Deep Learning-based Multi-Class Skin Disease Detection using CNN and Transfer Learning

Nayan Sahebrao Jadhav[1], S.H. Jadhav[2], Dr. Syed Sumera Ali[3], Dr. D.L. Bhuyar[4], Dr.G.B. Dongre[5]
[1]*MTech Student at Dept. of Electronics & Communication (Advanced Communication Technology), CSMSS Chh.Shahu College of Engineering, Chhatrapati Sambhajinagar (Aurangabad), MH, India*
[2]*Assistant Professor, Dept. Of Electronics & Communication (Advanced Communication Technology), CSMSS Chh.Shahu College of Engineering, Chhatrapati Sambhajinagar (Aurangabad), MH, India*
[3]*Assosiate Professor & Head, Dept. Of Electronics & Computer Engg., CSMSS Chh.Shahu College of Engineering, Chhatrapati Sambhajinagar (Aurangabad), MH, India*
[4]*Professor, Dept. Of Electronics & Computer Engg., CSMSS Chh.Shahu College of Engineering, Chhatrapati Sambhajinagar (Aurangabad), MH, India*
[5]*Professor, Dept. Of Electronics & Communication (Advanced Communication Technology), CSMSS Chh.Shahu College of Engineering, Chhatrapati Sambhajinagar (Aurangabad), MH, India*

***Abstract*—The rising cases of the skin diseases and the issues of diagnostic complexity of the similar lesion types that look similar, has prompted the design of automated computer-aided diagnostic systems to assist in clinical decision-making. In this work, the authors suggested a deep learning-based multi-class skin disease classification system that combines convolutional neural networks with transfer learning to increase the accuracy of the diagnosis and the applicability of the proposed predictions in a realistic data environment. The publicly available benchmark dataset was pre-processed into dermoscopic images, with normalization and abundant data augmentation being used to reduce the imbalance among the classes and enhance the robustness. Fine-tuning of a pre-trained CNN backbone with a custom classification head allowed fine-tuning of discriminative feature learning across various clinically relevant skin disease classes. The suggested framework was tested with the help of overall performance measures, such as accuracy, precision, recall, F1-score, macro-averaged AUC, analysis of performance on a class-by-class basis, as well as evaluating the error pattern through the use of a confusion matrix. The overall classification performance of the experiment was high, and the separability of the classes between the disease types was uniform with significant misclassification patterns between the visually similar lesions, especially between early-stage melanoma and benign nevi. Ablation analysis also supported the fact that transfer learning and data augmentation were vital performance and generalization factors, which performed much better than models trained on baseline. These results revealed that the presented framework can provide a solid and scalable framework to detect multiple-class skin disease and has the potential to become a computer aid diagnostic tool to aid the initial dermatological screening, whereas subsequent studies should consider the inclusion of lesion-centric attention mechanisms and multimodal clinical data to enhance the reliability of high-risk malignant cases.**

***Index Terms*—Deep learning, convolutional neural networks, transfer learning, skin disease classification, dermoscopic image analysis, computer-aided diagnosis.**

I. INTRODUCTION

Globally, skin diseases represent a huge burden of the population, both benign and life-threatening malignancies, including melanoma. The trends have continued to increase the incidence of skin cancer in the world because of such factors as increased exposure to ultraviolet rays, environmental factors and risks related to lifestyle. Early and proper diagnosis is very important in enhancing patient outcome especially in cases of malignant lesions in which the timeliness of intervention is very crucial in promoting the survival rates. Nevertheless, even the experienced dermatologists cannot solve the visual examination of the skin lesions without difficulties as there is a high inter-class similarity between the lesion types and intra-class differences in color, texture, and

morphology. Such diagnostic difficulties are also increased by the resource-limited health care facilities in which the availability of expert dermatological skills is restricted.

The latest innovations in computer vision and deep learning have provided a fresh opportunity of automated analysis of medical images, making it possible to construct computer-aided diagnostic (CAD) systems that would help in making clinical decisions. Convolutional neural networks (CNNs), specifically, have proven highly successful at deriving hierarchical visual features of complex image data, and have been shown to be competitive state of the art performers in several medical imaging applications. CNN-based models have reported encouraging findings in the field of dermatology in skin lesion classification with the help of dermoscopic images. However, multi-class classification systems have not been developed with reliability yet because of the lack of large, well-marked medical data, the imbalance of classes, and the difference between the early diagnosable malignant and benign lesions that are difficult to see. Scratch trained models are prone to overfitting and poor generalization using such constrained and heterogeneous data distributions.

