Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

Skin Lesion Analysis using Ensemble Transfer Learning

Sabaina Haroon*

[1]Department of Computer Science and Electrical Engineering, University of Central Florida, Orlando.

*To whom correspondence should be addressed.

Abstract

Motivation: Skin cancer is one of the deadliest and most prevalent cancer in countries like United States and Australia. There are around 500,000 cases of skin cancer diagnosed early in US, among which melanoma causes most deaths, but if it is diagnosed earlier, survival rate can increase by 95 percent. As we have seen Machine Learning paving its path in various fields, it has been making tremendous breakthrough findings in medical domain also. Convolutional Neural Networks (CNNs) play an important role in that where a machine learning algorithm learns a distribution based up on image data. Skin lesion types are visually remarkably like each other, even a practicing dermatologist can diagnose a skin cancer with around 60-80 percent accuracy only. This project focuses to research and develop an intelligent biomedical deep learning-based system that would be able to perform early diagnosis of skin cancer in assisting the dermatologists. Moreover, this deep learning system could be implemented in smartphones, which would give an easy access to everyone at getting a quick consult.
Results: This project performs transfer learning applied on different state-of-the-art CNN models using ImageNet pretrained weights to diagnose skin lesions. These models are then ensembled into one model that would predict cancer types for test data. Method introduced also uses semi-supervised learning approach with transfer learning. I was successfully able to improve results from the base model using the approaches introduced in this paper. Average recall for Ham10000 achieved was 0.788 and for ISIC 2018 test dataset it was 0.71.
Availability: The codes are available at –give github link here----.
Contact: sabainaharoon@knights.ucf.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancer is usually caused by abnormal number of mutations occurring in our DNA structures. Among different type of cancers, Skin cancer is the most prevalent cancer in United States producing around 5 million cases yearly. Skin cancers are diagnosed visually by initial clinical screening followed by dermatological analysis and a biopsy if suggested. There are different methods present to diagnose skin cancer which involve dermatological analysis. Hand crafted methods are the most popular one in which different features of a lesion are studied such as the most popular ABCDE rule. This rule diagnosis skin cancer based on Asymmetry, Color, Dermascopic structure and Evolving of a lesion. However, there are several challenges faced by dermatologists in detection of these deadly cancer types since there is a visual similarity between these lesions, even the cancerous and non-cancerous ones. Among different skin lesion types such as basal cell carcinoma, malignant melanoma, and benign tumors such as melanocytic nevus, seborrheic keratosis, and dermatofibroma, Melanoma is the deadliest one. But if melanoma is detected at an early stage, 5-year survival rate can increase up to 95 percent. International Skin Imaging Collaboration (ISIC) is an yearly challenge focusing on better understanding of different skin lesion and focuses towards their detection through automation. This paper focuses on the 3rd task of ISIC 2018 challenge that works towards diagnosing skin cancer types for HAM10000 dataset. There are seven different categories of skin lesions present in this dataset: Melanoma, Melanocytic Nevus, Basal cell carcinoma, Actinic Keratosis, Benign Keratosis (solar lentigo/ seborrheic keratosis/ lichen planus-like keratosis), dermatofibroma and Vascular Lesion. It is important for the dataset having dermascopic images to contain high intensity digital images taken under proper lightening for lesion diagnosis, have proper zoom, angle, and lightening. Some of the cancer types, might occur frequently

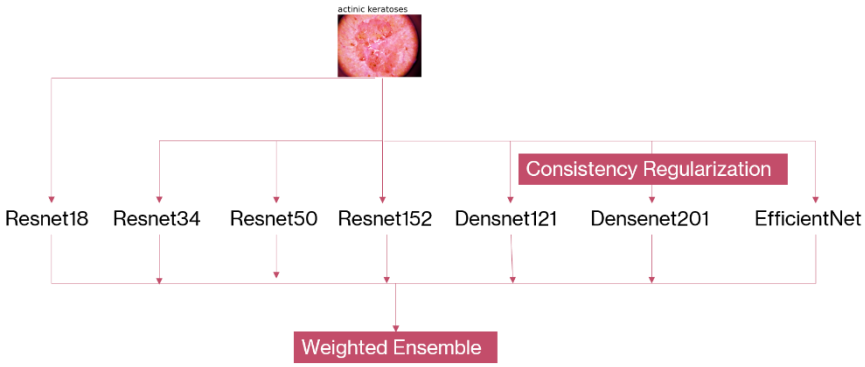

Figure 1: Proposed Architecture: Ensemble of different transfer learning models combined with transfer learning models also trained with Consistency Regularization on top of ImageNet pretrained weights. Weighted Ensemble was obtained by average prediction.

compared to some other types. Because of this reason, skin lesion datasets are highly skewed. Therefore, the major challenges faced during this project were the visual similarity between different lesion types and highly imbalanced classes data.

