Condorcet Jury Theorem-Based Ensemble for Skin Lesion Classification

Nitesh Pradhan[1], Harshit Jain[1], Diya Ghodasara[1]

[1]Department of CSE, The LNM Institute of Information Technology, Jaipur, 302031
[1]Department of CSE, The LNM Institute of Information Technology, Jaipur, 302031
[1]Department of CSE, The LNM Institute of Information Technology, Jaipur, 302031
E-MAIL: nitesh.pradhan943@gmail.com, 21ucs091@lnmiit.ac.in, 21ucs075@lnmiit.ac.in

Abstract:

Skin cancer, the most common cancer worldwide, requires early and accurate detection for effective treatment. Traditional methods like surgical excision are often painful and costly. Deep learning algorithms offer a faster, more affordable alternative for classifying skin cancer, but their performance varies across different models and datasets. In this research, the authors aimed to enhance skin cancer classification accuracy by ensembling the outputs of various Convolutional Neural Networks (CNNs). Fine-tuning was performed on pre-trained models, including EfficientNetB0, EfficientNetB1, ResNet50, and DenseNet121, on the HAM10000 dataset. To address the dataset's class imbalance, data augmentation techniques were employed, such as random rotations, translations, zooming, and flipping. Additionally, Condorcet's Jury Theorem (CJT) was applied to determine the majority voting ensemble score, ensuring that each model's contribution to the ensemble improves overall accuracy. Experimental results demonstrated that the proposed CJT-based ensemble method achieved an accuracy of 97.37%, sensitivity of 97.10%, and specificity of 99.51%, outperforming state-of-the-art individual models. The use of diverse CNN architectures and ensemble strategies, combined with effective data augmentation, significantly improved skin cancer classification performance.

Keywords:

Condorcet Jury Theorem; Ensemble Learning; HAM10000; Convolutional Neural Network; Skin Cancer

1. Introduction

Skin cancer is the most prevalent type of cancer worldwide. Skin cancer represents 33% of every cancer diagnosed in the United States [1]. Skin cancers are divided into two broad categories: 1) Melanoma skin cancer; 2) non-melanoma skin cancer (NMSC). Most skin cancers arise in epidermal cells. Ultraviolet (UV) rays are one of the most important causes of skin cancer [2]. In the past 30 years, the incidence of cutaneous melanoma has increased five times. NMSC is becoming an increasingly serious issue in health-care services. It is estimated that there are approximately 600,000 cases of NMSC each year. The incidence of NMSC is 18-20 times higher than melanoma [3].

Skin cancer treatment depends on the type and stage of the disease, the and place of tumor and personal medical history. Typically, the goal of treatment is to fully eliminate or kill the cancer. Most skin cancers can be cured if detected and treated early [4].The average annual cost for skin cancer annually mounts to $8.1 billion each year [5]. There are many treatments for NMSC and most appropriate of them is excision. Local treatment of melanoma is also radical excision [6]. While excision is the most common method, but it is very uncomfortable and painful process, also the cost of such tests is arbitrarily high.

Deep learning algorithms offer a quick, easy, and affordable way to diagnose the symptoms of skin cancer. Deep learning algorithms have been widely employed for skin cancer classification in recent years, as they do not require domain expertise or feature extraction [7]. These algorithms can efficiently categorize skin cancer and produce risk scores by recognizing patterns and characteristics suggestive of cancer. Major contributions in this research involve: -

1. Authors have performed fine tuning on the pre-trained models according to the HAM10000 dataset and trained the pre-trained models for HAM10000 dataset.

2. Authors have implemented Condorcet's Jury Theorem to determine the majority voting ensemble score based on individual classifier scores [8].

3. The authors handled class imbalance issues by augment-

ing data using various augmentation techniques to biased output towards certain classes.

The remaining paper can be outlined as follows: Section 2 discusses previous state of the art in skin cancer classification. Section 3 discusses the material and methods used in the proposed method. Furthermore, section 4 shows the details of experiment and results. At the end, work is concluded with comparative study and future possibilities around it in section 5 and 6 respectively.

