Skin Lesion Classification using Deep Learning

Aditya Bhardwaj[1] [adityabhartu@gmail.com], Priti P.Rege[2] [ppr.extc@coep.ac.in]

[1] Department of Technology, Savitribai Phule Pune University, Pune
[2] College of Engineering, Pune

Abstract. Skin cancer is a common disease and considered to be one of most prevalent form of cancer found in humans. Over the years various imaging techniques have shown improvement and reliability in diagnosis process of Skin Cancer. However, quite a few challenges are being faced in generating reliable and well-timed results as adoption of clinical computer aided systems is still limited. With the recent emergence of learning algorithms and its application in computer vision suggests a need for combination of sufficient clinical expertise and systems to achieve better results. Here we attempt to bridge the gap by mining collective knowledge contained in current Deep Learning Techniques to discover underlying principles for designing a neural network for skin disease classification. The solution is based upon merging of top-N performing models used as a feature extractor and a SVM to facilitate classification of diseases. Final model gave 86% accuracy on ISIC 2019 dataset along with high precision and recall values of 0.8 and 0.6 respectively.

Keywords: Deep Learning, CNN, Neural network

1 Introduction

Melanoma is currently one of the most dangerous types of cancer. The World Health Organization (WHO) indicates that Skin diseases are among the most common of all human health afflictions and affect almost 900 million people in the world at any time. Five common conditions account for over 80% of all skin diseases [23]. Several skin diseases are associated with long-term disfigurement, disability and stigma. The ability to diagnose skin diseases is often made at the late stages of disease development due to late medical attendances and as a result the chances of survival are falling short. The traditional system is based upon manual inspection of dermatoscopic images by experts by using certain imaging tools [10]. The experience of the expert dictates the accuracy of the diagnosis process and also the timeline of detection of the skin lesion.

Dermatoscopy is regarded as a technique in skin cancer screening which provides a higher diagnostic accuracy than the unaided eye. Prior literature infers that Dermascopy when used by expert field doctors improved the diagnostic accuracy in compari-

son to naïve photography. Without computer based assistance, the clinical accuracy of skin disease detection especially Melanoma and malignant forms is around 65-75 percent. Use of dermoscopy helps to increase the accuracy but still the differences between melanoma and benign types is still subtle. One should note that treatment for these diseases require different procedures. If diagnosis is incorrect or delayed it may be fatal. Early detection is imperative and hence gives rise to develop systems that can detect and identify such diseases at an early stage which can then in turn help to cure and save funds as it is a costly process It is therefore critical to have timely accurate diagnosis. For reasons suggested above a trained intelligent based system can assist physicians to detect and identify skin diseases. In this particular work we are interested to detect certain kind of malignant diseases, especially melanoma [8].

In this paper , a method for skin lesion classification using deep learning and computer vision is proposed , implemented and successfully test against a public challenge and dataset hosted by ISIC[22]. Besides it also presents the design and evaluation of models and features for objectively detecting diseases by applying deep feature extraction. It uses a CNN (Convolutional Neural Network) which comprises of stacked layers comprising of convolutional layers, pooling layers, rectified linear units, batch normalizers along with a decision layer for final output [2]. The CNN till now is considered the most prevalent architecture in several applications especially in image classification process. The approach involved 1) implementation and analyzing of different CNN based architectures to sort out best performing models , 2) incorporating multiple models to enhance overall classification ability.

2 Literature Survey

Previous literature show that prior efforts to apply machine intelligence to skin lesion detection and classification. Recent years a lot of image processing techniques were used along with Machine Learning Algorithms for extraction of features [3]. These techniques were mainly based upon Gabor, HSV filters [6, 18]. Over the years, significant advancement in GPU computations and hardware has been made along with dataset collection by ImageNet and Kaggle platforms.

Approach proposed by Barata et.al [19] for detection of melanoma is based on global and local features. The global form uses segmentation, histograms and other filters to help extract features such as texture, shape, color. These features are then feed forwarded to a binary classifier. He concludes that color features outperformed texture features. Coella et.al [15] utilizes hand coded feature extraction techniques.

With the emergence of Deep Learning, it has embedded development in medical imaging and proving to be of much better assistance .Kawahara et.al [1] suggests to idea of using pre trained ConvNets as feature extractor. With the use of filters it then classifies more than 3 classes of non-dermoscopic images.

Liao[16] attempted to construct a universal diseases classifier by applying transfer learning on a deep CNN. Esteva et.al [20] classified around 100,000 skin lesion

dataset using GoogleNet. Another work suggests using ResNets in parallel as part of modality fusion.

Adria Rmoero Lopez et.al [21] uses a VGG net along with transfer learning on skin images which achieves an average accuracy of 80%

Jordan yap et.al [5] discusses an idea of Multimodal models which incorporate different resource inputs to make a decision. In this particular paper, dermascopic images as well as macroscopic images were used in parallel and later their results were combined. It achieved and average accuracy of 82% but with greater loss.

