Ensemble Deep Learning Approach For Multi-Class Skin Cancer Classification

Original Scientific Paper

Ali Abdulameer*
Electrical Engineering Technical College, Middle Technical University
Baghdad, Iraq
bdc0063@mtu.edu.iq

Raaed Hassan
Electrical Engineering Technical College, Middle Technical University
Baghdad, Iraq
drraaed_alanbaki@mtu.edu.iq

Abbas Humadi
Electrical Engineering Technical College, Middle Technical University
Baghdad, Iraq
drabbas@mtu.edu.iq

*Corresponding author

Abstract *– Skin cancer is one of the most prevalent types of cancer, often caused by prolonged exposure to ultraviolet (UV) radiation, such as sunlight. This cancer is mainly categorized into benign and malignant lesions, where the latter could cause severe complications and even death. Traditional diagnostic methods, such as visual inspection and dermoscopy, often lack accuracy, while biopsy, though highly accurate, is invasive, time-consuming, and costly. This study aims to develop an automated deep learning model that leverages an ensemble of "Convolutional Neural Networks" (CNNs) to perform four-class classification of common skin lesions: Basal Cell Carcinoma (BCC), Benign Keratosis Lesion (BKL), Melanocytic Nevus (NV), and Melanoma (MEL). Seven widely used CNNs in medical imaging, GoogLeNet, InceptionV3, Xception, ResNet18, ResNet50, ResNet101, and DenseNet201, were evaluated for their performance in this classification task. The ISIC2018 and ISIC2019 datasets were employed, and data augmentation techniques were applied to address dataset imbalances. The analysis identified InceptionV3, Xception, and DenseNet201 as the top-performing networks. Therefore, they are utilized for the ensemble model. These networks were used as feature extractors, and their output features were combined and classified using a "Support Vector Machine" (SVM) algorithm. This approach demonstrates the potential of combining CNNs and SVM in an ensemble framework to enhance the accuracy and reliability of automated skin cancer classification. The proposed model achieved an accuracy of 94.46%, outperforming individual CNNs (93.27%) and existing ensemble methods such as Max Voting (94.12%) and hybrid models like DenseNet201 with Random Forest (91.28%).*

Keywords: *Skin Cancer, Deep Learning, CNN, Machine Learning, Artificial Intelligence*

Received: March 6, 2025; Received in revised form: June 12, 2025; Accepted: June 12, 2025

1. INTRODUCTION

Skin cancer is one of the most common cancers and was the fourth most diagnosed cancer in 2020 [1]. In the United States, 9,500 persons are diagnosed with skin cancer every day [2]. It is mainly caused by ultraviolet (UV) radiation from sources such as sunlight or tanning devices, leading to DNA mutations and consequently to abnormal growth of skin cells [3]. This type of carcinoma commonly occurs in fair-skinned individuals due to their lack of melanin pigmentation, which protects the skin from UV light [3]. There are many types of this cancer according to malignancy and origin. Melanoma is considered the most dangerous lesion, and early detection of this type is critical because it is highly malignant and can cause severe complications or even death. Statistics show that one person dies from every 7.8 melanoma cases [4]. Diagnosis typically begins with visual inspection of the lesion, but its accuracy is below 60%, even among experienced dermatologists [5]. Dermoscopy, a non-invasive procedure using a light source and magnifying lenses, was introduced to improve diagnostic accuracy by enhancing visualization

of tumor features achieving diagnostic accuracy between 75% and 84% [6]. Accurate diagnosis often requires a biopsy, an invasive procedure in which a skin sample is extracted and examined microscopically [7]. A biopsy is an expensive and time-consuming procedure, making early diagnosis of this serious disease difficult and sometimes inaccessible in rural or underserved areas [7].

Computer-Aided Diagnosis (CAD) systems were developed to assist dermatologists in accurately diagnosing skin lesions [8]. Early systems employed machine learning (ML) algorithms such as logistic regression, KNN, decision trees, random forest, SVM, and ANN, relying on manually inserted ABCDE features. This manual feature extraction was tedious and limited, especially since certain melanoma types, such as nodular melanoma, do not conform to the ABCDE criteria [9]. Later, convolutional neural networks (CNNs) were introduced as automated feature extractors, using convolutional filters trained via backpropagation to identify patterns[10]. The increased availability of GPUs in personal computers enabled researchers to efficiently train CNNs on dermoscopic images [11].

Bazgir *et al*. [12] conducted a binary classification to differentiate between benign and malignant skin lesions using a modified version of the InceptionV3 network. The modification involved adding an extra dense layers at the end of the original architecture. They achieved a maximum classification accuracy of 85.94%, which is relatively low for a binary classification task. We believe this is because they trained their modified network from scratch, without utilizing the pre-trained weights of InceptionV3.