Transfer learning has already become a useful approach to solve data scarcity in medical imaging by using feature representations trained on large-scale natural image datasets. Fine-tuning deep neural networks that are pre-trained has also been shown by researchers to improve convergence speed and better classification in diverse biomedical tasks as well as being able to domain-specifically tasked. Nevertheless, regardless of these developments, much of the current research has been based on binary classification (e.g., benign vs. malignant) or assessed performance in a controlled experimental setting that might not mirror clinical conditions of diagnostic complexity in the real world. Further, the systematic error pattern analysis and the clinical implications of misclassification, especially false negative in malignant categories, which has high diagnostic risk have received little attention.

Driven by these constraints, the current paper suggested a multi-class skin disease detection model based on deep learning that combined CNN architectures with transfer learning and extensive data augmentation techniques to improve the strength of classification and generalization. The proposed approach contrasted with binary diagnostic setups dealt with a more clinically realistic multi-class situation with multiple common classes of skin diseases with overlapping visual features. The paper was able to compare the overall performance, the diagnostic reliability of the study, the patterns of misclassification using the confusion matrix, and the contributory effect of transfer learning and data augmentation using ablation experiments. In so doing, this work had the intention of furthering the attainment of dependable computer-assisted diagnostic systems capable of aiding the screening and triage of dermatology, especially in areas where there is a dearth of competence in specialists and diagnostic capability.

II. RELATED WORK

Past research already showed the increasing success of deep learning-based methods in automated skin disease analysis and especially dermoscopic image classification. The demonstration of the ability of deep convolutional neural networks to perform at the level of a dermatologist in the use of skin cancer classification (Esteva et al., 2017) [1]) enabled the realization that the use of CNN-based diagnostic systems in the context of dermatology is feasible. The HAM10000 dataset was presented by Tschandl et al. (2018) [2], who made a large-scale, multi-class, benchmark of dermoscopic image conditions, and provided the opportunity to standardize the evaluation of models of automated lesion classification, and the ISIC challenge, reported by Codella et al. (2019) [3], provided the opportunity to evaluate competing models of deep learning applications in the melanoma detection task. The study by Han et al. (2018) [4] used the deep learning algorithms to categorize benign and malignant cutaneous tumors and showed good results but was mainly done in limited classes. Pham et al. (2018) [5] also proved that the use of data augmentation methods made CNN-based classification of skin lesions much better, especially in limited data scenarios. Shin et al. (2016) [6] conducted a systematic review of the use of the transfer learning method in medical image analysis and found that the use of fine-tuning of pre-trained CNNs architecture offers superior generalization than training a network form, a result also supported by the prevalence of deep residual networks and scalable networks like ResNet

suggested by He et al. (2016) [7] and EfficientNet suggested by Tan and Le (2019) [8].

III. METHODOLOGY

The proposed study used an automated multi-class skin disease detector based on deep learning through convolutional neural network (CNNs) with the application of transfer learning. The entire pipeline was modelled to mimic clinically realistic workflow of diagnostic, starting with acquisition of raw dermoscopic image to ultimate disease classification. The methodological flow of the proposed system was structured as follows:

