Article

Multiclass Skin Cancer Classification Using Ensemble of Fine-Tuned Deep Learning Models

Nabeela Kausar [1], Abdul Hameed [1], Mohsin Sattar [2], Ramiza Ashraf [3], Ali Shariq Imran [4,*], Muhammad Zain ul Abidin [1] and Ammara Ali [5]

1   Department of Computing and Technology, Iqra University, Islamabad 44000, Pakistan;
    nabeela.kausar@iqraisb.edu.pk (N.K.); hameed@iqraisb.edu.pk (A.H.);
    zainamjad23354@gmail.com (M.Z.u.A.)
2   MIS Division, Pakistan Institute of Nuclear Science and Technology (PINSTECH), Islamabad 44000, Pakistan;
    mohsin08@gmail.com
3   School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu (UMT),
    Terengganu 21030, Malaysia; ramizaashraf2010@gmail.com
4   The Norwegian Colour and Visual Computing Laboratory (ColorLab), Department of Computer Science,
    Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway
5   Department of Circulation and Medical Imaging, Faculty of Medicine and Health Sciences,
    Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway;
    ammaraa@stud.ntnu.no
*   Correspondence: ali.imran@ntnu.no

Abstract: Skin cancer is a widespread disease associated with eight diagnostic classes. The diagnosis of multiple types of skin cancer is a challenging task for dermatologists due to the similarity of skin cancer classes in phenotype. The average accuracy of multiclass skin cancer diagnosis is 62% to 80%. Therefore, the classification of skin cancer using machine learning can be beneficial in the diagnosis and treatment of the patients. Several researchers developed skin cancer classification models for binary class but could not extend the research to multiclass classification with better performance ratios. We have developed deep learning-based ensemble classification models for multiclass skin cancer classification. Experimental results proved that the individual deep learners perform better for skin cancer classification, but still the development of ensemble is a meaningful approach since it enhances the classification accuracy. Results show that the accuracy of individual learners of ResNet, InceptionV3, DenseNet, InceptionResNetV2, and VGG-19 are 72%, 91%, 91.4%, 91.7% and 91.8%, respectively. The accuracy of proposed majority voting and weighted majority voting ensemble models are 98% and 98.6%, respectively. The accuracy of proposed ensemble models is higher than the individual deep learners and the dermatologists' diagnosis accuracy. The proposed ensemble models are compared with the recently developed skin cancer classification approaches. The results show that the proposed ensemble models outperform recently developed multiclass skin cancer classification models.

Keywords: skin cancer; deep learning; ensemble classifier; multiclass skin cancer; classification model; ensemble models

1. Introduction

Cancer can cause death if not diagnosed and treated in a timely fashion and can start almost anywhere in the human body. Skin cancer is a common type of cancer, as more than three million Americans are diagnosed with skin cancer each year (https://www.skincancer.org/skin-cancer-information/skin-cancer-facts, accessed on: 24 October 2021). If skin cancer is diagnosed early, it can usually be treated. There are eight categories of skin cancers: melanoma (MEL), melanocytic nevi (NV), basal cell carcinoma (BCC), benign keratosis lesions (BKL), actinic keratosis (AK), dermatofibroma (DF), squamous cell carcinoma (SCC), and vascularl Lesions (VASC) [1]. MEL is the most dangerous type

of cancer, as it spreads to other organs very rapidly. It develops in body cells called melanocytes. MEL is not common as compared to other categories of skin cancer. NV are pigmented moles and vary in different colors of skin tones. It mostly develops during childhood and early adult life as the number of moles increases up to the age of 30 to 40. Thereafter, the number of naevi tends to decrease. BCC develops in cells of the skin called basal cells. The basal cell performs the functionality of producing new skin cells as old ones die off. AK is a pre-cancer that develops on skin affected by chronic exposure to ultraviolet (UV) rays. BKL is one of the common benign neoplasms of the skin. DF occur at all ages and in people of every ethnicity. It is not clear if DF is a reactive process or a neoplasm [2]. The lesions are made up of proliferating fibroblasts. Vascular lesions are relatively common abnormalities of the skin and underlying tissues. SCC is the most accruing form of skin cancer after melanoma and usually results from exposure to UV rays.

In literature, machine learning approaches such as support vector machine (SVM) [2], neural networks [3], naïve Bayes classifier [4], and decision trees [5] have been used for skin cancer classification. The problem with machine learning approaches is the requirement of human-engineered features. In the last decade, deep learning approaches, such as convolutional neural networks (CNN) became popular due to their ability with regard to automatic feature extraction [6–9], and have been extensively used in research [10–13]. Dorj et al. [14] worked on skin cancer classification using deep CNN. Romero et al. in [15] performed melanoma cancer classification with the dermoscopy images using CNN. The technique has an accuracy of 81.3% on the International Skin Imaging Cancer (ISIC) archive dataset. Jinnai et al. [16] carried out pigmented skin lesion classification using the clinical images and faster region-based CNN. The classification accuracy of the method was compared with the ten-board certified dermatologist diagnosis accuracy. Esteva et al. [11] performed multiclass skin cancer classification using dermoscopy images with the different variants of CNN. Adegun et al. in [17] developed a probabilistic model to achieve the better performance of a fully convolutional network-based deep learning system for the analysis and segmentation of skin lesion images. The probabilistic model achieved an accuracy of 98%.

Recently, researchers have proposed ensemble methods to enhance classification performance [18–20]. Bajwa et al. in [21] developed ensemble model using ResNet-152 [22], DenseNet-161 [22], SE-ResNeXt-101 [23], and NASNet [23] for the classification of seven classes of skin cancer using the ISIC dataset and achieved an accuracy of 93%. The ensemble is a machine learning method that combines the decision of several individual learners to increase classification accuracy [24]. The ensemble model exploits the diversity of individual models to make a combined decision; therefore, it is expected that the ensemble model increases classification accuracy [25,26]. The binary class skin cancer classification has been performed in [15,27–29], but many researchers could not address multiclass classification with better results. The recent approaches developed in [11,19,30–32] for multiclass skin cancer classification also failed to achieve higher accuracy. In this research, improved performance heterogeneous ensemble models are developed for multiclass skin cancer classification using majority voting and weighted majority voting. The ensemble models are developed using diverse types of learners with various properties to capture the morphological, structural, and textural variations present in the skin cancer images for better classification. The proposed ensemble methods perform better than both the individual deep learning models and deep learning-based ensemble models proposed in the literature for multiclass skin cancer classification.

The following contributions are made in this research work:

- Five pre-trained models are developed, and their decision is combined using majority weighting and weighted majority voting for the classification of eight different classes of skin cancer.
- The pre-trained models with different structural properties are trained to capture the morphological, structural, and textural variations present in the skin cancer images with the following idea: residual learning, extraction of more complex features, improvement in the declined accuracy caused by the vanishing gradient, feature

invariance through the residual learning, and extraction of the fine detail present into the image.
- The proposed ensemble methods perform better than the expert dermatologists and previously proposed deep learning-based ensemble models for multiclass skin cancer classification.
- A comparative study is conducted for the performance analysis of five fine-tuned deep learning models and their ensemble models on the ISIC dataset to determine the model with better performance.
- In our proposed method, no extensive pre-processing has been performed on the images, and no lesion segmentation has been carried out to make the work more generic and reliable.

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 describes the proposed method. Ensemble methods are discussed in Section 4, whereas Section 5 discusses different deep neural network models followed by individual models in Section 6. The quality measures used to measure the performance of the proposed study are presented in Section 7. Section 8 discusses the results, followed by the conclusion.

2. Related Work

Skin cancer is usually diagnosed with the physical examination of the skin or with the help of biopsy. The detection of skin cancer through the physical examination requires a great degree of experience and expertise, and biopsy-based examination is a tedious and time consuming task as it also requires expert pathologists. Currently, the macroscopic and dermoscopy images are used by the dermatologist during the detection procedure of skin cancer. But even with the dermoscopy images, accurate skin cancer detection is a challenging task, as multiple skin cancers may appear similar in initial appearance. Furthermore, even the expert dermatologists have limited studies and exposure experience to different types of skin cancer through their lifetimes. Expert dermatologists have a skin cancer detection accuracy of 62% to 80% [33,34]. The reports on diagnostic accuracy of clinical dermatologists have shown 62% accuracy with the clinical experience of three to five years. However, the dermatologists with more than 10 years of experience have a diagnostic accuracy of 80%, and diagnostic performance falls even more for dermatologists with the less experience [34]. Therefore, dermoscopy in the hands of an inexperienced dermatologist may reduce diagnostic accuracy [33,35,36].

