn represent the number of true negative, false negative, true positive, and false positive, respectively as in Equation (4).

$$\mathrm{AC} = \frac{\mathrm{N_{tp}} + \mathrm{N_{tn}}}{\mathrm{N_{tp}} + \mathrm{N_{tn}} + \mathrm{N_{fp}} + \mathrm{N_{fn}}} \tag{4}$$

2) Precision

Precision is a metric that measures the model's ability to predict data samples of the positive class. It calculates the proportion of positive samples that were actually predicted correctly as in Equation (5).

$$\mathrm{Prec} = \frac{\mathrm{N_{tp}}}{\mathrm{N_{tp}} + \mathrm{N_{fp}}} \tag{5}$$

3) Recall

The recall is a metric called Sensitivity. It calculates the proportion of positive samples that were actually predicted correctly to all samples in the actual class as in Equation (6).

$$\mathrm{Recall} = \frac{\mathrm{N_{tp}}}{\mathrm{N_{tp}} + \mathrm{N_{fn}}} \tag{6}$$

4) F1 Score

F1 Score is the weighted average function of Precision and Recall. It is useful when we use an unbalanced dataset. It takes into account both false negative and false positive as in Equation (7).

$$\mathrm{F1\ Score} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{7}$$

4. Experimental results

4.1 Dataset

We use the International Skin Imaging Collaboration (ISIC) 2019 dataset [7][8] which contains 25331 dermoscopic training images of nine skin cancer types. The nine classes of the ISIC2019 training dataset are: melanocytic nevus (NV), vascular lesion (VASC), melanoma (MEL), Squamous cell carcinoma (SCC), actinic keratosis (AK), basal cell carcinoma (BCC), dermatofibroma (DF), benign keratosis (BKL), and UNKNOWN. We diagnose eight classes from nine classes of skin cancer types because the ninth class is an unknown skin cancer type and not found in the training dataset. The number of samples in each class is shown in Table I. It can be clearly noticed that this dataset suffers from a severe class imbalance problem. This dataset has other challenges as unclear tumors, dense hair, and incomplete tumors. We split this dataset into 70% as a training set (17731 images), 10% as a validation set (3800 images), and 20% as a testing set (3800 images).

Table 1. *Distribution of dataset diagnosis*

CLASS NAME	MEL	NV	BCC	AK	BKL	DF	VASC	SCC	UNK
NUM IMAGES	4522	12875	3323	867	2624	239	253	628	0
RATIO	0.178	0.508	0.13	0.03	0.1	0.009	0.01	0.024	0

4.2 Experimental Results

We fine-tuned resnet18, densnet121, and inception v4 as transfer learning models on ISIC 2019 dataset. We split ISIC 2019 training dataset into 70% as a training set, 10% as a validation set, and 20% as a testing set. We fine-tuned all models on 100 epochs with a batch size of 32 and used Adam [28] optimizer with a start learning rate of 5e-5. We evaluated the deep leaner models and ensemble models by using accuracy, balanced accuracy, recall, precision, and f1 score measures.

Table II shows the performance of individual base learners of ResNet18, DenseNet121, Inception V4, and the ensemble approaches. The single Inception V4 learner achieved higher performance than other individual learners because its ability to learn more feature than ResNet18 and DenseNet121. The proposed ensemble learning models were developed to improve the learning system's generalization. Therefore, from Table II, it is observed that the proposed ensemble learning models achieved higher performance than individual base learners in terms of accuracy, recall, precision, and f1 score. The individual learner's accuracy of ResNet18, DenseNet121, and Inception V4 is respectively, 79.5%, 81.2%, and 82.6%. However, the Average ensemble method accuracy is 83.2%. SVM and Random Forest were developed to enhance the performance of the average ensemble method by combining the outputs of the base learners as the ensemble classification. SVM and Random Forest accuracy are respectively, 84.9% and 86.2%.

Table III shows the accuracy of each class of comparative classification of individual base learners of ResNet18, DenseNet121, Inception V4, and proposed ensemble learning models. The training ISIC 2019 dataset contains eight classes which are melanocytic nevus (NV), vascular lesion (VASC), melanoma (MEL), Squamous cell carcinoma (SCC), actinic keratosis (AK), basal cell carcinoma (BCC), dermatofibroma (DF) and benign keratosis (BKL). These diagnosis classes have different numbers of samples. Therefore, the single base learners and proposed ensemble models achieved high accuracy when diagnosing the melanocytic nevus (NV) class than other classes because the melanocytic nevus class has more samples than other classes. Finally, SVM and Random Forest diagnosed the eight classes with high accuracy than other single base learners and average ensemble methods. Especially for melanoma, proposed ensemble models achieved higher accuracy with 71% than other techniques because melanoma is the most harmful skin cancer type. Only three classes VASC, DF and AK out of eight classes achieved worse accuracy results with respectively, 82.8%, 53% and 51.9% for Random Forest and 82.3%, 59.4% and 46.4% for SVM than the other classes but for the other, the majority the difference in results is high. So, in the future, we will make a deep analysis of this problem to investigate what is the reason for this.

