Official Journal of Erciyes University Faculty of Medicine

DOI: 10.14744/cpr.2025.27623
J Clin Pract Res 2025;47(3):235–245

Investigation of Binary and Multiclass Classification Performance of Skin Cancer Images Using Transfer Learning Methods

Ferdi Güler,[1] Melih Ağraz[1]

[1]Department of Statistics, Giresun University Faculty of Arts and Sciences, Giresun, Türkiye

ABSTRACT

Objective: This study addresses the skin cancer classification problem using transfer learning, comparing different learning architectures and investigating the effects of pre-processing techniques and hyperparameter tuning on model performance.

Materials and Methods: Two datasets were used for binary and multiclass classification tasks. For binary classification, the International Skin Imaging Collaboration (ISIC) 2019 and 2020 datasets were utilized, while the ISIC 2019 dataset was used for multiclass classification. Pre-processing steps such as DullRazor, Histogram Equalization, and Gamma Correction were applied, along with techniques like data augmentation, early stopping, and learning rate reduction.

Results: In binary classification, the ResNet50 model achieved the highest performance with an accuracy of 0.8869 before hyperparameter tuning, while the Visual Geometry Group 16 (VGG16) model outperformed others with an accuracy of 0.9017 after tuning. For multiclass classification, DenseNet121 initially showed the best accuracy of 0.8271 without hyperparameter adjustments. However, after tuning, the VGG16 model again delivered the best performance, achieving an accuracy of 0.9292. Additionally, models such as ResNet50 and MobileNetV2 also demonstrated strong results, confirming the critical role of both pre-processing and hyperparameter optimization in enhancing accuracy.

Conclusion: This study demonstrated the effectiveness of transfer learning models combined with pre-processing techniques and hyperparameter tuning for skin cancer classification. Both classification tasks showed significant performance improvements using these methods. The VGG16 model achieved the highest accuracy in both scenarios, highlighting its potential for further development in dermoscopy systems to assist dermatologists in diagnosing skin cancer. Future research should explore a broader range of datasets and refine pre-processing techniques.

Keywords: Classification, deep learning, hyperparameter settings, pre-processing, skin cancer, transfer learning.

Cite this article as:
Güler F, Ağraz M. Investigation of Binary and Multiclass Classification Performance of Skin Cancer Images Using Transfer Learning Methods. JClinPractRes2025;47(3):235–245.

Address for correspondence:
Melih Ağraz.
Department of Statistics, Giresun University Faculty of Arts and Sciences, Giresun, Türkiye
Phone: +90 505 218 88 04
E-mail: melih.agraz@giresun.edu.tr

Submitted: 18.10.2024
Revised: 09.12.2024
Accepted: 05.04.2025
Available Online: 17.06.2025

Erciyes University Faculty of Medicine Publications - Available online at www.jcpres.com

INTRODUCTION

The skin is one of the body's most vital organs, constantly exposed to external factors. Therefore, maintaining healthy skin and ensuring timely intervention against diseases are of vital importance. Skin cancer results from the uncontrolled proliferation of skin cells with damaged DNA structures and is one of the most common cancers worldwide.[1] If detected early, skin cancer can be permanently treated, making early diagnosis crucial for successful outcomes. However, the traditional method of diagnosing skin cancer by visually examining lesions with the naked eye reduces detection accuracy.[2] In contrast, the increasing use of computer-aided diagnostic systems has improved diagnostic accuracy.[3] Consequently, with today's advancing technology, the use of deep learning techniques has become increasingly important, and image-based classification methods for diagnosing skin lesions are becoming more widespread.

Machine learning is advancing rapidly in genomics and clinical prediction, with models such as Machine Learning Genomics Analysis Platform (ML-GAP), which utilize autoencoders and data augmentation to improve classification performance.[4] Artificial intelligence (AI) also contributes to predicting severe hypoglycemia in type 2 diabetes through multi-view co-training techniques.[5] Additionally, tools like ChatGPT-Enhanced ROC Analysis (CERA) are used to optimize biomarker classification.[6] In dermatology, deep learning, particularly convolutional neural networks (CNNs), has significantly enhanced the classification and segmentation of skin lesions.[7] These AI-driven advancements are becoming increasingly vital for improving the accuracy of skin cancer diagnosis. The most effective method for obtaining images of skin lesions is through dermoscopy.[8] Dermoscopic imaging reduces or filters reflections, provides a clearer and enlarged view of the lesion, and produces high-resolution images.[8] Consequently, dermoscopic images are widely used with deep learning techniques for skin cancer classification.

Matsunaga et al.[9] normalized images by applying brightness and color balance adjustments, as well as color constancy techniques to their dataset, resulting in improved accuracy. Innani et al.[10] applied several images preprocessing steps, including segmentation, to the PH2 dataset and examined the classification results, observing positive effects on model performance. Yalçın and Gürsel[11] enhanced the images by applying preprocessing techniques such as median filtering and black-and-white filtering on their dataset, subsequently analyzing the performance of deep learning models. Their results showed that the models achieved high accuracy. Zebari[12] applied several preprocessing methods to improve image quality, reduce noise, and normalize images in the International Skin Imaging Collaboration (ISIC) datasets they used. By incorporating feature extraction techniques such as Gabor filters and wavelet transformations, they significantly boosted model performance.

KEY MESSAGES

- Image preprocessing techniques and hyperparameter tuning improved the classification performance of transfer learning models on skin cancer data.
- The VGG16 transfer learning model, achieving an accuracy of 90.17% in binary classification and 92.92% in multiclass classification, demonstrated the most effective results for skin cancer classification and holds strong potential for clinical applications.
- Model performance could be further enhanced with larger datasets and more advanced preprocessing techniques, offering new opportunities for improving skin cancer diagnosis.

UdriȘtoiu et al.[13] utilized the HAM10000 dataset for seven-class classification and demonstrated that data augmentation had a positive effect on model accuracy. Hosny et al.[14] classified skin cancer data using pre-trained deep learning models and showed that data augmentation techniques effectively improved the model's generalization performance. Wu et al.[15] performed classification using transfer learning with deep learning models on the ISIC 2019 dataset and applied data augmentation to address data imbalance. As a result, they achieved high accuracy rates. Naqvi et al.[16] focused on improving model accuracy by employing preprocessing methods and data augmentation techniques, demonstrating that these procedures enhanced model performance.