In early 1980s computer-aided diagnosis (CAD) systems were developed to assist the dermatologist to meet the challenges faced during the process of diagnosis [37]. Initially, CADs were developed using a dermoscopy image for binary classification of melanoma and benign [37]. Since then, much research has been carried out to solve this challenging problem. Several studies [37–39] have been performed using the manual evaluation methods developed using the ABCD method proposed by Nachbar et al. in [40]. Moreover, machine learning classifiers have been developed with handcrafted features. These classifiers include support vector machine (SVM) [2], naïve Bayes [4], k-nearest neighbor [41], logistic regression [42], and artificial neural networks [3]. But the existence of high intra-class and low interclass variations in MEL have caused the unsatisfactory performance of handcrafted features-based cancer detection [43].

With the advent of deep learning, CNN caused a breakthrough for the solution of many problems, including skin cancer classification. CNN gave higher detection accuracy and it also reduced the burden of hand-crafted feature extraction by automatically extracting the feature [44]. As CNNs require huge datasets for better feature leaning at abstract levels [45], transfer learning has been introduced to meet the limitation of huge dataset requirements [20,32], where a model trained for a task is reused for another task. Khalid et al. proposed a skin lesion detection technique by developing a transfer learning-based AlexNet model for the classification of three skin lesions in [46]. To develop the proposed method, data augmentation is applied to enhance the dataset to achieve an accuracy of 98.61%. Kawahara et al. developed a skin cancer detection technique using dataset

consisting of 1300 images to construct a linear classifier using the features extracted by the CNN in [47]. The technique does not require preprocessing or skin lesion segmentation. They carried out classifications of five and 10 classes and achieved an accuracy of 85.8% and 81.9%, respectively. In [32], a novel CNN architecture has been developed. The architecture relies on multiple tracts to perform the skin lesion classification. The authors used a pretrained CNN for single resolution and retained it for multi-resolution on publicly available datasets and obtained an accuracy of 79.15% for the ten classes. In [48], Ali et al. developed a skin lesion classification approach using deep convolutional neural networks (DCNN) to classify benign and malignant skin lesions. To develop the proposed method, the authors carried out the preprocessing consisting of noise removal by applying the filter, input normalization, and data augmentation steps and achieved an accuracy of 91.93%.

Ensemble methods are developed to enhance classification performance. Ensemble methods exploit the diversity in individual models to obtain their higher accuracy. Recently, multiclass skin cancer classification techniques have been developed in the literature using ensemble approaches. Harangi et al. [18] proposed how an ensemble of CNNs models can be developed for enhancement of skin cancer classification accuracy and developed an ensemble model for three classes of skin cancer and achieved an accuracy of 84.2%, 84.8%, 82.8%, and 81.4% for the models of GoogleNet, AlexNet, ResNet, and VGGNet, respectively. The authors enhanced the accuracy of 83.8% with the ensemble model of GoogleNet, AlexNet, and VGGNet. In [20], Nyiri and Kiss developed different ensemble methods using CNNs. To develop the proposed technique, the authors performed the preprocessing on ISIC2017 and ISIC2018 datasets using different preprocessing methods and got an accuracy of 93.8%. In [49] Shahin et al. carried out skin lesion classification using ensemble of deep learners and developed an ensemble by aggregating the decision of ResNet50 and Inception V3 models to carry out the classification of seven skin cancer classes with an accuracy of 89.9%. In [19], Majtner et al. developed the ensemble of VGG16 and GoogleNet architectures using the ISIC 2018 dataset. To develop the proposed ensemble methods, the authors carried out the data augmentation and colour normalization on the dataset. The proposed method achieved an accuracy of 80.1%. [50] Rahman et al. developed a multiclass skin cancer classification approach using a weighted averaging ensemble of deep learning approaches using ResNeXt, SeResNeXt, ResNet, Xception, and DenseNet as individual models to develop the ensemble for the classification of seven classes of skin cancer with an accuracy of 81.8%.

Previous work for skin cancer classification based on dermoscopy images not only lacks the generality but also has lower accuracy for multiclass classification [11,19,32]. In this paper, we propose a multiclass skin cancer classification using diverse types of learners with various properties to capture the morphological, structural, and textural variations present in the skin cancer images for better classification. The proposed ensemble models perform better than both the individual deep learning models and deep learning-based ensemble models proposed in the literature for multiclass skin cancer classification.

3. Proposed Methodology

The proposed work is performed in two stages. In the first stage, we have developed five diverse deep learning-based models of ResNet, Inception V3, DenseNet, InceptionResNet V2, and VGG-19 using transfer learning with the ISIC 2019 dataset. The selection of five pre-trained models with different structural properties is made to capture the morphological, structural, and textural variations present in the skin cancer images with the following idea: residual learning, extraction of more complex features, improvement in the declined accuracy caused by the vanishing gradient, feature invariance through the residual learning, and extraction of the fine detail present into the image. At the second stage, two ensemble models have been developed. For ensemble model development, the decisions of deep learners have been combined using majority voting and weighted majority voting to classify the eight different categories of skin cancer. Figures 1 and 2 shows the overall block diagram of the proposed system.
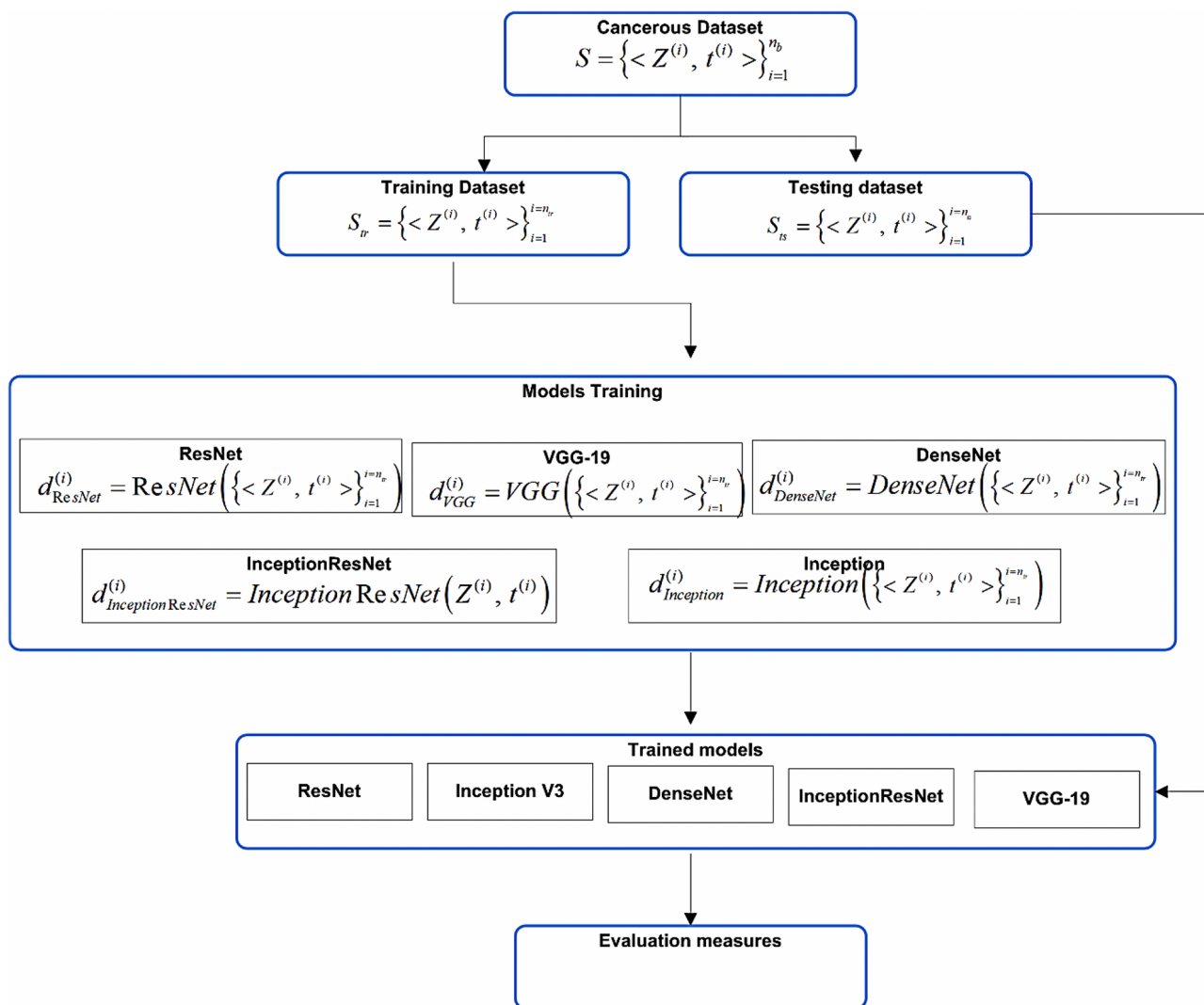
Figure 1. Block diagram of individual models.

