Melanoma classification using EfficientNets and Ensemble of models with different input resolution

Sagar Karki
School of CET
MIT-WPU
India
sagarkarki136@gmail.com

Pradnya Kulkarni
School of CET
MIT-WPU, India
Federation University,
Australia
pradnya.kulkarni@mitwpu.edu.in

Andrew Stranieri
School of Engineering,
Information Technology and
Physical Sciences
Federation University,
Australia
a.stranieri@federation.edu.in

ABSTRACT

Early and accurate detection of melanoma with data analytics can make treatment more effective. This paper proposes a method to classify melanoma cases using deep learning on dermoscopic images. The method demonstrates that heavy augmentation during training and testing produces promising results and warrants further research. The proposed method has been evaluated on the SIIM-ISIC Melanoma Classification 2020 dataset and the best ensemble model achieved 0.9411 area under the ROC curve on hold out test data.

KEYWORDS

CNN, EfficientNet, Kaggle, Skin lesion classification, Deep learning, Melanoma

1 Introduction

Rates of diagnosis for Melanoma have increased dramatically over the past three decades, outpacing almost all other cancers [1]. Today, it is one of the most common cancers found among young adults in the United States and is predicted that about 100,350 new melanomas will be diagnosed (about 60,190 in men and 40,160 in women) in the United States in 2020 [1]. The lifetime risk of getting melanoma is about 2.6% (1 in 38) for whites, 0.1% (1 in 1,000) for Blacks, and 0.6% (1 in 167) for Hispanics[2].

One of the methods to identify melanoma is with parameters known as Asymmetry, Border, Color and Diameter (ABCD) [3]. Melanoma lesions are typically symmetrical with irregular borders and larger than 6mm diameter. More than one color is normally present in these lesions. However, the identification of melanoma is labor intensive and requires skilled analysis so there is great benefit in automating the process.

Research on skin cancer using artificial intelligence (AI) has progressed rapidly in recent years which has led to faster and better diagnosis of skin cancer cases. This paper provides a method to classify the dermoscopic images into two classes melanoma (malignant) or benign. Convolutional neural networks known as EfficientNets [4] were trained on TPUs on different resolutions were used for experimenting on the SIIM-ISIC Melanoma Classification 2020 dataset. The results are promising with an area under the ROC curve of 0.9411. The research indicates that heavy augmentation during training the model and test time augmentation during evaluation boosts performance.

2 Related works

Skin cancer has been widely studied by the AI research community. The methods for automated identification and analysis of images can be categorised as Traditional methods (conventional machine learning models) and convolutional neural networks. Related works in each category are described below.

2.1 Traditional method

Traditional machine learning was performed by Ballerini [5] using a hierarchical classification system based on the K-Nearest Neighbors (K-NN) model used color and texture features extracted from skin lesion images. The method archived an overall classification accuracy of 74 % over five common classes of skin lesions, including two non-melanoma cancer types. Similar but much exhaustive experiments were performed in [6] examining the role played by color features only, by texture features only, and by combining both of them in the final classification. The research concluded that over a dataset of 176 dermoscopy images from Hospital Pedro Hispano, Matosinhos, color features outperform texture features when used alone.

Dreiseitl et al [7] compared a number of traditional machine learning approaches on the task of classifying pigmented skin lesions to conclude that conventional ANNs and SVMs performed on about the same level, with *k*-nearest neighbors and decision trees performing worse. Classification accuracies were improved with an ensemble of four classifiers namely, support vector machine, random forest, logistic model tree, and hidden naive Bayes applied on a set of 289 dermoscopy images (114 malignant, 175 benign) [8]. The method achieved an accuracy of 91.26% and area under the curve value of 0.937 when 23 features were used.
Kawahara et al [9] combined traditional with convolutional approaches by training linear classification models with features extracted from convolutional neural networks. The method achieved an accuracy of 81.8% over the entire 10-class dataset of 1300 images captured from a standard (non-dermoscopic) camera.