Xinyuan Zhang et.al [4] classified four skin diseases by using deep learning algorithms. Dataset was dermatoscopic level images. A hierarchical structure along with domain knowledge. An accuracy of 85% on 1067 images was a the result of the experimentation and the hierarchical structure helped to implement a better computer aided support system.

3 Data

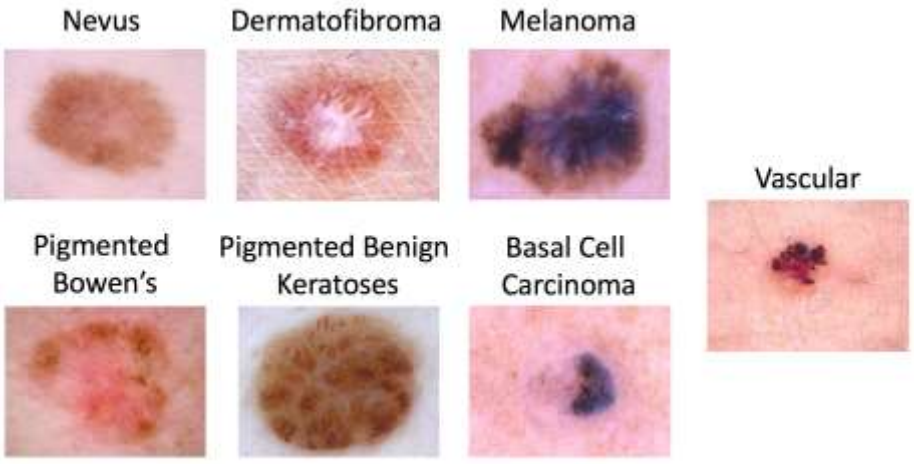


Fig. 1. ISIC 2019 Dataset

The input data comprises of dermoscopic images in jpeg format. These images are part of the ISIC 2019 challenge and all come from HAM1000 dataset [7]. The response data is a csv file consists of a binary classification for each of the 8 disease states. The diagnosis ground truth were established by one of the following methods namely Histopathology, Reflectance confocal microscopy, Lesion did not change during digital dermatoscopic follow up over two years with at least three images. Consensus of at least three expert dermatologists from a single image[22].

Table 1 ISIC HAM1000 dataset

Disease	No of images
AK (Actinic keratosis)	867
BCC (Basal cell carcinoma)	3,323
BKL (Benign keratosis)	2,624
DF (Dermatofibroma)	239
MEL (Melanoma)	4,522
NV (Melanocytic nevus)	12,875
SCC (Squamous cell carcinoma)	628
VASC (“Vascular lesion)	253

4 Proposed System

The proposed methods first insists on experimenting with various models and analyze them to find out a set of best performing ones as we know that Deep Learning models are data specific. This involved studying different Deep Learning Architectures and developing one best suited for AI field. In short we followed an iterative process (Fig 2) to gain better insights about each model. Evaluation of every model was done primarily on accuracy and loss curves. Precision, recall and F1 score was compared to get more insights on how good the model is generalizing itself i.e. performance on an unseen data. Various hyper-parameter tuning methods were applied according to the evaluation parameters [26]. Our approach also consisted of visualizing feature maps and activation maps at each layer to understand and assess how our model processes images at every layer [27].This also directed the study to venture into Neural Network Design patterns and cater way outside the black box testing [24].

4.1 Model

A final algorithm was developed based on an ‘Average of N-models’ which uses InceptionV along with Deep-CCN and MobileNet. Layers giving insufficient or redundant information were freezed during training process. Multiple networks were trained using additive strategy and corresponding adjustments were made according to the samples. A softmax layer was adjusted according to the number of inputs as used from our datasets. Secondly, experiments were also conducted by replacing the last softmax function with a Support Vector Machine (SVM) which is acts like a classifier by introducing a planar surface between the set of data points in a 2-D scenario [9].

