HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 3 (2022)

Comparing Skin Cancer Diagnosis between Manual, 4 Classical, and 1 Deep Machine Learning Algorithms

Maya Kasbekar[1] and Catherine Phillips[1#]

[1]Shrewsbury High School
[#]Advisor

ABSTRACT

Skin cancer incidence has increased significantly with approximately 1.2 million melanoma cases diagnosed each year globally. An experienced dermatologist's visual inspection has only been able to achieve a maximum accuracy of 78%. Recently, convolution neural networks algorithms have successfully outperformed human detection. The study aim was to compare the accuracy and computational time of one deep and four shallow machine learning classification algorithms and demonstrate their statistical superiority over human detection. One deep and four shallow learning algorithms were used to compare the detection accuracy and computational time. Using verified skin lesion images from the ISIC Database, the deep and shallow machine learning algorithms were trained, validated, classified and tested to detect a cancerous skin mole. The results showed that the accuracy of the deep algorithm (VGG-16- 98.86%) was superior to both the shallow algorithms (SVMs – 88.29%, decision trees-88.62%, logistic regression -88.26%, neural network -88.59%) as well as human detection (78.3%). The shallow algorithms were also superior to human detection. Deep Learning algorithms, specifically convolution neural networks, can reduce the false negatives and false positives during melanoma detection and could also be used for early detection of skin cancer at home, saving lives and significant costs.

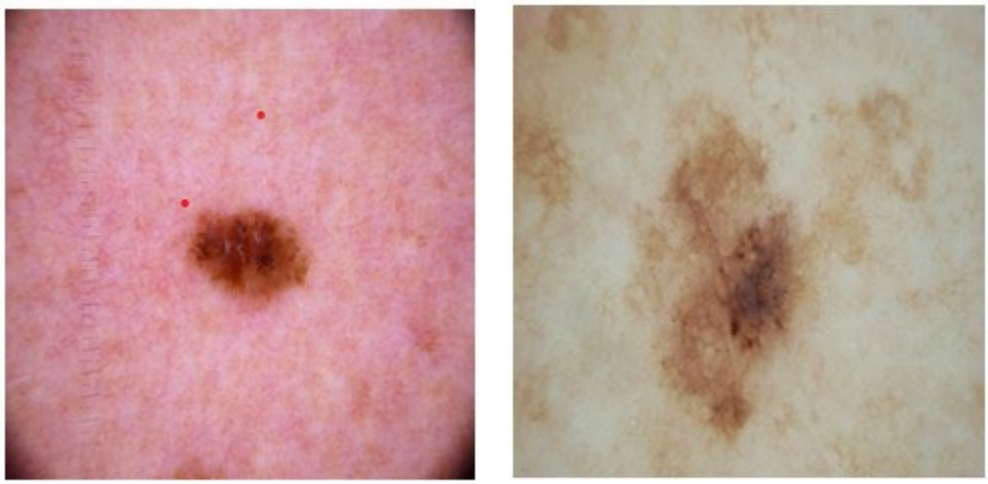
Introduction

Skin cancer is the most common cancer in the United States [1]. According to dermatological experts, twenty percent of Americans will develop skin cancer in their lifetime [American Academy of Dermatology,2020]. Approximately 9500 people in the U.S. are diagnosed with skin cancer everyday [1]. Skin cancer consists of melanoma skin cancer (MSC) and non-melanoma skin cancer (NMSC). Melanoma skin cancer afflicts over one million Americans, while non melanoma skin cancer that includes basal cell carcinoma (BCC) and squamous cell carcinoma (SCC), affects more than 3 million Americans a year [1]. It is estimated that 197,700 new cases of melanoma, 97,920 noninvasive (in situ) and 99,780 invasive, will be diagnosed in the U.S. in 2022. Invasive melanoma is projected to be the fifth most diagnosed cancer for both men (57,180 cases) and women (42,600 cases) in 2022 [1]. Globally, 1.2 million cases of skin cancer were reported in 2020 [18]. Skin cancer can affect anyone, regardless of skin color.