ISIC developed an international repository of dermoscopy images known as the ISIC Archive (https://www.isic-archive.com, accessed on 24 October 2021) for technical research. The ISIC 2019 repository contains a training dataset consisting of 25,331 dermoscopy images across eight different categories. Details of dataset and the distribution of data samples for each class have been shown in Table 1. It is observed from Table 1 that distribution of data samples across different classes varies. For example, the melanocytic nevi(NV) class consists of 12,875 images. Similarly, the melanoma class consists of 4522 images, and basal cell carcinoma(BCC) consists of 3323 images. To prepare the dataset for the development of the proposed ensemble models, 1500 images have been randomly selected from each of the NV, BCC, Melanoma, and BKL classes. From the rest of the four classes, all available images in the ISIC repository have been added into the dataset. Thus, the dataset has been formed with 7487 images. Then it has been splitted into two parts: training and test dataset. The training dataset consists of 5690 images and the test dataset has been formed by taking 25% of the total dataset. Thus, the test dataset consists of 1797 images. Figure 3 shows the sample images of eight different classes of skin cancer. In the proposed approach, images have been resized to $224 \times 224 \times 3$.
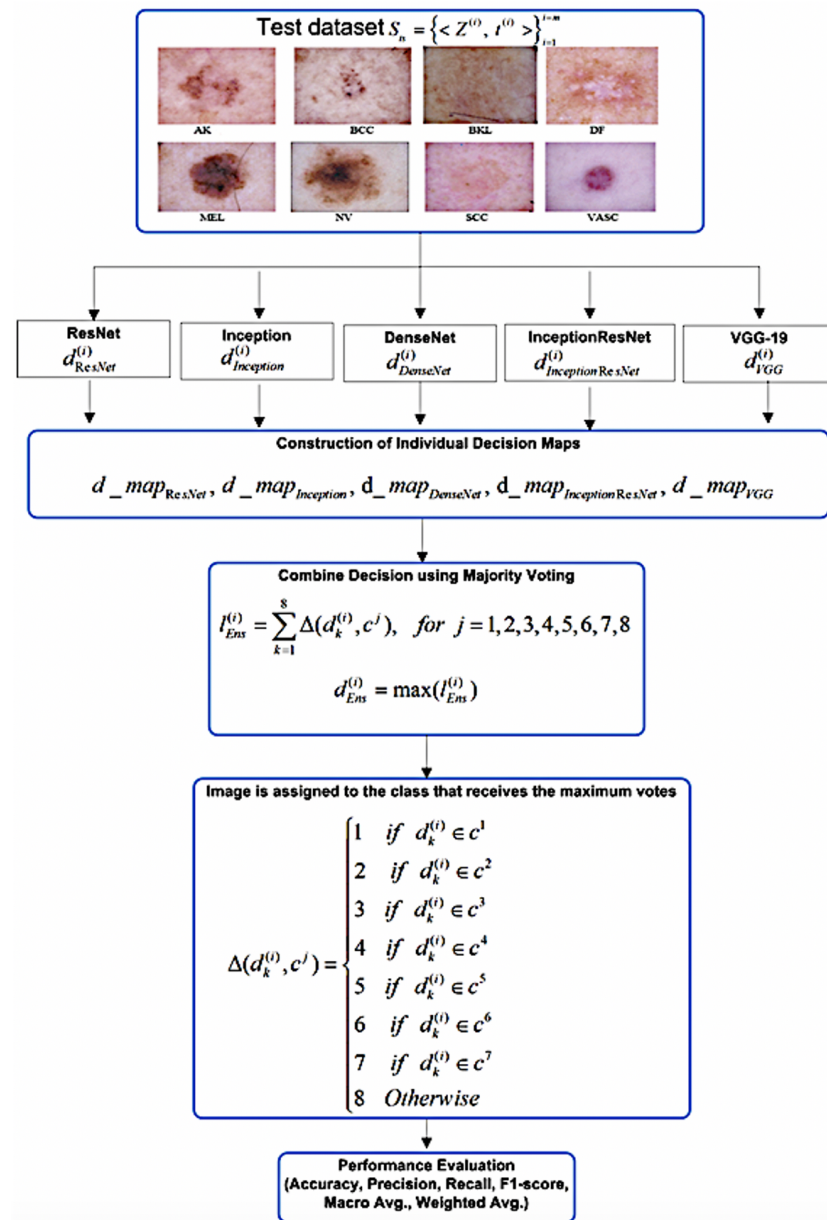
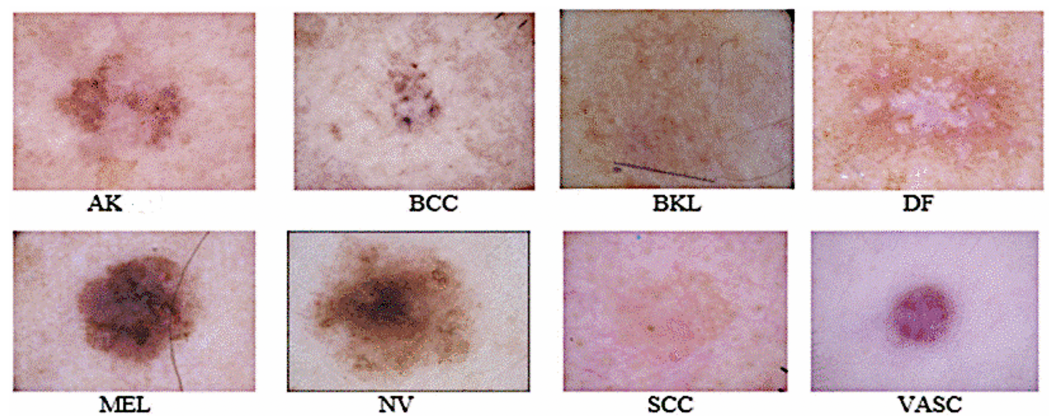**Figure 2.** Block diagram of ensemble model.

Figure 3. Sample images of eight skin diseases from the ISIC-2019 dataset.

Table 1. Detail of distribution of images across different classes in ISIC 2019 training dataset.

Class Label	Abbreviation	Class	Number of Images
1	AK	Actinic keratosis	867
2	BCC	Basal Cell Carcinoma	3323
3	BKL	Benign keratosis	2624
4	DF	Dermatofibroma	239
5	MEL	Melanoma	4522
6	NV	Melanocytic Nevi	12,875
7	SCC	Squamous cell carcinoma	628
8	VASC	Vascular Lesions	253
	Total		25,331

4. Ensemble Methods

The motivation behind the development of ensemble models with diverse leaner is to deal with the complexity of multiclass problem by utilizing the pattern extraction capabilities of CNNs and improving the generalization of multiclass problems with the help of ensemble systems. In the machine learning model, as the number of classes increase, the complexity of the model increases, resulting in a decrease in accuracy. Ensemble methods combine the results of individual learners to enhance accuracy by exploiting their diversity and improving the generalization of the learning system. Machine Learning models are bounded by their hypothetical spaces due to some bias and variance. Ensemble techniques aggregate the decision of individual learners to overcome the limitation of a single learner that may have a limited capacity to capture the distribution (variance) of data. Therefore, making a decision by aggregating the multiple diverse learners may improve the robustness as well as reduce the bias and variance. Ensemble learning employs various techniques to generate a robust and accurate combined model by aggregating the base learners. The combining strategies may consist of voting, averaging, cascading or stacking. Voting strategies consist of majority voting and weighted majority voting whereas, averaging strategy consists of averaging and weighted averaging. In this work, we have developed an ensemble model using majority voting, weighted majority voting, and weighted averaging strategies. The basis of ensemble learning is diversity. The ensemble model may fail to achieve better performance if there is no diversity in base learners [51]. There are various methods to incorporate the diversity, such as (1) the exploitation of features spaces, (2) taking subsamples from the training dataset, and (3) the selection of diverse base/individual learners [52]. In the proposed ensemble systems, the diversity is incorporated by combing the individual learners with various properties to capture the morphological, structural, and textural variations present in the skin cancer images for better classification.

5. Deep Neural Network Models

To develop the proposed ensemble models, five deep neural network models, namely ResNet, InceptionV3, DenseNet, ResNetInceptionV2, and VGG-19, have been developed by fine-tuning the model parameters. Short details of the models are described below.