Dahdouh *et al*. [13] conducted a seven-class skin cancer classification using the HAM10000 dataset. They integrated a convolutional neural network (CNN) with reinforcement learning (RL), where a Q-network replaced the CNN's dense layer. Preprocessing and segmentation steps were also applied. The model achieved a classification accuracy of approximately 80%. However, the proposed CNN architecture was not reported, and the specific contribution of the Q-network remains unclear, as the results section did not present the CNN's performance without RL integration.

Dogan and Ozdemir [14] developed a hybrid model to distinguish benign lesions from melanoma by evaluating multiple pre-trained CNNs, ResNet152V2, VGG16, Xception, InceptionV3, MobileNetV2, DenseNet201, InceptionResNetV2, and EfficientNetB2, in combination with machine learning algorithms such as K-Nearest Neighbors (KNN) and Random Forest (RF). The best performance was achieved using DenseNet201 as a feature extractor combined with Random Forest as the classifier, yielding an accuracy of approximately 91.28%. This result highlights the effectiveness of using a hybrid approach that integrates CNN-based feature extraction with traditional ML classifiers.

Natha *et al*. [15] conducted a seven-class skin cancer classification using three machine learning algorithms: Random Forest (RF), Multi-layer Perceptron Neural Network (MLPN), and Support Vector Machine (SVM), which were used as classifiers and combined using the Max Voting method. Color and texture features were extracted using basic image processing techniques, and a genetic algorithm was employed to optimize the feature vector by selecting the most relevant features extracted from the ISIC2018 dataset. The study achieved a classification accuracy of approximately 94.70%, demonstrating the effectiveness of an ensemble approach that combines multiple algorithms.

Researchers achieved relatively high diagnostic accuracy compared to traditional methods, but they remain far from achieving 100% accuracy. This challenge arises due to the high visual similarity between different skin lesions. For instance, melanocytic nevi look similar to melanoma, but the two lesion types differ in their malignancy. Another limitation is the lack of sufficient and balanced dermoscopic image datasets. This research aims to address these limitations by introducing the following contributions:

- Eliminating rare skin lesion types from the dataset to increase the diagnostic performance on the common types, BCC, BKL, MEL, and NV.
- Introducing a new data balancing approach that combines external dermoscopic images from multiple datasets with targeted augmentation techniques to address the class imbalance problem inherent in dermatological datasets.
- Utilizing a distinctive feature-level fusion method that directly concatenates high-dimensional deep features extracted from three diverse CNN architectures without dimensionality reduction, preserving the complete feature information for classification.
- Utilization of a computationally efficient SVM classifier that leverages the concatenated feature vector rather than traditional voting-based ensemble methods or prediction score fusion approaches.

The rest of the paper is organized as follows: Section 2 describes the methodology, including the dataset, preprocessing techniques, CNN model, support vector machine model, and ensemble model. Section 3 presents the results, and Section 4 provides the conclusion, discussion, and directions for future work.

2. METHODOLOGY

2.1. DATASET

The ISIC2018 dataset [16] was used to train the CNN models. It contains seven lesion classes: vascular lesion (VASC), dermatofibroma (DF), "benign keratosis lesion" (BKL), "actinic Keratoses and intraepithelial carcinoma" (AKIEK), "Melanocytic Nevus" (NV), melanoma (MEL), and "basal cell carcinoma" (BCC). The ISIC2018 dataset is divided into training, validation, and test sets containing 10,015; 193; and 1,512 dermoscopic images, respectively. This dataset suffers from class imbalance,

with the NV class comprising 67% of the total images, while the DF class accounts for only 1%.

This imbalance is addressed in this work through two main contributions. The first approach involves two steps: first, excluding rare lesion types from the classification task due to their uncommon occurrence. This results in focusing on four common classes: BCC, BKL, NV, and MEL.

Secondly, additional images are imported from the ISIC2019 dataset. Since the ISIC2019 dataset includes all images from ISIC2018, the import process ensures that no duplicate images are included in the combined dataset. The distribution of the resulting dataset, referred to as "ISIC2018^{+}" for simplicity, is illustrated in Fig. 1.

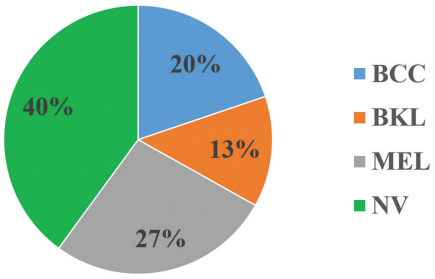


Fig. 1. The percentage of classes in the ISIC2018^{+} dataset

The second contribution involves data augmentation to increase the number of images for underrepresented classes. The augmentations include image scaling in the range of 1 to 1.2, image flipping along the x-axis and y-axis, and image rotation from -90° to 90°. All these augmentations are applied with random parameters to balance the dataset equally across the four classes. As a result, the augmented dataset achieves an equal distribution, with each class representing 25% of the data.