Malo et al.[17] proposed a model based on the VGGNet-16 architecture, using 2,460 skin cancer images from the ISIC dataset. Their model achieved strong results, with an accuracy of 0.876. Yıldız[18] introduced a deep neural network model called C4Net, and comparisons with other deep learning and traditional machine learning models showed that C4Net was more successful. Sun et al.[19] applied data processing techniques such as image preprocessing, data augmentation, and noise reduction in their study. They then classified the data using models like CNN, AlexNet, GoogleNet, ResNet, and VGGNet, and concluded that deep learning models are an effective method for skin cancer classification. Ayan and Ünver[20] compared the accuracy of several deep learning models using the transfer learning method and found that the ResNet-34 model provided the best results.

Based on the literature review, the following observations can be made:

i) Deep learning algorithms and transfer learning methods are frequently used and have proven to be effective approaches for the classification of skin cancer image data.

ii) Applying preprocessing techniques to image data has a positive impact on the classification performance of models.

iii) Due to the limited size of datasets used in skin cancer studies and to address data imbalance, the use of data augmentation techniques has been shown to improve model performance.

In this study, the classification performance of convolutional neural networks for skin cancer was examined using the transfer learning method. Two different datasets were employed for binary and multi-class (seven-class) classification. ISIC 2019 and 2020 data were used to train and test the network for binary classification, while ISIC 2019 data were used for seven-class classification. The study focused on classifying both binary (benign and malignant) and multi-class skin cancer data, specifically the categories of melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis-like lesions (BKL), dermatofibroma (DF), and vascular lesions (VASC). The main goal of this study was to compare different learning architectures and investigate the impact of applying specific preprocessing steps and adjusting hyperparameters on model performance. To achieve this, preprocessing methods such as DullRazor, Histogram Equalization, and Gamma Correction, along with techniques like data augmentation, early stopping, and learning rate reduction, were applied to the data. In this study, transfer learning models including VGG16,[21] ResNet50,[22] InceptionV3,[23] Xception,[24] DenseNet121,[25] and MobileNetV2,[26] as well as a custom convolutional neural network model, were used.

MATERIALS AND METHODS

Dataset

In this study, the ISIC 2019 and 2020 datasets were used for binary (benign vs. malignant) and multi-class (seven categories) classification. For binary classification, ISIC 2019 and 2020[27] were used, while ISIC 2019[28] was utilized for multi-class classification. Kırğıl and Erdaş[29] achieved an accuracy of 0.835 using ISIC 2019, which was also employed in this study. The binary dataset includes 8,100 benign and 6,597 malignant samples, as shown in Figure 1. The multi-class dataset consists of 12,875 melanocytic nevus, 4,522 melanoma, 3,323 basal cell carcinoma, 2,624 benign keratosis, 867 actinic keratosis, 253 vascular lesions, and 239 dermatofibroma samples (Fig. 2). The binary dataset was randomly split into 80% training, 15% testing, and 5% validation sets (Table 1). For multi-class classification, due to data imbalance, augmentation techniques[30] were applied to underrepresented classes (VASC, DF, AK) to increase each to 2,000 samples, while oversampled classes (MEL, NV, BCC, BKL) were downsampled to 2,000. The resulting dataset was then split into 80% training, 15% testing, and 5% validation sets (Table 2). This study evaluates both binary and multi-class classification performance, addressing a gap where most previous studies have focused solely on multi-class classification.

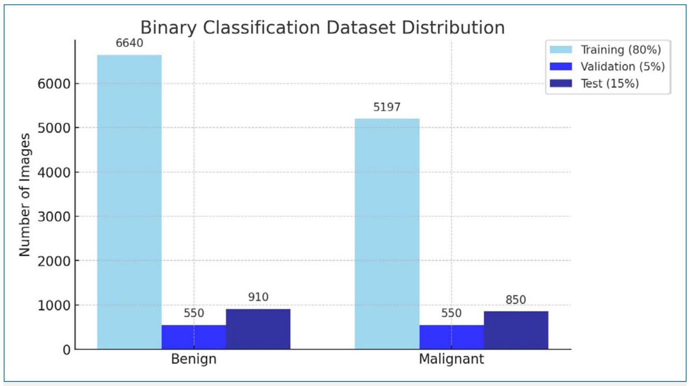

Figure 1. Distribution of the skin cancer binary classification dataset.

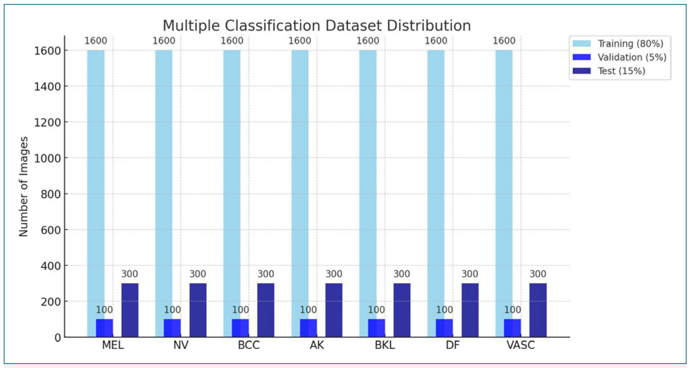

Figure 2. Distribution of the skin cancer multiclass classification dataset.

Table 1. ISIC 2019 and 2020 skin cancer dataset breakdown for binary classification

Cancer type	Total	Training (80%)	Validation (5%)	Testing (15%)
Benign	8,100	6.640	550	910
Malignant	6.597	5.197	550	850

ISIC: International Skin Imaging Collaboration.

