UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

MASTER DEGREE

Research and Development

MASTER 1 INTERNSHIP REPORT

By

PHAM GIA PHUC

M23.ICT.010

Information and Communication Technology

Title:

SKIN CANCER CLASSIFICATION USING DEEP LEARNING

Supervisor – Dr. TRAN Giang Son

Laboratory – USTH ICTLab

Hanoi, September 2024

ACKNOWLEDGEMENTS

My supervisor, Dr. TRAN Giang Son, has been nothing short of genuine and supportive, and I want to convey my sincere gratitude for that. I found encouragement from Dr. TRAN Giang Son as I struggled to complete this study. He is the ideal role model and the embodiment of leadership. Without his direction from the first stage of the research, I would not have been able to complete this report and have a grasp of the topic. He set up some incredible experiences for me, and I'm grateful for the chances he gave me to advance professionally. Learning from Dr. TRAN Giang Son is an honour.

I am appreciative of my parents' unwavering love and encouragement, which keep me inspired and self-assured. They believed in me, and as a result, I achieved success. My sincere gratitude goes out to my siblings for helping to keep me grounded, reminding me of what's essential in life, and encouraging me on all of my experiences. I will always be grateful for the unwavering love and support I received daily and throughout the report-writing process.

Contents

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BN	Batch Normalization
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
Conv	Convolution
CPU	Central Processing Unit
DWConv	Depthwise Convolution
FC	Fully Connected
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
ICT	Information and Communication Technology
ISIC	International Skin Imaging Collaboration
ISIC	International Skin Imaging Collaboration
Max	Maximum
MBConv	Mobile inverted Bottleneck Convolution
MFLOPS	Mega Floating-point Operations per Second
Min	Minimum
Pool	Pooling
RAM	Random Access Memory
ReLU	Rectified Linear Unit
RGB	Red Green Blue
TN	True Negative
TP	True Positive
USTH	University of Science and Technology of Hanoi
UV	Ultraviolet

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Skin is the largest organ in the body, serves as an important barrier since it protects internal organs and muscles. As part of the integumentary system, skin contains multiple layers of ectodermal tissue that safeguard the underlying structures.

Skin cancer is one of the most common cancer in the world, and early detection can help improve survival rates. However, diagnosis can vary significantly depending on the clinician's expertise; hence, the development of advanced and reliable diagnostic tools using robust algorithms, like those presented in deep learning methods, is necessary.

This study presents a state-of-the-art approach for automatic skin cancer classification using deep learning techniques. By using the advance of transfer learning with pre-trained convolutional neural networks, together with ensemble methods, this approach is expected to enhance diagnostic accuracy in dermatological image analysis, and provide a reliable automatic tool for skin cancer early identification.

Key words: *Skin cancer, early detection, deep learning, transfer learning, ensemble, skin cancer classification, diagnostic accuracy, reliable, automatic tool.*

CHAPTER 1 – INTRODUCTION

1.1. Context and Motivation

Skin cancer is one of the most common types of cancer and is characterized by uncontrolled reproduction of skin cells that form malignant tumours. It is usually manifested in melanoma and non-melanoma types of skin cancers, which further include basal cell carcinoma and squamous cell carcinoma. Although less common, melanoma is particularly feared because of its aggressiveness and high potential for metastasis; non-melanoma skin cancers are generally less dangerous but can cause tremendous morbidity and disfigurement if not treated early [1].

This disease is brought about by numerous factors, which include increased exposure to ultraviolet radiation from the sun, aging of the global population, and extensive use of tanning beds. While the superiority of this sickness keeps growing, so too does the vital for early and reliable diagnosis.

Figure 1. Some skin cancer symptoms [1]

Skin cancers have to be early diagnosed because the results will bring much better results to the patients [2]. For example, melanomas can be treated so much earlier, hence saving a great number of lives and avoiding the reduction of the survival rate when the cancers have already travelled to other parts of the body [3]. In cases of non-melanoma skin cancers, it is easy to avoid heavy tissue destruction and further invasive surgery [4]. With all the importance of early detection, though, many cases are still diagnosed only in

[1] *https://www.skincancer.org/skin-cancer-information/skin-cancer-pictures/*

advanced stages, where treatment alternatives become greatly limited and the prognosis is not as good. This underlines the urgent need for better diagnostic techniques.

The traditional methods of diagnosis of skin cancer involve a dermatologist undertaking a physical examination [5], where the skin is observed for some lesions that would raise suspicion. A biopsy would be carried out on the lesion suspected of being concerning to rule out malignancy. Generally, these kinds of measures are effective; however, they entail various limitations. However, visual examination strongly depends on the dermatologist's experience, and even then, the most experienced clinicians may misinterpret benign as malignant or vice versa. This can be even more challenging for atypical lesions or in situations where patients have multiple lesions needing evaluation. While biopsies are entirely diagnostic, they are invasive procedures and carry some degree of discomfort for the patients, so there is also a risk of scarring. Moreover, the whole process-from initial examination to biopsy and final diagnosis-may take time, which can be used by the cancer to further progress.

In those situations where challenge is experienced, there is also the issue of accessibility. In many parts of the world, specifically in low-resource areas, the lack of trained dermatologists may lead to late diagnosis and treatment [6], hence increasing all the risks related to skin cancer. That means more dermatological treatments, and the rising incidence of skin cancers is putting greater demands on healthcare systems [6]. It is in this context that there is an increasing need for diagnostic technology not only to be accurate and reliable but also scalable and accessible to a large population.

Such challenges would, therefore, require the development of superior diagnostic technology. It should support clinicians in the early detection of skin cancer while minimizing the need for invasive procedures, such as biopsies, and enhance diagnostic confidence. It is even desirable that such tools be further developed in a user-friendly manner so that their use could easily be carried out in primary care settings or even by patients themselves as part of periodic pores and skin checks. This can seriously reduce

the burden of skin cancer by making early diagnosis more available and much more accurate, improving outcomes, and saving lives.

1.2. Objective

This study's primary objective is the development of a deep learning model for skin cancer classification that utilizes transfer learning techniques with pre-trained models. The implemented model aims to assist healthcare professionals in accurately diagnosing various skin lesions, providing a reliable and automated tool for the early detection of skin cancer.