5.1. ResNet

In this deep neural network model, residual learning is introduced and was chosen as a component model of the ensemble. It constructs a deep network with a large number of layers that keep learning residuals to match the predicted labels with the actual labels. The essential components of the model are the convolution and pooling layers that are fully connected and stacked one over the other. The identity connection among the layers of the residual network differentiates between the normal network and the residual network. The residual block of the ResNet is shown in Figure 4. To skip one or more layers in the ResNet,

it introduces the "skip connection" and "identity shortcut connection" in the model. The residual block $F(X)$ of the ResNet model can be represented mathematically by Equation (1).

$$Y = F(X, \{W_i\}) + X \tag{1}$$

where X and Y are the input and output, respectively, and F is the function applied on the input given to the residual block.
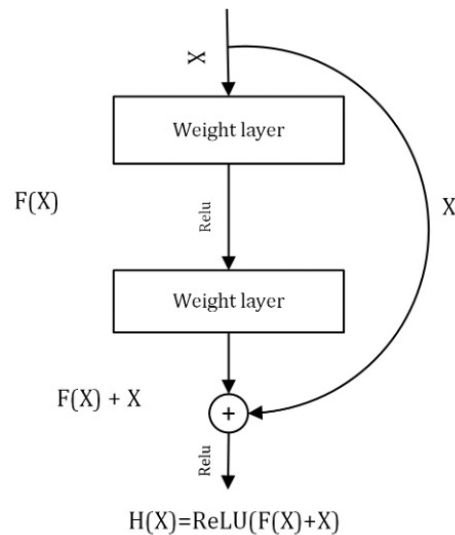

Figure 4. Residual block of deep Residual Network.

5.2. Inception V3

The motivation for choosing the inception neural network as a component model of the ensemble is the inception module that consists of 1×1 filters followed by the convolutional layers of different sizes. Due to this, the inception neural network is able to extract more complex features. Inception V3 is inspired by GoogleNet. It is the third edition of Google Inception built from symmetric and asymmetric blocks, including convolution, average pooling, max pooling, dropouts, and fully connected layers. The batch normalization is used extensively throughout the model architecture.

5.3. DenseNet

DenseNet is chosen as a component model of the ensemble due to improvement in the declined accuracy caused by the vanishing gradient. In neural networks, the information may vanish before it reaches the last layer due to the longer path between the input and output layers. In the DenseNet model, every layer receives additional information from the preceding layers and then passes its feature maps to all subsequent layers. Concatenation of information is performed in the model and each layer gets a "collective knowledge" from all preceding layers.

5.4. ResNetInception V2

This is a variant of the Inception V3 model developed on the basis of the main idea taken from the ResNet model. It has simplified the ResNet block, which facilitates the development of the deeper network. The study in [53] shows that the residual connections play an essential role in accelerating the training of the inception network.

5.5. VGG-19

VGG-19 was developed by the Visual Geometry Group, and the number 19 stands for the number of layers with trainable weights. It is a simple network, as the model is made up of sixteen convolutional layers and three fully connected layers. VGG uses very small

size filters (2×2 and 3×3). It uses max pooling for downsampling. The VGG-19 model has approximately 143 million parameters learned from the ImageNet dataset.

6. Development of Individual Models

The brief architecture of the five deep learning models is given in Section 5. In this section, the training and fine-tuning details of the individual models are provided. First, the training dataset $S_{tr} = \left\{ < Z^{(i)}, t^{(i)} > \right\}_{i=1}^{i=n}$ has been used to train and optimize the parameters of the individual models, where Z represents the image. The training dataset consists of $n = 5690$ images. It is used to develop the classifiers of ResNet, InceptionV3, ResNetInceptionV2, DenseNet, and VGG-19. To train and fine-tune the ResNet model, global average pooling (GlobalAvgPool2D) is applied to downsample the feature maps so that all the spatial regions may contribute to the output. Moreover, a fully connected layer containing eight neurons with the SoftMax activation function are added to classify eight different classes. The ResNet model is trained with 50 epochs, adaptive moment estimation (Adam) optimizer for the quick optimization of the model, learning rate of 1e-4, and categorical cross-entropy loss function. Inception V3 is fine-tuned by applying GlobalAvgPool2D to downsample the feature maps, adding two dense layers at the end containing 1028 and eight neurons with a rectified linear unit (ReLU) and SoftMax activation functions, respectively. The model is trained using 50 epochs, a learning rate of 0.001, and an RMSprop optimizer, as it uses plain momentum. Additionally, RMSprop maintains a moving average of the gradients and uses that average to estimate the variance.

DenseNet is fine-tuned by adding a fully connected layer containing eight neurons with SoftMax activation function to classify the eight classes of skin cancer. It is trained using 50 epochs, an Adam Optimizer, and a learning rate of 1e-4. InceptionResNetV2 is fine-tuned by adding two dense layers containing 512 and eight neurons with ReLU and SoftMax activation functions, respectively. GlobalAvgPool2D pooling is applied to downsample the feature map. Moreover, the model is trained with 50 epochs, a stochastic gradient descent (SGD) optimizer, and a learning rate of 0.001 with a batch size of 25. VGG-19 is fine-tuned by applying GlobalAvgPool2D to downsample the feature maps and adding two dense layers containing 512 and eight neurons with ReLU and SoftMax activation functions, respectively. The model is trained with 50 epochs, a learning rate of 1e-4, an SGD optimizer, and a categorical cross-entropy loss function. After retraining and fine-tuning individual models, the test dataset $S_{ts} = \left\{ < Z^{(i)}, t^{(i)} > \right\}_{i=1}^{i=m}$, ($m = 1797$) is used to validate the trained component models.

Development of Ensemble Models for Skin Cancer Classification

In this stage, individual models trained using different parameters are combined using different combination rules. The details of different combination rules can be found in [54]. Many empirical studies show that simple combination rules, such as majority voting and weighted majority voting, show remarkably improved performance. These rules are effective for the construction of ensemble decisions based on class labels. Therefore, for the current multiclass classification, majority voting, weighted-majority voting, and weighted averaging rules are applied to combine the decision of individual models. For the weighted averaging ensemble, the same weights are assigned to every single model. The final softmax-based results from all the learners are averaged by $\frac{\sum p_i}{N}$, where N is number of learners. For weighted-majority voting weights of each model can be set proportional to the classification accuracy of each learner on the training/test dataset [55]. Therefore, for the weighted majority-based ensemble, weights are empirically estimated for each learner ($W_{ResNet}, W_{Inception}, W_{DenseNet}, W_{InceptionResNet}, W_{VGG}$) with respect to their average accuracy on the test dataset. The obtained weights $W_k; k = 1, \dots, 5$ are normalized so that they add up to 1. This normalization process will not affect the decision of the weighted majoring-based ensemble.

The ensemble decision map is constructed by stacking the decision values of the individual learners for each image Z in the test dataset, i.e., $d_{ResNet}^{(i)} = ResNet(Z^{(i)})$, $d_{Inception}^{(i)} = Inception(Z^{(i)})$, $d_{DenseNet}^{(i)} = DenseNet(Z^{(i)})$, $d_{InceptionResNet}^{(i)} = InceptionResNet(Z^{(i)})$ and $d_{VGG}^{(i)} = VGG(Z^{(i)})$. The ensemble decision values are obtained for two well-known ensemble methods of majority voting and weighted majority voting. For each image the vote given to the jth class is computed using indicator function $\Delta(d_k^{(i)}, c^j)$; which matches the predicted value of the kth individual model with the corresponding class label as in Equation (2).

$$\Delta(d_k^{(i)}, c^j) = \begin{cases} 1 & if\ d_k^{(i)} \epsilon c^1 \\ 2 & d_k^{(i)} \epsilon c^2 \\ 3 & d_k^{(i)} \epsilon c^3 \\ 4 & d_k^{(i)} \epsilon c^4 \\ 5 & d_k^{(i)} \epsilon c^5 \\ 6 & d_k^{(i)} \epsilon c^6 \\ 7 & d_k^{(i)} \epsilon c^7 \\ 8 & Otherwise \end{cases} \tag{2}$$

The total votes $votes_j(i)$ received from individual models for jth class are obtained using majority voting as in Equation (3).

$$votes_j^{(i)} = \sum_{k=1}^{8} \Delta(d_k^{(i)}, c_j^{(i)}), for\ j = 1\ to\ 8 \tag{3}$$

However, with the weighted majority voting rule the votes for jth class are obtained for the learners $k = 1$ to 5 as in Equation (4).

$$votes''_j^{(i)} = \sum_{k=1}^{5} k = d_k(i) = c_j(i)\ w_k\ , for\ j = 1\ to\ 5 \tag{4}$$

The ensemble decision class values, $\acute{l}_{Ens}^{(i)} \epsilon \acute{M}_{Ens}$ and $l''_{Ens}^{(i)} \epsilon M''_{Ens}$ are obtained using majority voting and weighted majority voting rules as in Equations (5) and (6).

$$\acute{l}_{Ens}^{(i)} = max(votes_j^{(j)}) \tag{5}$$

$$l''_{Ens}^{(i)} = max(votes''_j^{(i)}) \tag{6}$$

The image is assigned to the class that receives the maximum votes.