2. Related Work

Recent advancements in deep learning have significantly improved skin cancer classification, particularly using datasets like HAM10000 and ISIC2019. Several notable methods have been proposed, employing different strategies to enhance accuracy and robustness. The following works highlight these key contributions.

One of the novel method was designed by Yang et al. It uses transformer network structure to capture the disease area on the dermoscopy image and reduce the interference from healthy area and noise on HAM10000 dataset. The transformer model achieved the accuracy of 94.1% [9]. Another notable method was proposed by Ahmad et al. They used Dual Attention Mechanism with novel loss function which combined complementary and cross entropy. It also used training samples from non-labelled classes to train a model robust to incorrect class information and handle unbalanced class distributions. It was able to achieve accuracy of 93.86% on HAM10000 dataset and 94.24% on ISIC2019 dataset [10]. In 2023, Wu et al. proposed a wavelet down-sampling feature reconstruction method-based convolution neural network to extract features for classification. They also employed data augmentation and hair removal algorithm to pre-process the data in HAM10000 dataset. Model was able to achieve accuracy of 95.84% and F1-Score of 95.96%.[11].

In 2021, Datta et al. proposed a method based on soft attention mechanism along with baseline InceptionResnet version 2 (IRv2). They were able to achieve accuracy of of 90.4%, sensitivity of 91.6% and specificity of 71.1% [16]. In year 2022, authors performed multiclass skin cancer classification using series of EfficientNets. They employed EfficientNets B0-B7 on HAM10000 dataset to investigate performance of different models. They found B4 and B5 as top performing models with accuracy of 87.91% [19]. An ensemble method was proposed by Popsecu et al. where the authors used collective intelligence of multiple neural networks by taking weighted decision fusion from multiple models leveraging each model's property and performance to classify. The authors were able to achieve the accuracy of 86.71% [20].

These works highlight various strategies such as transformers, attention mechanisms, ensemble methods, and CNN-based approaches to address challenges like class imbalance, noise interference, and model robustness, forming the basis for further improvements in skin cancer classification.

3 Methodology

3.1 Data Preprocessing

The dataset is structured in a CSV manner whereby images and class information are contained. The images are first gotten from the CSV and then turned into a caption of 32 x 32 x 3 (RGB) to suit CNNs which requires 3 channels for Red, Green and Blue Images. Since the data is imbalance as shown in figure 1, it is important to note that the classifier will not function efficiently for all lesions of the skin. To alleviate this problem, authors employed upsampling, containing the process of synthetic expansion of the underrepresented classes to the level of the overrepresented class. Authors used Image Data Generator to apply different transformations like Random Rotating, Translating, Zooming, and Horizontal Flipping. This helped prepare the data by improving the small classes of the dataset through the generation of more varied samples for the small classes of training making it more conducive for training.

To avoid any inconveniences with CNN's input requirements, all images are resized to 32 x 32 pixels. This kind of resizing helps to train the models with different input size combinations like the EfficientNet, DenseNet, and ResNet on a similar set of pictures. Image Resizing decreases memory usage as well, which is very important when the datasets size is huge, or the models have to be implemented on devices with limited resources.

3.2 Model Architecture

Author's approach involves numerous recent CNN designs, which have all been trained on ImageNet, to leverage transfer learning. The models are modified by incorporating classification layers to customize their designs to the unique characteristics of the task, i.e., skin lesions. The models include EfficientNetB0, EfficientNetB1, ResNet50, and DenseNet121.