1.3. Report organisation

As the first part of this report is the introduction of the study, the rest is structured as follows:

- **Chapter 2: Materials** describes the materials used in the problems of this study.
- **Chapter 3: Method**: describes the methodology of solving this study's problem.
- **Chapter 4: Evaluation**: shows the experimental results of work in chapter III.
- **Chapter 5: Conclusion**: presents the conclusion and future works of this study.

CHAPTER 2 – MATERIALS

The resources needed to implement the skin lesion classification problem are described in this session.

2.1. Dataset

All data are downloaded from the ISIC 2019 Challenge [2] provided by the International Skin Imaging Collaboration (ISIC). The ISIC 2019 Challenge dataset consists of 25,331 dermoscopic images gathered from the following main sources: HAM10000 [7], BCN_20000 [8], and MSK [9]. Each of these sources contributes distinct characteristics to the dataset: the HAM10000 dataset consists of 12,413 images that are 600 × 450 pixels

[2] https://challenge.isic-archive.com/data/#2019

in size, with lesions centred and cropped; the BCN_20000 dataset includes 10,015 images with a higher resolution of 1024 × 1024 pixels each, uncropped, with higher variability in the location, scale, and orientation of the lesion; and the MSK dataset adds 819 images of different sizes to further enrich the diversity.

A notable feature of the ISIC 2019 dataset is the strongly pronounced class imbalance. The dataset is divided into eight diagnostic categories as shown in Table 1, the most numerous classes correspond to Melanocytic Nevi, comprising 12,875 images and thus over 50% of the total number of shots in this dataset. In contrast, a Dermatofibroma class is represented by only 239 images. Such a problem is crucial for developing robust classification models, because it might result in biased predictions toward more frequent classes.

Label	Number of images	Ratio (%)
Melanocytic Nevi (NV)	12875	50.83
Melanoma (MEL)	4522	17.85
Basal Cell Carcinoma (BCC)	3323	13.12
Benign keratosis-like lesions (MEL)	2624	10.36
Actinic keratoes (AK)	867	3.42
Squamous Cell Carcinoma (SCC)	628	2.25
Vascular lesions (VASC)	253	1.00
Dermatofibroma (DF)	239	0.94

Table 1. Description of ISIC 2019 Challenge Dataset [7] [8] [9]

In addition, images in the dataset, as shown in Figure 2, have diverse lesion characteristics, such as irregular boundaries, varied thickness, and the presence of dark edges or thick hair, which enhances complexity in the classification tasks. Such diversity in the dimensions of the images and the features of the lesions represents real-world

variability in dermatology and positions this dataset to be holistic for advancing skin lesion classification techniques.

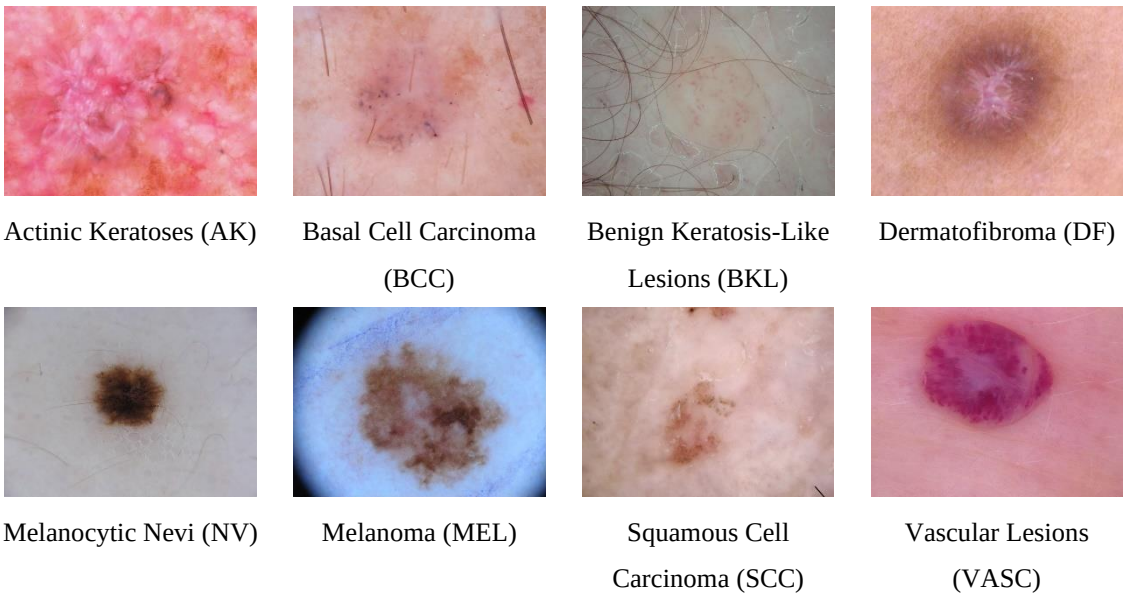

Figure 2. Examples of ISIC 2019 skin lesion images provided by the dataset [7] [8] [9]

2.2. Hardware infrastructure

Since graphics cards play a significant role in quick, effective, and high-performance processing, they must be mentioned while discussing the development of machine learning. A high complexity problem can be solved much more quickly because of the daily improvements in GPU processing speed. The training experiments for this internship were carried out using the high-performance computer infrastructure at USTH ICTLab, and the specifics are as follows:

- CPU: Intel (R) Xeon (R) CPU E5-2620 v3 @ 2.40Ghz
- GPU: Tesla K80
- RAM: 128 GB

CHAPTER 3 – METHODOLOGY

The methodology needed to implement the skin lesion classification problem are described in this session.