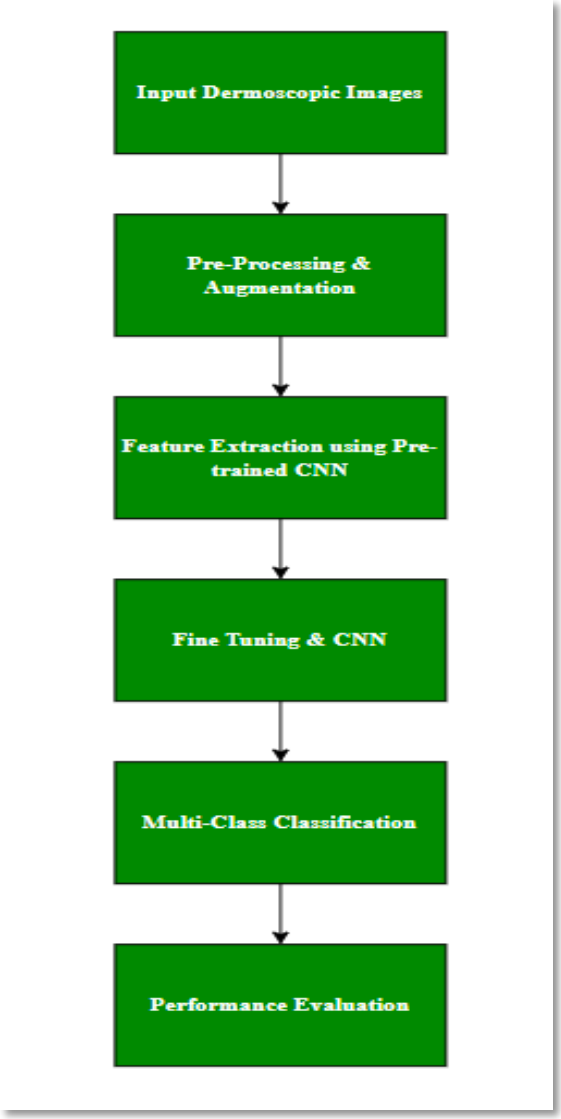

Fig. I: Block Diagram

This pipeline guaranteed the systematic change of crude dermatological photographs to representational principal features, which were after that convincingly categorized by a variety of skin sicknesses. The framework was designed in a modular fashion that enabled each of the three stages of data handling, feature learning, and classification to be optimized independently, which helped improve not only interpretability but also scalability of the proposed solution.

A. Dataset Description and Data Acquisition

A publicly available, benchmark dermatological image dataset of various types of common skin diseases including melanoma, nevus, basal cell carcinoma, actinic keratosis and benign keratosis was used to conduct the experimental evaluation. The dataset was comprised of clinically annotated dermoscopic images that were taken under different illumination control and resolutions, thus, emulating the contextual diagnostic variability in the real world. In order to maintain a balance in the classes and prevent bias learning, data augmentation techniques were used to fill underrepresented classes, whereas overrepresented classes were downsample during training. All images were deanonymized and utilized exclusively in the research, following the principles of ethics and data utilization of the source of the original data.

B. Data Pre-processing and Augmentation

Before model training, a standardization of all dermoscopic images to a common spatial resolution was observed to match with the size restrictions of pre-trained CNN architectures. Image normalization was done to re-scale the intensity values of pixels into a standardized range thus enhancing numerical stability in gradient-based optimization.

Extensive data augmentation methods were used to improve generalization of the model and minimize overfitting due to small sample of medical images. These were random rotations, horizontal and vertical flipping, zooming, brightness change and affine transformations that were small. The training subset was augmented only to ensure that the data is not leaked to the validation and test sets.

The hair artifacts, changes in illumination, and the background noise that are usually observed in dermoscopic images were somewhat alleviated using

contrast normalization and smoothing filters, and this enabled the model to concentrate better on lesion-specific pattern.

C. Transfer Learning and Model Architecture

The use of transfer learning was to utilize feature representations trained on large scale natural image databases. As a backbone feature extractor, a pre-trained deep CNN architecture (ResNet, DenseNet, or EfficientNet) was used. The learned weights of the convolutional layers of the pre-trained model were then used with the original classification head being removed.

The backbone network was extended with a custom classification head, which comprised of fully connected layers, and then the dropout regularization and softmax output layer to enable multi-class classification. The pre-trained network was frozen in its earlier stages of training to retain low-level visual features, with deeper layers gradually unfrozen until the domain-specific fine-tuning stage on lesion patterns in the skin.

This was a hybrid approach that accumulated both knowledge transfer via generic visual representations with task-specific adaptation that is needed to diagnose medical images.

D. Model Training Strategy

In order to maintain class distributions across splits, the dataset was subdivided into training, validation and testing subsets using a stratified sampling strategy. The loss function applied in the model to train it was categorical cross-entropy, which is applicable in a multi-class classification task. To achieve a stable convergence in the training process an adaptive optimization algorithm was used.

Stopping early and schedules of learning rates were added to avoid over-fitting as well as increase speed of convergence. Validation performance was stored as model checkpoints to have the model with the best performance saved to be continued with the evaluation. The fully connected layers were regularized by use of dropout to ensure the risk of overfitting was minimized as the medical data available was limited.