For several decades until recently, the method of diagnosis for melanoma has been a physician's visual inspection of images, predominantly MRI images. However, a physician's accuracy, even with experience, plateaus at 78.3%. Haenssle et al. [9] compared a Convolution Neural Network's (CNN) diagnostic performance with a large international group of 58 dermatologists, including 30 experts. Most dermatologists were outperformed by the CNN. The dermatologists achieved an overall accuracy of 78.3% (sensitivity – 88.9% and specificity – 75.7%). The CNN Receiver Operating Curve (ROC) area under curve (AUC) was greater than the mean ROC AUC of dermatologists (0.86 versus 0.79, $P < 0.01$).

Over the last two decades, machine learning has made significant advances in classification of skin cancer images and has surpassed human performance in image classification and pattern recognition, primarily due to advances in new deep learning algorithms and improvement in training methods [9]. Machine learning algorithms exploit

the symmetry, shape, color, size and density between a benign and malignant skin mole. A benign skin mole (as shown in figure 1a) is symmetrical, round and uniform in density with one consistent color throughout; even though the size may vary [4]. In contrast, a malignant skin mole, as seen in figure 1b, is asymmetrical, irregular, and non-uniform in density with multi-colored borders, areas or spots; with a size usually larger than 6 mm [4]. These cancerous cells can multiply at a faster rate than the immune system can handle, causing a detrimental effect [16]. When these tumors grow too quickly, they spread to other organs and tissues through a process called metastasis and become life threatening [2]. Many cancers, specifically skin cancer can be cured or significantly prolonged if detected early and treated effectively [18], improving the quality of life as well as healthcare costs.



1a. Benign 1b.Malignant

Figure 1: 1a. Benign and 1b. Malignant Skin Mole Images

Background

There is a daily increase in the number of people who pass away due to this disease, even though skin cancer is widely regarded as one of the most lethal forms of cancer globally [19,20]. In addition, it is also one of the types of cancer that can spread the quickest [21]. Nevertheless, if it is caught in its initial stages, there is a chance that it can be treated [22]. Recent research has shown that twenty percent of cases of skin cancer have progressed to the point where survivability is no longer a possibility due to the advancement of the disease [23]. Around the world, around 50,000 people lose their lives to skin cancer annually [23,24]. This accounts for approximately 0.7 per cent of the overall death rate caused by the disease [24]. It is impossible to treat this condition considering the estimated cost of around 30 million US dollars, which makes therapy impossible [21].

Several machine learning classification algorithms, both shallow and deep, have been used on MRI and other images for skin cancer detection with good results. While Artificial Neural Networks have been utilized [27] to deal with this challenge, several other machine learning algorithms have been employed in studies [28, 29] to accomplish the same objective. Approaches based on computer vision have been extremely important in many earlier bodies of literature. These methods rely heavily on conventional computer vision techniques to extract various features, including shape, color, and texture [31, 32]. These are then fed into a classifier such as a support vector machine [23]. Jantu et al. [11] used several machine learning classifier algorithms including decision trees, support vector machines and neural networks to train and classify images as benign or malignant based on texture analysis of features extracted from soft tissue tumors in T1-MRI images. Using support vector machines, they were able to achieve an accuracy of

93% as opposed to a trained radiologist accuracy of 90%. Qing Guan et al. [6] have successfully used a deep learning VGG-16 convolution neural network to classify an image as a papillary thyroid carcinoma or a benign thyroid module. Using characteristics of the cell nucleus such as contours, perimeter, area and mean pixel intensity, the authors achieved an accuracy of 97.66%.

An aggregation of classic classification methods has also been used, such as support vector machine (SVM) and k-nearest neighbors, that are fed by these characteristics (KNN). In their study [25], Jain et al. examined 6 different transfer machine learning to classify multiclass lesions. Nevertheless, the results they published were contingent on the augmented dataset to achieve a larger size. The purpose of augmentation is often to make changes to the image pixels without duplicating them. Therefore, producing numerous enhanced duplicates of the same picture in the data may result in biased findings that do not accurately reflect the actual performances [26].