7. Performance Measures

The classification performance of the five deep learners and proposed ensemble models has been evaluated using the following quality measures.

7.1. Accuracy

Accuracy is a performance measure that indicates the overall performance of classifier as the number of correct predictions divided by the total number of predictions. It shows the ability of the learning models to correctly classify the images data samples. It is computed as in Equation (7).

$$\frac{TP + FP}{TP + FP + TN + FN} \tag{7}$$

where *TP* is true positive, *FP* is false positive, *TN* is true negative, and *FN* is false negative.

7.2. Precision

Precision is a performance measure that shows how accurately a classification model predicts the same result when a single sample is tested repeatedly. It evaluates the ability of the classifier to predict the positive class data samples. It is calculated as in Equation (8).

$$\frac{TP}{(TP + FP)} \tag{8}$$

7.3. Recall

Recall is a classification measure that shows how many truly relevant results are returned. It reflects the ratio of all positive class data samples predicted as positive by the learner. It is calculated as in Equation (9).

$$\frac{TP}{(TP + FN)} \tag{9}$$

7.4. F1 Score

F1 score is calculated based on precision and recall. It can be considered as the weighted average of precision and recall. Its value range between [0, 1]. The best value of F1 score is 1 and the worst is 0. It is computed as in Equation (10).

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

8. Results and Discussion

The performance of the proposed models have been evaluated using the measures of accuracy, precision, recall, f1-score, and support. Tables 2 and 3 show the comparative classification performance of individual deep learners of ResNet, InceptionV3, DenseNet, InceptionResnetV2, VGG-19, and the proposed ensemble model. It is observed from the table that the ensemble model outperforms the individual models in terms of precision, recall, f1 score, and accuracy. The accuracy of individual learners of ResNet, InceptionV3, DenseNet, InceptionResnetV2, VGG-19 is 92%, 72%, 92%, 91%, and 91%, respectively.

However, the accuracy measures of the majority voting, weighted averaging based ensemble, and weighted majority voting-based ensemble models are 98%, 98.2%, and 98.6%, respectively. Figure 5 shows that the accuracy of the ensemble approach is much higher than the individual models.

Figure 5. Accuracy comparison of individual learners and their ensemble decision.

Table 2. Performance comparison of individual learners with the ensemble approach.

Classification Report of ResNet				
Skin Cancer Type	**Precision**	**Recall**	**F1-Score**	**Support**
AK	0.85	0.95	0.89	261
BCC	0.91	0.94	0.93	292
BKL	0.95	0.92	0.93	306
DF	0.95	0.84	0.89	63
MEL	0.90	0.90	0.90	325
NV	0.93	0.94	0.93	305
SCC	0.96	0.82	0.88	173
VASC	0.93	0.97	0.95	72
Accuracy			0.92	1797
Macro Avg.	0.92	0.91	0.91	1797
Weighted Avg.	0.92	0.92	0.92	1797
Classification Report of DenseNet				
Skin Cancer Type	**Precision**	**Recall**	**F1-Score**	**Support**
AK	0.95	0.88	0.91	261
BCC	0.89	0.94	0.91	292
BKL	0.92	0.95	0.93	306
DF	0.95	0.89	0.92	63
MEL	0.91	0.90	0.90	325
NV	0.92	0.95	0.94	305
SCC	0.89	0.86	0.87	173
VASC	0.99	0.96	0.97	172
Accuracy			0.92	1797
Macro Avg.	0.93	0.91	0.92	1797
Weighted Avg.	0.92	0.92	0.92	1797
Classification Report of VGG-19				
Skin Cancer Type	**Precision**	**Recall**	**F1-Score**	**Support**
AK	0.95	0.87	0.91	261
BCC	0.88	0.92	0.90	292
BKL	0.96	0.87	0.92	306
DF	0.90	0.90	0.90	63
MEL	0.87	0.91	0.89	325
NV	0.92	0.95	0.94	305
SCC	0.87	0.94	0.90	173
VASC	0.95	0.96	0.95	72
Accuracy			0.91	1797
Macro. Avg	0.91	0.92	0.91	1797
Weighted Avg.	0.91	0.91	0.91	1797
Classification Report of Inception V3				
Skin Cancer Type	**Precision**	**Recall**	**F1-Score**	**Support**
AK	0.65	0.80	0.72	261
BCC	0.87	0.72	0.79	292
BKL	0.87	0.72	0.79	306
DF	0.69	0.43	0.53	63
MEL	0.73	0.72	0.73	325
NV	0.59	0.98	0.73	305
SCC	0.90	0.42	0.57	173
VASC	1	0.40	0.57	72
Accuracy			0.72	1797
Macro Avg.	0.79	0.65	0.68	1797
Weighted Avg.	0.77	0.72	0.72	1797
Classification Report of Inception ResNet V2				
Skin Cancer Type	**Precision**	**Recall**	**F1-Score**	**Support**
AK	0.98	0.93	0.91	261
BCC	0.92	0.93	0.93	292
BKL	0.91	0.93	0.92	306
DF	0.98	0.90	0.89	63
MEL	0.89	0.90	0.89	325
NV	0.96	0.93	0.95	305
SCC	0.87	0.84	0.86	173
VASC	0.99	0.92	0.95	72
Accuracy			0.91	1797
Macro Avg.	0.93	0.91	0.92	1797
Weighted Avg.	0.91	0.91	0.91	1797

Table 3. Performance of proposed ensemble models.

Classification Report of Majority Voting Ensemble				
Skin Cancer Type	**Precision**	**Recall**	**F1-Score**	**Support**
AK	0.96	0.99	0.97	261
BCC	0.98	0.99	0.98	292
BKL	0.99	0.99	0.99	306
DF	1.00	0.97	0.98	63
MEL	0.98	0.97	0.97	325
NV	0.97	0.99	0.96	305
SCC	1.00	0.93	0.96	173
VASC	1.00	0.99	0.99	72
Accuracy			0.98	1797
Macro Avg.	0.98	0.98	0.98	1797
Weighted Avg.	0.98	0.98	0.98	1797
Classification Report of Weighted Averaging Ensemble				
Skin Cancer Type	**Precision**	**Recall**	**F1-Score**	**Support**
AK	0.96	0.99	0.97	261
BCC	0.98	0.98	0.98	292
BKL	0.99	0.99	0.99	306
DF	1.00	0.97	0.98	63
MEL	0.97	0.97	0.97	325
NV	0.97	0.99	0.96	305
SCC	1.00	0.94	0.96	173
VASC	1.00	0.99	0.99	72
Accuracy			0.982	1797
Macro Avg.	0.98	0.98	0.98	1797
Weighted Avg.	0.98	0.98	0.98	1797
Classification Report of Weighted Majority Ensemble				
Skin Cancer Type	**Precision**	**Recall**	**F1-Score**	**Support**
AK	0.96	0.99	0.98	261
BCC	0.98	0.99	0.99	292
BKL	0.99	0.99	0.99	306
DF	1.00	0.98	0.98	63
MEL	0.98	0.98	0.98	325
NV	0.97	0.99	0.98	305
SCC	1.00	0.95	0.97	173
VASC	1.00	0.99	0.99	72
Accuracy			0.986	1797
Macro Avg.	0.99	0.99	0.99	1797
Weighted Avg.	0.99	0.99	0.99	1797

Table 4 shows the performance comparison of the individual deep learning models developed in [10–12,19,31,32,47,56,57] and the deep learning models developed in the proposed work for eight classes of skin cancer. It is observed from the table that the individual fine-tuned deep learning models perform better than the individual deep learning models developed in [13,32,47,57]. Table 4 shows classification results with different numbers of classes. Usually, in machine learning models, as the number of classes increases the classification accuracy decreases due the increased model complexity. It is shown in the Table 4 that the individual models developed for the eight classes can perform in comparison to the models developed for the lesser number of classes. The comparison has been made with the classification model that uses the ISIC or HAM

dataset that has been used in the ISIC 2018 challenge (Task 3) and is available on (https://challenge2018.isic-archive.com/, accessed on: 10 October 2021).

Table 4. Performance comparison with other deep learning-based models.