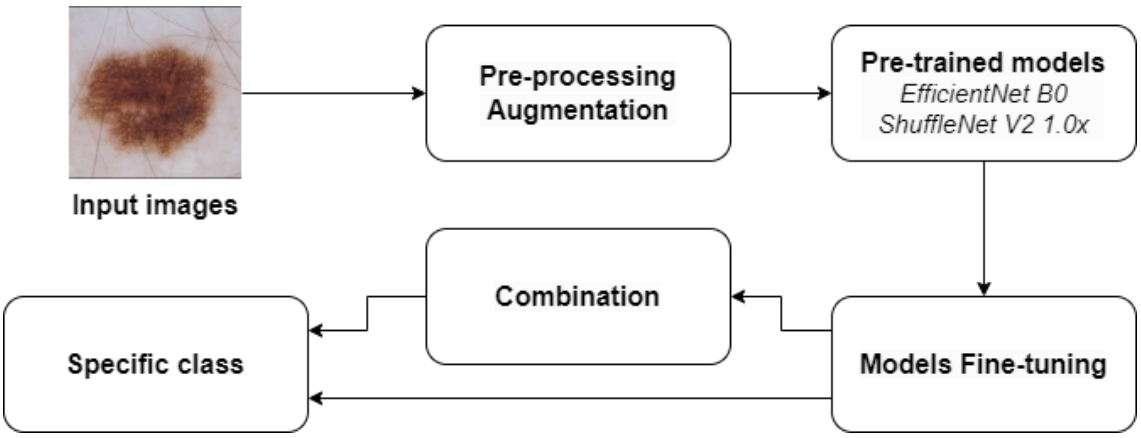

Figure 3. This study's classification system of skin cancer images

3.1. Data preparation

Deep learning generally requires a large and balanced dataset to make sure the models are highly accurate and robust. As per Table 1, the ISIC 2019 Challenge Dataset is highly imbalanced in the distribution of skin lesion classes. Such imbalance could let the model bias to classes that are more prevalent. To avoid this problem, different data augmentation techniques were applied increase the samples of minority classes such that every class becomes equally represented in the training set.

The following data augmentation methods are utilized on the original training set with 17,048 images:

- Resizing images to 224 × 224 pixels.
- Randomly flips images.
- Randomly rotates images from 0° to 180°.
- Randomly crops images within the range of 90% to 110%.
- Randomly changes 20% brightness, contrast and saturation.

After augmentation, the number of images in the training set became 62,368, with all classes have an equal number of samples. Also, it is important to note that both the test and validation sets each contain 2,131 original images from the ISIC 2019 dataset, and do not include any duplicated images from the training set.

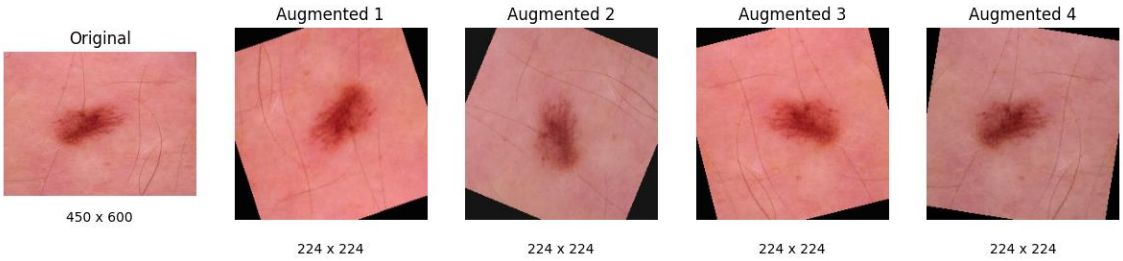

Figure 4. Example of pre-processing

3.2. Transfer Learning

With the rapid development of deep learning, various classification models have been pre-trained on large-scale and diverse datasets, such as CIFAR-100 or ImageNet. These pre-trained models are publicly available with their source codes and weights and, thus, could be reused for several tasks along with their architecture. If a new model reuses anything from these pre-trained models, this new model is called a "transferred model," while such learning is referred to as "transfer learning" [10].

Recently, transfer learning has become an increasingly popular methodology in deep learning classification task [11], since the computation is expensive and enormous data are required for training a new model from scratch. It enables model performance improvement with reduced extensive re-training by borrowing huge knowledge encoded in pre-trained models. The advantage of this approach is that the target task can profit from limited training data through leveraging the feature representations learned during the pre-training phase.

Basically, transfer learning involves two major steps:

- Pre-trained model selection: Select the pre-trained model that best fits the target task. For this study, EfficientNet B0 and ShuffleNet V2 1.0× was selected due to their performance and efficiency in image classification tasks.
- Pre-trained model fine-tuning: Upon further steps, the architecture of the model will be analysed to perform fine-tuning. This means it is necessary to decide which part of the chosen models will be kept, and which part needs retraining to learn new features from the new dataset, in this case, ISIC 2019 Challenge.

With the use of transfer learning, it is able to efficiently adapts existing models to a new application, demonstrating the effectiveness of reusing and refining pre-trained models in order to solve specific challenges and improve practical applications.

3.3. Pre-trained model architecture

3.3.1. Pre-trained EfficientNet B0

EfficientNet was proposed by Google AI scientists in 2019 with a baseline model called EfficientNet B0; it is a family of CNN models designed for high accuracy with efficiency in performing the image classification task [12]. The basic principle behind EfficientNet B0 is the use of compound scaling, which uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients.

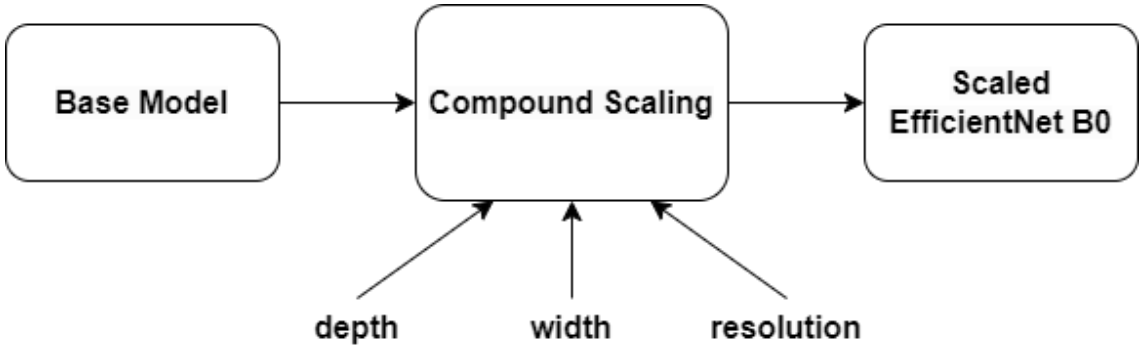

Figure 5. Visualization of scaling factors

The key innovation of EfficientNet B0 is its approach to compound scaling method. Whereas prior methods arbitrary scale network dimensions such as width, depth, or

resolution, EfficientNet B0 scales all dimensions uniformly with a compound coefficient [12], as illustrated in Figure 5, and can be expressed mathematically as:

- $Depth = \alpha^{\varphi}$
- $Width = \beta^{\varphi}$ such that: $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$
- $Resolution = \gamma^{\varphi}$ $\alpha \geq 1, \beta \geq 1, \gamma \geq 1$

Where α, β, γ are constant coefficients determined by a grid search, and φ is the compound coefficient that uniformly scales network.