A comprehensive review of these publications revealed that no comparison exists in literature that compares the commonly used classical machine learning algorithms with a state-of-the-art deep learning algorithm. Our hypothesis was that the deep learning convolution neural network algorithm would be significantly better than the manual method as well as the classical machine learning methods.

Design

In this five-arm assessment on 460 images used for training and 350 images used for testing and obtained from the ISIC database, the accuracy of determination of a malignant skin mole was the primary endpoint. Four shallow machine learning algorithms (logistic regression, support vector machines, decision tree and artificial neural network) were compared to each other and to one deep machine learning algorithm (VGG-16 Convolution neural network) and all the five arms were then compared to the historical rate of the accuracy of detection performed by an experienced physician.

Research Methodology

Sample Size Selection

A sample size of 460 images, of which 230 were benign and 230 malignant were preprocessed and randomly selected from the ISIC database as shown in figure 2a and 2b. An independent set of 350 images of which 300 were benign and 50 malignant were randomly selected for testing the five algorithms; this ratio of 6 to 1 benign to malignant images was chosen to replicate the actual prevalence of malignant to benign moles found in literature [1].

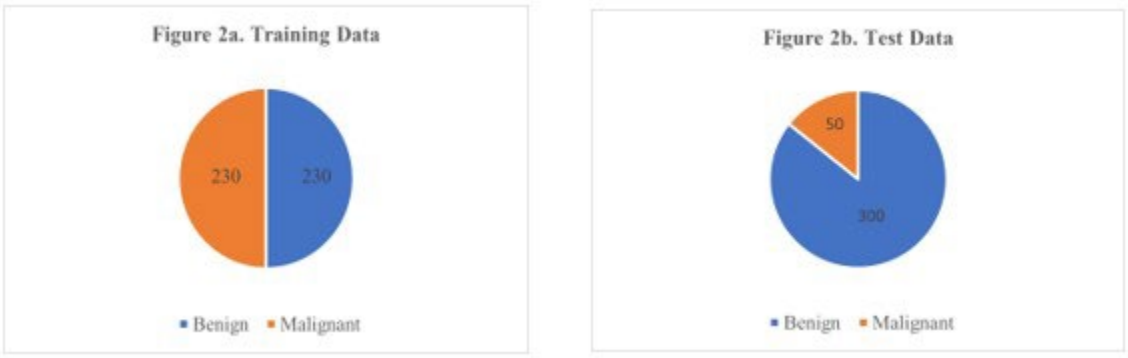


Figure 2: Number of Images Used for Training and Testing in classical and deep learning models

Exclusion Criteria

The 460 images used for training and 350 images used for testing were randomly chosen from over 2000 images in the database. The only criterion for exclusion was if the cell was devoid of a nucleus.

Ethics

Since the images were anonymized by the ISIC database, there were no concerns related to the anonymity of data in accordance with US HIPAA or GDPR regulations.

Procedure

The images were obtained from the ISIC Database in jpeg format for preprocessing. An 8th generation Intel Core-i5-8250U CPU was used to implement the shallow algorithms using MATLAB Version 2019. The classification learner app in MATLAB was used for implementing the Decision tree, Support Vector Machine, Logistic Regression and Artificial Neural Network algorithms. The VGG-16 deep learning algorithm was implemented on Google Colab server using Python version 3.7 in Keras.

Classical Machine Learning Algorithm Architecture

The decision tree algorithm uses nodes and sub-nodes for making decisions using probability classes. Logistic regression uses linear weights and biases along with a sigmoid activation function for classification. Support vector machines separate data using decision boundaries and look for the most optimal distance along the support vectors, while the artificial neural network uses multiple hidden layers in addition to an input and an output layer as well as an activation function. The 'tansig' activation function was used for decision making in the neural network. All four shallow algorithms use forward and backward propagation and the labeled data to implement a supervised algorithmic solution. The test images were then used for prediction using the predict function and a center cutoff threshold to classify images as 'benign' or 'malignant'. The resulting predictions were compared to labelled data to estimate the accuracy, sensitivity, specificity and the confusion matrix.