Table 2. ISIC 2019 skin cancer dataset breakdown for multiclass classification

Cancer type	Total	Classification data	Training (80%)	Validation (5%)	Testing (15%)
MEL	4.522	2.000	1.600	100	300
NV	12.875	2.000	1.600	100	300
BCC	3.323	2.000	1.600	100	300
AK	867	2.000	1.600	100	300
BKL	2.624	2.000	1.600	100	300
DF	239	2.000	1.600	100	300
VASC	253	2.000	1.600	100	300

ISIC: International Skin Imaging Collaboration; MEL: Melanoma; NV: Melanocytic nevus; BCC: Basal cell carcinoma; AK: Actinic keratosis; BKL: Benign keratosis lentigo; DF: Dermatofibroma; VASC: Vascular lesion.

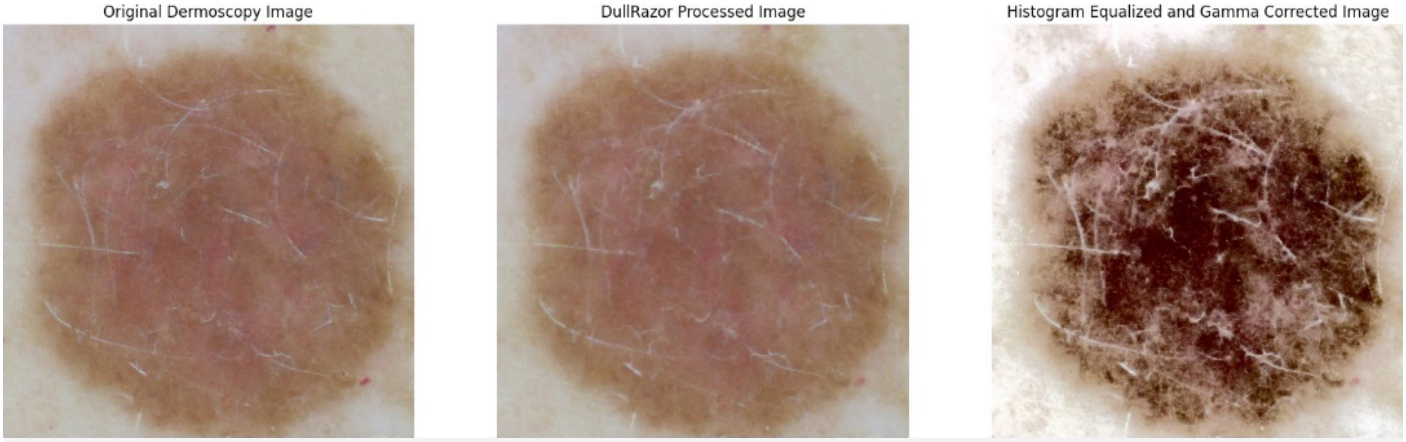

Figure 3. Pre-processing steps applied to dermoscopy images: 1) Original dermoscopy image; 2) DullRazor-processed image; 3) Histogram-equalized and gamma-corrected image.

Pre-Processing

In this study, we examine the importance of preprocessing methods and hyperparameter settings to compare the performance of transfer learning models on two different datasets, each consisting of 80% training, 5% validation, and 15% testing sets. Hyperparameter optimization refers to adjusting the model parameters used in this study, as detailed at the end of the preprocessing section. Akyel and Arıcı[31] applied techniques such as noise masks, adaptive thresholding, and median filtering as preprocessing methods in their study, improving the quality of the data. Similarly, Hermosilla et al.[32] used contrast enhancement and noise reduction techniques to improve image quality in their work with the ISIC 2018 and 2019 datasets. These processes were shown to positively impact model classification performance. In our study, we employed DullRazor, Histogram Equalization, and Gamma Correction as preprocessing methods for the images.

DullRazor is an effective method used to remove thick and dull hairs from images.[33] Histogram Equalization is a commonly used process to enhance contrast or adjust the brightness distribution of an image.[34] Gamma Correction is applied to define the relationship between the numerical values of pixels and their actual brightness.[35] Figure 3 shows the processed image obtained after applying DullRazor, Histogram Equalization, and Gamma Correction to the original image. After preprocessing, hyperparameter settings were applied to the images. The setting "image_size = 128" resized the width and height of each image to 128 pixels. The setting "num_channels = 3" specified that the images were three-channel color images (RGB: Red, Green, Blue). The "batch_size = 32" setting defined the number of images processed by the model in each step during training. The "initial_lr = 1e-4" setting indicated the learning rate used during model training. The "num_epochs = 30" setting specified the number of times the model would scan the training set from start to finish. Finally, "dropout_rate = 0.5" set the probability of randomly disabling a portion of neurons during each training step to help prevent overfitting.

Table 3. Similarities and differences between the classification algorithms used

Model	Advantages	Disadvantages
VGG16	- High accuracy after hyperparameter tuning. - Performs well on balanced datasets.	- Computationally expensive due to high parameter count (~138M). - Prone to overfitting on small datasets.
ResNet50	- Residual connections prevent vanishing gradients. - Good performance even without tuning.	- Requires fine-tuning for optimal accuracy. - Higher inference time due to deep architecture.
DenseNet121	- Compact architecture with fewer parameters (~8M). - Strong performance on imbalanced datasets due to dense connections.	- Longer training time due to feature reuse. - More challenging to fine-tune compared to other models.
MobileNetV2	- Highly efficient, designed for mobile and low-resource environments. - Requires less computation power (~3.5M parameters).	- Lower accuracy without tuning. - May struggle with complex feature extraction compared to deeper networks.
InceptionV3	- Excellent feature extraction for complex patterns. - Multi-scale convolution filters improve generalization.	- Prone to overfitting on small datasets. - Computationally more expensive than MobileNetV2.
Xception	- Efficient handling of complex data structures. - Uses depthwise separable convolutions to improve efficiency.	- High computational cost due to deep layers. - Requires large datasets for optimal performance.
Non-transfer learning	- Fully customizable for specific tasks. - Can be optimized for the dataset used.	- Lower performance in multi-class classification compared to transfer learning models. - Requires extensive training to achieve high accuracy.

VGG16: Visual Geometry Group 16; ResNet50: Residual Network 50; DenseNet121: Densely Connected Convolutional Networks 121.