As a result, the ISIC2018^{+} dataset contains 16,787 images, while the ISIC2018^{+}(augmented) comprises 26,819 images. Fig. 2 shows an example of augmented images generated by the augmentation process. The validation and test sets remained unchanged after excluding the underrepresented categories, AKIEC, DF, and VASC, from the original dataset. This study adheres to using the official test set provided by the ISIC archive to ensure the validity and comparability of evaluation results. Randomly selecting a test subset from the resulting dataset is discouraged, as it does not meet strict evaluation standards.

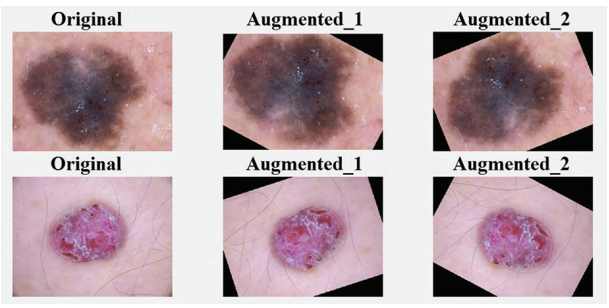


Fig. 2. Example of image augmentation process

2.2. PEPROCESSING TECHNIQUES

The preprocessing procedure in this work includes image resizing and normalization. Image resizing is essential to adapt dermoscopic images to the required input dimensions of the CNN model. For example, GoogLeNet requires input images of size 224×224 pixels. Normalization converts pixel intensity values from the [0,255] range to the [0,1] range. This process helps prevent issues such as exploding gradients and ensures that all features contribute equally to gradient updates. By scaling pixel intensities, brighter pixels do not overpower dimmer ones, which stabilizes and accelerates the training process. Normalization is performed using Eq. (1) [10].

$$x_{normalized} = \frac{x}{255} \tag{1}$$

Here, x represents the pixel intensity of the input image.

2.3. CNN MODEL

This research utilizes CNN as an automated feature extractor. CNN captures image patterns at various layers, extracting features ranging from low-level details to high-level abstractions. Filters (kernels) with initially random weights are updated during training to minimize the error between true labels and predicted probabilities [10]. For classification tasks, the "Cross-Entropy Loss Function" is preferred over "Mean Squared Error" (MSE) because it penalizes confident incorrect predictions more heavily, enabling faster convergence during training. "Cross-Entropy Loss Function" is computed in Eq. (2) [10].

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}(\log \hat{y}_{i,c}) \tag{2}$$

Here, N is the total number of samples. C is the number of classes. $y_{i,c}$ is the true label for the i^{th} sample and c^{th} class. $\hat{y}_{i,c}$ is the predicted probabilty for (i, c) indices.

The convolution process is mathematically expressed in Eq. (3) [10].

$$y_{[i,j,c]} = \sum_{m=1}^{K}\sum_{n=1}^{K}\sum_{d=1}^{D} x_{[i+m-1,j+n-1,d]} \cdot w_{[m,n,d,c]} + b_{[c]} \tag{3}$$

Here, $x_{[i,j,d]}$ represents the pixel value at indices (i, j) in the d^{th} channel of the input feature map. Similarly, $y_{[i,j,c]}$ represents the pixel value at indices (i, j) in the c^{th} channel of the output feature map.

The weight tensor, $w_{[m,n,d,c]}$, corresponds to the filter indexed by (m, n), which defines the kernel dimensions, and (d, c), which specifies the input and output channels of the feature maps. K is the kernel size, D is number of input channels, $b_{[c]}$ is the bias term which is added to each output channel.

The spatial size of the output feature map of every convolution process is governed by Eq. (4) [10].

$$Z = \frac{I-K+2P}{S} + 1 \tag{4}$$

Here, Z is the spatial size of the output feature map. I is the spatial size of the input matrix. K is the kernel

size. P is the padding applied around the input matrix. S is the stride value.

In this research, seven commonly used CNNs are employed for transfer learning to perform skin cancer classification: GoogLeNet, InceptionV3, Xception, ResNet18, ResNet50, ResNet101, and DenseNet201. Their selection is justified by their architectural diversity, varying depths, and proven success in medical imaging applications [17, 18]. These models represent distinct design philosophies in deep learning, each offering unique strengths. The CNN models are customized to perform a four-class classification task. This customization involves replacing the dense layer with one having four outputs, and the classification layer with one that outputs four class probabilities. The properties of those seven models are detailed in Table 1.