Table IV shows the comparison between the proposed ensemble models and the recent ensemble models in [29,30,31,32,33,34,35] Although the recent ensemble models used the normal ensemble algorithms but some of them achieved higher accuracy than proposed ensemble learning model because using extra dataset and ensemble with more classification deep learner models lead to increasing the accuracy result. The recent ensemble model in [30,31] achieved accuracy higher than the proposed ensemble models because they ensemble more CNN models that is lead to complexity in the process and used an extra dataset. The recent ensemble model in [33] achieved accuracy higher than the proposed ensemble models because they implemented their experiments on part of dataset by selecting randomly images from each class. The recent ensemble model in [34] achieved accuracy higher than the proposed ensemble models because they ensemble five models to diagnose seven classes of skin lesions types by only using seven classes from ISIC 2019 like HAM dataset and the classification accuracy decrease when the number of diagnostic classes increases.

Figures (8,9) show the confusion matrices of individual base learners and proposed ensemble learning models. It is observed from the figure that most melanocytic nevus (NV) samples are classified as melanoma (MEL) and benign keratosis (BKL) samples are classified as melanoma (MEL) and melanocytic nevus (NV). This is because there are high similarities between NV, MEL, and BKL classes in terms of size, color, and shape.

Table 2. Final evaluation results for six models

	ACC	Balanced ACC	Recall	Precision	F1 score
Resnet18	0.795	0.646	0.646	0.737	0.697
Densnet121	0.812	0.675	0.676	0.767	0.716
Inception v4	0.826	0.682	0.682	0.762	0.722
Normal Average Ensemble	0.832	0.712	0.714	0.824	0.765
Learnable Ensemble Using SVM	0.849	0.734	0.734	0.805	0.769
Learnable Ensemble Using Random Forest	0.862	0.749	0.749	0.833	0.789

Table 3. Final accuracy results for each class

	MEL	NV	BCC	AK	BKL	DF	VASC	SCC
Resnet18	0.695	0.910	0.814	0.50	0.592	0.495	0.76.2	0.447
Densnet121	0.721	0.925	0.869	0.456	0.585	0.662	0.778	0.431
Inception v4	0.727	0.934	0.872	0.471	0.662	0.541	0.866	0.472
Normal Average Ensemble	0.727	0.931	0.894	0.539	0.669	0.569	0.801	0.49
Learnable Ensemble Using SVM	0.734	0.964	0.919	0.464	0.699	0.594	0.823	0.59
Learnable Ensemble Using Random Forest	0.740	0.972	0.894	0.519	0.742	0.53	0.828	0.551

Table 4. Performance comparison with other deep learning-based ensemble models.

Ref	Ensemble	No. of epochs	No. of Classes	Accuracy %	Dataset
[29]	efficientnet-b2+ efficientnet-b4+densenet121	90	eight	75%	ISIC 2019
[30]	SENet +PNASNet + InceptionV4, ResNet-50/101/152 + DenseNet-121/169/201 + MobileNetV2 + GoogleNet + and VGG-16/19	١٥٠	eight	89%	ISIC 2019
[31]	inceptionV3+ xinception+ inception_resnet_V2+ DenseNet-121/169/201 + resnext101+ resnet101+ resnet101V2+ resnet152+ resnet152V2+ nasnet	unknown	nine	91%	ISIC 2019 + SD
[32]	SENet154 SS+ ResNeXt-101 WSL-32x8d/32x16d+ EN B0-SS/RR+ EN B1- SS/RR+ EN B2- SS/RR+ EN B3- SS/RR+ EN B4- SS/RR+ EN B5- SS/RR+ EN B6- SS/RR+	100	eight nine	74.2% as balanced accuracy	ISIC 2019 + in-house
[33]	ResNet+InceptionV3+DenseNet+InceptionResNetV2+VGG 19	50	eight	98%	ISIC 2019
[34]	ResNeXt+ SeResNeXt+ DenseNet+ Xception+ ResNet	unknown	seven	88%	ISIC 2019 + HAM10000
[35]	EfficientNet-B5 + SE-ResNeXt-101(32x4d) + EfficientNet-B4 + Inception-ResNet-v2	32	nine	63.4%	ISIC 2019 + other five
Proposed Learnable Ensemble Model	ResNet18+ DenseNet121+ Inception V4	١٠0	eight	86.2%	ISIC 2019