Deep learning is being used now to automate almost every field, from being able to win Atari games, Alpha Go, designing self-driving cars to designing intelligent robots. With the incremental breakthrough research ongoing in the field of Machine Learning and Computer vision, different works have been done for skin lesion classification. Approaches used ranged from using SVM, Random Forests, Deep learning models, fast recognition convolutional neural networks (FRCNN) for lesion localization and segmentation to several other methods. A popular research undertaken by Stanford university presents dermatologist-level classification of skin lesions using deep neural networks. They have used 129,450 clinical images, consisting of 2032 different diseases. Their method works by dividing dataset for the detection of first binary classification between most frequent skin cancers and deadliest one. These binary classification CNN models then further have different children classes depending on their categories such as benign or malignant cancer types, or other features. Google Inception v3 was used as the CNN model in their work. This work performed the most extensive research in skin diseases diagnosis. Their dataset can be used with our dataset for better performance in future. Another work done by Md Aminur et al. uses Atrous or dilated convolution in their CNN models together with transfer learning models for skin diagnosis. Barata et al used both local and global features with bag of words to detect Melanoma using dermascopy images. She et al, used manual features of darmascopic images such as color, geometry, border, diameter and several other features of the lesion for diagnosis. Deep learning models whereas would learn these features on their own without any preprocessing. Several other works used transfer learning for skin lesion diagnosis such as Esteva et al. implemented pre-trained inception v3 for classification of nine different skin lesion categories.

In proposed method, I trained different transfer learning models for classification of HAM10000 dataset. These models when trained with best parameters were ensembled into one model using average of predictions from all the models. Secondly, I researched on integrating the rising field of self and semi-supervised learning with ensemble transfer learning. We are aware that all kinds of data is available in today's age in abundance and with easier access. But it requires tremendous amount of human labor to annotate and label the data, which makes it useless in methods involving machine learning. Semi-supervised and self-supervised learning manages to utilize machine learning methods for dataset that are not labeled. Semi-supervised methods use both labeled and unlabeled dataset, learn a model from labeled dataset and use it to annotate the unlabeled dataset. Whereas self-supervised can work with complete unlabeled dataset by defining and classifying certain pretext tasks. The most prominent of them is self-supervised rotation pretext task, in which an input image is rotated at four different angles and a model is trained to classify degree of rotation applied to an input image. Both these unsupervised learning techniques enhance the capability of a model to learn better representations of a distribution. Although HAM10000 is fully labeled but some of the categories possess exceedingly small number of instances for training such as dermatofibroma class which has only around 100 images out 10k images available for training and testing. This project aims to enhance the representation learning of such classes using the consistency regularization approach from semi-supervised learning between original and new instances of input data generated by flipping original data. Given the limited time available for this project and time constraint attached, this project still has a lot of room of improvement with these methods. Whereas here I aimed to direct our attention towards the demanding unsupervised learning field and how in future by using proper resources and time available we can enhance the results further by maturing these methods. For instance, we can use dermascopic unlabeled data and can easily combine consistency regularization with self-supervised learning by simultaneously learning whether a flip was applied to an input image or not.

2 Data

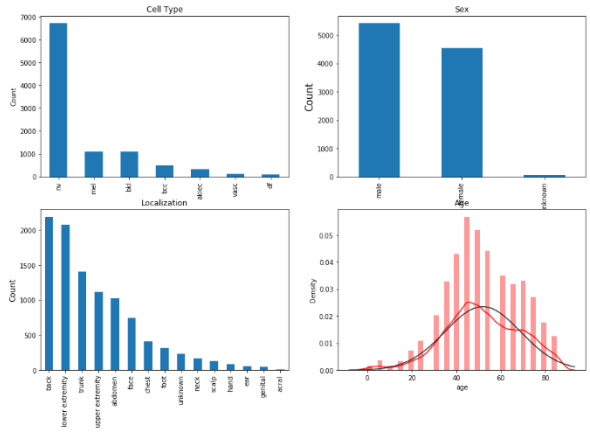

Figure 2: Some Insights into the nature of Skin Lesion Data distribution for seven different categories of HAM10000

This project uses publicly available HAM10000, Human against machine with 10000 images, dataset from ISIC 2018 archive. The International

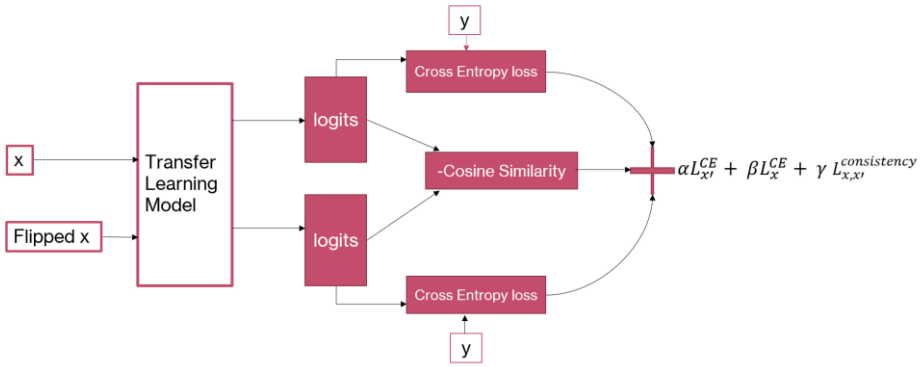

Figure 3: Conceptual diagram of how consistency regularization work. Goal of using this method is to make model learn different representations of a distribution and make them consistent with each other