Table 1. Properties of CNN models

Network name	Depth	Input size	Parameter memory	No. of parameters (millions)
GoogLeNet	22	224×224	27 MB	7.0
InceptionV3	48	299×299	91 MB	23.9
Xception	71	299×299	88 MB	22.9
DenseNet201	201	224×224	77 MB	20.0
ResNet18	18	224×224	45 MB	11.7
ResNet50	50	224×224	98 MB	25.6
ResNet101	101	224×224	171 MB	44.6

2.4. SUPPORT VECTOR MACHINE MODEL

Support Vector Machine (SVM) is a widely used machine learning algorithm primarily designed for binary classification tasks. It can also be extended to multi-class classification by employing a one-vs-all approach. The core concept involves using a hyperplane to separate data points into two classes within a high-dimensional space. The hyperplane is represented by Eq. (5) [19].

$$W^T \cdot X_i + b = 0 \tag{5}$$

Here, W is the weight vector normal to the hyperplain. X_i is the feature vector for the i^{th} point. b is the bias term.

The weight vector is optimized to maximize the margin, defined as the perpendicular distance between the hyperplane and the nearest data points. This margin is symmetric on both sides of the hyperplane and is given by $(2/\|W\|)$, as illustrated in a 2-dimensional perspective in Fig. 3.

To maximize the margin, the denominator term $(\|W\|)$ of the margin should be minimized. For mathematical convenience, the term $((1/2)\|W\|^2)$ is used instead for two key reasons. First, the squared term simplifies optimization by enabling the use of "Convex Optimization Techniques," Second, the constant factor (1/2) makes the calculation of derivatives with respect to (W) more efficient. To train the model, each data point should be considered with its true label (y_i), where $y_i \in \{1,-1\}$ represents the class labels for binary classification. For a correctly classified point, the following constraint must be satisfied as expressed in Eq. (6) [19].

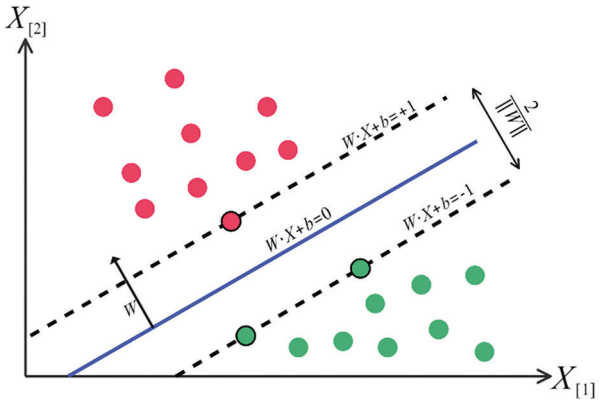


Fig. 3. Two-dimensional perspective of the hyperplane of SVM

$$y_i(W^T \cdot X_i + b) \geq 1 \tag{6}$$

The "Lagrange Multipliers" method is utilized to optimize the model and update weights, where the loss function L is introduced by Eq. (7) [19].

$$\mathcal{L}_{(W,b,\alpha)} = \frac{1}{2}\|W\|^2 - \sum_{i=1}^{n} \alpha_i (y_i(W^T \cdot X_i + b) - 1) \tag{7}$$

Where the first term $((1/2)\|W\|^2)$ measures the value responsible for maximizing the margin, while the second term adds a penalty for violating the constraint in Eq. (6). This penalty is scaled by the Lagrange multipliers (α_i). This method focuses on points near the hyperplane, known as support vectors while excluding other points. This property makes it computationally efficient during the optimization process. As a result, a new data point (X) is classified based on the sign of the hyperplane equation as expressed in Eq. (8) [19].

$$y(X) = sign(W^T \cdot X + b) \tag{8}$$

2.5. ENSEMBLE MODEL

In this work, an ensembled model has been proposed which includes three main steps. First, seven CNN architectures are trained on the ISIC2018$^+$ dataset to identify the top-performing models. Second, the best-performing models from the previous step are trained on the ISIC2018$^+$(augmented) dataset individually. Finally, The output features from the top-performing models in the first two steps are concatenated into a single matrix with full dimensionality, which serves as the input to an SVM model for generating the final predictions. In essence, the proposed ensemble model leverages CNNs as feature extractors and utilizes an SVM for classification. The CNN features are extracted from the inputs to the dense layers of each network. The complete methodology for this work is illustrated in Fig. 4, where the chosen models are justified in the results section.