The authors chose categorical cross-entropy loss as the objective function to train each of the model. Categorical cross-entropy is ideal for multi-class classification tasks, where the goal is to minimize the difference between the predicted prob-

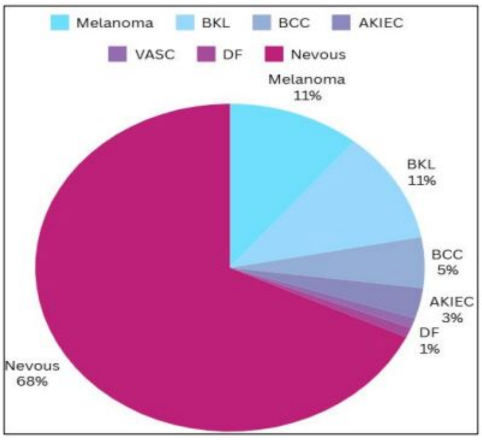

FIGURE 1. HAM-10k Dataset

ability distribution and the true labels. The loss function is defined as shown in equation 1.

$$L = -\sum_{j=1}^{7} y_j log(p(y = j|x)) \quad (1)$$

where y_j is the true class label and p (y = jx) is the predicted probability for class j.

For optimization, authors used the Adam optimizer with an initial learning rate of 1 x 10^{-4}. Additionally, ReduceLROnPlateau was applied to dynamically reduce the learning rate when the validation accuracy plateaued.

The Adam Optimizer id defined as shown in equation 2 and 3.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (2)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) {g_t}^2 \quad (3)$$

Where g_t is the gradient, m_t are the first moment estimates, and v_t are the second moment estimates.

3.3 Condodcet Jury Theorem

The premise of the theorem resides in the belief which states that if every individual in a group is able to make a right decision with a probability higher than random chance (more than 50% accuracy), it follows that the probability that the majority verdict in the group is right improves with the learning, the Condorcet Jury Theorem (CJT) seeks to offer a rationale for ensemble methods where a number of different models (or "jurors") work on one decision.

Let's assume a group of N jurors (or models) each have a probability p of making a correct decision independently. According to CJT, if:

- Each juror makes decisions independently.
- Each juror is more likely to be correct than incorrect $p > 0.5$.
- The group uses majority voting to make a collective decision.

Then, as the number of jurors N rises, the likelihood that the group's majority vote is accurate rises as well. Formally, the probability P_N that the majority vote is correct is given by equation 4.

$$P_N = \sum_{k=N/2}^{N} NkP^k(1 - P)^{N-k} \quad (4)$$

where Nk is the binomial coefficient, representing the number of ways to get exactly k correct votes out of N total votes, and [N/2] represents the minimum number of correct votes required for a majority.
Figure 2 shows the model architecture, at first the imbalanced dataset is augmented and upsampled to create a balanced dataset, which is then used to train EfficientNetB0, EfficientNetB1, DenseNet, and ResNet50 models. Then the individual model's scores are ensembled using the CJT method followed by a SoftMax layer for final classification.

4 Experiment Results

4.1 About Dataset

Authors have used the HAM10000 dataset. The HAM10000 ("Human Against Machine with 10000 images") dataset is a huge collection of dermatoscopic images used to train machine learning models for computer-aided diagnosis of pigmented skin lesions. This is a prominent class imbalance dataset. It includes 10,015 photos from various populations and major diagnostic classes: 1) Actinic keratoses and intraepithelial carcinoma/Bowen's disease (akiec). 2) Basal cell carcinoma (bcc) 3) benign keratosis-like lesions (solar lentigines, seborrhoeic keratoses, and lichen-planus-like keratoses (bkl)), 4) Dermatofibroma (df). 5) Melanoma (mel), 6) Melanocytic Nevi (nv), and 7) Vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc) [12]. Sample images of each category can be seen in figure 3.