VGG-16 Network architecture

The architecture of VGG-16 is shown in table 1 below. It uses 13 convolution layers and 3 fully connected layers. The convolutional layers in VGG-16 are all 3×3 convolutional layers with a stride size of 1 and the same padding, and the pooling layers are all 2×2 pooling layers with a stride size of 2. The default input image size of VGG-16 is 224×224. After each pooling layer, the size of the feature map is reduced by half. The last feature map before the fully connected layers is 7×7 with 512 channels and it is expanded into a vector with 25,088 (7×7×512) channels.

Table 1: Architecture of VGG-16 network

Layer	Patch size	Input size
conv×2	3×3/1	3×224×224
pool	2×2	64×224×224
conv×2	3×3/1	64×112×112
pool	2×2	128×112×112
conv×3	3×3/1	128×56×56
pool	2×2	256×56×56
conv×3	3×3/1	256×28×28
pool	2×2	512×28×28
conv×3	3×3/1	512×14×14
pool	2×2	512×14×14
fc	25088×4096	25088
fc	4096×4096	4096
fc	4096×2	4096
Softmax	classifier	2

Note: The architecture of VGG-16 is shown in Table 1 above; it uses 13 convolutional layers and 3 fully connected layers. The convolutional layers in VGG-16 are all 3×3 convolutional layers with a stride size of 1 and the same padding, and the pooling layers are all 2×2 pooling layers with a stride size of 2. The default input image size of VGG-16 is 224×224. After each pooling layer, the size of the feature map is reduced by half. The last feature map before the fully connected layers is 7×7 with 512 channels and it is expanded into a vector with 25,088 (7×7×512) channels.

For shallow machine learning methods, the procedure consisted of the data preparation phase, where the training and test images were converted to a double precision format with the “image to double precision” function in MATLAB. The double precision training data were then trained and validated using in built five-fold cross validation in the classification apps.

For deep learning VGG-16 method, the data preparation phase consisted of converting the images to the 224x224x3 format, the standard input for the inbuilt VGG 16 function (keras in Tensorflow). The model was then compiled and fitted, before being given inputs. The training and validation were then implemented using the convolution filtering and pooling process.

Using a model predict function, the test image data were input to get the predicted condition of benign or malignant. The resulting predictions were compared to labeled data to estimate the accuracy, sensitivity, sensitivity and the confusion matrix.

Statistical Methods

The primary endpoint of the study was accuracy of malignant skin mole detection. The confusion matrix, sensitivity and specificity of the four shallow machine learning algorithms and the VGG-16 deep learning algorithm were also

computed. These were compared to each other and to the gold standard historical rate of 78.3% for accuracy of detection performed by an experienced physician. P-values were used to determine significant difference between the various methods. A disease prevalence rate of 16% [1] was assumed based on published data for rate of skin cancer existence. An exploratory endpoint of computational time was used. The computational time was compared between the five methods for convenience of use in a home or professional setting.

Results

Population Summary

The 460 images used for training obtained from the ISIC database were blinded for male and female populations. An equal number of benign and malignant images (230 each) were used to ensure training is effective for both the outcomes. The 350 images used for training obtained from the ISIC database were also blinded for male and female populations. A six to one ratio (300:50) of benign to malignant independent images were chosen to simulate the actual prevalence rate.

Accuracy, Sensitivity and Specificity Comparison

Table 2 below shows the confusion matrix with the number of moles detected correctly (true positives, true negatives) and moles not detected correctly (false positives, false negatives).