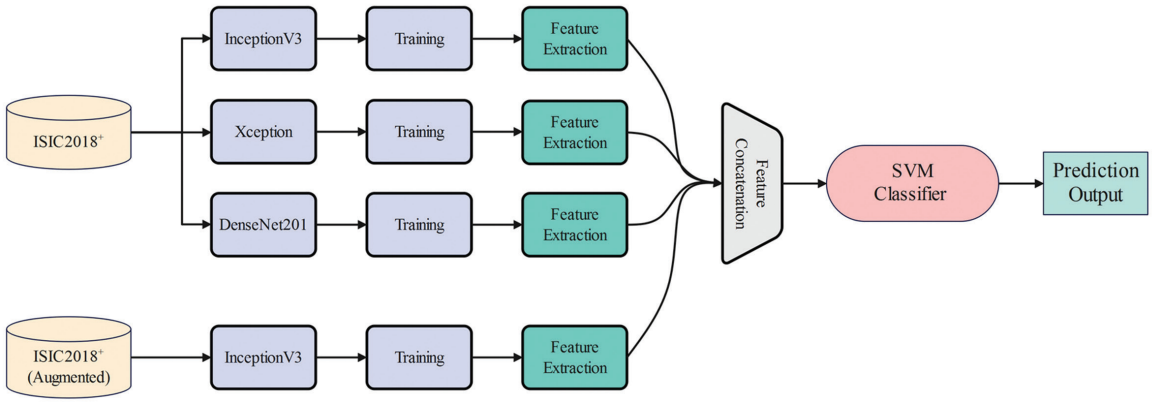


Fig. 4. The proposed ensemble model architecture

2.6. TRAINING PROCESS

The training is conducted in a MATLAB environment using the Deep Learning Toolbox (version 14.3) to customize and train the CNN architectures. Additionally, the Statistics and Machine Learning Toolbox (version 12.2) is employed to implement the SVM classifier. All experiments are performed on a "Dell Precision 7740" laptop equipped with an "NVIDIA Quadro RTX 5000" featuring 16 GB of GDDR6 VRAM. The CNN training process utilizes a scheduled learning rate, which starts at 0.01 and decays by a factor of 1/10 every 10 epochs, over a total of 30 epochs. Training is carried out using the "Stochastic Gradient Descent with momentum" (SGDM) optimizer, configured with a momentum value of 0.9. All hyperparameters are detailed in Table 2.

Table 2. Training hyperparameters

Hyperparameter	Value
Learning rate	(0.01), (0.001), (0.0001)
epochs	30
Optimizer	SGDM
Batch size	64
Momentum	0.9
L2 Regularization	0.0001

The same augmentation techniques used in the ISIC2018+(augmented) dataset are employed during the training of the CNN models to mitigate overfitting and enhance the model's generalization. However, the augmentation process is explained in details in Section 2.1.

2.7. EVALUATION METRICS

In this task, evaluation metrics are essential to rate the performance of the deep learning model. The test set is imported from the ISIC archive, which has 1512 dermoscopic images for seven classes. After the elimination of the minor classes, a test set is achieved with four classes, BCC, NV, BKL, and MEL, forming 1390 images in total. In this section, accuracy, recall, precision, F1-score, ROC ("Receiver Operating Characteristic") curve, AUC ("Area Under Curve"), and the confusion matrix are explained.

Accuracy is the main evaluation metric used in a deep learning context. It is simply the ratio of the correct predictions to the total number of predictions. It's expressed further in Eq. (9) as muti-class accuracy.

$$Multi\ Class\ Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

Here, TP is true positive predictions. TN is true negative predictions. FN is false negative predictions. FP is false positive predictions.

Since multi-class accuracy could be misleading for imbalanced datasets, another accuracy metric is considered in this research, the mean accuracy, which is expressed in Eq. (10).

$$Accuracy_{class(k)} = \frac{TP_{class(k)}}{Total\ number\ of\ prediction\ per\ class(k)} \quad (10)$$

Precision is another metric that expresses the ratio of the positive predictions over all positive predictions. as expressed in Eq. (11).

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

Recall gives the indication of how the actual positive predictions are correctly identified as expressed in Eq. (12).

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

F1-Score balances precision and recall by taking the harmonic mean as expressed in Eq. (13).

$$F1\ Score = 2 \times \frac{Precision \cdot Recall}{Precision+Recall} \quad (13)$$

A confusion matrix summarizes true positives, true negatives, false positives, and false negatives to evaluate classification performance. The ROC curve shows the trade-off between the true positive and false negative rates across thresholds, while the AUC represents the area under the ROC curve.

3. RESULTS

3.1. PERFORMANCE EVALUATION

This section outlines the evaluation process of this work, which is divided into three steps. The first step

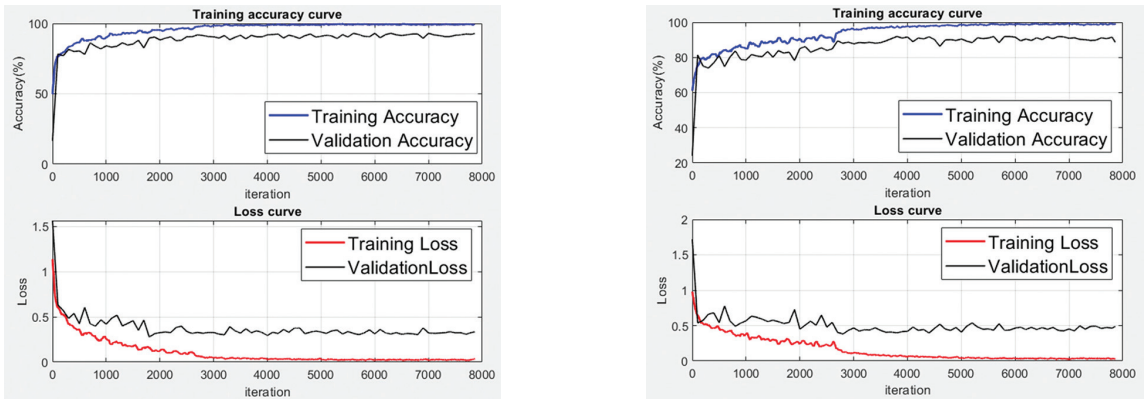
evaluates the performance of seven CNN models with softmax classifier trained on the ISIC2018⁺ to identify the best-performing networks. In the second step, only the top-performing models from the previous step are trained individually on the ISIC2018⁺(augmented) dataset. This approach avoids the need to train all seven models on the larger dataset, saving time and effort. The third step involves selecting the best models from the previous steps to be used as feature extractors for the ensemble model. Table 3 summarizes the evaluation results. InceptionV3, Xception, and DenseNet201 achieved the highest accuracy.