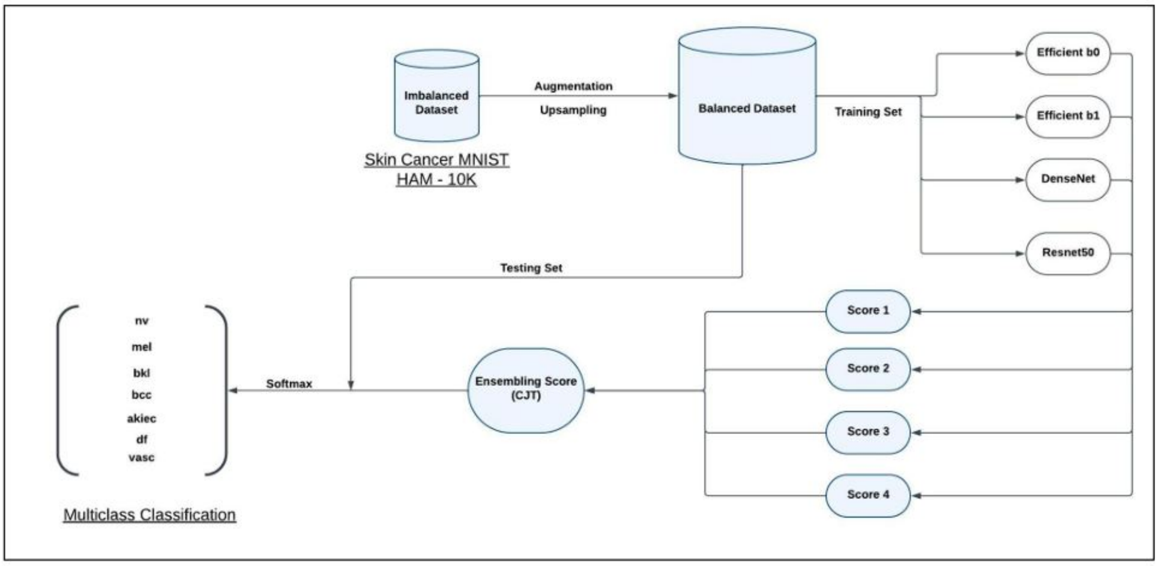

FIGURE 2. Model Architecture

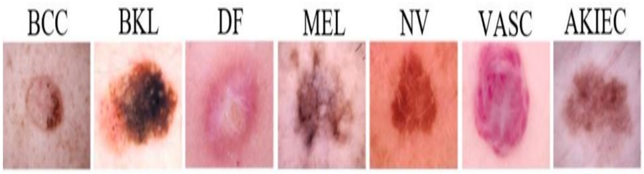

FIGURE 3. Dataset Classes

4.2 Experiments

A sparse categorical cross-entropy loss, the Adam optimizer and learning rate reduction callbacks based on validation accuracy to dynamically adjust learning rates are used to train each model independently. Training takes place for a period of 25 epochs selecting 128 batch size. After each epoch, the model is evaluated on the test set. The experiments held has been implemented in Python making use of TensorFlow framework. The preliminary experiments and the fine-tuning of the model were carried out on Google Colab with the help of a Tesla T4 GPU with a 16 GB GPU RAM. Each set of models in the ensemble was fit for 100 epochs and a batch size of 32 was employed, using the Adam optimizer to train the networks. Also, the Colab T4 GPU shortened the image processing and model convergence times because the training and validation processes were much quicker. The accuracy, sensitivity, and specificity are calculated to measure the model's performance. These metrics are defined as shown in equation 5, 6 and 7.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

4.3 Results

The performance of each individual CNN model (EfficientNetB0, EfficientNetB1, ResNet50, DenseNet121) and the CJT-based ensemble method was evaluated on the HAM10000 dataset. The loss and accuracy curves for each model, shown in figure 5, demonstrate the learning behavior during training and validation.

The confusion matrices for each individual model shown in figure 6 provide insight into their classification performance across different classes. Notably, the confusion matrix for the CJT ensemble figure 4 shows a significant reduction in misclassifications, highlighting the effectiveness of the ensemble approach.