The baseline architecture of EfficientNet B0 is built upon mobile inverted bottleneck convolution (MBConv), similar to MobileNetV2 [12], but with the addition of squeeze-and-excitation optimization. The network consists of several stages, each comprising multiple identical layers. Table 2 shows the breakdown of the EfficientNet B0 architecture [12]:

Stage	Operator	Output size	Output Channels	Number of Layers
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x4	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

Table 2. EfficientNet B0 baseline network [12]

The network architecture in Table 2 can be conceptually divided into three main parts: entry flow, middle flow, and exit flow. Entry flow (stage 1 and 2) takes the 224x224x3 input images, and processes it through initial convolution and early MBConv blocks. This part is responsible for capturing low-level features and gradually increasing channel depth while reducing spatial dimensions. The middle flow (stage 3 to 8), consists of repeated MBConv blocks with varied channel depths and sometimes with a stride-2 operation for further spatial reduction, is where the network develops its most complex and abstract features. Finally, the exit flow (stage 9) is completed with a convolution, global average pooling, and a fully connected layer for classification.

By using the pre-trained EfficientNet B0 as the base model for the fine-tuning process, this study aims to benefit from its powerful feature extraction capabilities with an efficient architecture in this particular context of the skin lesion classification task.

3.3.2. Pre-trained ShuffleNet V2 1.0×

ShuffleNet V2, proposed by researchers at Megvii in 2018 [13], is a family of lightweight CNN models which provide both high accuracy and efficiency for mobile and embedded vision applications. The "1.0×" in ShuffleNet V2 1.0× refers to the variant of the baseline model with 1.0× output channel in this family, with a width multiplier of 1.0. The channel split and shuffle operations are keys behind ShuffleNet V2 1.0×, which allow for sufficient information flow between channels [13].

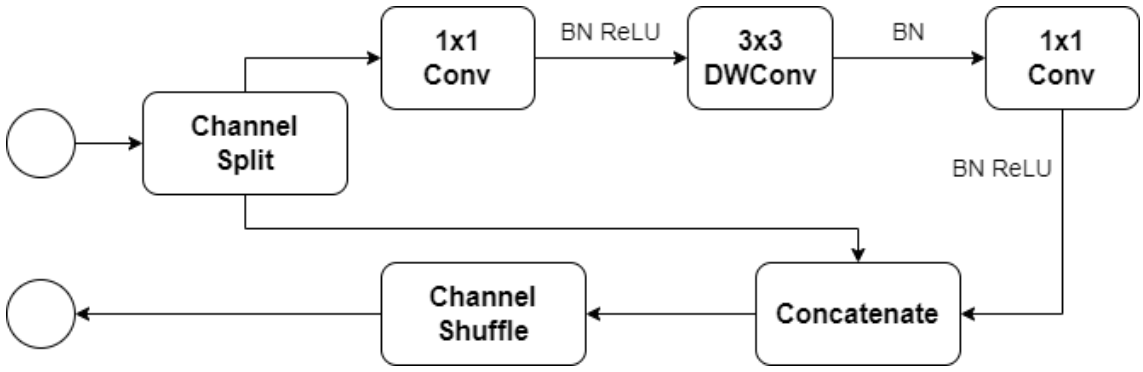

Figure 6. Building block of ShuffleNet V2 1.0×

The basic building block of ShuffleNet V2 1.0× is ShuffleNet V2 unit [13], structured as in Figure 6, where input feature channels are split into two branches: one branch remain unchanged; the other undergoes 1x1 convolution, 3x3 depthwise convolution, and another 1x1 convolution. Afterward, the two branches are concatenated and apply channel shuffle operation to enable information communication. The network consists of several stages, which is presented in Table 3 [13]:

Stage	Operator	Output size	KSize	Stride	Repeat	Output channels
1	Conv1	112 × 112	3 × 3	2	1	24
	MaxPool	56 × 56	3 × 3	2		
2	ShuffleNet V2 Unit	28 × 28		2	1	116
		28 × 28		1	3	
3	ShuffleNet V2 Unit	14 × 14		2	1	232
		14 × 14		1	7	
4	ShuffleNet V2 Unit	7 × 7		2	1	464
		7 × 7		1	3	
5	Conv5	7 × 7	1 × 1	1	1	1024
6	GlobalPool	1 × 1	7 × 7			
7	FC					1000

Table 3. ShuffleNet V2 1.0× baseline network [13]

It is able to conceptually divide the ShuffleNet V2 1.0×'s architecture into three main sections: the entry flow (stage 1) takes the input images, processes it through initial convolution and max pooling to capture low-level features and reduce spatial dimensions; the middle flow (stage 2 to 4) consists of repeated ShuffleNet V2 units with varying channel depths and occasional stride-2 operations for further spatial reduction, which help the network develops its most complex and abstract features; the exit flow (stage 5 to 7)

culminates a final convolution, global average pooling, and a fully connected layer for classification.

In addition to EfficientNet B0, this study also makes use of the pre-trained ShuffleNet V2 1.0× as the base model for fine-tuning, leveraging its compact architecture to efficiently extract features for skin lesion classification.

3.4. Weighted voting ensemble

In many cases, decision-making is benefit by aggregating several expert opinions. This principle extends itself in the domain of deep learning through the methods of ensemble learning.

A weighted ensemble can be considered a more advanced technique that weights models based on their performance on the validation set [14] [15]. For classification tasks, for instance, this study's problem, the final prediction would be made through a weighted vote, where the models with better performance have greater influence [16].