Skin Imaging Collaboration (ISIC) archive consists of most extensive pigmented skin lesion digital images. Before HAM10000 dataset was introduced, data available at ISIC covered mainly two types of diseases in skin lesion, melanoma, and melanocytic nevi. Methods proposed using such dataset would perform binary classification for lesion diagnosis, which worked in experimental settings but not performed that well in clinical settings. HAM10000 whereas proposed dataset for multi-class predictions of dermatoscopic images instead of clinical images. Dermascopic images are obtained by removal of different surface reflections on skin providing a deeper level of visualization for the lesions. This dataset covered multiple different skin diseases other than nevi and melanoma. HAM10000 consists of 10015 dermatoscopic images belonging to seven different categories of benign and malignant skin lesions. The data distribution can be seen from figure 2. The bar chart in top right shows how the instances of all the categories are imbalanced. Melanocytic Nevi consists of around 67 percent of the total datasets. Whereas classes like Dermatofibroma, vascular lesions, actinic Keratosis consists of 1.15%, 1.42% and 3.26% of the instances from the total dataset. Metadata further provides information about the localization, Age and sex group of the lesions occurring in the patients. Images for HAM10000 were collected from the repositories of Department of Dermatology at the Medical University of Vienna, Austria and skin cancer practice of Cliff Rosendahl in Queensland, Australia. Both these sites collected these images over a span of 20 years. Ground truth for more than 50 percent of the data was obtained through pathology confirmation whereas the rest was obtained using different approaches such as consensus of at least 3 experts for a single lesion, follow-up, or confirmation by in-vivo confocal microscopy. Dataset was split into 64% of training, 20% of testing and 16% of validation images using stratified sampling.

For a better performance analysis of the models, I have used held-out data for testing from the datasets available for task 3 of ISIC 2018 Challenge under the category of validation and test sets. Ground truth for 193 validation images was provided, whereas for 1,512 images of test data it was not available. I have used both small datasets for evaluation and are not seen by the model during training.

3 Method

This project aims at solving the task of skin lesion classification into seven different categories. Deep learning has shown promising contributions in classifying different distributions but for model to learn an accurate prediction it needs large amount of data for training purposes. Although machine learning in medical domain has advanced over the years but it still struggles with the availability of enough data compared to other fields. To cater such a problem transfer learning is used. Transfer learning transfers the knowledge learnt from training of a deep learning model on a large dataset to acquire predictions for small dataset. I have used pretrained weights obtained from transfer learning to categorize skin lesions present in HAM10000 dataset. Method section is divided into three sections, in first section I will talk about preprocessing stage of data, second section will illuminate upon ensemble of transfer learning models and last section will describe semi-supervised learning with consistency regularization and how it was combined with transfer learning.

3.1 Data Preprocessing

As it has been discussed in the data section that HAM10000 dataset is highly imbalanced since some cancers occur frequently compared to others. Datasets having class imbalance create a problem in model learning if not catered, by making the model biased towards frequently appearing classes during the training. For instance, for a dataset consisting of two classes and 1000 images in total. Suppose 900 images belong to Class A and 100 images belong to class B. Our model can give us an accuracy of 90% in scenarios even where all the images of class A are classified correctly and all the images belonging to Class B are misclassified. But the model has only learnt 50% of the distribution. And loss calculated normally for such models would stick at a local minimum considering that loss has decreased over time and has minimized for over 90 percent of the data. To diminish the biasness of model towards majority classes different methods have been researched over the years. One of the methods uses equalizing sampling to balance data in classes. Classes having smaller instances would get their data duplicated to match to the size of class having the largest data instances. After duplication process, all the classes will have on average same number of instances. Alternative method is to give more weightage to the loss calculated against minor classes compared to that of major classes. This method uses focal loss and calculates class weightages. I have experimented with both the methods in this project but used focal loss method over equalizing sampling as it gave performance edge.

Class weightages for focal loss are calculated using Median Frequency Balancing method. I record the frequency of each class in a set and calculate its median. w_i weightage of class i is then calculated by taking ratio of median with total count of class i.

$$w_{i_{frequency}} = \frac{median\ frequency}{count_i}$$

For classes having larger count, w_i becomes small because of being inversely proportional to class size. I also performed experiments by calculating weights using balanced class weight calculation method.

$$w_{i_{balanced}} = \frac{total_{count}}{count_i * total_{classes}}$$

however, results were not much different from frequency-based method, therefore I kept using frequency-based method for further experimentations.

In second step of data preprocessing, HAM10000 dataset was divided into training, validation, and test sets using stratified sampling. Medical data needs special attention before sampling it into different sets, it should be kept in mind that images obtained from patient A should not be shared in different sets. This might give us false impression of model performing well on validation dataset, but would it fail once tested on held out dataset. Since some of the classes have only around 1 percent of total instances, simple splitting technique might send all the instances for minor class into one of the sets. Whereas Stratified sampling makes sure that data from minor class is distributed with specified percentage in all three sets. For instance, if a class has 10 images, stratified sampling will sample 7 images of this class for training set, 2 for test set and 1 for validation set. This was implemented using StratifiedShuffleSplit function from sklearn library.