The features used for traditional methods require segmenting the lesions. A systematic overview of recent border detection methods is shown in [10] indicates the feasibility of the approach and the problems faced while applying the discussed methods. Iyotomi et al [11] discuss web services designed using a highly accurate dermatologist-like tumor area extraction algorithm. The system achieved a sensitivity of 85.9% and a specificity of 86.0% on a set of 1258 dermoscopy images. Celebi et al [12] describe a segmentation method to segregate the lesion from the background skin. Using color and texture related features, the image is divided into various clinically significant regions using the Euclidean distance transform and finally, optimal features are selected using an optimization framework. The method achieved a specificity of 92.34% and a sensitivity of 93.33% on a set of 564 images

2.2 Convolution neural network method

Early convolutional networks including Lopez et al VGG model [13] and Simonyan [14] using RMSProp optimizer trained with 3 different training methods and comparison between the proposed methods concluded that the fine-tuning method worked the best. The models applied on datasets from the ISIC archive [15] achieved 78.66% sensitivity. In other research by Milton et al [16] PNASNet-5-Large, InceptionResNetV2, SENet154, InceptionV4 models trained on dermoscopic images post preprocessing and augmentation over the 2018 ISIC challenge dataset [15] were compared. The research concluded that the PNASNet-5-Large model performed better than other models scoring 0.76 on the dataset. Liao [17] investigated the feasibility of a universal skin disease diagnosis system using deep convolutional neural networks (CNN) by further back-propagating. The system achieved 73.1% Top-1accuracy and 91.0% Top-5 accuracy when testing on the Dermnet dataset. On the OLE dataset, the system achieved Top-1 and Top-5 accuracies as 31.1% and 69.5% respectively. Codella et al [18] studied segmentation and classification approaches in ensembles to show these performed better than human graders in terms of accuracy and specificity with similar sensitivity using the dataset of ISIC 2016 (ISBI 016). El-Khatib et al [19] suggests a global fusion-based decision system that uses the results obtained by three different methods to establish the fusion weights. Method 1 used a neural network for classification. Method 2 used fine-tuned CNN and method 3 used SVM. The fusion method achieved an accuracy of 95% on the PH^2 database and on the ISIC 2019 database accuracy of 93%. Research in [20] proposed DermoNet, which can reuse information from preceding layers to ensure high accuracy in later layers using densely connected convolutional blocks and skip connections. similar to Densenet[21].The method was evaluated on the ISBI 2016, ISBI 2017, and the PH2 dataset, and in runtime performance of DermoNet with two other related architectures, that are fully convolutional networks and U-Net, Dermonet turned out to be faster and well suitable for practical application.
Li et al [22] proposed methods to tackle all three tasks of ISIC 2017 i.e. lesion segmentation (task 1), lesion dermoscopic feature extraction, and lesion classification. The researchers proposed a deep learning framework consisting of two fully-convolutional residual networks (FCRN) to simultaneously produce the segmentation result and the coarse classification result. The classifier is further refined using a lesion index calculation unit (LICU) and a straight-forward CNN is proposed for the dermoscopic feature extraction task. The method achieved 0.718 for segmentation, 0.833 for feature extraction, and 0.823 for lesion classification. In Unver et al [23], a pipeline for skin lesion segmentation in dermoscopic images combining a deep convolutional neural network named as You Only Look Once (YOLO)[24] and the GrabCut algorithm is explained. These methods achieved a 90% sensitivity rate on the ISBI 2017 datasetThere has been successful attempt at classifying skin lesions from HAM10000 dataset using simple CNN model with modified Adam optimizer that gave 78% accuracy [29].

3 **Methodology**

3.1 Dataset

Models are trained on data from ISIC 2020 competition plus melanoma data from ISIC 2019 and 2018 data. SIIM-ISIC Melanoma Classification 2020 dataset [25] has a training set of 33,000 examples including 584 malignant examples. Some sample images from this dataset (Melanoma and Benign lesions) are shown in Figure 1. Data used for training is in tfrecord format with a color image of 1024x1024 resolution. To tackle data imbalance we also use malignant examples from ISIC archives and previous competition dataset to increase the total count of distinct malignant examples to 2000 (all having the same resolution). The test set has 11,000 images with different data distribution then the training set. Public leaderboard is scored on the 30 % of the test data. 70 % of the test dataset is a hold out data set (referred to as test dataset for private leaderboard in the Analysis and Discussion section of this paper) to test the robustness of the models.