Table 2: Confusion Matrix

True Positives	False Negatives
DT(16)	DT(34)
SVM(16)	SVM(34)
LR(15)	LR(35)
NN(15)	NN(35)
VGG-16 (44)	VGG-16 (6)
True Negatives	**False Positives**
DT(268)	DT(32)
SVM(267)	SVM(33)
LR(267)	LR(33)
NN(268)	NN(32)
VGG-16 (297)	VGG-16 (3)

Note 1: DT- Decision Tree, SVM- Support Vector Machine, LR- Logistic Regression, NN- Neural Network, VGG-16 – Deep Learning

Note 2: The prevalence was assumed at 1.25%, based on 3 M cases of skin cancer[1] in the US population over 300 M giving a prevalence rate of 1.25%.

Accuracy is defined as the rate of correct detection of both kinds of moles, malignant and benign. The accuracy of the shallow learning algorithms ranged in the 88 to 89 percent range as shown in table 2 and figure 3, while the accuracy of the VGG-16 CNN was significantly higher (98.86%, $p < 0.0001$) than both the shallow learning algorithms as well as the physician detection rate of 78.3%. The 95% confidence range varied between 84.42% to 91.75%

for the shallow algorithms and between 97.11% to 99.69% for the VGG16 deep learning algorithm as shown in table 3.

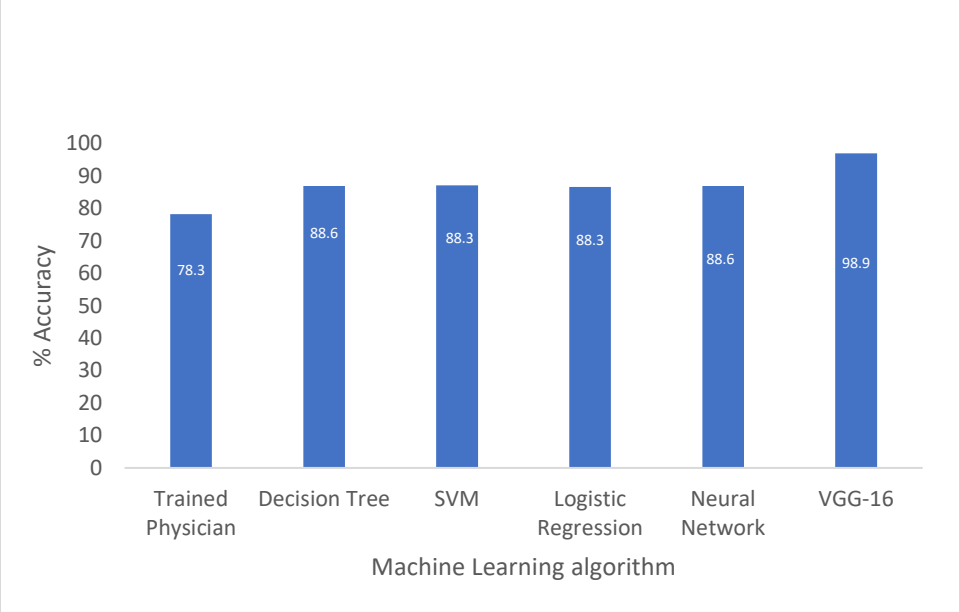


Figure 3: P-value for significance in statistical difference of Accuracy between CNN (VGG-16) and manual detection was $p<0.05$.

Sensitivity is defined as the rate of correct detection of malignant moles, while specificity is defined as the rate of correct detection of benign moles. The sensitivity of the shallow learning algorithms ranged in the 30.00 to 32.00 percent range as shown in table 3, while the sensitivity of the VGG-16 CNN was significantly higher (88%, $p <0.0001$) than the shallow learning algorithms. The specificity of the shallow learning algorithms ranged in the 89.00% to 89.33% percent range as shown in table 3, while the specificity of the VGG-16 CNN was significantly higher (99%, $p <0.05$) to the shallow ML methods. The 95% confidence range varied between 84.90% to 92.59% for the shallow methods, while it was between 97.11% and 99.79% for the deep learning algorithm as shown in table 4.