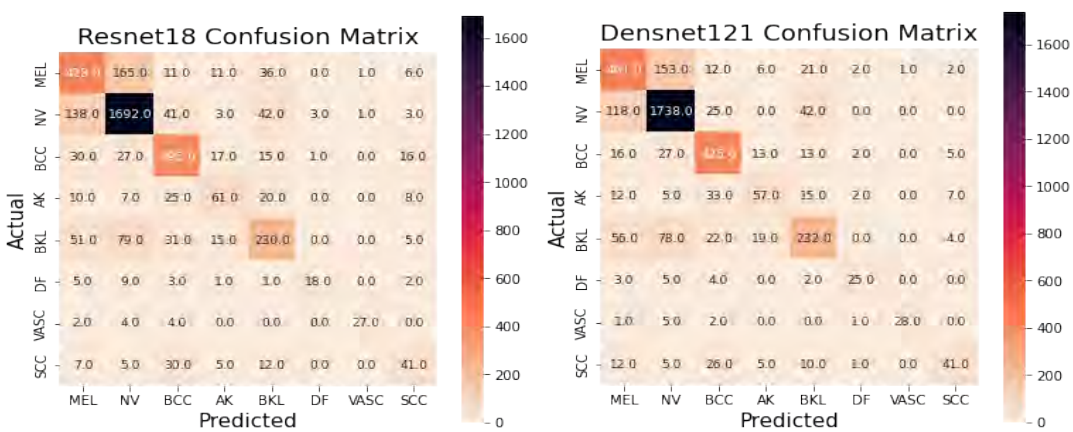

Fig.8. The confusion matrix of CNN models (ResNet18, DenseNet121)

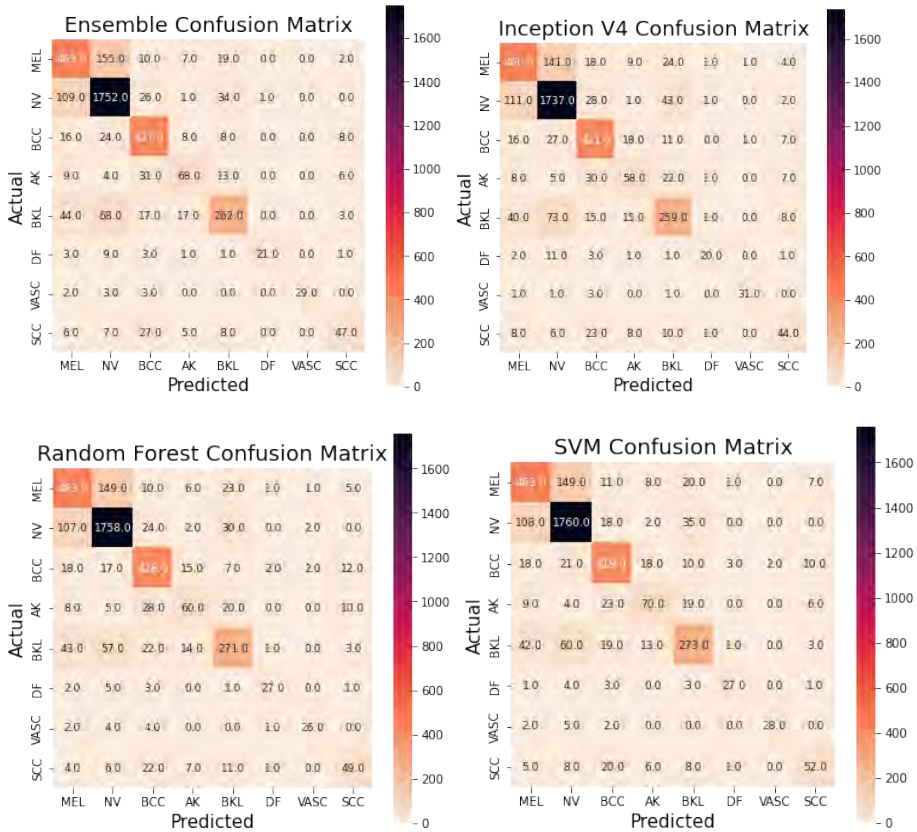

Fig.9. The confusion matrix of Inception V4 and ensemble learning models.