3.2 Preprocessing and Augmentation

Images were randomly rotated, sheared, zoomed, and shifted. Random horizontal flip, contrast, saturation, brightness and hue was also performed. Images with hair posed a problem as most malignant examples had hair and most benign did not and to help the model generalise better we added hair augmentation using real hair images. We also added coarse drop out i.e. removing random boxes of size 0.1 times the dimension of the image. Figure 2 shows the augmentation on one image. The augmented hair is visible from the difference between the image in 2nd last row and 6th column and other images. The other augmentation effect is also clearly visible in Figure 2

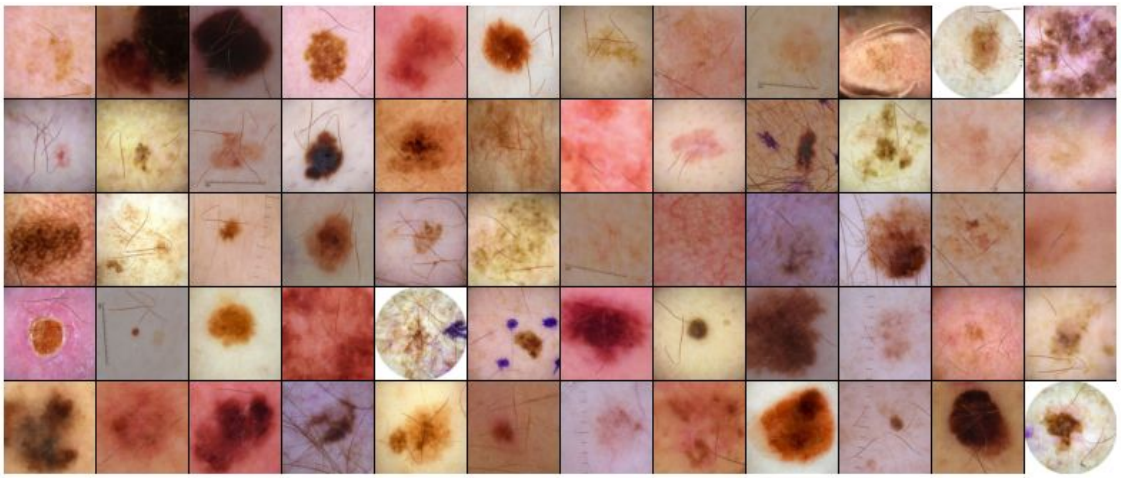

Figure 1. Images from dataset

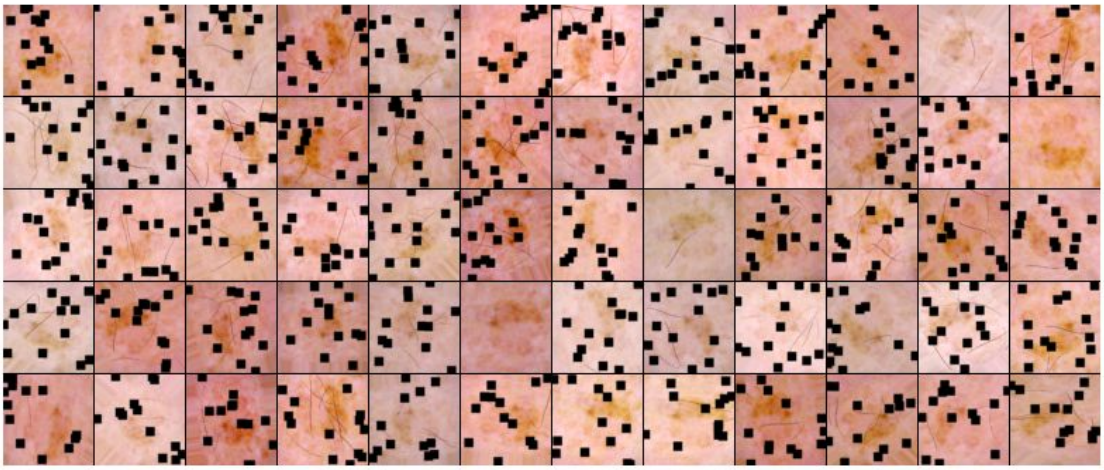

Figure 2. Data Augmentation on one image

3.3 Training and Testing

All the models are trained on Kaggle TPUs [26]. This hardware reduced the training time drastically providing more time for experiments. Models are trained using 5 fold cross-validation. The data is divided in such a manner that the ratio of malignant to benign images in each fold is equal to the ratio of malignant to benign in the overall dataset. This ensures equal data distribution in the training and validation set of each fold. All the members of EfficientNet are trained on different resolutions(shown in Table 1 column 2) and the best models based on the best out of fold prediction's area under ROC curve score are selected initially (before the competition ended) and later the best performing models are used for the ensemble. Adam optimizer with a custom learning rate scheduler is used while training. Heavy test time augmentation is applied. Each image is analyzed 25 times with augmentation applied during training with low probabilities for dropout and hair augmentation. The test time augmentation has proved to be helpful in various computer vision problem statements and it proved to be helpful in experiments mentioned in this paper as well. Models are trained for 15 epochs. The training is done using TensorFlow[27] and Keras[28].