Their superior performance is attributed to architectural strengths: InceptionV3 and Xception combine depth and width to capture multi-scale lesion features, while DenseNet201's dense connections promote feature reuse and reduce redundancy.

Table 3. The evaluation metrics for the experienced CNN models.

Dataset	CNN model	Precision	Recall	F1-Score	AUC	Mean Accuracy	Multi-Class Accuracy
ISIC2018⁺	ResNet18	77.47%	75.71%	76.50%	95.17%	92.09%	84.17%
	ResNet50	75.25%	74.42%	74.73%	93.94%	91.55%	83.10%
	ResNet101	75.54%	75.47%	75.33%	94.48%	91.40%	82.80%
	GoogLeNet	77.42%	74.52%	75.79%	94.48%	92.13%	84.25%
	InceptionV3	77.97%	81.01%	79.20%	95.65%	92.56%	85.10%
	Xception	80.00%	83.31%	80.95%	96.27%	92.82%	85.54%
	DenseNet201	80.45%	78.75%	79.19%	95.59%	92.95%	85.90%
ISIC2018⁺ (augmented)	InceptionV3	80.78%	80.80%	80.50%	95.75%	93.27%	86.55%
	Xception	77.94%	80.95%	78.98%	96.11%	92.30%	84.60%
	DenseNet201	78.74%	77.94%	78.16%	95.75%	92.77%	85.54%

The ISIC2018⁺(augmented) dataset proved to be beneficial in improving the performance of InceptionV3 but did not yield similar enhancements for the other two networks. Subsequently, the three best-performing models from the first dataset, along with the top-performing model from the second dataset, were selected as feature extractors for the ensemble model. Due to space constraints, results charts are provided only for the two best-performing models from the first dataset. Training accuracy and loss curves are shown in Fig. 5.



(a) (b)

Fig. 5. The accuracy and loss curves for (**a**) Xception. (**b**) DenseNet201.

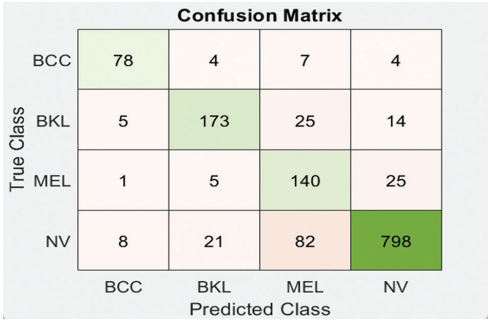


Fig. 6. The confusion matrix for Xception

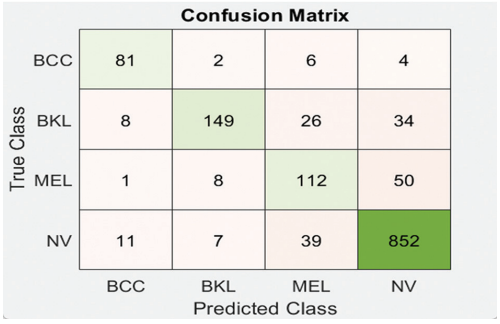
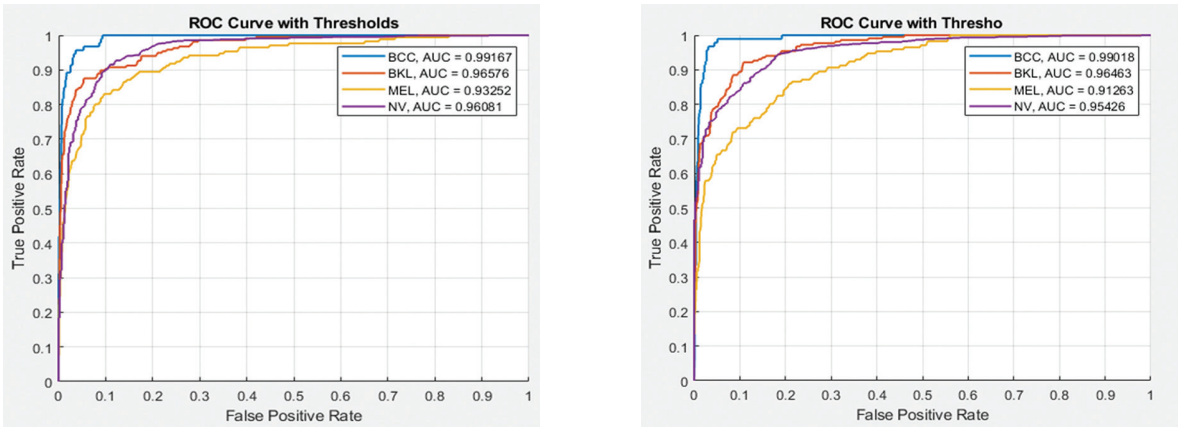