Ref	Method	Number of Classes	Accuracy	Precision%	Recall%
[18]	VGGNet	Two	81.3	79.74	78.66
[47]	Fully Convolutional Network	Five	85.8	N/A	N/A
		Ten	81.8		
[32]	Multi-tract CNN	Ten	79.15	NA	N/A
[11]	CNN	Three	69.14	N/A	N/A
	CNN-PA	Nine	72.1		
	CNN		48.9		
	CNN-PA		55.4		
[20]	VGG16	Seven	75.6	N/A	N/A
	ResNet50		86.6		
	DenseNet121		89.2		
	Xception		90.1		
	Inception V3		74.3		
	DenseNet161		88.7		
[19]	VGG16	Seven	80.1	N/A	N/A
	GoogleNet		79.7		
[49]	ResNet50	Seven	87.1	78.6	77.0
	InceptionV3		89.7	84.9	80.0
[56]	MObileNet	Seven	83.1	89.0	83.0
[19]	InceptionResNetV2	Seven	70.0	N/A	N/A
	PNASNet-5-Large		76.0		
	SENet154		74.0		
	InceptionV4		67.0		
[57]	Triple-Net+CAM-BP	Two	82.0	N/A	N/A
[13]	Dilated VGG 16	Seven	87.42	87.0	87.0
	Dilated VGG 19		85.02	85.0	85.0
	Dilated MobileNet		88.22	89.0	88.0
	Dilated InceptionV3		89.81	89.0	89.0
[10]	IRRCNN	Seven	87.0	N/A	N/A
[12]	CNN	Seven	77.0	N/a	N/A
	CNN (one vs All)		92.90		
[31]	InceptionV3	Seven	91.56	89.0	89.0
	ResNetXt101		93.20	88.0	88.0
	InceptionResNEtV2		93.20	87.0	88.0
	Xception		91.47	89.0	88.0
	NASNetLarg		91.11	86.0	86.0
Proposed	ResNet	Eight	92.0	0.92	0.92
	InceptionV3		72.0	0.79	0.65
	DenseNet		92.0	0.93	0.91
	InceptionResNetV2		91.0	0.93	0.92
	VGG 19		91.0	0.91	0.92

Table 5 shows the performance comparison of the proposed ensemble model with the recent deep learning-based ensemble models proposed in [18,19,21,31,49]. It is observed from the table that the majority voting, weighted averaging, and weighted majority ensemble models have an accuracy of 98%, 98.2%, and 98.6%, respectively, which is much higher than the ensemble models proposed in [18,19,21,31,49]. It is observed from the literature that in classification, as the number of classes increases, the classification accuracy decreases. The previous works carried out in [18,19,31,49] have lower accuracy as compared to the proposed ensemble models. Our ensemble models have outperformed both the dermatologists and the recently developed deep learning-based models for multiclass skin cancer classification without extensive pre-processing. Figure 6 shows the training accuracy of individual deep learning models. Confusion matrices of individual and ensemble models are shown in Figure 7. The motivation for adopting the ensemble learning models is that they improve the generalization of the learning systems. Machine learning models are bounded by the hypothetical spaces that have bias and variance. The ensemble models combine the decision of individual weak learners to overcome the problem of the single learner that may have a limited capacity to capture the distribution (causing variance error) present in the data. Our results show that making a final decision by consulting multiple diverse learners may help in improving the robustness as well as reducing the bias and variance error.

Table 5. Performance comparison with other deep learning-based ensemble models.

| Ref.	Ensemble	No. of Classes	Accuracy%	Weighted Average		
				Precision%	Recall%	F1-Score%
[18]	AlexNet + VGGNet	Three	79.9	-	-	-
	GoogleNet + AlexNet		80.7	-	-	-
	GoogleNet + VGGNet		81.2	-	-	-
	GoogleNet + AlexNet + GoogleNet		83.8	-	-	-
[19]	VGG16+GoogleNet	Seven	81.5	-	-	-
[49]	ResNet50 + InceptionV3	Seven	89.9	-	-	-
[31]	InceptionV3 + Xception		91.56	82	84	83
	Inception ResNetv2+ ResNetTx101		88.66	80	82	81
	Inception RESnETv2+ ResNetTx101		92.83	83	84	84
	InceptionResNetV2+ ResNetTx101+ ResNetTx101		89.66	83	85	84
[21]	ResNet-152, +DenseNet-161, SE-ResNeXt-101, and NASNet	Seven	93	-	-	-
	Proposed Ensemble Majority Voting	Eight	98	98	98	98
	Proposed Weighted Averaging Ensemble		98.2	98	98	98
	Proposed Weighted Majority Voting		98.6	99	99	99
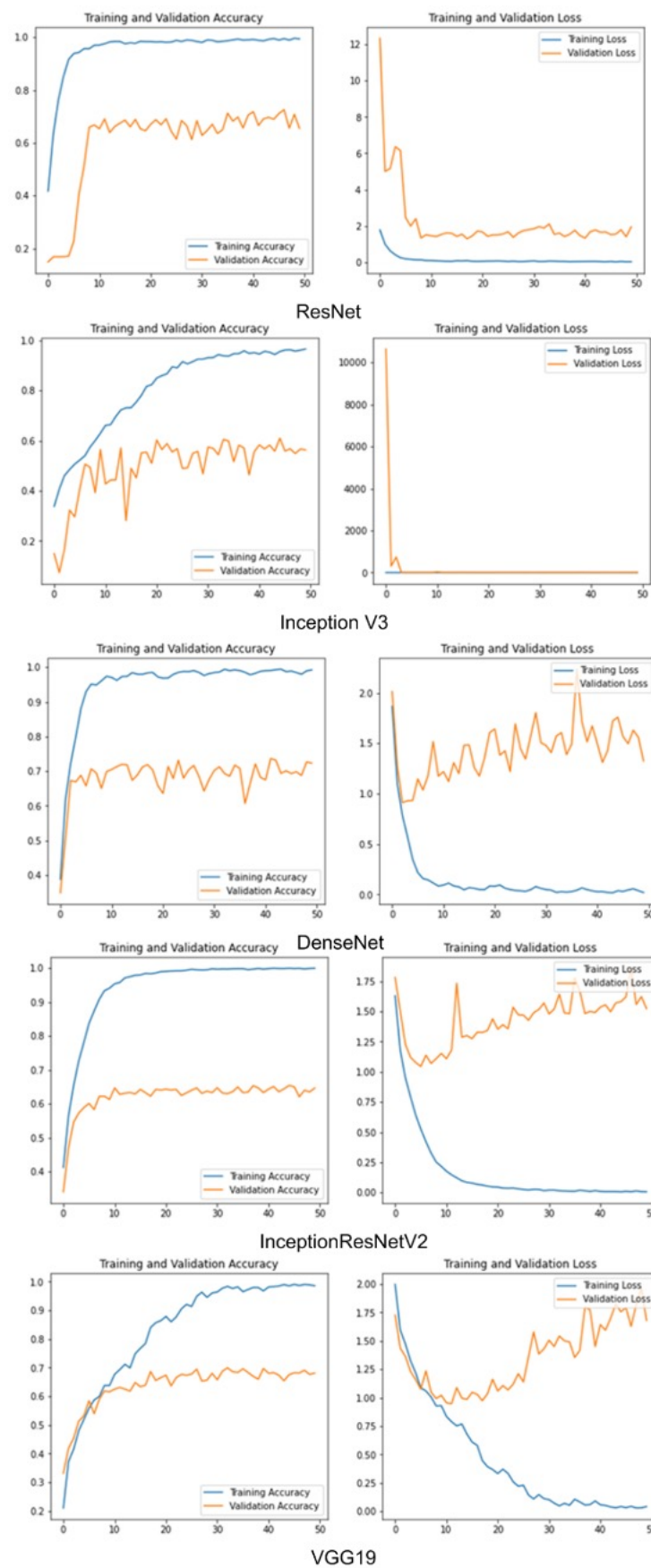
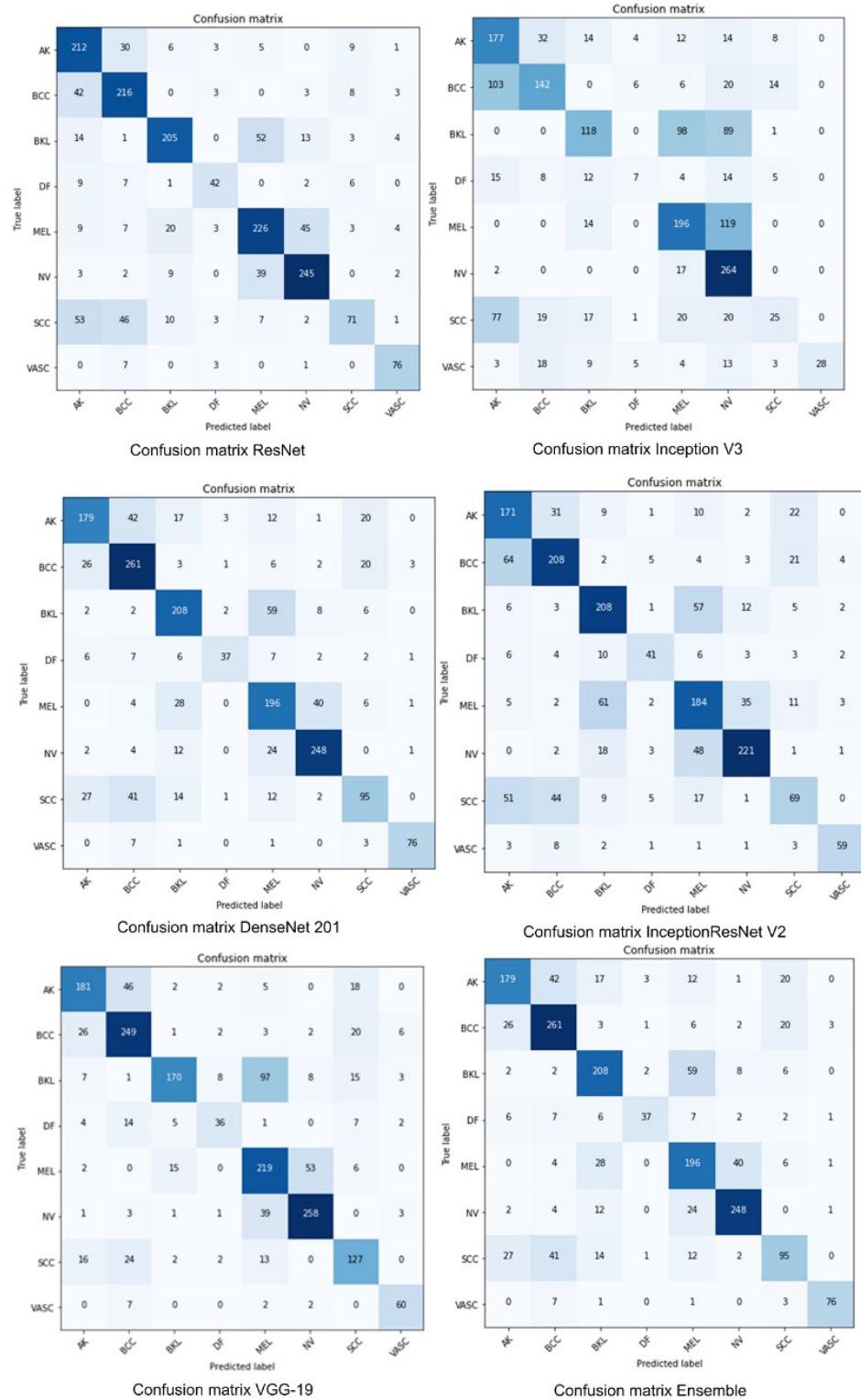**Figure 6.** Training and validation accuracy vs. loss.