4.3 Ablation study

To study the effect of data augmentation and class imbalance handling on the results, we performed three different experiments. In first experiment, we built the proposed ensemble approach from three base learners, performed data augmentation and a weighted loss function was utilized to handle the class imbalance problem. In second experiment, we used the same base learners and weighted loss function and trained all models using the original data without applying any data augmentation. In the third experiment, we used the same learners again and performed data augmentation, but unweighted categorical cross entropy was utilized in this experiment. From the results of these experiments shown in Table V and Table VI, we can notice that these processes enhance the accuracy as well as the F1 score.

Table 5. The effective of data augmentation on results

	ACC With applying Data Augmentation	ACC Without applying Data Augmentation	F1 score With applying Data Augmentation	F1 score Without applying Data Augmentation
Resnet18	0.795	0.779	0.697	0.677
Densnet121	0.812	0.805	0.716	0.712
Inception v4	0.826	0.812	0.722	0.714
Normal Average Ensemble	0.832	0.825	0.765	0.748

Table 6. The effective of handling the class imbalance on results

	ACC Before handling class imbalance problem	ACC After handling class imbalance problem	F1 score Before handling class imbalance problem	F1 score After handling class imbalance problem
Resnet18	0.778	0.795	0.667	0.697
Densnet121	0.799	0.812	0.698	0.716
Inception v4	0.812	0.826	0.710	0.722
Normal Average Ensemble	0.819	0.832	0.729	0.765
Learnable Ensemble Using SVM	0.838	0.849	0.742	0.769
Learnable Ensemble Using Random Forest	0.848	0.862	0.745	0.789

5. Conclusion

We developed different ensemble methods to enhance the classification accuracy of classifying eight skin cancer types by transfer learning of deep CNN models such as resnet18, densnet121, and inception v4. We used ISIC 2019 dataset which has more challenges and augmented it after preprocessing it. We trained all models but Inception v4 achieved better accuracy than others. Ensemble methods enhance the accuracy of individual classifiers. So, we perform the average ensemble model by combining the prediction of all three models. The average ensemble model achieved accuracy was 83.2%, which more accurately than classifies the individual CNN deep learner models. Finally, we enhanced the performance of the average ensemble model by using other ensemble models of SVM and Random Forest. SVM and Random Forest achieved the best accuracy by concatenating the probability predictions of resnet18, densnet121, and inception v4 as an input vector of SVM and Random Forest classifiers. SVM and Random Forest achieved accuracy was 85% and 86.2% respectively.

References

[1] Ballerini, Lucia, Robert B. Fisher, Ben Aldridge, and Jonathan Rees. "A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions." In Color medical image analysis, pp. 63-86. Springer, Dordrecht, 2013.

[2] Arnold, M., M. Kvaskoff, A. Thuret, P. Guenel, F. Bray, and I. Soerjomataram. "Cutaneous melanoma in France in 2015 attributable to solar ultraviolet radiation and the use of sunbeds." Journal of the European Academy of Dermatology and Venereology 32, no. 10: 1681-1686, 2018.

[3] Arnold, M. "Vries E de." Whiteman DC, Jemal A, Bray F, Parkin DM, Soerjomataram I : 1305-1314, 2018.

[4] Parkin, D. M., D. Mesher, and P. Sasieni. "13. Cancers attributable to solar (ultraviolet) radiation exposure in the UK in 2010." British journal of cancer 105, no. 2: S66-S69, 2011.

[5] Pluta, Ryszard M., Alison E. Burke, and Robert M. Golub. "Melanoma." Jama, 305(22), 2368-2368, 2011.

[6] Elwood, J. M., J. A. H. Lee, S. D. Walter, T. Mo, and A. E. S. Green. "Relationship of melanoma and other skin cancer mortality to latitude and ultraviolet radiation in the United States and Canada." International journal of epidemiology 3, no. 4: 325-332, 1974.

[7] Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions." Scientific data 5, no. 1: 1-9, 2018.

[8] Combalia, Marc, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera et al. "Bcn20000: Dermoscopic lesions in the wild." arXiv:1908.02288, 2019.

[9] Kawahara, Jeremy, Aicha BenTaieb, and Ghassan Hamarneh. "Deep features to classify skin lesions." 2016 IEEE 13th international symposium on biomedical imaging (ISBI). IEEE, 2016.