Transfer Learning

In our study, transfer learning is used to solve similar problems by utilizing models that have been previously trained on millions of datasets.[36] Instead of starting the training process from scratch, the model begins with patterns learned during prior training. Çetiner[37] performed classification using transfer learning with MobileNet models in their study. Magsood and Damaševičius[38] applied contrast enhancement to the data and performed classification using transfer learning, achieving very high accuracy rates, such as 0.9898. This widely used method in the literature has been shown to significantly improve model performance. Demir[26] achieved an accuracy rate of 92.2% using the MobileNetV2 model in their study. In this study, transfer learning models commonly used for skin cancer classification, such as VGG16, ResNet50, InceptionV3, Xception, DenseNet121, and MobileNetV2, were compared with a non-transfer learning model (CNN) specifically designed for the task. The similarities and differences between these models, based on their respective advantages and disadvantages, are summarized in Table 3.

Table 4. Confusion matrix

	True value	
	Positive	Negative
Estimated value		
Positive	TP	FP
Negative	FN	TN

TP: True positives; FP: False positives; FN: False negatives; TN: True negatives.

Performance Metrics

In this study, accuracy was used as the primary metric to evaluate classification results. The confusion matrix, an important tool for measuring classification performance, provides a tabular visualization of the model's predictions compared to the reference labels. In the method shown in Table 4, the accuracy value given in Equation 1 reflects the overall classification performance of the model by measuring how many observations, both positive and negative, were classified correctly.

$$Accuracy = (TN + TP) / (TN + TP + FN + FP) \quad (1)$$

Table 5. Binary and multiclass classification results

Algorithm	Binary classification	Multiclass classification
Performance of transfer learning algorithms on datasets without preprocessing and hyperparameter tuning		
VGG16	0.8022	0.7724
ResNet50	**0.8869**	0.5567
InceptionV3	0.8414	0.7850
Xception	0.8494	0.7996
DenseNet121	0.8823	**0.8271**
MobileNetV2	0.8339	0.8167
Non-transfer learning model	0.8249	0.4916
Performance of transfer learning algorithms on datasets with preprocessing and hyperparameter tuning		
VGG16	**0.9017**	**0.9292**
ResNet50	0.8869	**0.9024**
InceptionV3	0.8500	0.8789
Xception	0.8954	0.8871
DenseNet121	0.8892	0.8296
MobileNetV2	0.8613	**0.9028**
Non-transfer learning model	0.8380	0.6082

VGG16: Visual Geometry Group 16; ResNet50: Residual Network 50; DenseNet121: Densely Connected Convolutional Networks 121.

RESULTS

In our study, the VGG16, ResNet50, InceptionV3, Xception, DenseNet121, and MobileNetV2 transfer learning models, along with a custom convolutional neural network model, were used to classify skin cancer images. For binary classification, the ISIC 2019 and 2020 dataset, consisting of 14,697 images categorized as benign or malignant, was used. For seven-class classification, the ISIC 2019 dataset, consisting of 14,000 images categorized as MEL, NV, BCC, BKL, VASC, DF, and AK, was utilized. The main aim of this study was to examine the effect of preprocessing and hyperparameter tuning on model performance. Therefore, the models were tested in two ways: first, to observe the results without any modifications, and second, to observe the results after applying hyperparameter tuning. Table 5 presents the test results, including the accuracy rates of the transfer learning models before and after preprocessing and hyperparameter tuning. Figure 4 shows examples of correct classification results for binary classification, while Figure 5 presents examples for seven-class classification.

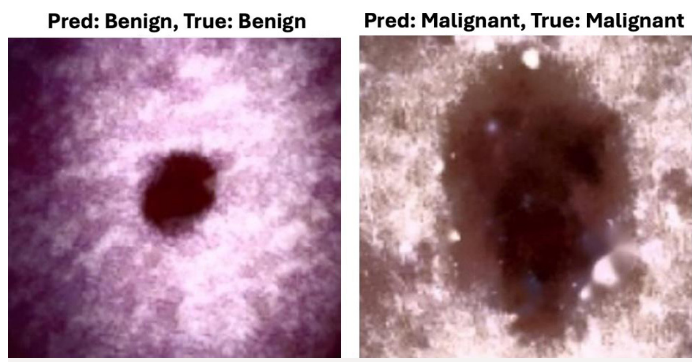

Figure 4. Example of a correct classification result for the binary classification test case in the model with hyperparameter settings.

In Figure 6, the effect of preprocessing and hyperparameter tuning on model performance is graphically presented using the accuracy metric. According to Table 5, the following conclusions were drawn from the binary and seven-class classification test results for all transfer learning models and the custom convolutional neural network model, both with and without hyperparameter tuning:

i) For binary classification performance, ResNet50 showed the best accuracy at 0.8869 when no hyperparameter tuning was applied. However, after hyperparameter tuning, the VGG16 model achieved the best performance with an accuracy of 0.9017.

ii) For multi-class (seven-class) classification, DenseNet121 demonstrated the best performance with an accuracy of