The evolution of the classification accuracy achieved by single models and proposed ensemble methods on the HAM10000

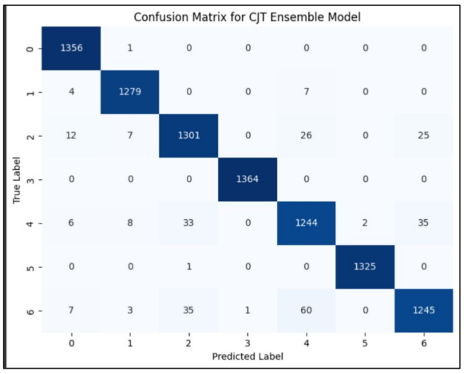

FIGURE 4. Confusion Matrix For CJT Ensembling

dataset is presented in the Table 1. Among individual models, the highest accuracy of 96.39% was demonstrated by EfficientNetB1, followed closely by EfficientNetB0 (96.11%) and DenseNet121 (94.93%). The worst performance was shown by ResNet50, which scored an accuracy of only 89.18%, thus illustrating its weakness for this type of multi-class classification problem.The Weighted Ensemble method, which mixes in all the models proportionally to how well they performed on the validation set, reports an accuracy of 97.01%. The CJT-Inspired Ensemble, which is based on the Condorcet Jury Theorem principle of most simple majority voting, increased it even further to 97.37%.

TABLE 1. Different Classifier Accuracy

Different Classifiers	*Accuracy*
EfficientNetB0	0.9611
Dense Net	0.9493
EfficientNetB1	0.9639
ResNet50	0.8918
Weighted Ensemble	0.9701
CJT-Inspired Ensemble	0.9737

5 Comparative Study

Skin cancer classification has been the subject of many studies, which have used several datasets that are already in the public domain. It is worth noting, however, that researchers have executed and claimed the effectiveness of their models on different datasets. Some of them have even included several different datasets in assessing their approaches. This, however, results in differences in evaluation measures and distributions of datasets. Hence, it becomes difficult to compare our model with existing research work due to different experimental designs, data preprocessing used and test class distributions differences.In this regard, authors restricted ourselves to studies, which either analyzed skin lesions classification using the HAM10000 dataset. The proposed deep learning based on Condorcet Jury Theorem outperformed all other models in terms of accuracy and resilience to problems like data imbalance and interference of other classes. Besides the accuracy improvement witnessed in our approach when compared to the best model and ensembles specified in other papers. Table 2 shows the caoparative study with various other researchers.

6. Conclusions

This research describes the efforts to develop an ensemble model for skin cancer detection using the HAM10000 data set based on the Condorcet Jury Theorem. By combining several high-performing architectures, namely EfficientNetB1, EfficientNetV2S, ResNet50 and DenseNet121, authors managed to enhance the classification results and the stability of the model. The approach taken demonstrated the possible advantages of ensemble learning in complicated imaging issues, particularly in improving classification of different skin lesions which were class-imbalanced. By means of upsampling methods as well as careful processing of the raw data, which had the effect of carrying out considerable data augmentation, the problem of imbalance was tackled effectively, thus easing the variances of the model performance across all the classes. The test results indicated that individual models were inferior to the proposed ensemble paradigm, thus offering attractive prospects for clinical use in diagnosing skin cancers. One caveat of this study is the input images size which was set to a relatively low 32x32 mainly for ease of computation. It is probable that the use of larger sized images for example, images of the different skin lesions could provide more information in that the finer details corresponding to the various skin lesions may be captured by the models.

References

[1] Jerant AF, Johnson JT, Sheridan CD, Caffrey TJ. Early detection and treatment of skin cancer. Am Fam Physician. 2000 Jul 15;62(2):357-68, 375-6, 381-2. PMID: 10929700.

[2] Lacy, K. and Alwan, W., 2013. Skin cancer. Medicine, 41(7), pp.402-405.

[3] Reichrath, J., Leiter, U., Eigentler, T. and Garbe, C., 2014. Epidemiology of skin cancer. Sunlight, vitamin D and skin cancer, pp.120-140.