3.2 Ensemble of Transfer Learning Models

In Transfer learning, pretrained weights for ImageNet dataset are learnt for state-of-the-art models. ImageNet consists of 1.2 million images from 1000 different categories. We assume that having such large number of categories, ImageNet would learn representations for most of the general distributions. And knowledge from these learnt representations is then transferred to the same state-of -the-art model but for prediction on smaller dataset. HAM10000 only consists of 10k images, with transfer learning we will transfer weights learnt from 1.2 million images to our skin lesion dataset. After transferring the weights, to properly align the weights with nuances involved in new dataset we either finetune the network by adding extra Convolutional or linear layers at the end of the model and then training the new model on top of frozen layers from transfer learning model, or we perform feature extraction by extracting weights from transfer learning model and training it over last linear layer by replacing number of classes with total number of skin lesions in HAM10000.

I have used different pretrained models such as different versions of ResNet, DenseNet, also trained EfficientNet and Inception v3 model. These models were trained using 6409 images from HAM10000 categories. And validated on 1603 images. After training of the model, I tested it on HAM10000 test set and recorded metrics for each one of the models. These models were further finetuned by changing the learning rate and optimizers. Final optimizer selected was Adam. For optimization categorical cross entropy loss was used as focal loss by incorporating class weights into it, thereby giving more weightage to the loss of classes having smaller number of instances and vice versa.

After this, I tested the performance of each of the trained transfer learning model on held-out test sets from ISIC 2018 archive. Looking into the performance metrics from leatherboard, in next step, I combined the models into an ensemble by taking average of all the models. Ensembling the models in this way helped me analyze that combination of models for the ensemble matter when taking average prediction. Since a model might decrease the overall performance by bringing the average down. This can be solved in future by assigning weights to each model according to their performance rather than simple averaging. Ensemble of transfer learning models was successfully able to enhance the performance of individual transfer learning models.

3.3 Consistency Regularization based Semi-supervised Transfer learning.

In second phase of the project, I researched on how to combine transfer learning with semi-supervised learning and analyze If they can work together or not. For the ensemble of transfer learning models, some of the models were trained using method defined in previous section whereas some other models were trained using combination of transfer learning and consistency regularization. Figure 3 shows the working of such model. A lesion input is horizontally flipped, original image and its flipped version are passed through a transfer learning model to obtain logits for both the inputs. Logits from both the inputs are used for consistency regularization i.e., F1 cosine similarity loss is calculated between them. This loss enforces the model to make the representations of input and it is flipped version to be closer to each other in the latent space. The idea here is that any new version of the input that is created by augmenting the input will also belong to the same distribution. For instance, an image of lesion taken from smartphone instead of dermatoscopy could be rotated to any degree, might have different illumination, or zoom. In such cases a model trained in this way would be robust and would have learnt more about the distribution compared to simple transfer learning model. An important thing to notice here is that claim made here depends upon the type of augmentation applied to the input image. Due to time constraint associated with this project, I only tested this new technique for flipped version of the input. In my opinion results would have further increased if along with flip other augmentations have been applied to the input, such as puzzle pretext task. Puzzle pretext task in self-supervised domain, crops the original image into smaller patches and then shuffle those patches at random location. The model then learns to predict the random shuffling applied to an input image. This helps in boosting the model performance by making it learn more about a certain distribution, along with just learning the ground truth labels of the data. I could also have used rotated versions of the input image for consistency regularization like rotation pretext task. But it might not have worked on lesions that were symmetrical in nature or would have circular shape. A recent paper, presented in NeurIPS 2018 by Feng et al (Self Supervised Representation Learning by Rotation Feature Decoupling), tackles this problem by decoupling the impact of such images in model training that have no effect to rotation, i.e., are symmetrical or circular. It was beyond the scope of this project, currently but it can be explored in future and must provide important insights to the impact of applying such technique for skin lesion datasets.

Apart from enhancing the learning capability of model by training it on different versions of the input data, another motivation to use this technique was to improve the performance of skin lesion classes that have very small number of instances such as dermatofibroma, akiec or vasc; Since this method will generate extra input samples for such classes by augmentation. From the experimentation, I have analyzed that claim made in previous statement, would work more better in future if I would decrease the number of samples from melanocytic nevus (nv) class as it has around 6700 images. Sampling out this class for a smaller number of instances, and then letting consistency regularization method to generate more instances of these classes would be more impactful. Currently by using flip technique, instances of smaller classes have been doubled. But so, that of nv class has also doubled (around 15000). I accommodate this problem by