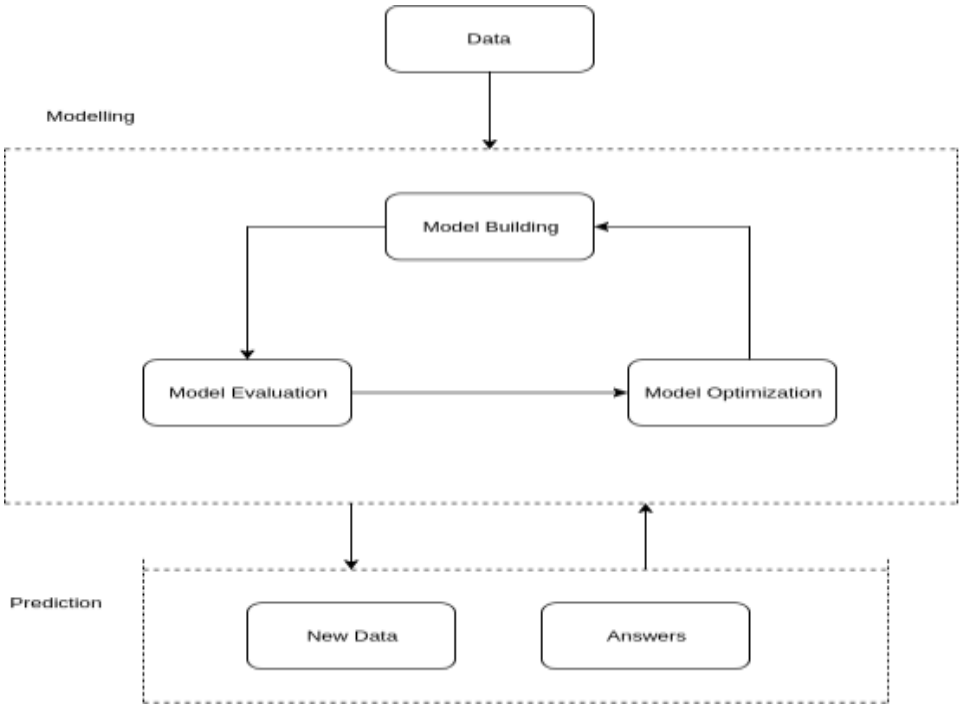


Fig.2 Basic Methodology Overview

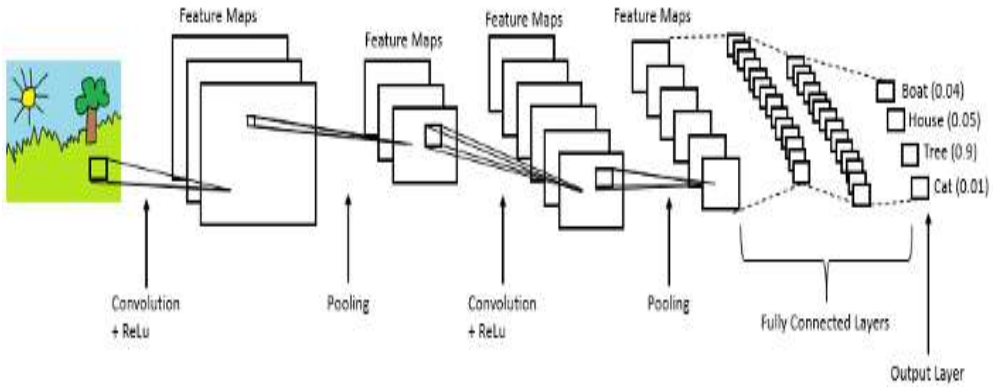


Fig.3 Feature Maps in CNN

4.2 Working

Each input after it goes through a combination of CONV, POOL and activation layers is basically under the feature extraction process. After applying a linear kernel filter and then passing it through non-linear activation function, feature maps are obtained at each layer basically it extracts generic components of an image (Fig 3). The first few convolution layers extract low-level features or high resolution feature maps are created which map out edges, lines, corners. As we deeper in the network higher level convolutions semantically strong features or higher level features. With grouped convolutions we are able to build wider networks. Basically, replicate convolutions filters without using excess memory. Networks with different sizes of convolution filter when applied on a particular feature map or input might help us gain more information as more information will shared among them(Fig 4)[13].

Pooling layers make the input representations smaller and more manageable. The non-linear layers act as trigger function and signal out distinct features in each hidden layer. Intuitively, the network will learn filters that activate when they see some type of visual feature such as an edge of some orientation or a blotch of some color on the first layer.

The prediction values were compared with actual labels. Their difference was being calculated which was then used to update final layers weights respectively. A loss function is continuously being calculated for the overall model at each epoch and its derivative is being back propagated to tune the learning process. The calculated parameters from the training set where then examined by applying it on the validation set to see if its fits well. The classification results by a neural network are based on probability summing function where probability of each adds up to 1. On the same intuition, a softmax layer is used which on the back end uses the same math but result is in form of a vector where the 1 in the vector directs to the class of disease it belongs to. Besides, the softmax, a SVM classifier was also used to predict the final decision.