4 **Analysis and discussion**

The results are evaluated on the area under the ROC curve between the predicted probability and the observed target. This metric depends mainly on the ranking. So the area under the ROC curve value remains the same until the ranking of the data points is the same. Therefore we tried 4 different methods for ensemble. Rank data ensemble, normal average, log ensemble, and power ensemble to ensemble the model at each fold and final ensemble of different architectures. It was observed from the experiments that the log ensemble worked best.

$$\text{log ensemble} = \exp\left(\frac{\sum_{i=1}^{5} log_2(x_i)}{5}\right)$$

$$\text{Average} = \frac{\sum_{i=1}^{5} x_i}{5}$$

Where x_i represents malignant probability predicted by the model at fold 'i' for an image. The results of the Area under the ROC curve for the best ensemble model for each architecture at specific input resolution are depicted in Table 1. Table 2 provides a description of the 1st row of Table 1.

Models are trusted based on their cross-validation score. EfficientNet B3, B2, B2, B1, B0 did not give a much promising score during cross-validation with the proposed training method and therefore were not included in the final ensemble. The evaluation metric Area Under the ROC curve (AUC) is used to assess the performance of the ensemble technique. Ensemble of the best 13 architectures (shown in Table 1) gives a 0.9400 score on test dataset for private leaderboard. Ensemble of EfficientNet B6 models with different input sizes give 0.9409. The best ensemble score 0.9411 is given by ensemble of all EfficientNet B6 models plus a EfficientNet B5 (input size 384*384) on test dataset for private leaderboard.

Although the performance of the ensemble of the 13 architecture is less on the private dataset but the diversity that it has to offer is better than what is given by ensembles of all EfficientNet B6 models and an EfficientNet B5.

Table 1 Area Under Curve for Different Image Resolutions

Mode	Inp_size	AVG_private	AVG_public	LogAvg_priv	LogAvg_pub.
B6	512	.9362	.9313	.9369	.9337
B6	456	.9354	.9356	.9368	.9381
B6	384	.9366	.9426	.9373	.9397
B6	300	.9355	.9258	.93370	.9262
B6	256	.9338	.9264	.9332	.9364
B5	512	.9334	.9320	.9364	.9353
B5	456	.9310	.9308	.9336	.9306
B5	384	.9358	.9310	.9363	.9345
B5	300	.9307	.9272	.9321	.9275
B5	256	.9256	.9236	.9259	.9232
B4	512	.9272	.9286	.9300	.9306
B4	456	.9308	.9313	.9319	.9313
B4	384	.9351	.9343	.9347	.9292
B4	300	.9363	.9227	.9347	.9258
B4	256	.9294	.9271	.9304	.9287

.

Table 2. Description of Acronyms

Model	Model name of Efficientnet family eg. B6 is 5 Efficient Net 'B6' models from 5 fold of validation whose predictions have been averaged or log ensemble
Inp_size	Input resolution for the architect ture at each fold
AVG_private	Average of the fold predictions from 5 fold cross-validation and its Area under the ROC curve score on test dataset for private leaderboard
AVG_public	Average of the fold predictions from 5 fold cross-validation and its Area under the ROC curve score on public leaderboard
LogAvg_pub	log ensemble of the fold predictions from 5 fold cross-validation and its Area under the ROC curve score on public leaderboard
LogAvg_priv	log ensemble of the fold predictions from 5 fold cross-validation and its Area under the ROC curve score on the test dataset for private leaderboard

Figure 3 represents the data from Table 1 to compare the performance of each model at different resolutions with different ensemble methods.