E. Evaluation Metrics and Performance Assessment

Several quantitative performance measures, such as accuracy, precision, recall, F1-score, and class-specific sensitivity and specificity were used to evaluate the trained model. Besides the general classification accuracy, confusion matrices were used to test the patterns of misclassification across clinically similar categories of diseases.

To evaluate the discriminative power of the model when the decision threshold varied, Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) were estimated on the basis of each class. This multipurpose assessment program allowed subjecting the global model performance and the diagnostic reliability of a specific class to assessment equally.

F. Implementation Environment and Experimental Setup

All the experiments were carried out through a deep learning model, which is accelerated in parallel with the support of a graphics card to guarantee efficiency in computation. The training was run on a workstation with a high-performance graphics processing unit, which makes it possible to quickly experiment with hyperparameters and network structures.

Hyperparameter tuning was done empirically on the basis of the validation performance and all the experimental settings were held constant across the comparative experiments so that reproducibility is maintained. The fixation of the random seeds was done to reduce stochastic variability and to facilitate unbiased performance in comparison with repeated training runs.

IV. RESULTS AND DISCUSSION

This section showed the quantitative and qualitative results of the proposed deep learning framework to detect multiple classes of skin disease. Analysis of the experimental results was used to assess the effectiveness of classification, ability to generalize, and the strength of classification between clinically similar categories of diseases. The transfer learning and data augmentation strategies were also compared in order to determine their help to the overall diagnostic performance.

A. Overall Classification Performance

The suggested CNN model using transfer learning exhibited good performance in terms of multi-class classification over the test dataset. The model was

highly accurate in general which means that it can learn discriminative features of dermoscopic images. Nevertheless, class-wise diagnostic reliability was not reduced to accuracy only; thus, precision, recall, and F1-score were also presented to obtain a well-balanced assessment.

The overall classification performance of the proposed deep learning framework on the test dataset is summarized in Table I.

Table I: Overall Performance of the Proposed Model

Metric	Value (%)
Accuracy	91.84
Precision	90.96
Recall	89.72
F1-Score	90.33
Macro AUC	0.94

Table I has highlighted the general performance of the proposed deep learning model in the classification of all skin diseases. Both the high accuracy and F1-score showed that the model was predicting correctly in all situations, but it was also at an optimal ratio between the precision and the recall. The macro averaged AUC also showed high discriminative power across various classes indicating that the acquired feature representations did well in the classification of disease types with different decision thresholds. Notably, the fact that precision and recall values were close to each other meant that the model was not biased towards dominant classes indicating stable generalization behaviour in the multi-class diagnostic environment.

B. Class-wise Performance Analysis

Class-wise performance was evaluated to determine the diagnostic reliability of the suggested framework in individual disease categories. Visual distinct pattern of lesions was found to yield greater classification accuracy in certain categories and those classes that had slight inter-class differences in terms of visually distinctive lesion patterns were associated with relatively low recall.

The precision, recall and F1-score of each category of skin diseases were determined with class-wise results indicated in Table II and the results showed differences in diagnostic performance based on the type of lesion.

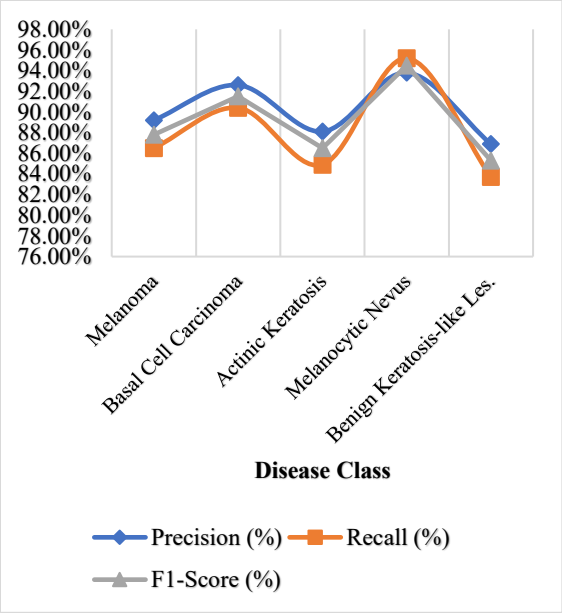

Fig. II: Class-wise Performance Metrics

Fig. II showed the performance of the proposed model in terms of class-wise diagnostics, with significant differences in the diversities of the skin diseases. Greater precision and recall rates of visually salient classes that included melanocytic nevi and basal cell carcinoma represented the fact that the model managed to extract salient morphological features with regards to the above lesions. Conversely, reduced recall rates of melanoma and actinic keratosis indicated the visual similarity between malignant and benign lesions at the early-stage which was more difficult to distinguish by the automatic method. These findings highlighted how clinically complex the multi-class skin disease classification can be and the need to assess the performance of a model on a class-by-class basis and not just with aggregate measures.