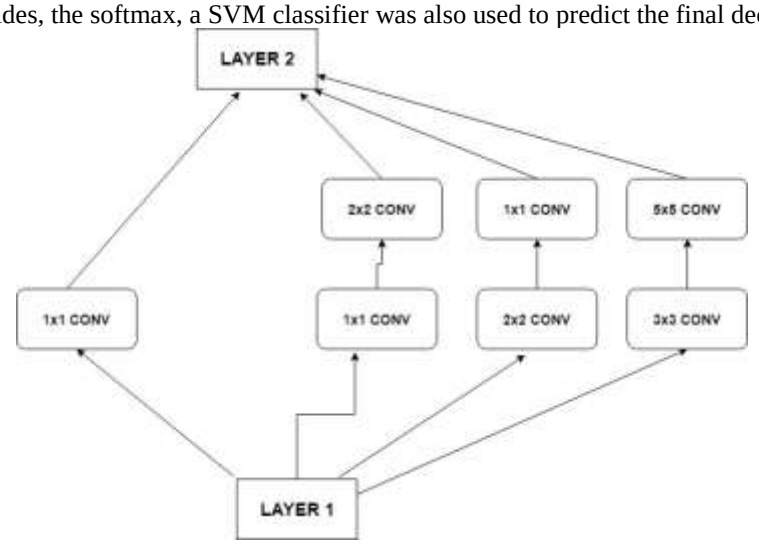


Fig. 4.Grouped Convolutions

5 Conduct

Experiments were run using datasets A and B separately. Data set B included a binary set of all the categories while A comprised of 7 categories. Each data set was split into a 80:20 with the larger share towards the training set and lower share for the validation set. Our parallel experiments included one v/s all approach to specifically tackle melanoma classification against all. For the later one we replace our classification layer to SVM. Average training was done for 50 epochs. Tensorflow and Keras were the frameworks used. Implementation was done on Google Colab, an open source cloud platform to run ML based applications.

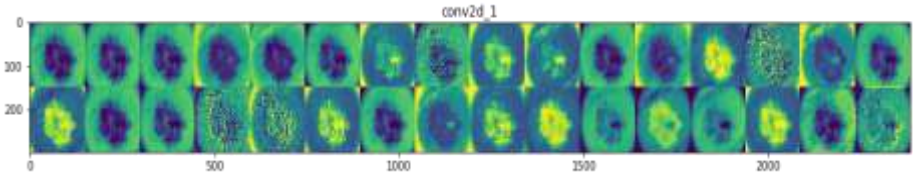
Parameters: Accuracy, Loss, loss function, Optimizing Algorithm, Learning Rate, batch size, Hidden Layers.

6 Results and Discussion

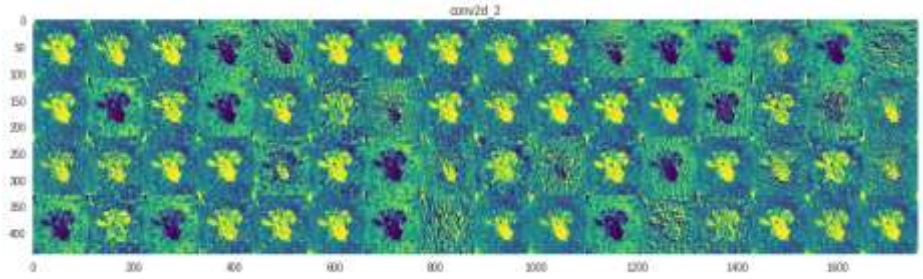
Loss function and loss metrics can be considered as the primary parameter along with accuracy to test the models on a given dataset. For binary classification, 'binary-cross entropy' proved to be most effective and 'categorical-cross entropy' for more than 2 classes. A deviation of +3% in accuracy was seen when loss function was switched to binary-cross entropy.

Adam optimizer gave the best results throughout all datasets and architectures. Reason maybe be stated as it realizes benefits of both AdaGrad and RMSprop i.e. tackling properties based on both sparse gradients and noisy problems. Adam outperforms RMSprop towards the end of the problem as gradients become sparse and has been set as a benchmark for this particular experiment.

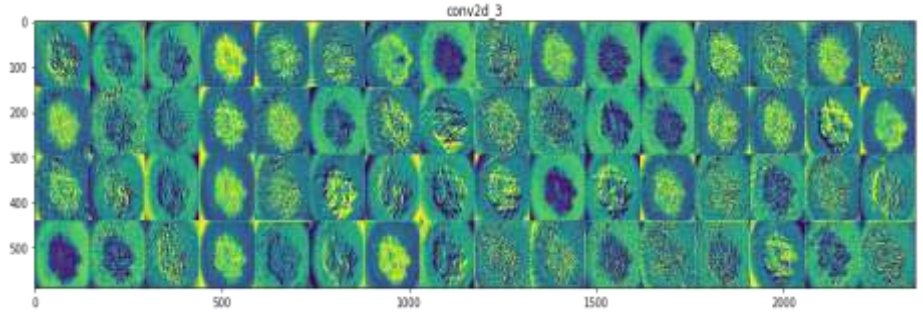
A regular update in learning rate over time gave better results. A high learning rate signifies higher kinetic energy and hence more chaotic making it difficult to settle down for deeper and narrower loss function. If the learning rate is set too low, training will progress very slowly as you are making very tiny updates to the weights in the network. Hence, a decay function was added to the learning cycle with patience set =2.However, at times a 'saddle point' was encountered and a decay learning rate scheduler proved unrewarding. A solution suggests that jumping up the learning rate can help the network get over the saddle point inferring towards adding a 'cyclical' functionality to the decay learning scheduler.