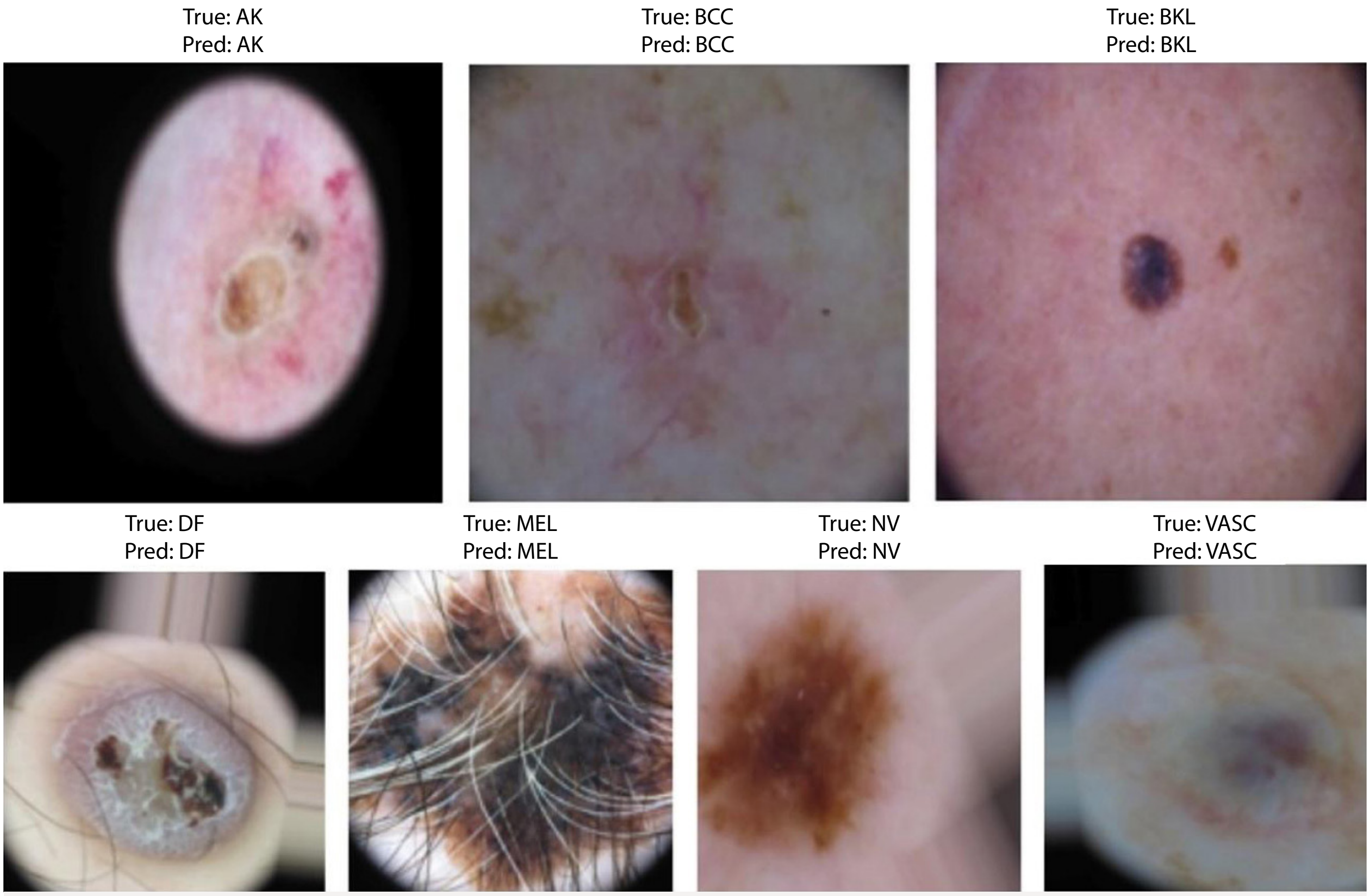

Figure 5. Example of a correct classification result for the multiclass classification test case in the model with hyperparameter settings.

0.8271 without hyperparameter tuning. After tuning, the VGG16 model achieved the highest performance with an accuracy of 0.9292. Additionally, both ResNet50 (0.9024) and MobileNetV2 (0.9028) showed strong performance.

iii) Examining the custom convolutional neural network model, it achieved a solid performance of 0.8249 for binary classification without hyperparameter tuning, which improved to 0.8380 after tuning. In seven-class classification, its initial performance was lower at 0.4916 without tuning, but after tuning, the accuracy significantly improved to 0.6082.

From the test results, it was observed that applying preprocessing and hyperparameter optimization improved performance across all models, with a particularly significant impact on the multi-class classification task. For both binary and seven-class classification, the VGG16 model achieved over 90% accuracy after hyperparameter tuning and preprocessing, making it the best-performing model and well-suited for the data.

When analyzing the multi-class classification results, it was observed that certain classes exhibited high visual similarities in the misclassified examples. Additionally, the difficulty in learning from classes with low sample sizes due to class imbalance negatively impacted classification performance.

i) In multi-class classification, the most misclassified classes were "Actinic Keratosis (AK)" and "Basal Cell Carcinoma (BCC)." It was observed that the model struggled to distinguish between these two classes due to the high visual similarity of their dermatological images.

ii) Classes with a low number of samples were more prone to misclassification. In particular, the error rate was higher for the Vascular Lesion (VASC) and Dermatofibroma (DF) classes compared to the others.

When analyzing the test results in Table 5, the positive effects of preprocessing and hyperparameter optimization on model performance were clearly observed.