Sr No.	Model	Akiec	bcc	bkl	DF	MEL	NV	VASC	MCA	RECALL
1	Resnet18	0.56	0.78	0.80	0.52	0.50	0.95	0.85	0.71	0.86
2	Resnet34	0.75	0.76	0.68	0.60	0.58	0.95	0.89	0.74	0.859
3	Resnet50	0.69	**0.85**	0.69	0.60	0.79	0.86	0.85	0.76	0.83
4	ResNext101	0.67	0.79	0.77	0.73	0.66	0.90	0.89	0.77	0.85
5	DenseNet121	0.64	0.79	0.80	0.56	0.65	0.81	0.89	0.73	0.78
6	DenseNet201 with consistency TL	0.58	0.75	0.71	0.69	0.63	0.95	0.75	0.72	0.86
7	ResNet152 with consistency TL	0.67	0.75	0.77	0.69	0.65	0.94	0.89	0.76	0.86
8	EfficientNet with consistency TL	0.69	0.85	0.71	0.60	0.72	0.89	0.85	0.75	0.84
9	InceptionV3 with consistency TL	0.66	0.66	0.64	0.47	0.69	0.92	0.74	0.68	0.83
	Ensemble of (1,2,3,4,5,6,7)	**0.78**	**0.83**	0.76	0.65	**0.72**	0.95	0.89	**0.80**	**0.89**
	Ensemble of 4, 7, 5	0.72	0.81	**0.80**	**0.73**	0.68	0.95	0.89	**0.80**	**0.89**

Table 1. Results for Different Transfer Learning models. 3rd column to 7th show recall for individual classes, last two columns report mean class accuracy and recall.

using focal loss with class weightages but sampling of nv class plus focal loss will help in future.

Two other losses as shown in figure 3 are, Cross Entropy Focal Loss calculated between original image x and ground truth label y, and Cross entropy focal loss between flipped input x' and ground truth label y. Note here, logits are converted into softmax predictions unlike cosine similarity loss. Focal loss makes sure to put equal focus to optimization for instances of minor classes by using class weightages.

3.3.1 Weight Scheduling Scheme

I have used ramp functions to calculate weights for scheduling scheme. All three weights α, β, γ are different for every iteration. They are calculated by dividing total training iterations into three phases. In first phase ramp function generates largest value for α, training the model to learn ground truth labels for original input images using L_x^{CE} : cross entropy focal loss for input image x. In second phase, value of alpha drops, value of beta maximizes, and model is trained now to learn predictions of ground truth label against flipped input using $L_{x'}^{CE}$. In last phase, model enforces consistency between logits learnt for original and flipped input. This weight scheduling scheme makes sure that in case if our idea of consistency regularization fail to work, model would still learn predictions for input distributions in the first two phases of training.

$$L = \alpha L_x^{CE} + \beta L_{x'}^{CE} + \gamma L_{x,x'}^{consistency}$$

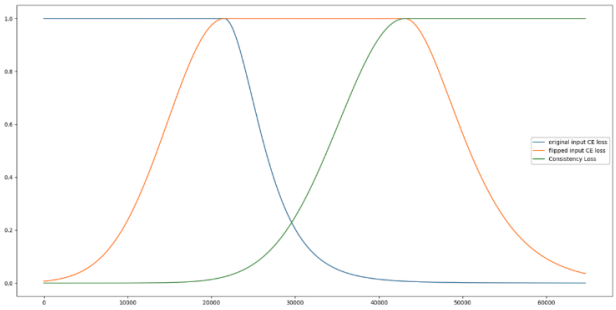

Figure 4: Weight Scheduling Scheme. Y-axis shows values of weights α, β, and γ. X-axis shows number of iterations.

4 Experimentation

All experiments for this project are performed on Google Collab Pro, using 16GB Tesla P100-PCIE GPU. HAM10000 dataset was downloaded from Kaggle as google collab provides functionality for exporting Kaggle datasets to collab repository. Dataset size was around 7.74 gigabytes. Pytorch and torch-vision libraries were used for model training and evaluations. scikit-learn library was used for data preprocessing. Transfer learning models used were pretrained on ImageNet dataset. Models trained were, ResNet18, ResNet34, ResNet50, Resnet101, Resnet154, ResNext101, DenseNet121, Densenet201, EffiicientNet and InceptionV3.

ResNet and DenseNet models are implemented for almost all transfer learning problems as they create a deeper model by adding residual networks or skip connections. These skip connections feed input from an earlier in model to later layer and hence helps in vanishing gradient problem.

EffecientNet was introduced in ICML 2019 by Google AI. This model uniformly scales the depth, width and resolution of a baseline architecture to achieve higher accuracy and efficiency on ImageNet. This model is built upon on existing networks such as MobileNet and ResNet.