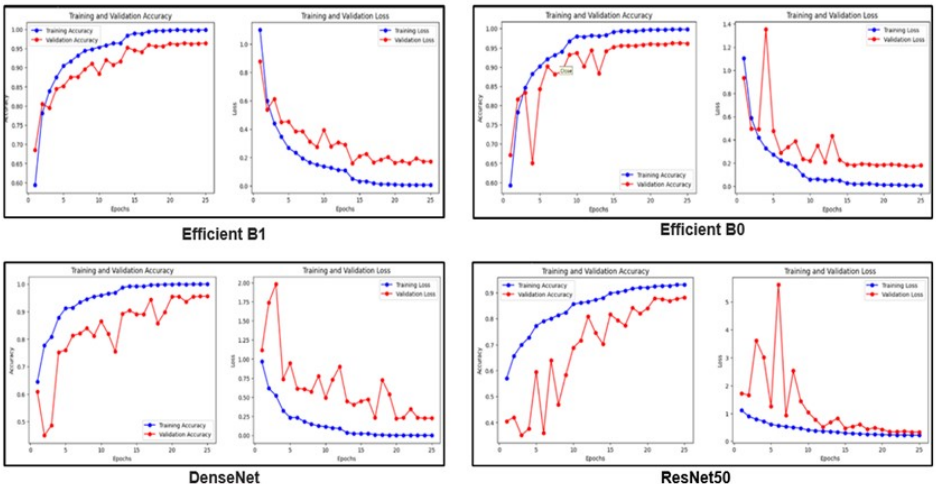

FIGURE 5. Accuracy And Loss Curves Of Different Models

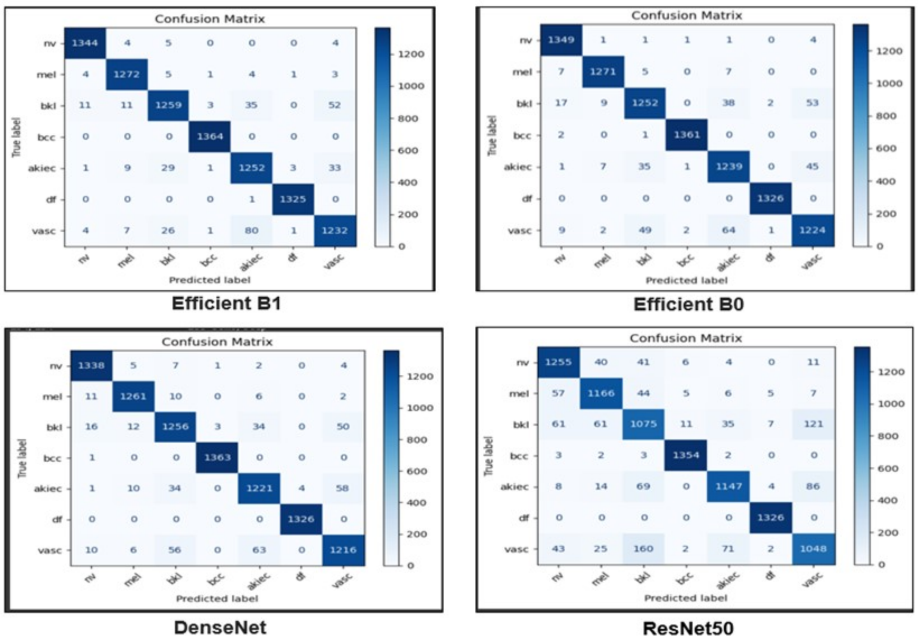

FIGURE 6. Confusion Matrix of Different Classifiers

[4] Ahmed, K., Jesmin, T. and Rahman, M.Z., 2013. Early prevention and detection of skin cancer risk using data mining. International Journal of Computer Applications, 62(4), pp.1-6.

[5] Guy Jr, G.P., Machlin, S.R., Ekwueme, D.U. and Yabroff, K.R., 2015. Prevalence and costs of skin cancer treatment in the US, 2002 2006 and 2007 2011. American journal of preventive medicine, 48(2), pp.183-187.