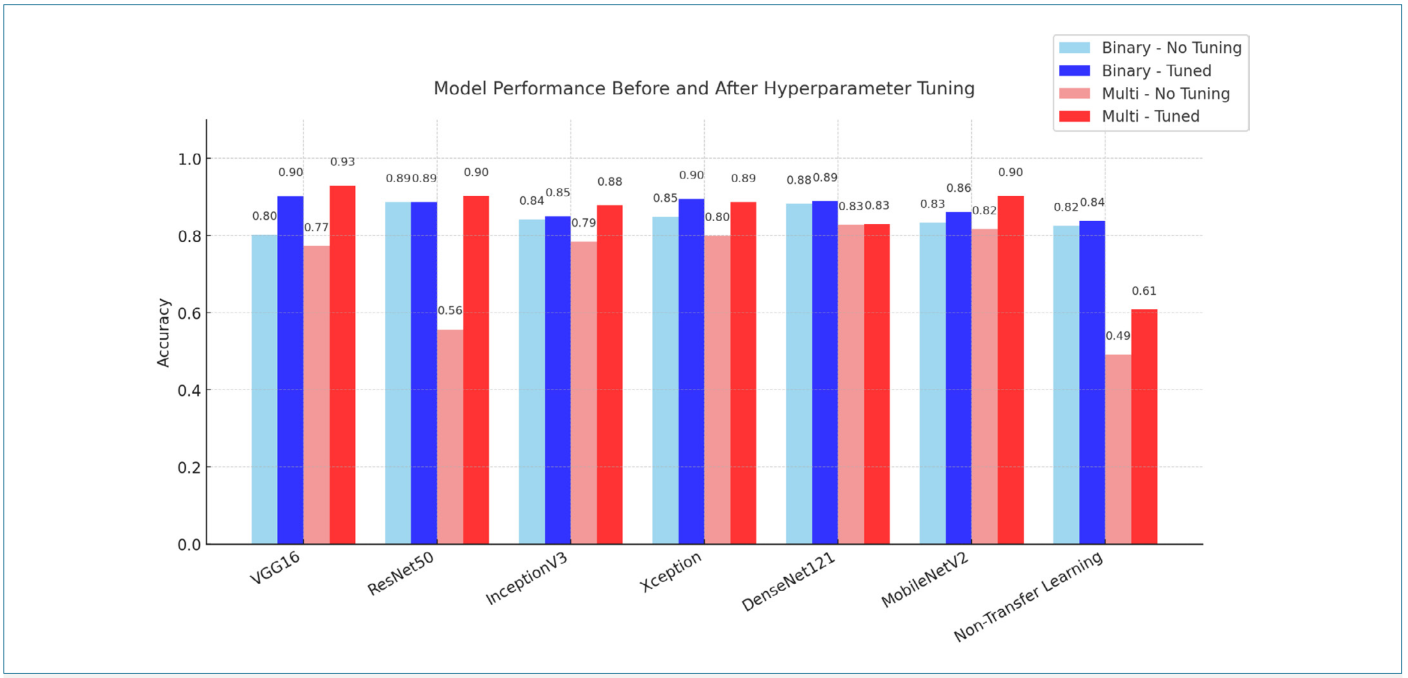

Figure 6. Comparison of transfer learning model results before and after preprocessing and hyperparameter tuning.

i) Preprocessing techniques (DullRazor, Histogram Equalization, and Gamma Correction) enhanced the contrast of skin lesions, enabling the model to capture visual differences more effectively.

ii) Hyperparameter optimization, particularly through learning rate reduction and early stopping techniques, helped reduce overfitting and improved model accuracy.

iii) The custom-designed CNN model showed lower performance in multi-class classification compared to transfer learning models (49.16% → 60.82%). This indicates that transfer learning models can better generalize complex features learned from large pre-trained datasets.

Finally, hypothesis tests were conducted to determine whether the obtained results were statistically significant.

i) The independent sample t-test results indicated that the VGG16 model performed significantly better after hyperparameter tuning compared to other models ($p<0.05$).

ii) The analysis of variance (ANOVA) confirmed that the effect of preprocessing methods on model performance was statistically significant.

These statistical test results demonstrate that the model's optimization process is not a random improvement but a systematic enhancement.

DISCUSSION

The main objective of this study is to investigate the effect of various preprocessing steps and hyperparameter tuning on model performance, in order to compare different transfer learning architectures. For this purpose, a custom-designed CNN model, along with transfer learning models such as VGG16, ResNet50, InceptionV3, Xception, DenseNet121, and MobileNetV2, was used to perform both binary (benign vs. malignant) and seven-class (MEL, NV, BCC, BKL, VASC, DF, AK) skin cancer classification. Preprocessing techniques such as DullRazor, Histogram Equalization, and Gamma Correction were applied, while data augmentation, early stopping, and learning rate reduction were used as additional techniques. Hyperparameter tuning involved adjustments such as "image_size = 128", "num_channels = 3", "batch_size = 32", "initial_lr = 1e-4", "num_epochs = 30", and "dropout_rate = 0.5".

In this study, transfer learning methods for skin cancer classification were compared, and the effects of preprocessing techniques and hyperparameter tuning on model performance were examined. The results indicate that transfer learning models can be effectively used for skin cancer diagnosis and that accuracy rates can be improved with appropriate data processing techniques. As stated in the Results section, transfer learning models demonstrated substantial accuracy improvements through preprocessing and hyperparameter optimization. The obtained results were evaluated by comparing them with the accuracy rates reported in the existing literature.

Table 6. Comparison of the accuracies of similar studies

Model	Researchers	Accuracy in their studies	Accuracy in this study
VGG16	Malo et al.[17]	87.60%	90.17% (binary), 92.92% (multiple)
ResNet50	Ayan and Unver[20]	89.50%	88.69% (binary), 90.24% (multiple)
InceptionV3	Naqvi et al.[16]	85.10%	85.00% (binary), 87.89% (multiple)
Xception	Kirgil and Erdas[28]	86.60%	89.54% (binary), 88.71% (multiple)
DenseNet121	Ergun and Kilic[25]	78.40%	88.92% (binary), 82.96% (multiple)
MobileNetV2	Demir[26]	92.20%	86.13% (binary), 90.28% (multiple)

The key contributions of this study are:

i) Preprocessing techniques (DullRazor, Histogram Equalization, and Gamma Correction) enhanced model accuracy.

ii) Hyperparameter optimization significantly improved the performance of models such as VGG16 and MobileNetV2.

iii) When compared with similar studies in the literature (as presented in Table 6), the high accuracy rates obtained highlighted the importance of preprocessing and hyperparameter tuning in transfer learning models, as well as their potential for skin cancer diagnosis.

iv) By addressing both binary and multi-class classification tasks and providing a comprehensive analysis of the most widely used transfer learning models in the literature, this study filled a significant gap in the field.

This study evaluated the effectiveness of transfer learning models in skin cancer classification, with a particular emphasis on the impact of preprocessing and hyperparameter adjustments on model performance. Although many previous studies have explored skin cancer using deep learning approaches, the achieved accuracy rates have varied considerably depending on whether preprocessing techniques were applied, and in many cases, the effects of such techniques were not evaluated at all. However, this study demonstrates that these techniques led to a 5–10% improvement in model accuracy. Furthermore, when compared with similar studies in the literature, the results of this study surpass previous research, particularly in multiclass classification. Considering that earlier studies often struggled to achieve high accuracy due to imbalanced datasets and the absence of data augmentation, the findings presented here are significant. Table 6 summarizes the results of similar studies using the models evaluated in this work.