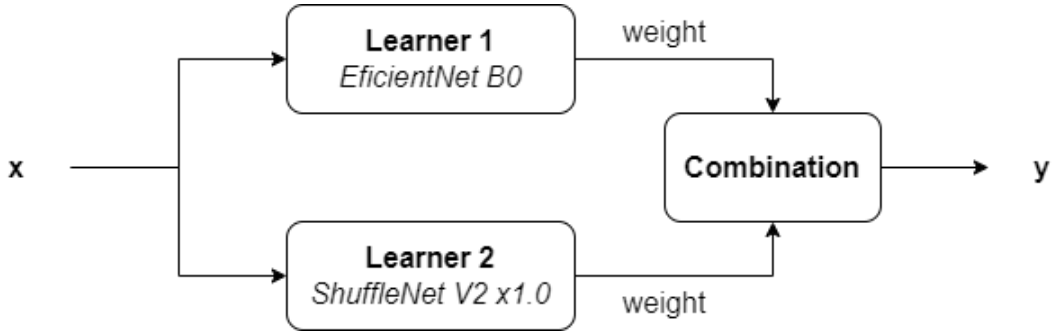

Figure 7. Weighted voting ensemble model architecture

As illustrated in Figure 7, the weighted voting ensemble typically follows these steps:

- Training base models: Individual models (in this study, EfficientNet B0 and ShuffleNet V2 1.0×) are trained independently on the same dataset.
- Assigning weights: These models are evaluated on a validation set, with weights assigned based on their performance metrics.

- Combining prediction: For a new input, all base model predictions are combined using the assigned weights, where the class with the highest weighted sum of votes becomes the final prediction

This approach is expected to leverages the strengths of each individual model while mitigating their weaknesses, especially when different model architecture excels at different aspects of the task.

In this study, the resulting metrics of both fine-tuned EfficientNet B0 and ShuffleNet V2 1.0× are used to calculate the weight given to each model, with higher-performing models receiving larger weight. The final weighted voting ensemble prediction is subsequently made by combining both base models' outputs, with more biased toward the one that has greater weight.

3.5. Metrics

This study employs a multi-dimensional approach to evaluate the deep learning models for the task of multi-classes skin cancer classification, utilizing metrics derived from the confusion matrix alongside cross-entropy loss.

The primary metric employed is cross-entropy loss, which quantifies the disparity between predicted probability distributions and actual target distributions:

$$Loss = -\sum_{i=0}^{N-1} y_i \log(\hat{y}_i)$$

Where N is the number of classes y_i is the true label and $\hat{y}_i$ is the predicted probability for class i.

Central to the evaluation process is the confusion matrix, which, in the context of multi-class classification, is an $N \times N$ matrix as shown in Table 4, with N representing the number of classes ($N = 8$ in this study case).

		Predicted		
Actual	**TP**	FN	FN	...
	FP	TN	TN	...
	FP	TN	TN	...
	...	...	...	...

Table 4. $N \times N$ Confusion matrix

In the case of a multi-class classifier, the number of N is not 2, and there are neither positive nor negative cases. However, the effectiveness of the classifier on each class can be measured. To do this, for each class $\{C_i, i = 1 \ ... \ N\}$, consider a binary classification in the test set: C_i as positive class whereas the rest $\{C_i, j \neq i\}$ as negative class. As such, for each class C_i, it is able to calculate TP_i, FP_i, FN_i and TN_i values based on the prediction output, where TP, FP, FN, and TN refer to true positive, false positive, false negative, and true negative, respectively.

To measure the validity of the models, a set of more statical metrics are used:

- $Accuracy = \frac{\sum_0^{N-1}(TP_i+TN_i)}{\sum_0^{N-1}(TP_i+FP_i+FN_i+TN_i)}$
- $Precision = \frac{\sum_0^{N-1} TP_i}{\sum_0^{N-1}(TP_i+FP_i)}$
- $Sensitivity\ (or\ Recall) = \frac{\sum_0^{N-1} TP_i}{\sum_0^{N-1}(TP_i+FN_i)}$
- $Specificity = \frac{\sum_0^{N-1} TN_i}{\sum_0^{N-1}(FP_i+TN_i)}$
- $F1\ Score = 2 \times \frac{Precision \times Sensitivity}{Precision+Sensitivity}$

These metrics provides insights into the model's overall performance and its effectiveness for individual classes, offering a nuanced understanding of any class-specific weaknesses or biases.

CHAPTER 4 – EVALUATION

This section focuses on assessing the effectiveness and accuracy of the work.

4.1. Model fine-tuning and weighted voting

For this study, with a survey performed, PyTorch was selected as the deep learning framework. It offers a flexible and efficient environment for developing and training deep learning models. Besides, it has plenty of pre-trained models, such as EfficientNet B0 and ShuffleNet V2 1.0×, which accelerated the processes of building the model due to its fast convergence.

In order to effectively classify multiple skin cancer classes, the pre-trained models EfficientNet B0 and ShuffleNet V2 1.0× are fine-tuned with the similar configuration as follow:

- Batch size: 32 for both training and validation.
- Model adaptation: For both model, top layer is removed and replaced with a dense layer for 8-class classification.
- Dropout rate: 0.25 to prevent overfitting.
- Optimizer: Adam, with an initial learning rate of 10^{-3}.
- Activation function: Softmax for the prediction layer.
- Loss function: Categorical cross-entropy.
- Transfer learning approach:
  - EfficientNet B0: Re-train new dense layer.
  - ShuffleNet V2 1.0×: Re-train new dense layer.
- Training duration: 50 epochs for each model.

To optimize the training process, ModelCheckpoint was set to save the best-performing model. Besides that, the learning rate is reduced by a factor of 0.1 after 3 epochs without any improvement.

Afterward, the weighted voting ensemble model combines the two saved fine-tuned models in the previous steps. It then assesses their performance metrics and assigns weights to each model using the following formula:

$$weight_i = \frac{\sum normalized\_metric_i}{\sum k(\sum normalized\_metric_k)}$$

Where i is the index of a specific model, and k refers to the index used in the summation to iterate over all component models.

This leads to the following weight distribution between the two fine-tuned EfficientNet B0 and ShuffleNet V2 1.0× as shown in Table 5:

Fine-tuned model	Weight
EfficientNet B0	0.5176
ShuffleNet V2 1.0×	0.4824

Table 5. Weight distribution between the two fine-tuned models

This slightly higher weight given to EfficientNet B0 denotes that it has a better margin in its performance, and thus has greater impact on the final prediction output of the weighted voting ensemble model.