These models were downloaded with ImageNet pretrained weights (pth file) using torchvision. Last layer of these models was replaced with 7 classes of lesion for HAM10000 instead of 1000 classes from ImageNet. ResNet18, ResNet34, ResNet50, DenseNet121 were trained using conventional transfer learning approach using Pytorch. Whereas rest of the models were trained with consistency regularization. There is no specific reason used behind selection of models, as which models should be trained with transfer learning and consistency regularization and which models should only be trained with transfer learning. Data loaders to load dataset into a deep learning model was designed using imagedataloader class from torchvision. This class reads metadata file and arranges input into different categorical folders in file explorer. And then reads batch of images during iteration. For testing of held out dataset since ground truth was not available so I had to write different data loader function, which would work without arranging files into categorical folder since we have no ground truth available. Different transforms applied to the input image on random before feeding it into the network were, Input image was resized to 224x224 to match with the image size of ImageNet dataset. Random horizontal flip was applied to the image, along with random rotation of 60 degree. In summary different experimentations performed were;

Table 2. Results Obtained from leaderboard of task3 for ISIC test dataset.

Sr No.	Model	ISIC 2018 Test set
1	EfficientNet with sampling	0.443
2	ResNext with sampling	0.465
3	Resnet101 with class weightage	0.554
4	DenseNet with weight change	0.553
5	inceptionV3	0.606
6	Resnet34	0.626
7	ResNet18	0.636
8	Resnet50	0.634
9	DenseNet201 with consistency learning	0.634
10	EfficientNet with weightage	0.664
11	DenseNet101 with weightage	0.665
12	ResNext101 with weightage	0.655
	Ensemble of 8 models	0.676
	ResNet18+34+50+101+Resnext101+Dense-net121	**0.715**

(1) Resnet18, Resnet34, Resnet50 transfer learning models were each trained using 70 epochs and learning rate (lr) = 1e-6

(2) Resnet101, DenseNet121 transfer learning model was trained using 100 epochs and lr = 1e-6

(3) ResNext101 was trained for 200 epochs and for lr=1e-4 and 1e-6.

(4) Resnet154, DenseNet201, EfficientNet were trained using both transfer learning and consistency regularization, but without weight scheduling. They were trained for 70, 300, 70 epochs and lr = 1e-5, 1e-6 and 1e-4 respectively. For EfficinetNet, I added three extra linear layers of size 512, 256 and last layer equal to the number of classes in HAM1000, at the end of original transfer learning model.

(5) Inceptionv3 was trained using transfer learning and consistency regularization with weight scheduling, for 140 epochs and lr=1e-6

(6) An experimentation was made between DenseNet121 trained with transfer learning and DenseNet121 trained with consistency Regularization based transfer learning. All the hyperparameters were kept same for both trainings.

(7) Different combinations of ensemble were tested and tried to analyze which ensemble would work best. This also proved that increase in number of transfer learning models in an ensemble does not guarantee better performance.

(8) To test which class imbalance removal method was better, I reproduced results for a GitHub repository code that used balanced sampling to classify HAM1000 dataset, I performed this experiment on DenseNet121

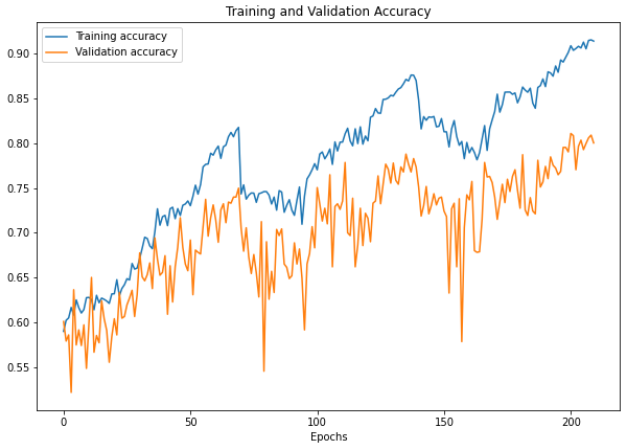

Figure 5: Training Performance for DenseNet121, trained with consistency Regularization and Weight Scheduling Scheme.

5 Results

I first collected results for individual models and saw their class performance. Class performance of individual transfer learning models showed that every model has it's own benefit in sense them some models were good in predicting benign classes, some were good in predicting dermatofibraoma and similarly melanoma. Almost all the models were able to perform well for melanocytic nevi class, since their instances were larger. From individual transfer learning models it can be seen that Resnet34, ResNet50 and Efficient perform the best for Actinic Keratosis class. ResNet50 and Efficientnet performed best for basal cell carcinoma class. DenseNet121 performed best for pigmented benign keratosis (bkl) class. ResNext101 performed best for dermatofibroma class. ResNet50 performed best for melanoma class. ResNet152 trained with consistency regularization, Resnet34, Resnext101 and DenseNet121 performed best for vascular lesions class.