[6] Orthaber, K., Pristovnik, M., Skok, K., Perić, B. and Maver, U., 2017. Skin cancer and its treatment: novel treatment approaches with emphasis on nanotechnology. Journal of nanomaterials, 2017(1), p.2606271

TABLE 2. Comparative Analysis of Existing Techniques and Methods/Models Proposed by Various Researchers

Authors	*Method Used*	*Year*	*Accuracy (%)*	*Sensitivity(%)*	*Specificity(%)*
Ye, Y.; Zhou, H [13]	CNN + OVA (one-versus-all)	2020	92.90	-	-
Polat, K.; Koc, K.O. [14]	MobileNet+LSTM	2021	91.86	88.24	92.0
Ahmad, B., Usama, M [10]	Dual Attention-Guided Compact Bilinear CNN (DACBNet)	2024	92.86	-	-
Srinivasu, P.N.; SivaSai, J.G [15]	Modified MobileNetV2	2022	85.34	91.09	92.66
Datta SK, Shaikh MA [16]	Inception ResNet with Soft Attention	2021	93.4	91.6	83.3
Yang, G., Luo, S [9]	ViT for skin cancer detection-Large	2022	94.1	-	-
Owida, Hamza Abu [17]	Custom CNN	2024	95.23	95.91	95.3
Ali, K., Shaikh, Z.A [19]	EfficientNetB4	2024	87.91	-	88
Azeem, Muhammad, et al. [18]	SkinLesNet	2023	92	-	-
	Proposed (CJT Model)	2025	97.37	97.10	99.51

[7] Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R. and Zhao, S., 2022. Skin cancer classification with deep learning: a systematic review. Frontiers in Oncology, 12, p.893972.

[8] Srivastava, G., Pradhan, N. and Saini, Y., 2022. Ensemble of Deep Neural Networks based on Condorcet's Jury Theorem for screening Covid-19 and Pneumonia from radiograph images. Computers in Biology and Medicine, 149, p.105979.

[9] Yang, G., Luo, S. and Greer, P., 2023. A novel vision transformer model for skin cancer classification. Neural Processing Letters, 55(7), pp.9335-9351

[10] Ahmad, B., Usama, M., Ahmad, T., Saeed, A., Khatoon, S. and Chen, M., 2024. DACB-Net: Dual Attention Guided Compact Bilinear Convolution Neural Network for Skin Disease Classification. arXiv preprint arXiv:2407.03439.

[11] Wu, Q.E., Yu, Y. and Zhang, X., 2023. A skin cancer classification method based on discrete wavelet down-sampling feature reconstruction. Electronics, 12(9), p.2103.

[12] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M. and Kittler, H., 2019. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic).

[13] Ye, Y.; Zhou, H.; Yu, H.; Hu, R.; Zhang, G.; Hu, J.; He, T. An Improved EfficientNetV2 Model Based on Visual Attention Mechanism: Application to Identification of Cassava Disease. Comput. Intell. Neurosci. 2022, 2022, 1569911.

[14] Polat, K.; Koc, K.O. Detection of Skin Diseases from Dermoscopy Image Using the combination of Convolutional Neural Network and One-versus-All. J. Artif. Intell. Syst. 2020, 2, 80–97.

[15] Srinivasu, P.N.; SivaSai, J.G.; Ijaz, M.F.; Bhoi, A.K.; Kim, W.; Kang, J.J. Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM. Sensors 2021, 21, 2852.

[16] Datta SK, Shaikh MA, Srihari SN (2021) Soft Attention Improves Skin Cancer Classification Performance. InInterpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data. Springer, Cham, pp 13–23.

[17] Owida, Hamza Abu, et al. "Improved deep learning architecture for skin cancer classification." Indonesian Journal of Electrical Engineering and Computer Science 36.1 (2024): 501-501.

[18] Azeem, Muhammad, et al. "SkinLesNet: Classification of Skin Lesions and Detection of Melanoma Cancer Using a Novel Multi-Layer Deep Convolutional Neural Network." Cancers 16.1 (2023): 108.

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

[20] Popescu, D., El-Khatib, M. and Ichim, L., 2022. Skin lesion classification using collective intelligence of multiple neural networks. Sensors, 22(12), p.4399.