4.2. Learning performance

Figure 8 and Figure 9 illustrates the learning performance of pre-trained EfficientNet B0 and ShuffleNet V2 1.0× model overtime. The training process involved several steps, including data pre-processing, augmentation, fine-tuning and model training with early stopping and regularization techniques. The results are depicted in line graphs showing training and validation loss and accuracy.

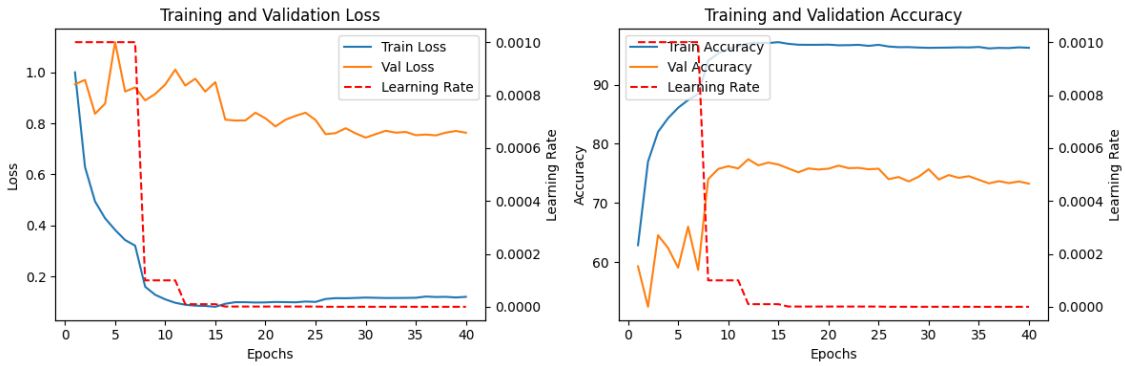

Figure 8. The learning curves of model EfficientNet B0

The learning performance of EfficientNet B0 approach is illustrated in Figure 8. Although initially, both training and validation losses exhibit a rapid decrease, indicating that the model quickly learns to distinguish between different classes of skin lesions. The training loss continues to go down linearly and reaches the minimum value of about 0.1. However, the validation loss drops steeply and reaches a local minimum, then starts increasing gradually. Such divergence signifies overfitting, where the model memorizes patterns in the training data rather than generalizes to unseen examples.

Moreover, accuracy curves have confirmed these observations. While the training accuracy goes up rapid to almost 100% toward the end of training, the validation accuracy increases sharply and then just stays at around 75%, failing to match the continual improvement seen in training accuracy. Along with a slight decline in later epochs, this form further confirmation of overfitting.

On the other hand, ShuffleNet V2 1.0× exhibits a more balanced learning trajectory as shown in the Figure 9:

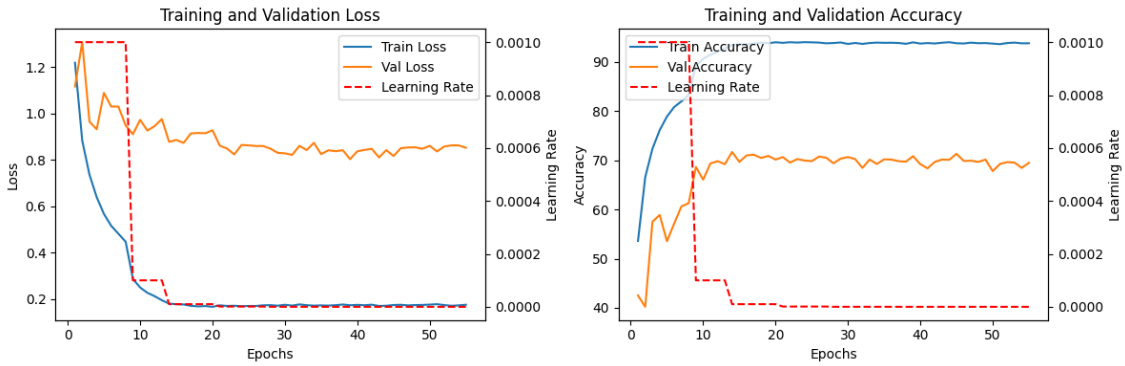

Figure 9. The learning curves of model ShuffleNet V2 1.0×

After an initial rapid decrease, both training and validation losses stabilize, with a constant gap between them. The training losses settles around 0.1, while the validation loss stabilizes higher at 0.8; training accuracy plateaus at around 95%, while the validation accuracy stabilizes at around 70%. Such behavior suggests better generalization compared to EfficientNet B0, albeit with lower peak performance.

One point to note is that, while EfficientNet B0 shows evident signs of severe overfitting, that can be seen from the increasing gap between training and validation metrics, ShuffleNet V2 1.0× also represents a gap; however, it manages to keep the difference quite consistent throughout training, hence balancing learning and generalization better. This phenomenon may be due to various causes, such as:

- Model Complexity: Higher model complexity in EfficientNet B0 likely contributes to its pronounced overfitting, whereas the efficiency-oriented design is more suitable in ShuffleNet V2 1.0× for this dataset.
- Regularization Effectiveness: Possibly, more effective regularization techniques were used for ShuffleNet V2 1.0×, or its architecture provides certain built-in regularization benefits.
- Data Characteristics: Sensitivity to class imbalance and limited diversity of the ISIC 2019 dataset might be high for both models; however, ShuffleNet V2 1.0× seems more robust against mentioned issues.

4.3. Classification results

To further assess the trained models on different skin lesion classes, confusion matrices provide insights into their performances. These class-specific performance indicators offer a more comprehensive view of the models' effectiveness across different skin lesion types.

The two figures 10 and 11 are the confusion matrices of the two models EfficientNet B0 and ShuffleNet V2 1.0×, respectively, utilized in this study:

Figure 10. Confusion matrix of EfficientNet B0

Figure 11. Confusion matrix of ShuffleNet V2 1.0×

The EfficientNet B0's confusion matrix shown in Figure 10 provides a quite fair idea about its performance among the different classes, whereas accuracies range from a low of more than 50% with Squamous Cell Carcinoma to a high of 81% in the Basal Cell Carcinoma class.

Similarly, as demonstrated in Figure 11, ShuffleNet V2 1.0× performs well on Basal Cell Carcinoma with an accuracy of 85%, and struggles with Actinic Keratoses, Dermatofibroma, and Squamous Cell Carcinoma where accuracy is just around 50%.