Figure 7. Confusion matrix-based performance of individual and proposed ensemble model.

9. Conclusions

Various research has been performed for the classification of skin cancer, but most of them could not extend their study for the classification of multiple classes of skin cancer with high performance. In this work, better-performing heterogeneous ensemble models were developed for multiclass skin cancer classification using majority voting and weighted majority voting. The ensemble models were developed using diverse types of learners

with various properties to capture the morphological, structural, and textural variations present in the skin cancer images for better classification. It is observed from the results that the proposed ensemble models have outperformed both dermatologists and the recently developed deep learning methods for multiclass skin cancer classification. The study shows that the performance of convolutional neural networks for the classification of skin cancer is promising, but the accuracy of individual classifiers can still be enhanced through the ensemble approach. The accuracy of the ensemble models is 98% and 98.6%, which shows that the ensemble approach classifies the eight different classes of skin cancer more accurately than the individual deep learners. Moreover, the proposed ensemble models perform better than recently developed deep learning approaches for multiclass skin cancer classification.

Author Contributions: Conceptualization, N.K.; Data curation, M.S. and M.Z.u.A.; Formal analysis, M.S. and R.A.; Funding acquisition, A.S.I.; Investigation, A.H., M.Z.u.A. and A.A.; Methodology, A.H.; Project administration, A.S.I.; Supervision, N.K.; Validation, A.H.; Visualization, A.A.; Writing—original draft, N.K., A.H. and R.A.; Writing—review & editing, A.S.I. and A.A. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

MEL	melonama
NV	melanocytic nevi
BCC	basal cell carcinoma
BKL	benign keratosis lesions
AK	actinic keratosis
DF	dermatofibroma
SCC	squamous cell carcinoma
VASC	vascular lesions
ISIC	international skin imaging collaboration
CNN	convolutional neural network
SVM	support vector machine

References

1. Minagawa, A.; Koga, H.; Sano, T.; Matsunaga, K.; Teshima, Y.; Hamada, A.; Houjou, Y.; Okuyama, R. Dermoscopic diagnostic performance of Japanese dermatologists for skin tumors differs by patient origin: A deep learning convolutional neural network closes the gap. *J. Dermatol.* **2021**, *48*, 232–236. [CrossRef] [PubMed]
2. Celebi, M.E.; Kingravi, H.A.; Uddin, B.; Iyatomi, H.; Aslandogan, Y.A.; Stoecker, W.V.; Moss, R.H. A methodological approach to the classification of dermoscopy images. *Comput. Med. Imaging Graph.* **2007**, *31*, 362–373. [CrossRef]
3. Iyatomi, H.; Oka, H.; Saito, M.; Miyake, A.; Kimoto, M.; Yamagami, J.; Kobayashi, S.; Tanikawa, A.; Hagiwara, M.; Ogawa, K.; et al. Quantitative assessment of tumour extraction from dermoscopy images and evaluation of computer-based extraction methods for an automatic melanoma diagnostic system. *Melanoma Res.* **2006**, *16*, 183–190. [CrossRef] [PubMed]
4. Maglogiannis, I.; Doukas, C.N. Overview of advanced computer vision systems for skin lesions characterization. *IEEE Trans. Inf. Technol. Biomed.* **2009**, *13*, 721–733. [CrossRef] [PubMed]
5. Celebi, M.E.; Iyatomi, H.; Stoecker, W.V.; Moss, R.H.; Rabinovitz, H.S.; Argenziano, G.; Soyer, H.P. Automatic detection of blue-white veil and related structures in dermoscopy images. *Comput. Med. Imaging Graph.* **2008**, *32*, 670–677. [CrossRef]
6. Rai, H.M.; Chatterjee, K. Detection of brain abnormality by a novel Lu-Net deep neural CNN model from MR images. *Mach. Learn. Appl.* **2020**, *2*, 100004. [CrossRef]
7. Lu, S.; Lu, Z.; Zhang, Y.D. Pathological brain detection based on AlexNet and transfer learning. *J. Comput. Sci.* **2019**, *30*, 41–47. [CrossRef]

8. Zhang, Y.D.; Pan, C.; Chen, X.; Wang, F. Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling. *J. Comput. Sci.* **2018**, *27*, 57–68. [CrossRef]

9. Sajjad, M.; Khan, S.; Muhammad, K.; Wu, W.; Ullah, A.; Baik, S.W. Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *J. Comput. Sci.* **2019**, *30*, 174–182. [CrossRef]

10. Alom, M.Z.; Aspiras, T.; Taha, T.M.; Asari, V.K. Skin cancer segmentation and classification with improved deep convolutional neural network. In *Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications*; International Society for Optics and Photonics: Bellingham, WA, USA, 2020; Volume 11318, p. 1131814.

11. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]

12. 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. [CrossRef]

13. Ratul, M.A.R.; Mozaffari, M.H.; Lee, W.S.; Parimbelli, E. Skin lesions classification using deep learning based on dilated convolution. *BioRxiv* **2020**, 860700. [CrossRef]

14. Dorj, U.O.; Lee, K.K.; Choi, J.Y.; Lee, M. The skin cancer classification using deep convolutional neural network. *Multimed. Tools Appl.* **2018**, *77*, 9909–9924. [CrossRef]

15. Lopez, A.R.; Giro-i Nieto, X.; Burdick, J.; Marques, O. Skin lesion classification from dermoscopic images using deep learning techniques. In Proceedings of the 2017 13th IASTED International Conference on Biomedical Engineering (BioMed), Innsbruck, Austria, 20–21 February 2017; pp. 49–54.