Comparison of Similar Studies and Their Results

Similarities

i) Malo et al.[17] achieved an accuracy of 87.60% using a VGG16-based model. In this study, the VGG16 model showed higher performance, achieving an accuracy of 92.92%. Nevertheless, both studies confirm the effectiveness of VGG16 for skin cancer classification.

ii) Ayan and Ünver[20] reported an accuracy of 89.50% using the ResNet34 model. In this study, the ResNet50 model demonstrated a similar performance with an accuracy of 90.24%, supporting the strong classification capability of the ResNet family.

iii) Naqvi et al.[16] achieved an accuracy of 85.10% using the InceptionV3 model. In this study, the InceptionV3 model reached an accuracy of 87.89%, showing a slight improvement while reinforcing the effectiveness of InceptionV3 in similar tasks.

iv) Kırğıl and Erdaş[29] reported an accuracy of 86.60% using the Xception model. In this study, the Xception model achieved a higher result with an accuracy of 88.54%, demonstrating that hyperparameter optimization and preprocessing techniques can lead to slight improvements in performance.

Differences

v) Ergün and Kılıç[25] reached an accuracy of 78.4% using the DenseNet121 model. In this study, the DenseNet121 model achieved a better result, with an accuracy of 88.92%. This difference can be attributed to the use of more effective preprocessing techniques (DullRazor, Histogram Equalization, Gamma Correction) and hyperparameter optimization (learning rate tuning, dropout adjustments).

vi) Demir[26] achieved an accuracy of 92.20% using the MobileNetV2 model. In this study, the MobileNetV2 model showed similar performance, with an accuracy of 90.28%. The slight decrease in performance could be due to differences in dataset composition, class imbalance, and variations in preprocessing approaches.

CONCLUSION

In summary, this study provides valuable insights into the use of transfer learning for skin cancer classification, demonstrating that preprocessing techniques and hyperparameter optimization can lead to significant improvements in model performance. Further work, including real-world clinical testing, will be crucial to fully realize the potential of these models in healthcare applications. Despite achieving high accuracy rates that contribute to the development of AI-assisted diagnostic systems and enable faster, more accurate decision-making in skin cancer diagnosis, this study has certain limitations. The following solutions are proposed for future similar studies.

Data Balance

i) In the multi-class dataset, some classes (such as VASC, DF, and AK) had a low number of samples, and data augmentation techniques were applied to balance them. More advanced generative neural networks, such as Generative Adversarial Networks (GANs), could be used for more effective data augmentation. However, to achieve better generalization with real-world data, experiments should be conducted with larger and more balanced datasets.

Optimization of Transfer Learning Models

i) Although VGG16 provided the highest accuracy, its high computational cost makes it inefficient for mobile applications. Therefore, an R Shiny web application could be developed to make the diagnostic system accessible to specialists and integrate it into the healthcare system.

ii) The integration of explainable AI techniques, such as Grad-CAM, could make the model's decision-making process more transparent. This would allow dermatologists to better understand the model's predictions and make more informed decisions regarding its reliability.

Medical Applicability and Clinical Testing

i) The model has not yet been tested in clinical settings for real-world applications. To enhance its clinical relevance, it should be tested by real doctors and validated on a patient basis.

Ethics Committee Approval: Ethics committee approval was not required for this study.

Informed Consent: Written informed consent was obtained from patients who participated in this study.

Conflict of Interest: The authors have no conflict of interest to declare.

Financial Disclosure: The authors declared that this study has received no financial support.

Use of AI for Writing Assistance: Not declared.

Author Contributions: Concept – FG, MA; Design – FG, MA; Supervision – MA; Resource – FG; Materials – FG; Data Collection and/or Processing – FG; Analysis and/or Interpretation – FG, MA; Literature Search – FG; Writing – FG, MA; Critical Reviews – FG, MA.

Acknowledgements: We would like to thank lecturer Selda Atalar for their invaluable contributions to the preparation of this manuscript. Their assistance in data collection and initial analysis was greatly appreciated. Additionally, we would like to thank Halil Ayar and Satı Sarı for their support.

Peer-review: Externally peer-reviewed.

REFERENCES

1. Rey-Barroso L, Peña-Gutiérrez S, Yáñez C, Burgos-Fernández FJ, Vilaseca M, Royo S. Optical technologies for the improvement of skin cancer diagnosis: A review. Sensors (Basel) 2021;21(1):252. [CrossRef]
2. Adanur Dedeturk B, Taşdemir K, Bakir-Gungor B. Cilt kanseri görüntü sınıflandırması için görüntü ön işlemenin evrişimsel sinir ağları performansı üzerindeki etkileri. Erciyes Univ Fen Bilim Enst Fen Bilim Derg [Article in Turkish] 2022;38:190-200.
3. Celebi ME, Kingravi HA, Uddin B, Iyatomi H, Aslandogan YA, Stoecker WV, et al. A methodological approach to the classification of dermoscopy images. Comput Med Imaging Graph 2007;31(6):362-73. [CrossRef]
4. Agraz M, Goksuluk D, Zhang P, Choi BR, Clements RT, Choudhary G, et al. ML-GAP: Machine learning-enhanced genomic analysis pipeline using autoencoders and data augmentation. Front Genet 2024;15:1442759. [CrossRef]
5. Agraz M, Deng Y, Karniadakis GE, Mantzoros CS. Enhancing severe hypoglycemia prediction in type 2 diabetes mellitus through multi-view co-training machine learning model for imbalanced dataset. Sci Rep 2024;14(1):22741. [CrossRef]
6. Agraz M, Mantzoros C, Karniadakis GE. ChatGPT-Enhanced ROC Analysis (CERA): A shiny web tool for finding optimal cutoff points in biomarker analysis. PLoS One 2024;19(4):e0289141. [CrossRef]
7. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115-8. Erratum in: Nature 2017;546(7660):686. [CrossRef]
8. Taghizadeh M, Mohammadi K. The fast and accurate approach to detection and segmentation of melanoma skin cancer using fine-tuned YOLOV3 and SegNet based on deep transfer learning. arXiv 2022;2210.05167.
9. Matsunaga K, Hamada A, Minagawa A, Koga H. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. arXiv 2017;1703.03108.
10. Innani S, Dutande P, Baheti B, Baid U, Talbar S. Deep learning based novel cascaded approach for skin lesion analysis. In: International Conf Comput Vis Image Process. Cham: Springer Nature Switzerland; 2022. p. 615-26. [CrossRef]
11. Yalçın T, Gürsel AT. Improving digital image quality for convolution neural network based computer-aided diagnosis (CNN-CAD) of skin cancer. Çukurova Univ Müh Fak Derg 2021;36:1099-110. [CrossRef]
12. Zebari NAM. Skin cancer diagnosis based on machine learning techniques. Master of Science Thesis: Harran University; 2022.