Table 3: Accuracy, Sensitivity and Specificity Using Absolute Values

	SVM	Logistic regression	Decision Tree	Neural network	VGG-16
Accuracy	88.29%	88.26%	88.62%	88.59%	98.86%
Sensitivity	32.00%	30.00%	32.00%	30.00%	88%
Specificity	89.00%	89.00%	89.33%	89.33%	99%

Table 4: Accuracy, Sensitivity and Specificity using 95% Confidence Intervals

	SVM	Logistic regression	Decision Tree	Neural network	VGG-16
Accuracy	84.45% to 91.46%	84.42% to 91.44%	84.81% to 91.75%	84.79% to 91.73%	97.11% to 99.69%
Sensitivity	19.52% to 46.70%	17.86% to 44.61%	19.52% to 46.70%	17.86% to 44.61%	75.69% to 95.47%
Specificity	84.90% to 92.31%	84.90% to 92.31%	85.28% to 92.59%	85.28% to 92.59%	97.11% to 99.79%

The computational time was the lowest for the decision tree method in comparison to the support vector machine, logistic regression of the neural network method. The VGG16 deep learning algorithm was run on a Google COLAB server and gave a computational time of 38.6 second which was comparable to the decision tree computational time of 34.3 seconds. The logistic regression and SVM had a computational time of over 180 seconds, while the neural network had a computational time of over 20 minutes. As the size of the training data increases, the computational time will play a major role in practical applicability.

Discussion

The four classical machine learning algorithms were used since they have well known architectures, have been found to be effective in image classification and have been extensively used in literature over the last several decades.

While the 95 % confidence range for specificity was in the high 84.90% to 92.59% range, the 95% confidence range for sensitivity varied between 17.86% to 46.70% for the classical methods as shown in table 4. The lower percent values in the 30% range for the classical methods was primarily due to the lower accuracy combined with a smaller sample size of only 50 for the malignant moles. A small sample size is very sensitive to a minute variation in accuracy. The improved accuracy of the VGG-16 deep learning algorithm increased the sensitivity in the 75.69% to 95.47% range, validating our hypothesis.

Conclusion

In summary, this research compared the accuracy of a VGG-16 deep learning algorithm to four classical machine learning algorithms as well as manual detection for automated detection of malignant skin cancer moles. The deep learning VGG-16 convolution neural network (CNN) algorithm was superior to the classical learning algorithms as well as to manual detection. The computational time for the CNN algorithm was also reasonable enough for practical use. The CNN could be used as a second method of diagnosis to increase confidence in the clinic. With the considerable increase in computation power in smart phones and other smart devices, a smart application could have a high utility value to do screening at home or for transmission of the image detection results to the user's physician using telemetry. With a 99% survival rate for early detection of patients with malignant moles, such a technology would not only increase the survival rate but could also reduce healthcare costs.

Limitations and future work

One limitation of this research is the relatively small sample size used for training of the skin cancer moles. A larger dataset consisting of thousands of images could further improve the accuracy of detection. These results should also be further validated in the future using a larger dataset and with the more recent RESNET convolution neural network architecture. Also, application of this method for other applications such as breast cancer and other skin conditions should also be explored.

Acknowledgements

I would like to thank my teacher Ms. Catherine Phillips at Shrewsbury High School and Rajesh Kasbekar for providing resources, guidance, and funding for this research.

References

[1] American Academy of Dermatology. Skin Cancer (2020). https://www.aad.org/media/stats-skin-cancer

[2] Brazier, Y. (2019). What are the different types of tumor? https://www.medicalnewstoday.com/articles/24291

[3] Brownlee, J. (2017). What is the Difference Between Test and Validation Datasets? https://machinelearningmastery.com/difference-test-validation-datasets/

[4] Diachiara, T. (2019). How Can You Tell If It's a Mole or Skin Cancer? https://www.verywell.com/moles-vs-melanoma-skin-cancer-identification-gallery-3010883

[5] Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn & TensorFlow. Sebastopol, CA: O'Reilly Media Inc.