(a)



(b)



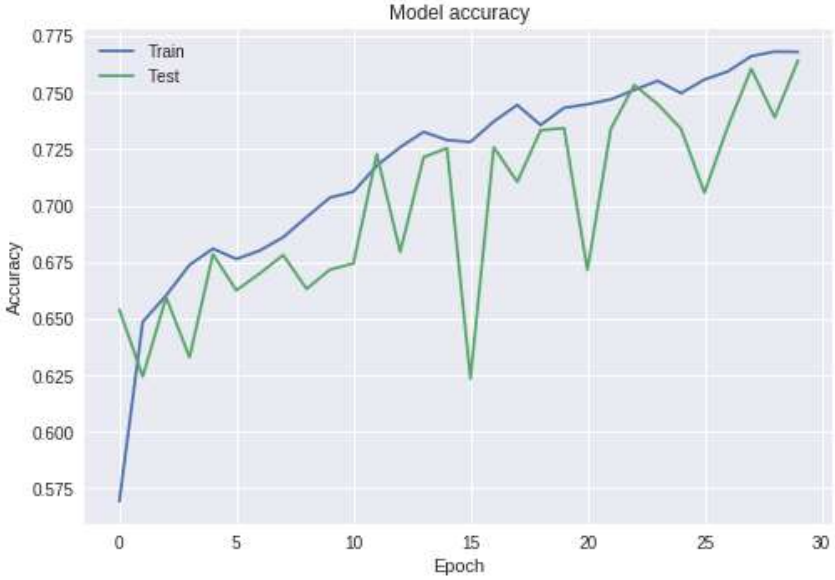
(c)

Fig 6.1 Visualization of Layers: (a) Inception Net: (b) AlexNet: (c) Average of 'N models'

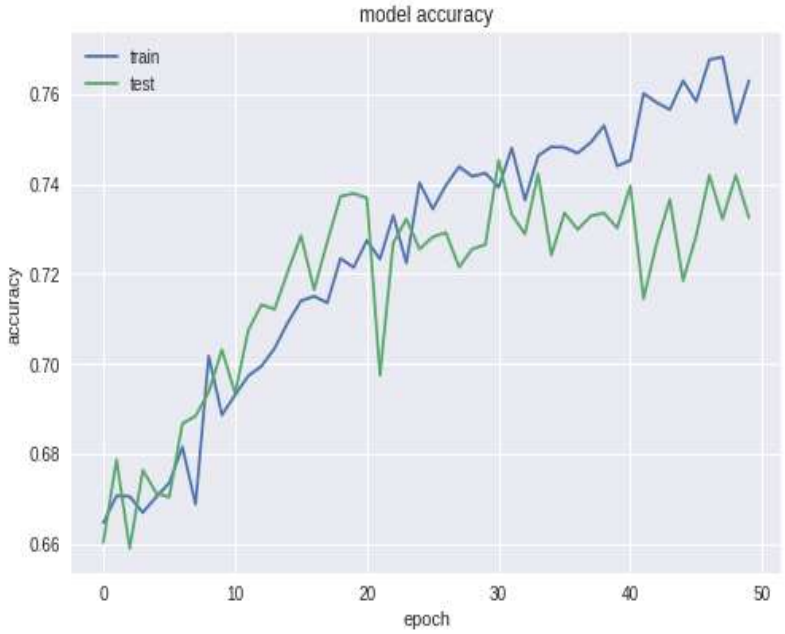
It was found that parallelism did improve results. There was a rise from a base accuracy of 71% to 75% using grouped convolutions and a further 3% rise by using regularization and tuning techniques (Table 2). An increase the number of kernels per layer will help to learn about more intermediate features, therefore increasing the number of channels in the next layer. This process of using different set of convolution filter groups on same image is called as grouped convolution (Fig 4). In short, create a deep network with some number of layers and then replicate it so that there is more than one pathway for convolutions on a single image. This particular idea can be first traced in Alex Net [25]. This concept can further be translated to Data parallelism and Model parallelism.