On the other hand, both models show a high misclassification rate of approximately 20% between the classes Actinic Keratoses, Squamous Cell Carcinoma, and Basal Cell

Carcinoma, which indicates that they face challenges at distinguishing these types of lesions.

The weighted voting ensemble model, combining the strengths of both EffficientNet B0 and ShuffleNet V2 1.0×, with the result shown in the confusion matrix at Figure 12, likely demonstrates improved overall performance.

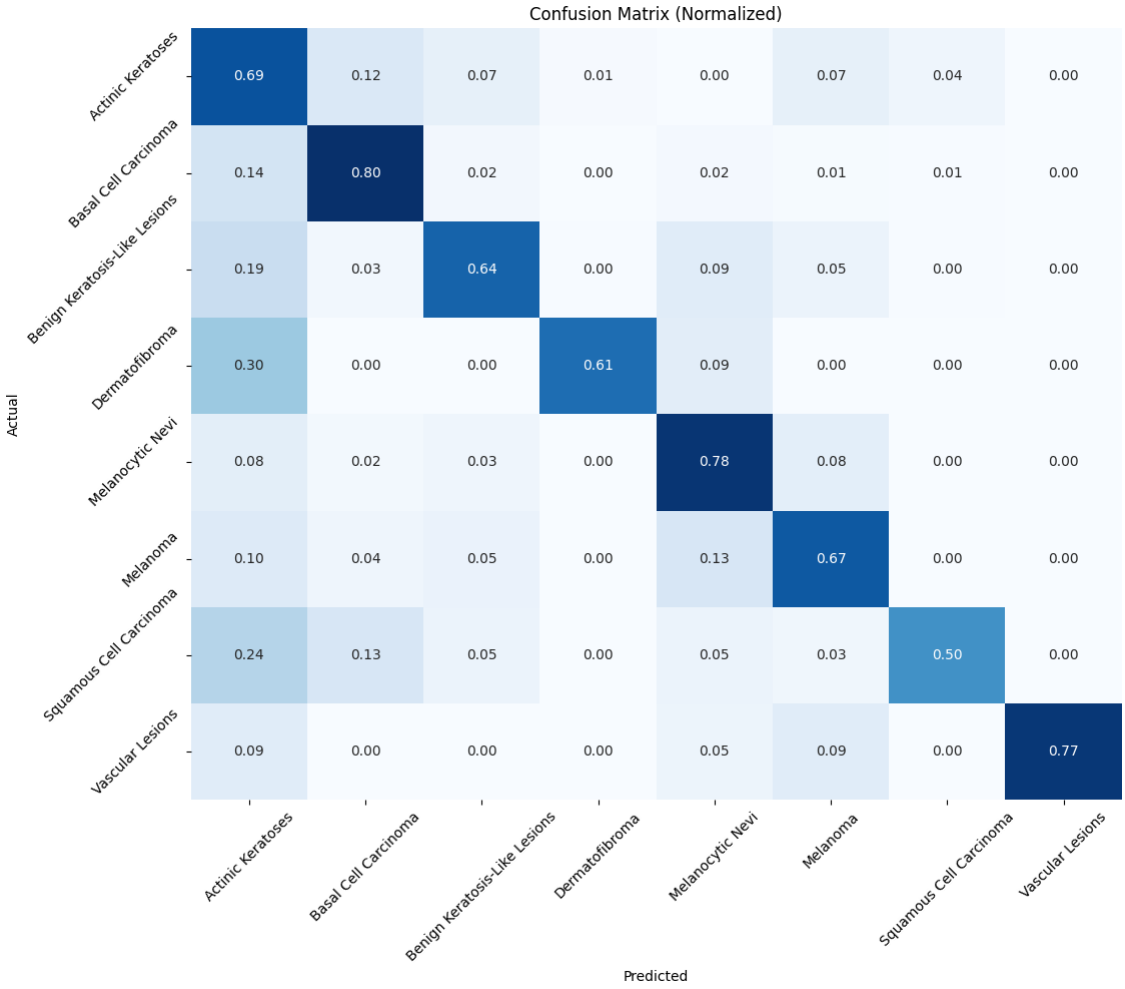

Figure 12. Confusion matrix of Weighted Voting Ensemble model

This model maintains strong performance in recognizing common lesions such as Basal Cell Carcinoma at 80% and Melanocytic Nevi at 78%, whereas improving the accuracy in Vascular Lesions to 77%, Actinic Keratoses to 69% and Dermatofibroma to 61%, surpassing both models. These increments show that indeed, the weighted voting ensemble model is able to pool the performance from its base models.

It worth noting that Squamous Cell Carcinoma remains an accuracy of 50%, meaning neither of the base models captures this class effectively. However, when it comes to multi-class classification problems, such as the one at hand, 50% accuracy in one class is not as bad it sounds.

Otherwise, the weighted voting ensemble model also significantly reduces misclassifications of similar lesion types, with errors such as those between Actinic Keratoses, Squamous Cell Carcinoma, and Basal Cell Carcinoma decreasing from about 20% to as low as 13-14%. This reduction is particularly important given the potential severity of misdiagnosing pre-cancerous and malignant conditions.

4.4. Discussion

This study explores the application of EfficientNet B0 and ShuffleNet V2 1.0× models and the weighted voting ensemble technique for skin lesion classification using the ISIC 2019 dataset. To objectivity and thoroughly evaluate the proposed model, it is essential to compare the results with those produced by existing models.

The Table 6 below summarizes the comparative study:

		Accuracy (%)	**Precision (%)**	**Sensitivity (%)**	**Specificity (%)**
Kassem et al. [17]		81	77	74	84
Gessert et al. [18]		63	–	73	–
Gong et al. [19]	*Decision Fusion*	99.5	–	98.3	99.6
This study	*EfficientNet B0*	75	76	75	95
	ShuffleNet V2 1.0×	72	73	72	95
	Weighted Voting Ensemble	73	80	73	94

Table 6. Comparative study

As can be seen from Table 6, firstly, the model by Kassem et al. [17] had obtained an accuracy of 81%, precision of 77%, sensitivity of 74%, and specificity of 84%; The model by Gessert et al. [18], on the other hand, has 63% accuracy and sensitivity of 73%; And the model by Gong et al. [19] with Decision Fusion method has the highest accuracy of 99.5%, with sensitivity and specificity of 98.3% and 99.6%, respectively.