When these results were combined in an ensemble overall performance of all the classes was increased thereby increasing total recall and mean class accuracy performance of the ensemble transfer learning model. Mean class accuracy is performance analysis metrics used to calculate model performance on datasets having imbalanced class data. For instance, Since melanocytic nevi class has 67 percent data of HAM10000, if model gives zero performance for all the classes but works only for melanocytic nevus, we will get 67% performance for the model. This is called micro accuracy or average recall. Whereas in actual here we know that our model is not performing well at all. And we need another performance measure to analyze it. Mean Class Accuracy is calculated by taking ratio of sum of recall of all the classes by total number of classes. In our example MCA for such a model would only be 9%. Therefore, for all the trainings I have calculated results for MCA along with average recall. Results show that average recall increases from 86 to 89 percent in ensemble and mean class accuracy increases from 77 to 80 percent.

$$\text{MCA: Mean Class Accuracy} : \frac{\sum recall_{class\_i}}{Total\ Classes}$$

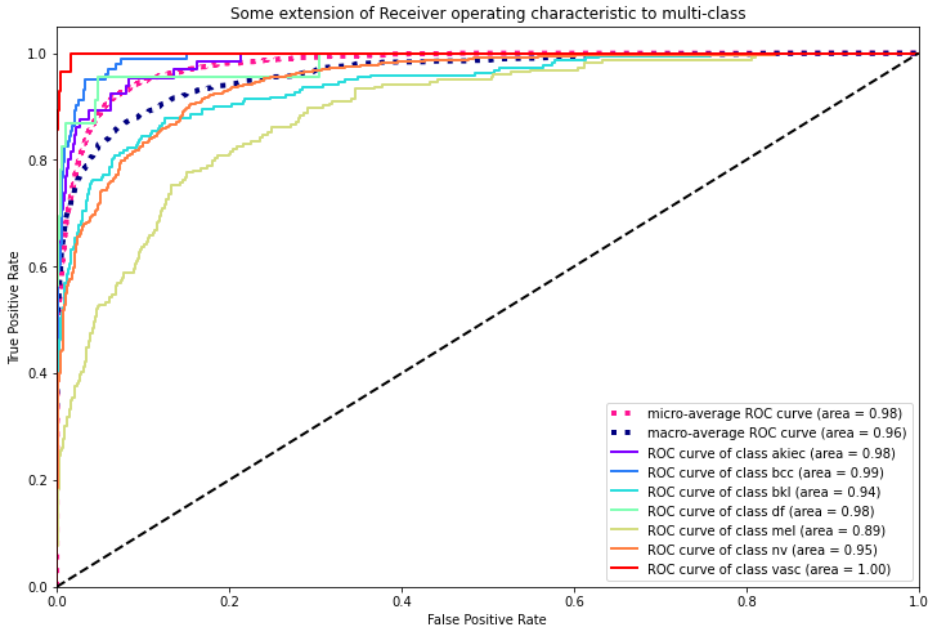

Figure 6: Multiclass ROC area curve for HAM10000 dataset tested with Ensemble of Resnext101, ResNet152 and DenseNet121. Black dotted line shows performance of a random model, with equal number of true positive rate and false positive rate. Akiec class has largest area under curve as it gives highest true positive rate for almost all the thresholds. Whereas Melanoma class has smallest area under curve.

EfficientNet was differently compared to other transfer learning models, as I added extra linear layers at the end of this model. Reason behind adding extra layers was to enhance the capability of consistency Regularization. In my opinion, Results have not given any extra benefit by addition of extra layers also, whereas it gave satisfactory results that were on par with other best performing models.

After training the models, I calculated their prediction for ISIC 2018 dataset in form of softmax predictions. These results were submitted to ISIC website since ground truth is not available for this data. As can be seen from Table 2. EffecientNet performed second best whereas ResNext101 performed the best. By changing the learning rate for ResNext101 from 1e-4 to 1e-6, I was able to increase it's mean class accuracy from 0.65 to 0.67 making ResNext101 the best performing model with ensemble. After ensemble of models, best mean class accuracy was achieved for the ensemble of different ResNet models , combined with DenseNet model and ResNext101. Mean class accuracy of this ensemble model was 0.71. It was interesting to find out that when I added EfficientNet to the ensemble of previous models, it's mean class accuracy decreased to 0.69. Which shows that combination of models matters in the performance.

Results from last two rows of Table 1, give an interesting insight since we can see that first ensemble was created using average prediction from 7 different models whereas second ensemble was calculated using average prediction from only 3 models. Both the ensembles perform equally well.

Figure 5, Shows a proof of concept for use of weight scheduling scheme in case of our three losses. The figure is a snapshot from the training process of DenseNet121 that was trained using transfer learning weights from ImageNet along with consistency regularization. This was trained using weight scheduling scheme. Its training process give an important insight that it can be seen during the start of the three phases of learning validation accuracy decreases and then increases gradually, whereas overall its performance would increase from the previous phase.

Figure 6, shows ROC curve for multiclass evaluation. ROC is generated by calculating false positive rate and true positive rate of all the classes against a given threshold of softmax. For instance if threshold is 0.5, if predicted class has probability greater than 0.5, it will be considered true positive. We see that melanoma class has smallest ROC curve compared to other classes, one probable reason could be that melanoma is usually confused with melanocytic nevi. So the softmax predictions for melanoma class must not be that peaky, nevus class must possess high probability value for such instances too, bringing confidence value of melanoma smaller. Which ultimately would give rise to more false positives. This insight could only be obtained through results of ROC as other performance measures do not show any considerable difference of this class performance with other 6 classes.