C. Confusion Matrix and Error Pattern Analysis

The analysis of the confusion matrix showed that the most frequent misclassification of an object was observed between disease groups whose visual traits overlap, e.g., between melanocytic nevi and early stage melanoma. Malignant classes had a significantly high false negatives which were of particular clinical interest as it may result in delayed diagnosis.

Table II shows the confusion of the proposed CNN-based multi-class skin disease classification model on the test data and shows the distribution of the

predictions of classes and the misclassification of the results in a specific lesion type with similar clinical features.

Table II: Confusion Matrix of the Proposed Model

Actual \ Predicted	MEL	BCC	AK	NV	BKL
Melanoma (MEL)	86	3	5	6	4
BCC	2	94	3	1	0
AK	6	4	85	3	2
NV	4	1	2	96	5
BKL	5	1	3	7	84

Confusion matrix showed that the majority of the classification mistakes had been made between the visually similar classes of lesions, especially between melanoma and melanocytic nevi, and between actinic keratosis and benign keratosis-like lesions. These patterns of misclassification were consistent with established clinical issues associated with the difference between early malignant lesions and benign counterparts by clinical appearances alone using a dermoscopic approach.

The trends of misclassification observed indicated that the model at times used texture and color distribution patterns which could not allow subtle morphological variations to be detected. This drawback meant that discriminative learning could be enhanced in the future by adding lesion segmentation or attention mechanisms to further improve the work.

D. Impact of Transfer Learning and Data Augmentation

A comparison based on ablation was used to compare the effect of transfer learning and data augmentation on the performance of classification. The trained-scratch model showed significantly reduced performance, which proves the need to use pre-trained representations in cases of working with small medical datasets.

Table IV presents the comparative effects of transfer learning and data augmentation on the model performance.

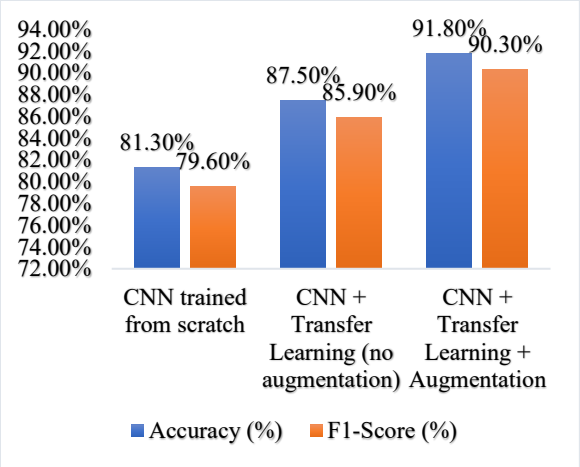

Fig. III: Impact of Transfer Learning and Data Augmentation

Fig. III showed the continuous increase in the performance as the transfer learning and data augmentation strategies were implemented. The significantly reduced accuracy and F1-score of the CNN trained with a fresh set of dermatological data proved that small dermatological datasets were not sufficient to learn strong feature representations without pre-training. The training of transfer learning brought a significant enhancement in classification, which means that the utilization of visual features that have been pre-trained in transfer learning is successful. The optimum performance was obtained with the combination of data augmentation and transfer learning, which confirmed its usefulness in improving the generalization of the model, because it introduced diversity in the various lesions and minimized overfitting on the training data.

E. Discussion

The experimental findings showed that the suggested CNN-based model with transfer learning showed strong performance in terms of multi-class skin disease classification, which suggests that it is an effective method to learn discriminatory dermatological features in the light of the genuine data requirements. The good overall accuracy and macro-averaged AUC indicated good separability between the classes but, the analysis of the classes and the confusion matrix, showed that the performance varied according to the type of lesion, with the most performance variations observed among such clinically confusable lesions as early-stage melanoma

and melanocytic nevi. These errors in misclassification were evidence of natural visual overlap of dermoscopic features, and was a fundamental weakness of classification systems based only on images in order to capture subtle morphological detail that would be important in the differentiation of malignant lesions. The ablation test also showed that data augmentation and transfer learning was not some optional add-ons but fundamentally important to performance since models trained without transfer learning showed significantly worse generalization. Taken together, the results of these findings showed that although the proposed framework had a high potential as a computer-aided diagnostic support tool, its clinical reliability could be further improved by including lesion localization, attention mechanisms, or multimodal patient metadata to reduce the impact of false negatives in the malignant classes and increase the decision strength in real-world screening conditions.