16. Jinnai, S.; Yamazaki, N.; Hirano, Y.; Sugawara, Y.; Ohe, Y.; Hamamoto, R. The development of a skin cancer classification system for pigmented skin lesions using deep learning. *Biomolecules* **2020**, *10*, 1123. [CrossRef] [PubMed]

17. Adegun, A.A.; Viriri, S.; Yousaf, M.H. A Probabilistic-Based Deep Learning Model for Skin Lesion Segmentation. *Appl. Sci.* **2021**, *11*, 3025. [CrossRef]

18. Harangi, B. Skin lesion classification with ensembles of deep convolutional neural networks. *J. Biomed. Inform.* **2018**, *86*, 25–32. [CrossRef]

19. Majtner, T.; Bajić, B.; Yildirim, S.; Hardeberg, J.Y.; Lindblad, J.; Sladoje, N. Ensemble of convolutional neural networks for dermoscopic images classification. *arXiv* **2018**, arXiv:1808.05071.

20. Nyíri, T.; Kiss, A. Novel ensembling methods for dermatological image classification. In Proceedings of the International Conference on Theory And Practice of Natural Computing, Dublin, Ireland, 12–14 December 2018; pp. 438–448.

21. Bajwa, M.N.; Muta, K.; Malik, M.I.; Siddiqui, S.A.; Braun, S.A.; Homey, B.; Dengel, A.; Ahmed, S. Computer-aided diagnosis of skin diseases using deep neural networks. *Appl. Sci.* **2020**, *10*, 2488. [CrossRef]

22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

24. Kiangala, S.K.; Wang, Z. An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. *Mach. Learn. Appl.* **2021**, *4*, 100024.

25. Fatemipour, F.; Akbarzadeh-T, M. Dynamic fuzzy rule-based source selection in distributed decision fusion systems. *Fuzzy Inf. Eng.* **2018**, *10*, 107–127. [CrossRef]

26. Chatterjee, A.; Roy, S.; Das, S. A Bi-fold Approach to Detect and Classify COVID-19 X-ray Images and Symptom Auditor. *SN Comput. Sci.* **2021**, *2*, 1–14. [CrossRef] [PubMed]

27. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv* **2019**, arXiv:1902.03368.

28. Yu, L.; Chen, H.; Dou, Q.; Qin, J.; Heng, P.A. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Trans. Med. Imaging* **2016**, *36*, 994–1004. [CrossRef]

29. Almaraz-Damian, J.A.; Ponomaryov, V.; Sadovnychiy, S.; Castillejos-Fernandez, H. Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures. *Entropy* **2020**, *22*, 484. [CrossRef] [PubMed]

30. Mahbod, A.; Schaefer, G.; Ellinger, I.; Ecker, R.; Pitiot, A.; Wang, C. Fusing fine-tuned deep features for skin lesion classification. *Comput. Med. Imaging Graph.* **2019**, *71*, 19–29. [CrossRef] [PubMed]

31. Hameed, N.; Shabut, A.M.; Hossain, M.A. Multi-class skin diseases classification using deep convolutional neural network and support vector machine. In Proceedings of the 2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), Phnom Penh, Cambodia, 3–5 December 2018; pp. 1–7.

32. Kawahara, J.; Hamarneh, G. Multi-resolution-tract CNN with hybrid pretrained and skin-lesion trained layers. In *International Workshop on Machine Learning in Medical Imaging*; Springer: Cham, Switzerland, 2016; pp. 164–171.

33. Kittler, H.; Pehamberger, H.; Wolff, K.; Binder, M. Diagnostic accuracy of dermoscopy. *Lancet Oncol.* **2002**, *3*, 159–165. [CrossRef]

34. Morton, C.; Mackie, R. Clinical accuracy of the diagnosis of cutaneous malignant melanoma. *Br. J. Dermatol.* **1998**, *138*, 283–287. [CrossRef] [PubMed]

35. Binder, M.; Schwarz, M.; Winkler, A.; Steiner, A.; Kaider, A.; Wolff, K.; Pehamberger, H. Epiluminescence microscopy: A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. *Arch. Dermatol.* **1995**, *131*, 286–291. [CrossRef]

36. Piccolo, D.; Ferrari, A.; Peris, K.; Daidone, R.; Ruggeri, B.; Chimenti, S. Dermoscopic diagnosis by a trained clinician vs. a clinician with minimal dermoscopy training vs. computer-aided diagnosis of 341 pigmented skin lesions: A comparative study. *Br. J. Dermatol.* **2002**, *147*, 481–486. [CrossRef]

37. White, R.; Rigel, D.S.; Friedman, R.J. Computer applications in the diagnosis and prognosis of malignant melanoma. *Dermatol. Clin.* **1991**, *9*, 695–702. [CrossRef]

38. Moura, N.; Veras, R.; Aires, K.; Machado, V.; Silva, R.; Araújo, F.; Claro, M. ABCD rule and pre-trained CNNs for melanoma diagnosis. *Multimed. Tools Appl.* **2019**, *78*, 6869–6888. [CrossRef]

39. Kasmi, R.; Mokrani, K. Classification of malignant melanoma and benign skin lesions: Implementation of automatic ABCD rule. *IET Image Process.* **2016**, *10*, 448–455. [CrossRef]

40. Nachbar, F.; Stolz, W.; Merkle, T.; Cognetta, A.B.; Vogt, T.; Landthaler, M.; Bilek, P.; Braun-Falco, O.; Plewig, G. The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions. *J. Am. Acad. Dermatol.* **1994**, *30*, 551–559. [CrossRef]

41. Ballerini, L.; Fisher, R.B.; Aldridge, B.; Rees, J. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 63–86.

42. Blum, A.; Luedtke, H.; Ellwanger, U.; Schwabe, R.; Rassner, G.; Garbe, C. Digital image analysis for diagnosis of cutaneous melanoma. Development of a highly effective computer algorithm based on analysis of 837 melanocytic lesions. *Br. J. Dermatol.* **2004**, *151*, 1029–1038. [CrossRef]

43. Yu, Z.; Ni, D.; Chen, S.; Qin, J.; Li, S.; Wang, T.; Lei, B. Hybrid dermoscopy image classification framework based on deep convolutional neural network and Fisher vector. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, VIC, Australia, 18–21 April 2017; pp. 301–304.

44. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]

45. Masood, A.; Al-Jumaily, A.A. Computer aided diagnostic support system for skin cancer: A review of techniques and algorithms. *Int. J. Biomed. Imaging* **2013**, *2013*, 323268. [CrossRef]

46. Hosny, K.M.; Kassem, M.A.; Foaud, M.M. Skin cancer classification using deep learning and transfer learning. In Proceedings of the 2018 9th Cairo International Biomedical Engineering Conference (CIBEC), Cairo, Egypt, 20–22 December 2018; pp. 90–93.

47. Kawahara, J.; BenTaieb, A.; Hamarneh, G. Deep features to classify skin lesions. In Proceedings of the 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), Prague, Czech Republic, 13–16 April 2016; pp. 1397–1400.

48. Ali, M.S.; Miah, M.S.; Haque, J.; Rahman, M.M.; Islam, M.K. An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models. *Mach. Learn. Appl.* **2021**, *5*, 100036.

49. Shahin, A.H.; Kamal, A.; Elattar, M.A. Deep ensemble learning for skin lesion classification from dermoscopic images. In Proceedings of the 2018 9th Cairo International Biomedical Engineering Conference (CIBEC), Cairo, Egypt, 20–22 December 2018; pp. 150–153.

50. Rahman, Z.; Hossain, M.S.; Islam, M.R.; Hasan, M.M.; Hridhee, R.A. An approach for multiclass skin lesion classification based on ensemble learning. *Inform. Med. Unlocked* **2021**, *25*, 100659. [CrossRef]

51. Wang, W. Some fundamental issues in ensemble methods. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 2243–2250.

52. Gomes, H.M.; Barddal, J.P.; Enembreck, F.; Bifet, A. A survey on ensemble learning for data stream classification. *ACM Comput. Surv.* **2017**, *50*, 1–36. [CrossRef]

53. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.

54. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; John Wiley & Sons: Hoboken, NJ, USA, 2014.

55. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **2006**, *6*, 21–45. [CrossRef]

56. Chaturvedi, S.S.; Gupta, K.; Prasad, P.S. Skin lesion analyser: An efficient seven-way multi-class skin cancer classification using mobilenet. In Proceedings of the International Conference on Advanced Machine Learning Technologies and Applications, Jaipur, India, 13–15 February 2020; pp. 165–176.

57. Ge, Z.; Demyanov, S.; Chakravorty, R.; Bowling, A.; Garnavi, R. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Quebec City, QC, Canada, 11–13 September 2017; pp. 250–258.