4.1 Comparison with Sampling based approach.

To compare performance of my datapreprocessing method combined with the transfer learning model and consistency Regularization, I reproduced results from GitHub repository, https://github.com/ishanrai05/skin-cancer-prediction. I was able to successfully reproduced results. This approach was only splitting HAM10000 in two portions, training and validation. Whereas class imbalance was catered by super sampling of instances belonging to minor classes using duplicates. Results calculated on validation set from their split gave around 90 percent average recall, but when I transferred these results to ISIC 2018 held out dataset. It gave only 0.44 mean class accuracy. Which was worst in all the models tested. As can be seen from Table 2. In my opinion, this has happened because they have not taken care of stratified sampling technique during splitting the dataset into two portions. This must have caused the duplicate instances to be present both in validation dataset and training dataset.

6 Discussion

Work done in this project, points out towards multiple important insights. By training of individual transfer learning model and comparing their

```
Accuracy of the network on the test images: 0.875648 %
Accuracy of the network on the test images: 0.809712 %
Accuracy of actinic keratoses :  75 %
Accuracy of basal cell carcinoma :  86 %
Accuracy of benign keratosis-like lesions :  72 %
Accuracy of dermatofibroma : 100 %
Accuracy of melanoma :  71 %
Accuracy of melanocytic nevi :  94 %
Accuracy of vascular lesions :  66 %
```

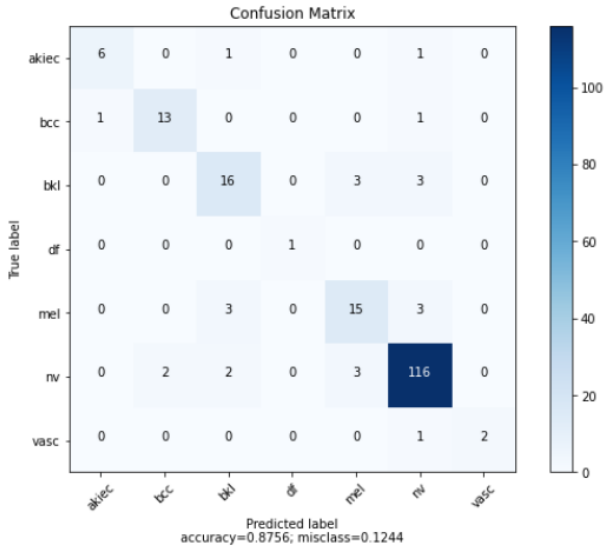

Figure 7: Evaluation for ISIC 2018 validation images

performance of Conventional transfer learning models with that models trained with both transfer learning and semi-supervised learning, We see that both the techniques perform well. But the combination of semi-supervised learning with transfer learning, did not give any considerable added benefit. Since, both techniques are trying to accomplish similar thing eventually, i.e., calculating pretrained weights that are fed into a network with a certain learning about the distribution instead of random weights. Other explanations for this could also be the nature of pretext task that I was using for consistency regularization. It might be the reason that a lesion does not change at all even after the flip, or the flip is not helping increase distance between different class distributions. We have also learnt from trying different combinations of models that if we are taking mean of predictions to create ensemble, one should be mindful of the individual performances of transfer learning model, and how much contribution a model added into an ensemble would give. In future, it would be better to use an Adaptive weight assigning method for Ensemble of models instead of using mean prediction. This adaptive weight can be designed such that more weightage is given to the prediction of transfer learning models that have higher performance.

7 Conclusion

In this project, we tried to classify different skin lesion classes, for one of the most prevalent cancer in United States, using Transfer Learning approach. Results have shown that transfer learning models have increased performance capability compared to an individual model. Ensemble based approach was successfully able to improve the performance. Results calculated on held-out test dataset gave 0.71 mean class accuracy, whereas state of the art MCA is 0.88. Considering the resources, available for this project, I think this performance is not that bad given the nature of class imbalance. We see from confusion matrices and ROC curve that melanoma class is mostly confused with melanocytic nevus class. In future we can add binary head to our transfer learning models alongwith multiclass classification of 5 other classes. Consistency Regularization approach can be made better in future by using unlabeled dataset available for unknown skin lesion types. Weights can be learnt using semi and self-supervised learning from these unlabeled datasets, and then can be further finetuned by combining them ImageNet weights and finetuning them for our seven class classification.

References

Skin Lesions Classification Using Deep Learning Based on Dilated Convolution Md. Aminur Rab Ratul, M. Hamed Mozaffari, Won-Sook Lee, Enea Parimbelli bioRxiv 860700; doi: https://doi.org/10.1101/860700

She, Z., Liu, Y., & Damatoa, A. (2007). Combination of features from skin pattern and ABCD analysis for lesion classification. Skin Research and Technology, 13(1), 25–33

Celebi, M. E., Wen, Q. U. A. N., Iyatomi, H. I. T. O. S. H. I., Shimizu, K. O. U. H. E. I., Zhou, H., & Schaefer, G. (2015). A state-of-the-art survey on lesion border detection in dermoscopy images. Dermoscopy image analysis, 10, 97–129.

Harangi, B. (2018). Skin lesion classification with ensembles of deep convolutional neural networks. Journal of biomedical informatics,