Fig. 7. The confusion matrices for DenseNet201

Fig. 8 shows the ROC curves for the designated models.



(a) (b)

Fig. 8. The ROC curves for (**a**) Xception. (**b**) DenseNet201

SVM classifier is applied to the best three CNNs individually to study its impact to boost the Performance. And lastly, the esemble model is restructed from the best models as previously illustrated in Fig. 4 to enhance the model further. Table 4 shows the evaluation metrics for the best-performing CNNs combined with SVM individually, along with the ensemble model that resulted from the combination of the four selected models. The ensemble model demonstrated improved performance by combining the strengths of the selected individual models into a single framework. Fig. 9 illustrates the confusion matrix of the ensemble model.

Table 4. The evaluation metrics for the best-performing CNNs combined with SVM classifier in comparison with the ensemble model

Model / [Dataset]	Precision	Recall	F1-Score	Mean Accuracy	Muti-Class Accuracy
InceptionV3 / [ISIC2018⁺ (augmented)]	81.72%	79.56%	80.33%	91.45%	86.50%
Xception / [ISIC2018⁺]	82.77%	82.12%	82.31%	93.89%	87.77%
DenseNet201 / [ISIC2018⁺]	79.79%	78.72%	78.91%	88.38%	85.76%
Ensemble	84.25%	83.01%	83.45%	94.46%	88.90%

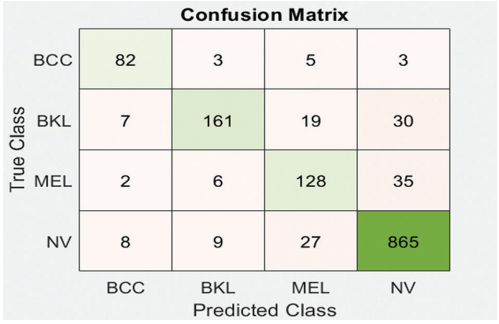


Fig. 9. The confusion matrix of the ensemble model

3.2. COMPARISON WITH OTHER WORKS

The methodology employed in this research contributed to improving the classification of common skin cancer types compared to other researchers' models. Table 5 presents a comparison of the results with previous works. However, this comparison is not strictly one-to-one, as some of the previous studies did not utilize the test set provided by the ISIC archive. Additionally, while some researchers performed multi-class classification on two, seven or eight classes, while this research focuses on only four classes.

Table 5. Results comparison with other researchers

Reference	Dataset	Model	Accuracy
Alwakid *et al.* [20]	HAM10000	ResNet50	86%
Jain *et al.* [21]	HAM10000	Multiple CNNs, with Xception as the top-performing model	90.48%
Alam *et al.* [22]	HAM10000	S2C-DeLeNet	91.03%
Dogan and Ozdemir [14]	ISIC archive	Hybrid model of DenseNet201 with Random Forest.	91.28%
Natha *et al.* [15]	ISIC2018	Max Voting ensemble method includes Random Forest, (MLPN), and SVM	94.12%
Proposed method	ISIC2018 and ISIC2019	Proposed ensemble model	94.46%

4. CONCLUSION

The ensemble model demonstrated performance enhancements over the best individual CNNs by leveraging the strengths of multiple architectures. Applying image augmentation techniques to balance the dataset proved beneficial for the InceptionV3 model, but no significant improvement was observed in the other two CNNs, Xception and DenseNet201.

Despite the high performance of this framework in the skin cancer classification task, its computational complexity is a notable limitation. The proposed model was trained exclusively on dermoscopic images captured under controlled conditions, with high clarity and specific lighting provided by dermoscopic equipment. This limitation may affect its performance when applied to real-world data captured under varying conditions. Furthermore, the dataset lacks diversity in terms of skin tones and age groups, potentially introducing biases in predictions and reducing the generalizability of the model. Additionally, the exclusion of the AKIEC, DF, and VASC classes further reduces the model's applicability across the full spectrum of skin lesion types. Nevertheless, the proposed ensemble model holds promise for integration into clinical decision support systems. It can be deployed in modest computing environments within dermatology clinics to assist practitioners or be incorporated into teledermatology platforms, extending diagnostic support to patients in remote or underserved areas. In the future, expanding the dataset with a greater number of image samples categorized by ethnicity and skin tone could enable the development of a two-level model. The first level would classify images by ethnicity, and the second would perform skin lesion classification within each group. This approach has the potential to mitigate skin color biases and improve classification accuracy.