Among these Inception Net performed pretty well on both the datasets while VGG was unable to learn anything from this data and generalize it. The concept of grouping and stacking of Convolution layers and subsequently pooling layers didn't work out. On the contrary we can infer that kernel filters of different sizes in the same network can be further studied especially in medical imaging. One can note in Fig 6.1(c) for Average of N models gave the best visualization and feature maps in comparison to Fig 6.1(a) and Fig 6.1(b). The feature maps developed in (Fig 6.1 a,b,c) infers that

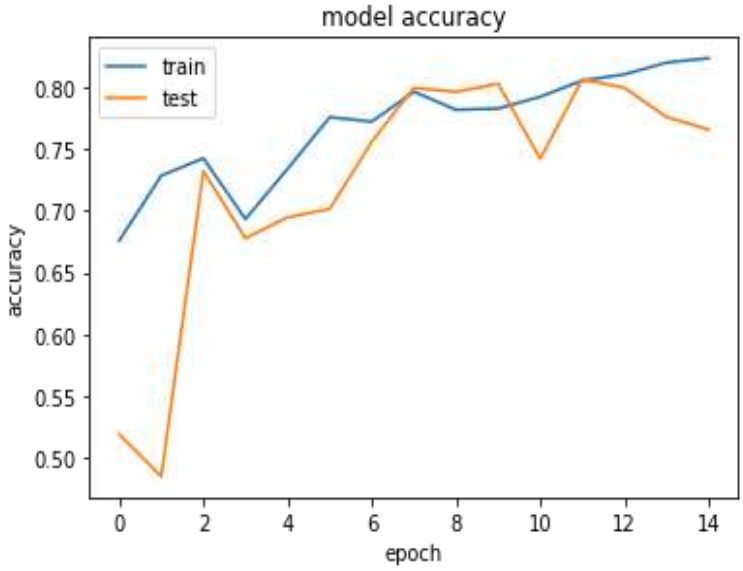
texture is a more prominent feature than just color and shape. This too suggests that architecture of the Neural Network plays an important role to develop models.



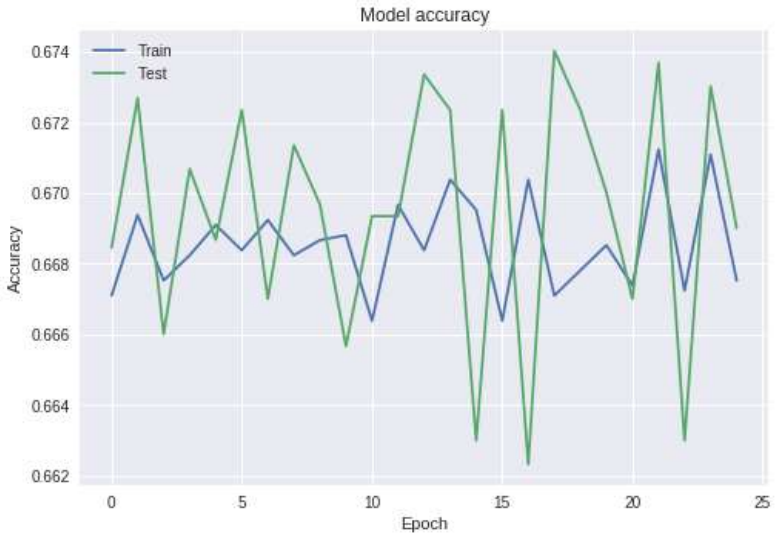
(a)


model accuracy
train
test
accuracy
epoch
0.66
0.68
0.70
0.72
0.74
0.76
0
10
20
30
40
50

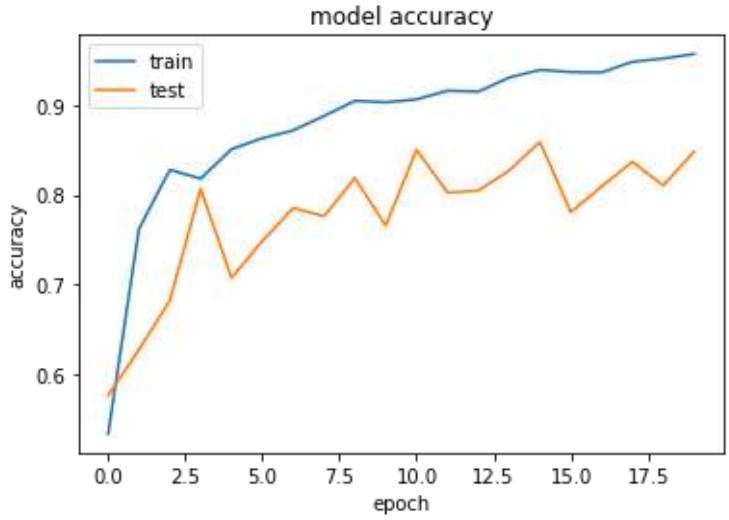
(b)


model accuracy
train
test
accuracy
epoch
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0
2
4
6
8
10
12
14