[10] Harangi, Balazs, Agnes Baran, and Andras Hajdu. "Classification of skin lesions using an ensemble of deep neural networks." In 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp. 2575-2578. IEEE, 2018.

[11] Gessert, Nils, Thilo Sentker, Frederic Madesta, Rüdiger Schmitz, Helge Kniep, Ivo Baltruschat, René Werner, and Alexander Schlaefer. "Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting." arXiv preprint arXiv:1808.01694, 2018.

[12] Mohamed, Ensaf Hussein, and Wessam H. El-Behaidy. "Enhanced skin lesions classification using deep convolutional networks." In 2019 ninth international conference on intelligent computing and information systems (ICICIS), pp. 180-188. IEEE, 2019.

[13] Hasan, M., Barman, S. D., Islam, S., & Reza, A. W. “Skin cancer detection using convolutional neural network”. In Proceedings of the 2019 5th international conference on computing and artificial intelligence pp. 254–258, 2019.

[14] Al-Masni, Mohammed A., Dong-Hyun Kim, and Tae-Seong Kim. "Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification." Computer methods and programs in biomedicine 190: 105351, 2020.

[15] Le, Duyen NT, Hieu X. Le, Lua T. Ngo, and Hoan T. Ngo. "Transfer learning with class-weighted and focal loss function for automatic skin cancer classification." arXiv preprint arXiv:2009.05977, 2020.

[16] Steppan, Josef, and Sten Hanke. "Analysis of skin lesion images with deep learning." arXiv preprint arXiv:2101.03814, 2021.

[17] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[18] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708, 2017.

[19] Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning." In Thirty-first AAAI conference on artificial intelligence, 2017.

[20] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20, no. 3 273-297, 1995.

[21] Belgiu, Mariana, and Lucian Drăguţ. "Random forest in remote sensing: A review of applications and future directions." ISPRS journal of photogrammetry and remote sensing 114: 24-31, 2016.

[22] Finlayson, Graham D., and Elisabetta Trezzi. "Shades of gray and colour constancy." In Color and Imaging Conference, vol. 2004, no. 1, pp. 37-41. Society for Imaging Science and Technology, 2004.

[23] Chen, Xue-Wen, and Xiaotong Lin. "Big data deep learning: challenges and perspectives." IEEE Access 2: 514-525, 2014.

[24] Ji, Qingge, Jie Huang, Wenjie He, and Yankui Sun. "Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images." Algorithms 12, no. 3: 51, 2019.

[25] Pham, Tri-Cong, Chi-Mai Luong, Muriel Visani, and Van-Dung Hoang. "Deep CNN and data augmentation for skin lesion classification." In Asian Conference on Intelligent Information and Database Systems, pp. 573-582. Springer, Cham, 2018.

[26]Aurelia, Jane Eva, et al. "Hepatitis classification using support vector machines and random forest." IAES International Journal of Artificial Intelligence 10.2: 446, 2021.

[27] Verikas, Antanas, Evaldas Vaiciukynas, Adas Gelzinis, James Parker, and M. Charlotte Olsson. "Electromyographic patterns during golf swing: Activation sequence profiling and prediction of shot effectiveness." Sensors 16, no. 4: 592, 2016.

[28] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.

[29] Zhou, Steven, Yixin Zhuang, and Rusong Meng. "Multi-category skin lesion diagnosis using dermoscopy images and deep CNN ensembles." DysionAI, Tech. Rep (2019).

[30] Pacheco, Andre GC, Abder-Rahman Ali, and Thomas Trappenberg. "Skin cancer detection based on deep learning and entropy to detect outlier samples." arXiv preprint arXiv:1909.04525 (2019).

[31] Juan Wang. " ISIC 2019 - Skin Lesion Analysis Towards Melanoma Detection." Delta Micro Technology Inc. Laguna Hills, CA 92653 (2019).
[32] Gessert, Nils, et al. "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data." MethodsX 7: 100864 (2020).
[33] Kausar, Nabeela, et al. "Multiclass Skin Cancer Classification Using Ensemble of Fine-Tuned Deep Learning Models." Applied Sciences 11.22: 10593 (2021).
[34] Rahman, Zillur, et al. "An approach for multiclass skin lesion classification based on ensemble learning." Informatics in Medicine Unlocked 25: 100659 (2021).
[35] Steppan, Josef, and Sten Hanke. "Analysis of skin lesion images with deep learning." arXiv preprint arXiv:2101.03814 (2021).