5. REFERENCES:

[1] H. Sung et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", CA: A Cancer Journal for Clinicians, Vol. 71, No. 3, 2021, pp. 209-249.

[2] R. L. Siegel, K. D. Miller, A. Jemal, "Cancer statistics, 2019", CA: A Cancer Journal for Clinicians, Vol. 69, No. 1, 2019, pp. 7-34.

[3] J. L. Bolognia, J. L. Jorizzo, J. V. Schaffer, "Dermatology", Elsevier Health Sciences, 2012.

[4] A. C. Geller et al. "Melanoma epidemic: an analysis of six decades of data from the Connecticut Tumor Registry", Journal of Clinical Oncology, Vol. 31, No. 33, 2013, pp. 4172-4178.

[5] H. Haenssle et al. "Reader study level-I and level-II Groups. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists", Ann Oncol, Vol. 29, No. 8, 2018, pp. 1836-1842.

[6] M. Vestergaard, P. Macaskill, P. Holt, S. Menzies, "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting", British Journal of Dermatology, Vol. 159, No. 3, 2008, pp. 669-676.

[7] T. J. Hieken, R. Hernández-Irizarry, J. M. Boll, J. E. J. Coleman, "Accuracy of diagnostic biopsy for cutaneous melanoma: implications for surgical oncologists", International Journal of Surgical Oncology, Vol. 2013, No. 1, 2013, p. 196493.

[8] L. F. di Ruffano et al. "Computer-assisted diagnosis techniques (dermoscopy and spectroscopy-based) for diagnosing skin cancer in adults", Cochrane Database of Systematic Reviews, Vol. 2018, No. 12, 1996.

[9] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning", Nature, Vol. 521, No. 7553, 2015, pp. 436-444.

[10] I. Goodfellow, "Deep learning", MIT Press, 2016.

[11] R. Brito et al. "GPU-enabled back-propagation artificial neural network for digit recognition in parallel", The Journal of Supercomputing, Vol. 72, 2016, pp. 3868-3886.

[12] E. Bazgir, E. Haque, M. Maniruzzaman, R. Hoque, "Skin cancer classification using Inception Network", World Journal of Advanced Research and Reviews, Vol. 21, No. 02, 2024, pp. 839-849.

[13] Y. Dahdouh, A. A. Boudhir, M. B. Ahmed, "A new approach using deep learning and reinforcement learning in healthcare: skin cancer classification", International Journal of Electrical and Computer Engineering Systems, Vol. 14, No. 5, 2023, pp. 557-564.

[14] Y. Doğan, C. Özdemir, "Enhancing Skin Cancer Diagnosis through the Integration of Deep Learning and Machine Learning Approaches", Bilişim Teknolojileri Dergisi, Vol. 17, No. 4, 2024, pp. 339-347.

[15] P. Natha, S. P. Tera, R. Chinthaginjala, S. O. Rab, C. V. Narasimhulu, T. H. Kim, "Boosting skin cancer diagnosis accuracy with ensemble approach", Scientific Reports, Vol. 15, No. 1, 2025, p. 1290.

[16] International Skin Imaging Collaboration (ISIC), https://www.isic-archive.com (accessed: 2024)

[17] M. Naqvi, S. Q. Gilani, T. Syed, O. Marques, H.-C. Kim, "Skin cancer detection using deep learning—a review", Diagnostics, Vol. 13, No. 11, 2023, p. 1911.

[18] Z. G. Hadi, A. R. Ajel, A. Q. Al-Dujaili, "Comparison Between Convolutional Neural Network CNN and SVM in Skin Cancer Images Recognition", Journal of Techniques, Vol. 3, No. 4, 2021, pp. 15-22.

[19] C. M. Bishop, N. M. Nasrabadi, "Pattern Recognition and Machine Learning", Springer, 2006.

[20] G. Alwakid, W. Gouda, M. Humayun, N. U. Sama, "Melanoma detection using deep learning-based classifications", Healthcare, Vol. 10, No. 12, 2022, p. 2481.

[21] S. Jain, U. Singhania, B. Tripathy, E. A. Nasr, M. K. Aboudaif, A. K. Kamrani, "Deep learning-based transfer learning for classification of skin cancer", Sensors, Vol. 21, No. 23, 2021, p. 8142.

[22] M. J. Alam et al. "S2C-DeLeNet: A parameter transfer based segmentation-classification integration for detecting skin cancer lesions from dermoscopic images", Computers in Biology and Medicine, Vol. 150, 2022, p. 106148.