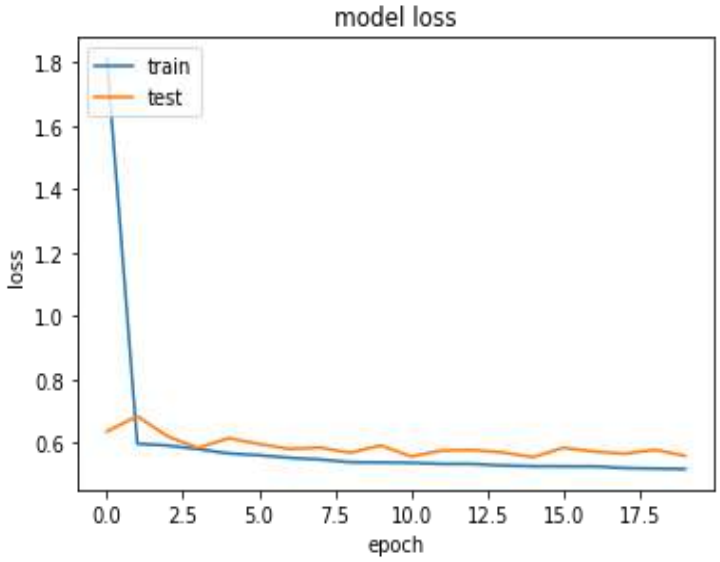
(c)


Model accuracy
Train
Test
Accuracy
Epoch
0.674
0.672
0.670
0.668
0.666
0.664
0.662
0
5
10
15
20
25

(d)


model accuracy
train
test
accuracy
epoch
0.9
0.8
0.7
0.6
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

(e)



(f)

Fig 6.2 Accuracy and Loss Plots over different Models: (a) AlexNet: (b) Mobile Net: (c) Inception Net: (d) VGG: (e) Average of 'N models' +SVM: (f) loss graph of SVM

Table 2 shows that the proposed system can produce high accuracy when applied to binary classification. The system was able to classify melanoma v/s others at a better accuracy and precision, recall, F1 score. InceptionNet performed the best Table 2 and Fig 6.2(c) on accuracy metrics but there was still a huge gap on precision and recall parameters. This suggested that model isn't able to generalize the knowledge which then got us to address the issue of imbalanced classes. The dataset was pretty imbalanced with NV having over 12,000 images along with minor representations of AK, DF, SVC ranging from 300-800. A skewed dataset didn't really help us with the precision and recall. A balanced dataset was tested and gave better results in P and R values [12]. Due to hardware constraints, data augmentation was not an efficient choice to address this issue. A comparison between Table 2 and Table 3 also proposes that specific models were not able to capture the statistical characteristics of each diseases and failing to apprehend the complexity of dataset. Two class classification performed very well for the same reason.

Table 2 Data Set A

Model	Accuracy
CNN	70.8%
VGG	63%

AlexNet	75%
AlexNet + fine tuning	**77.3%**

Table 3 Data Set B (melanoma v/s others)

Model	**Accuracy**	**Loss**	**Precision**	**Recall**
MobileNet	77.6%	0.55	0.4	0.3
MobileNet + fine tuning	84%	0.5	0.5	0.6
Inception + fine tuning	**88%**	0.43	0.5	0.4
Average+fine tuning	85.2%	0.34	0.5	0.5
Average+SVM	**86.3%**	**0.18**	**0.8**	**0.6**

The SVM classifier instead of the softmax performed way better Fig 6.2(e) and Table 3 as it is a more 'local' based classifier and not as soft as the softmax which calculates probabilities for each class. SVM instantly helped with better precision, recall and F1 scores. The metrics were double fold for binary classification. It is also important to note that on a ternary based classification the SVM performed better and gave balanced results on P and R values as well. We can conclude that SVM along with 'hinge-loss' as a classifier was able to handle data imbalance much better than other classifiers. SVM therefore encourages the correct class to have high scores and is hard with its classification methods on the other hand softmax has un-normalized log probabilities which interprets scores pretty softy and is 'never happy with' low differences in scores of classification

A predefined class weight improved detection towards melanoma. Besides, it was also found that loading ImageNet Data weights helped the model to train faster [11]. Training from scratch took around 800sec/epoch while pre trained weights from ImageNet took around half of the usual time.

7 Conclusion and Future Work

In this paper, an enhanced model based on 'Average of N models' has been proposed while using SVM as the base classifier for decision making. More specifically the system works well on two-class classifier. Dataset B consisting of two classes Melanoma v/s others achieved an accuracy of 86% while showing drastic improvement on the P and R values. Grouped convolutions seem to learn data better as every filter group is learning a unique representation of data. It can be further investigated how Proliferate paths, Incremental Feature Construction, Summation Joining aspects of neural network design can be used to make more robust models for medical imaging

applications. SVM as a classifier was able to tackle imbalance of dataset problem and gave better precision and recall values along with accuracy.

Using deep learning models to classify skin diseases would prove useful to doctors in reducing error and work in a capacity of a computer based assistance. It will help further to carry on diagnosis on remote patients therefore improving accessibility and also reducing costs.