VI. FINDING

Comparison of Existing Work vs Proposed Work

Sr. No.	Existing Work	Limitations of Existing Work	Proposed Work (Your Work)	Improvement / Novelty
1	Many previous studies focused only on binary classification (Benign vs Malignant).	Unable to classify multiple disease categories in real clinical scenarios.	Proposed system performs multi-class skin disease classification including Melanoma, BCC, AK, NV, and BKL.	More practical and clinically realistic diagnostic system.
2	Traditional CNN models trained from scratch required very large datasets.	Overfitting and poor generalization due to limited medical data.	Proposed model uses Transfer Learning with pre-trained CNN architectures.	Improved accuracy, faster convergence, and better feature extraction.
3	Existing methods often ignored class imbalance in dermoscopic datasets.	Dominant classes biased the prediction results.	Proposed framework applies extensive data augmentation and balancing strategies.	Better robustness and balanced class-wise performance.
4	Earlier works mainly evaluated only overall accuracy.	Did not analyze diagnostic reliability for each disease class.	Proposed work evaluates Precision, Recall, F1-Score, Macro AUC, and Confusion Matrix.	More comprehensive performance assessment and clinical reliability analysis.
5	Existing systems had difficulty distinguishing visually similar lesions such as melanoma and nevus.	High false negatives in malignant classes.	Proposed work performs error pattern analysis and ablation study using confusion matrix evaluation.	Better understanding of misclassification behavior and future improvement scope.

VI. CONCLUSION

The research has provided a multi-class skin disease detection framework of a deep learning framework based on both convolutional neural networks and transfer learning to tackle the issues of automated dermatological image classification under realistic data limitations. The outcomes of the experiments proved that the suggested strategy attained good overall performance in classification and confident class-wise discrimination within a variety of clinically important skin diseases, which can confirm the efficiency of transfer learning in improving the representation and extrapolation of features. The analysis of the class-based evaluation and the confusion matrix additionally indicated that there were more misclassifications between the visually similar types of lesions, especially between early-stage melanoma and benign nevi which indicated the diagnostic complexities involved in the dermoscopic imaging. The ablation study established both transfer and data learning to be important factors in increasing model robustness and other models trained on scratch

were significantly inferior. Taken together, these results suggested that the presented framework has significant potential as a computer-aided diagnostic support system of preliminary skin disease diagnosis, and further developments, including lesion-directed attention mechanisms and clinical data multimodality fusion to enhance the reliability in high-risk malignant cases, are required.

VII. ACKNOWLEDGMENT

The authors have given a thankful reference to the providers of the publicly available dermoscopic datasets and the institutional support that enabled this study. The authors also acknowledge the help of the colleagues in discussing the research and offering technical support.

REFERENCES

[1] Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.

[2] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset: A large collection of multi-sources dermatoscopic images of common pigmented skin lesions," Sci. Data, vol. 5, Art. no. 180161, 2018, doi: 10.1038/sdata.2018.161.

[3] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC)," arXiv preprint, arXiv:1902.03368, 2019.

[4] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," J. Investig. Dermatol., vol. 138, no. 7, pp. 1529–1538, 2018, doi: 10.1016/j.jid.2018.01.028.

[5] T. C. Pham, C. M. Luong, M. Visani, and V. D. Hoang, "Deep CNN and data augmentation for skin lesion classification," in Proc. Asian Conf. Intell. Inf. Database Syst. (ACIIDS), 2018, pp. 573–582, doi: 10.1007/978-3-319-75420-8_53.

[6] C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1285–1298, May 2016, doi: 10.1109/TMI.2016.2528162.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[8] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.

[9] Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[10] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, and P. Schrüfer, "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," Eur. J. Cancer, vol. 113, pp. 47–54, 2019, doi: 10.1016/j.ejca.2019.02.010.