13. UdriŞtoiu AL, Stanca AE, Ghenea AE, Vasile CM, Popescu M, UdriŞtoiu ŞC, et al. Skin diseases classification using deep leaning methods. Curr Health Sci J 2020;46(2):136-40.
14. Hosny KM, Kassem MA, Foaud MM. Skin cancer classification using deep learning and transfer learning. In: 2018 9th Cairo International Biomedical Engineering Conference (CIBEC). IEEE; 2018:90-3. [CrossRef]
15. Wu Y, Chen B, Zeng A, Pan D, Wang R, Zhao S. Skin cancer classification with deep learning: A systematic review. Front Oncol 2022;12:893972. [CrossRef]
16. Naqvi M, Gilani SQ, Syed T, Marques O, Kim HC. Skin cancer detection using deep learning-A review. Diagnostics (Basel) 2023;13(11):1911. [CrossRef]
17. Malo DC, Rahman MM, Mahbub J, Khan MM. Skin cancer detection using convolutional neural network. In: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC). IEEE; 2022:169-76. [CrossRef]
18. Yildiz O. Melanoma detection from dermoscopy images with deep learning methods: A comprehensive study. J Fac Eng Archit Gazi Univ 2019;34(4):2241-60.
19. Sun J, Yao K, Huang G, Zhang C, Leach M, Huang K, et al. Machine learning methods in skin disease recognition: A systematic review. Processes 2023;11(4):1003. [CrossRef]
20. Ayan E, Ünver HM. Data augmentation importance for classification of skin lesions via deep learning. In: 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT). IEEE; 2018:1-4. [CrossRef]
21. Kaya V, Akgül İ. Classification of skin cancer using VGGNet model structures. Gümüşhane Univ Fen Bilim Derg 2022;13(1):190-8. [CrossRef]
22. Tasci B. Ön eğitimli evrişimsel sinir ağı modellerinde öznitelik seçim algoritmasını kullanarak cilt lezyon görüntülerinin sınıflandırılması. Fırat Univ Müh Bilim Derg [Article in Turkish] 2022;34(2):541-52. [CrossRef]
23. Sagar A, Dheeba J. Convolutional neural networks for classifying melanoma images. Biorxiv 2020:110973. [CrossRef]
24. Kondaveeti HK, Edupuganti P. Skin cancer classification using transfer learning. In: 2020 IEEE Int Conf Advent Trends Multidiscip Res Innov; 2020; India. p. 1-4. [CrossRef]
25. Ergün E, Kılıç K. Derin öğrenme ile artırılmış görüntü seti üzerinden cilt kanseri tespiti. Black Sea J Eng Sci [Article in Turkish] 2021;4:192-200. [CrossRef]
26. Demir F. Derin öğrenme tabanlı yaklaşımla kötü huylu deri kanserinin dermatoskopik görüntülerden saptanması. Firat Univ Eng Sci [Article in Turkish] 2021;33:617-24. [CrossRef]
27. Kaggle Digital Repository. Skin cancer ISIC 2019 & 2020 malignant or benign. Available from: https://www.kaggle.com/datasets/sallyibrahim/skin-cancer-isic-2019-2020-malignant-or-benign/data. Accessed April 28, 2025.
28. Kaggle Digital Repository. Skin lesion images for melanoma classification. Available from: https://www.kaggle.com/datasets/andrewmvd/isic-2019. Accessed April 28, 2025.
29. Kırğıl ENH, Erdaş ÇB. Skin cancer diagnosis using deep learning techniques fed by dermoscopic images. In: 2023 Med Technol Congr TIPTEKNO; 2023; Cyprus. p. 1-4. [CrossRef]
30. Hussain M, Bird JJ, Faria DR. A study on CNN transfer learning for image classification. In: Lotfi A, Bouchachia H, Gegov A, Langensiepen C, McGinnity M, editors. Adv Comput Intell Syst Contributions Presented at the 18th UK Workshop on Computational Intelligence; 2018; Nottingham, UK. Cham: Springer; 2019. p. 191-202.
31. Akyel C, Arıcı N. Cilt kanseri görüntülerinde gürültü temizliği ve lezyonun dört sınıfa ayrılması. Afyon Kocatepe Univ Fen Muh Bilim Derg [Article in Turkish] 2024;24:284-93. [CrossRef]
32. Hermosilla P, Soto R, Vega E, Suazo C, Ponce J. Skin cancer detection and classification using neural network algorithms: A systematic review. Diagnostics 2024;14:454. [CrossRef]
33. Lee T, Ng V, Gallagher R, Coldman A, McLean D. DullRazor: A software approach to hair removal from images. Comput Biol Med 1997;27(6):533-43. [CrossRef]
34. Arısoy MÖ, Dikmen Ü. Manyetik belirti haritalarının histogram eşitleme yöntemi kullanılarak iyileştirilmesi. Yerbilimleri [Article in Turkish] 2014;35(2):141-68. [CrossRef]
35. Cao G, Huang L, Tian H, Huang X, Wang Y, Zhi R. Contrast enhancement of brightness-distorted images by improved adaptive gamma correction. Comput Electr Eng 2018;66:569-82. [CrossRef]
36. Agarwal N, Sondhi A, Chopra K, Singh G. Transfer learning: Survey and classification. In: Smart Innov Commun Comput Sci Proceedings of ICSICCS 2020. 2021. p. 145-55. [CrossRef]
37. Çetiner H. MobileNetV2 ve MobileNetV3 tabanlı derin öğrenme yaklaşımları ile cilt kanserlerinin sınıflandırılması. In: 3rd Int Conf Appl Eng Nat Sci 2022; Turkey. [CrossRef]
38. Maqsood S, Damaševičius R. Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare. Neural Netw 2023;160:238-58. [CrossRef]