For comparison, the proposed EfficientNet B0 obtains an accuracy of 75%, precision of 76% and sensitivity of 75%; The proposed ShuffleNet V2 1.0× obtains an accuracy of 72%, precision of 70% and sensitivity of 72%; And the proposed weighted voting ensemble method in this study had managed to obtain an accuracy of 73%, precision of 80%, sensitivity of 73%, and specificity of 94%, respectively.

In particular, this study proposed weighted voting ensemble model outperforms Kassem et al. [17]'s method in precision (80% versus 77%) and specificity (94% versus 84%), indicating its strength in correctly identifying the presence or absence of a particular disease; whereas the individual models EfficientNet B0 and ShuffleNet V2 1.0×, while not surpassing Kassem et al. [17]'s method, still outperform Gessert et al. [18]'s method in accuracy and offers comparable sensitivity. Although not matching the exceptional results of Gong et al. [19], this study proposed methods still show robust performance, particularly in precision and specificity.

In conclusion, most of the proposed models, especially the weighted voting ensemble approach, balance the skin lesion classification by contributing strong performance in correctly identifying positive and negative cases, and maintaining high precision. These results may indicate that there is possibly a clinical utility to the technique where the reduction of false positives is important.

CHAPTER 5 – CONCLUSION

5.1. Conclusion

Briefly, early diagnosis and proper treatment become vital with the rising incidences of skin cancers all over the world. The computer system will have a very significant role to play in supporting clinical diagnoses for enhancing the accuracy and thereby aiding health professionals to make more accurate decisions.

Based on the given scenario, this study proposes skin cancer classification methods using advanced machine learning techniques, such as Transfer Learning and Fine-tuning, with two pre-trained models, EfficientNet B0 and ShuffleNet V2 1.0×. Implemented in Python with the PyTorch framework the models were trained on data from the ISIC Archive, a large and expanding open-source public-access archive of skin images developed by the International Skin Imaging Collaboration (ISIC). The complexity of these dermoscopic images, as well as the limited number of labelled samples posed significant challenges, which were addressed through various pre-processing techniques.

After the processes of training, testing and validating the performance, among the two models EfficientNet B0 and ShuffleNet V2 1.0×, the former was the most accurate and sensitive, with its specificity being comparable to the later. Thus, the weighted voting ensemble model combined the strengths of both to yield the highest precision, hence giving quite a balanced performance on different metrics. However, there is still room for improvement, especially in enhancing accuracy and sensitivity further, imply that better optimization and experimentation could lead to even greater general performances for all models.

5.2. Future work

Due to time constraints, many adaptations, tests, and experiments have been left for future research. In the next work, extra pre-trained model testing will be included to find the best architecture for skin cancer classification. As more data becomes available from different sources, the results are foreseen to improve. A robust application is developed which would be able to assist in early detection and diagnosis of skin cancer, thereby saving lives.

References

[1] Linares, M.A., Zakaria, A. and Nizran, P., "Skin cancer," *Prim Care, 42(4),* pp. pp.645-659, 2015.

[2] Marks, R., "An overview of skin cancers," *Cancer, 75(S2),* pp. pp.607-612, 1995.

[3] Schadendorf, D., Van Akkooi, A.C., Berking, C., Griewank, K.G., Gutzmer, R., Hauschild, A., Stang, A., Roesch, A. and Ugurel, S., "Melanoma," *The Lancet, 392(10151),* pp. pp.971-984, 2018.

[4] Nguyen, T.H. and Ho, D.Q.D, "Nonmelanoma skin cancer," *Current treatment options in oncology, 3,* pp. pp.193-203, 2002.

[5] Ferrante di Ruffano, L., Dinnes, J., Deeks, J.J., Chuchu, N., Bayliss, S.E., Davenport, C., Takwoingi, Y., Godfrey, K., O'Sullivan, C., Matin, R.N. and Tehrani, H., "Optical coherence tomography for diagnosing skin cancer in adults," *Cochrane Database of Systematic Reviews, 2018(12),* 1996.

[6] Seth, D., Cheldize, K., Brown, D. and Freeman, E.E., "Global burden of skin disease: inequities and innovations," *Current dermatology reports, 6,* pp. pp.204-210, 2017.

[7] Tschandl, P., Rosendahl, C. and Kittler, H., "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data, 5(1),* pp. pp.1-9, 2018.

[8] Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S. and Malvehy, J., "Bcn20000: Dermoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288,* 2019.

[9] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, Allan Halpern, "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv:1710.05006,* 2017.

[10] Torrey, L. and Shavlik, J., "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010, pp. pp. 242-264.

[11] Shaha, M. and Pawar, M., "Transfer learning for image classification," in *2018 second international conference on electronics, communication and aerospace technology (ICECA) (pp. 656-660). IEEE,* 2018.

[12] Tan, M., "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946,* 2019.

[13] Ma, N., Zhang, X., Zheng, H.T. and Sun, J., "Shufflenet v2: Practical guidelines for efficient cnn architecture design," *Proceedings of the European conference on computer vision*

(ECCV), pp. pp. 116-131, 2018.

[14] Dogan, A. and Birant, D., "A weighted majority voting ensemble approach for classification," in *2019 4th International Conference on Computer Science and Engineering (UBMK)*, 2019.

[15] Rojarath, A. and Songpan, W., "Probability-weighted voting ensemble learning for classification model," *Journal of Advances in Information Technology Vol, 11(4),* 2020.

[16] Zhang, Y., Zhang, H., Cai, J. and Yang, B., "A weighted voting classifier based on differential evolution," *Abstract and applied analysis,* Vols. 2014, No. 1, no. Hindawi Publishing Corporation, p. p. 376950, 2014.

[17] Kassem, M.A., Hosny, K.M. and Fouad, M.M, "Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning," *IEEE Access, 8,* pp. pp.114822-114832, 2020.

[18] Gessert, N., Nielsen, M., Shaikh, M., Werner, R. and Schlaefer, A., "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX, 7,* p. p.100864, 2020.

[19] Gong, A., Yao, X. and Lin, W., "Dermoscopy image classification based on StyleGANs and decision fusion," *Ieee Access, 8,* pp. pp.70640-70650, 2020.