8 References

[1]. Kawahara, J., BenTaieb, A., and Hamarneh, G.:Deep features to classify skin lesions. IEEE International Symposium on Biomedical Imaging (IEEE ISBI), pp. 1397–1400.

[2]. He, K., Zhang, X., Ren, S., Sun, J.:IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June (2016), pp. 770-778.

[3]. Chang, Y., Stanley, R. J., Moss, R. H., and Van Stoecker, W. :A systematic heuristic approach for feature selection for melanoma discrimination using clinical images. Skin Res. Technol. 11(3):165–78, (2005)

[4]. Xinyuan Zhang, Shiqi Wang, Jie Liu and Cui Tao. :Towards improving diagnosis of skin diseases by combining deep neural network and human knowledge.Zhang et al. BMC Medical Informatics and Decision Making (2018), 18(Suppl 2):59

[5].Yap, J.,Yolland,W., Tschandl, P.: Multimodal skin lesion classification using deep learning. Exp Dermatol. (2018);27:1261–1267. https://doi.org/10.1111/exd.13777

[6]. Geert Litjens , Thijs Kooi , Babak Ehteshami Bejnordi.: A survey on deep learning in medical image analysis",ELSEVIER,(2018) exd13222

[7]. Philipp Tschandl, Cliff Rosendah, Harald Kittler.: Data Descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.www.nature.com/scientificdata

[8]. Sherin Youseff ,Waid Aildy .: Computer-Aided Model for Skin Diagnosis Using Deep Learning.Journal of Image and Graphics, Vol. 4, No. 2, December (2016)

[9]. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. Journal of Machine Learning Research, vol. 5, pp. 101–141, (2003).

[10]. Hamblin, M.R., Avci,P., Gupta,G.K.: Imaging in Dermatology. Academic Press,(2016).

[11]. ImageNet Large Scale Visual Recognition Competition (ILSVRC) [electronic resource], http://www.image_net.org/challenges/ LSVRC/ (accessed July 13, 2018).

[12]. Bartosz Krawczyk.: Learning from imbalanced data. Springer Series, Prog Artif Intell (2016) 5:221–232 ,DOI 10.1007/s13748-016-0094-0

[13]. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.: Going deeper with convolutions. CoRR abs/1409.4842 (2014).

[14]. Jerant,A.F.; Johnson, J.T.; Sheridan, C.D.; Caffrey, T.J.: Early detection and treatment of skin cancer.Am. Fam. Phys**. (**2000**)**, 62, 381–382.

[15]. Codella, N.; Cai, J.; Abedini, M.; Garnavi, R.; Halpern, A.; Smith, J.R.: Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images. In International Workshop on Machine Learning in Medical Imaging; Springer: Cham, Switzerland, (2015); pp. 118–126.

[16]. Liao, Y., Shen, L., Yu, S.: HEp-2 Specimen image segmentation and classification using very deep fully convolutional network. IEEE Trans. Med. Imaging (**2017)**, 36, 1561–1572.

[17]. Codella, N.C.F., Gutman, D., Celebi, E., Helba, B., Marchetti, A.M., Dusza, W.S., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). arXiv (**2017)**, arXiv:1710.05006.

[18]. Geert Litjens , Thijs Kooi , Babak Ehteshami Bejnordi et.al.: A survey on deep learning in medical image analysis. Elsevier Survey paper , Medical Image Analysis 42 (2017) 60–88

[19]. Barata, Catarina et al.: Two Systems for the Detection of Melanomas in Dermoscopy Images Using Texture and Color Features. *IEEE Systems Journal* 8 (2014): 965-979.

[20].Esteva, A., Kuprel, B., Novoa, R., Ko, J.: Dermatologist-level classification skin cancer with deep neural networks", *Nature* volume 542, pages115–118 (2017)

[21].Romero López, Adrià , Giró-i-Nieto, Xavier, Burdick, Jack, Marques, Oge. (2017).: Skin Lesion Classification from Dermoscopic Images Using Deep Learning Techniques. 10.2316/P.2017.852-053.

[22]. The International Skin Imaging Collaboration. https://challenge2018.isic-archive.com/

[23]. World Health Organization. https://www.who.int/

[24]. Smith, Leslie N. and Nicholay Topin.: Deep Convolutional Neural Network Design Patterns. ArXiv abs/1611.00847 (2017)

[25]. Krizhevsky,A., Sutskever,I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 1097-1105,(2012)

[26]. Feurer,M., Hutter, F.: Hyperparameter Optimization.Automated Machine Learning. The Springer Series on Challenges in Machine Learning. Springer, Cham(2019)

[27]. Zeiler,M.D., Fergus,R.: Visualizing and Understanding Convolutional Networks. ArXiv, abs/1311.2901. (2013).