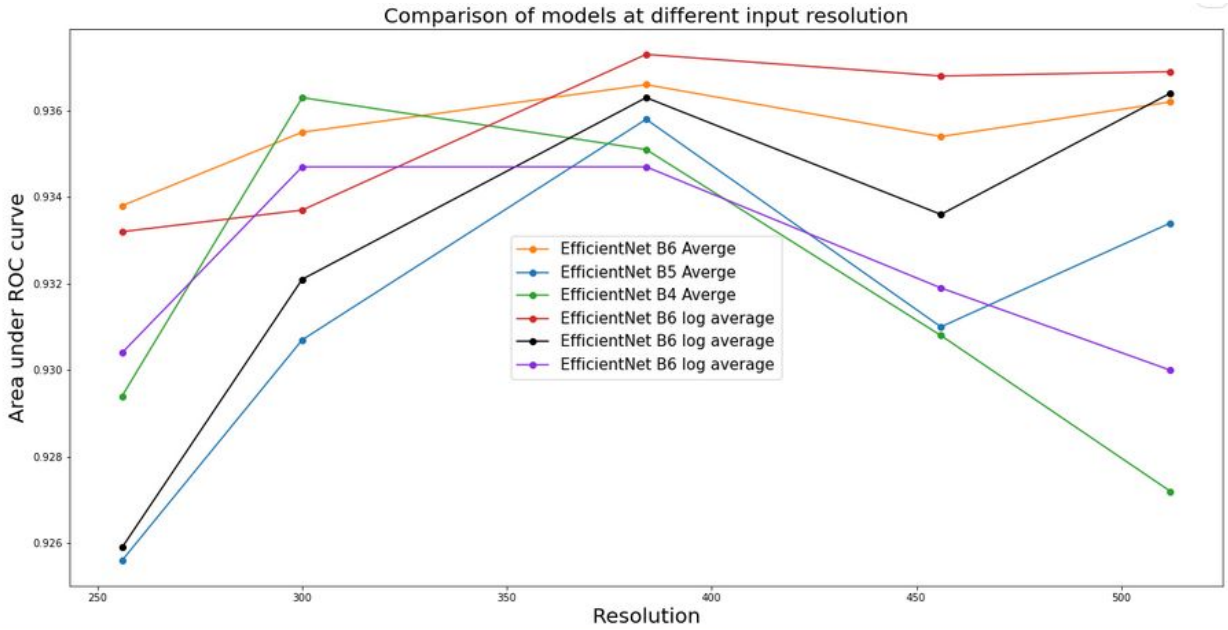

Figure 3. AUC score of models at different input resolution

5 **Conclusion**

Classification of skin cancer lesions into malignant and benign classes is a challenging as well as time-consuming job for human eyes. Considering the shortage of experts, automated diagnosis of skin cancer is essential. This paper has proposed an ensemble-based technique to classify the images into melanoma and benign lesions. Various augmentation techniques such as hair addition have been used as preprocessing to improve the classification performance. It has been observed that heavy test time augmentation averaged out the mistakes and helped in bringing out the best decision from the model. The model with depth and width seem to be better suited for the proposed training method for the identification of Melanoma. The area under the ROC curve was used to assess the performance of the ensemble models. The best result achieved on the SIIM ISIC 2020 dataset was 0.9411 using an ensemble of all EfficientNet B6 models plus one EfficientNet B5. The future research is planned to include more diversity in augmentations and training for more epochs over an expanded dataset. The use of unsupervised learning and the use of GANs for the classification can be clubbed with methods proposed in this paper to further aid in the diagnosis of melanoma.

References

1. Melanoma Statistics. [cited 5 Sep 2020]. Available: https://www.curemelanoma.org/about-melanoma/melanoma-statistics-2/

2. Melanoma Skin Cancer Statistics. [cited 5 Sep 2020]. Available: https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html

3. Soyer HP, Argenziano G, Zalaudek I, Corona R, Sera F, Talamini R, et al. Three-point checklist of dermoscopy. A new screening method for early detection of melanoma. Dermatology. 2004;208. doi:10.1159/000075042

4. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. 2019. Available: http://arxiv.org/abs/1905.11946

5. Ballerini L, Fisher RB, Aldridge B, Rees J. A Color and Texture Based Hierarchical K-NN Approach to the Classification of Non-melanoma Skin Lesions. Color Medical Image Analysis. Springer, Dordrecht; 2013. pp. 63–86.

6. Barata, C., Ruela, M., Francisco, M., Mendonça, T., & Marques, J. S. (2013). Two systems for the detection of melanomas in dermoscopy images using texture and color features. IEEE Systems Journal, 8(3), 965-979.

7. Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., & Binder, M. (2001). A comparison of machine learning methods for the diagnosis of pigmented skin lesions. Journal of biomedical informatics, 34(1), 28-36..

8. Garnavi, R., Aldeen, M., & Bailey, J. (2012). Computer-aided diagnosis of melanoma using border-and wavelet-based texture analysis. IEEE Transactions on Information Technology in Biomedicine, 16(6), 1239-1252.

9. Kawahara, J., BenTaieb, A., & Hamarneh, G. (2016, April). Deep features to classify skin lesions. In 2016 IEEE 13th international symposium on biomedical imaging (ISBI) (pp. 1397-1400). IEEE Conference Publication. [cited 4 Sep 2020]. Available: https://ieeexplore.ieee.org/document/7493528

10. Celebi ME, Iyatomi H, Schaefer G, Stoecker WV. Lesion Border Detection in Dermoscopy Images. 2010. doi:10.1016/j.compmedimag.2008.11.002

11. Iyatomi, H., Oka, H., Celebi, M. E., Hashimoto, M., Hagiwara, M., Tanaka, M., & Ogawa, K. (2008). An improved internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm. Computerized Medical Imaging and Graphics, 32(7), 566-579.

12. Celebi M, Kingravi HA, Uddin B, Iyatomi H, Alp Aslandogan Y, Stoecker WV, et al. A methodological approach to the classification of dermoscopy images. Comput Med Imaging Graph. 2007;31: 362.

13. Lopez, A. R., Giro-i-Nieto, X., Burdick, J., & Marques, O. (2017, February). Skin lesion classification from dermoscopic images using deep learning techniques. In 2017 13th IASTED international conference on biomedical engineering (BioMed) (pp. 49-54). https://ieeexplore.ieee.org/abstract/document/7893267

14. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://arxiv.org/pdf/1409.1556.pdf

15. ISIC Archive. [cited 4 Sep 2020]. Available: https://www.isic-archive.com/

16. Milton MAA. Automated Skin Lesion Classification Using Ensemble of Deep Neural Networks in ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection Challenge. 2019. Available: http://arxiv.org/abs/1901.10802

17. Liao, H. (2016). A deep learning approach to universal skin disease classification. University of Rochester Department of Computer Science, CSC.. Available: https://pdfs.semanticscholar.org/af34/fc0aebff011b56ede8f46ca0787cfb1324ac.pdf

18. Codella N, Nguyen Q-B, Pankanti S, Gutman D, Helba B, Halpern A, et al. Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images. 2016. Available: http://arxiv.org/abs/1610.04662

19. El-Khatib H, Popescu D, Ichim L. Deep Learning–Based Methods for Automatic Diagnosis of Skin Lesions. Sensors . 2020;20. doi:10.3390/s20061753

20. Baghersalimi S, Bozorgtabar B, Schmid-Saugeon P, Ekenel HK, Thiran J-P. DermoNet: densely linked convolutional neural network for efficient skin lesion segmentation. EURASIP Journal on Image and Video Processing. 2019;2019: 1–10.

21. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. 2016. Available: http://arxiv.org/abs/1608.06993

22. Li Y, Shen L. Skin Lesion Analysis towards Melanoma Detection Using Deep Learning Network. Sensors . 2018;18: 556.

23. Ünver HM, Ayan E. Skin Lesion Segmentation in Dermoscopic Images with Combination of YOLO and GrabCut Algorithm. Diagnostics (Basel, Switzerland). 2019;9. doi:10.3390/diagnostics9030072

24. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. 2015. Available: http://arxiv.org/abs/1506.02640

25. SIIM-ISIC Melanoma Classification. [cited 5 Sep 2020]. Available: https://kaggle.com/c/siim-isic-melanoma-classification

26. Tensor Processing Units (TPUs) Documentation. [cited 5 Sep 2020]. Available: https://www.kaggle.com/docs/tpu

27. TensorFlow. [cited 3 Oct 2020]. Available: https://www.tensorflow.org/

28. Keras Team. Keras: the Python deep learning API. [cited 3 Oct 2020]. Available: https://keras.io

29. Prasad T. and Siddhivinayak K. (2020, June). Skin Lesion Classification : A CNN Way .IJEAT . ISSN : 2249-8958 , Volume-9 Issue-5, June 2020