[6] Guan, Q., Wang, Y., Ping, B., Li, D., Du, J., Qin, Y., Lu, H., Wan, X., & Xiang, J. (2019). Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study. *Journal of Cancer, 10(20), 4876–4882. https://doi.org/10.7150/jca.28769*

[7] Guy GP, Thomas CC, Thompson T, Watson M, Massetti GM, Richardson LC. (2015). Vital signs: Melanoma incidence and mortality trends and projections—United States, 1982–2030. MMWR Morb Mortal Wkly Rep. 64(21):591-596.

[8] Guy, G., P., Machlin, S., R., Ekwueme, D., U., Yabroff, K., R. (2014). Prevalence and Costs of Skin Cancer Treatment in the U.S., 2002-2006 and 2007-2011. *American Journal of Preventive Medicine, 48, (183-187).*

[9] Haenssle, H., Fink, C., Schneiderbauer, R., Toberer, F., (2018). Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists, Annals of Oncology. 29(1) DOI:10.1093/annonc/mdy166

[10] Hassanien, A, E., Bhatnagar, R., Darwish, A., (2020). Advanced Machine Learning Technologies and Applications: Proceedings of Advanced Machine Learning Technologies and Applications (AMLTA). 2020. Singapore: Springer Singapore.

[11] Jantu, J. S., Sijbers, J., De Backer, S., Rajan, J., Van Dyck, D. (2010). Machine learning study of several classifiers trained with texture analysis features to differentiate benign from malignant soft-tissue tumors in T1-MRI images. *Journal of Magnetic Resonance Imaging, 31, 680-689.*

[12] Logistic Regression for Machine Learning and Classification. (2019). https://kambria.io/blog/logistic-regression-for-machine-learning/

[13] Long, M. (2019). Medical Imaging Glossary. https://www.aidoc.com/blog/medical-imaging-ai-glossary/

[14] Memorial Sloan Kettering Cancer Center (2020). Skin Cancer https://www.mskcc.org/cancer-care/types/skin

[15] Nnama, H. (2017). Terminal Stages of Cancer. https://healthfully.com/269063-terminal-stages-of-cancer.html

[16] Pietrangelo, E., K. (2019). Benign and Malignant Tumors: How Do They Differ? https://www.healthline.com/health/cancer/difference-between-benign-and-malignant-tumors

[17] Priya, R., Aruna, P. (2013). Diagnosis of Diabetic Retinopathy Using Machine Learning Techniques. *ICTACT Journal on Soft Computing, 3, 563-576.*

[18] World Health Organization. Cancer Fact Sheets (2020). https:// https://www.who.int/news-room/fact-sheets/detail/cancer

[19] Li, W., Raj, A. N. J., Tjahjadi, T., & Zhuang, Z. (2021). Digital hair removal by deep learning for skin lesion segmentation. *Pattern Recognition*, *117*, 107994.

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

[21] Zhang, X. (2017). Melanoma segmentation based on deep learning. *Computer assisted surgery*, *22*(sup1), 267-277.

[22] Oliveira, R. B., Mercedes Filho, E., Ma, Z., Papa, J. P., Pereira, A. S., & Tavares, J. M. R. (2016). Computational methods for the image segmentation of pigmented skin lesions: a review. *Computer methods and programs in biomedicine*, *131*, 127-141.

[23] Fraiwan, M., & Faouri, E. (2022). On the Automatic Detection and Classification of Skin Cancer Using Deep Transfer Learning. *Sensors*, *22*(13), 4963.

[24] Mahbod, A., Schaefer, G., Wang, C., Dorffner, G., Ecker, R., & Ellinger, I. (2020). Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification. *Computer methods and programs in biomedicine*, *193*, 105475.

[25] Jain, S., Singhania, U., Tripathy, B., Nasr, E. A., Aboudaif, M. K., & Kamrani, A. K. (2021). Deep Learning-Based Transfer Learning for Classification of Skin Cancer. *Sensors*, *21*(23), 8142.

[26] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1-48.

[27] Eurosurveillance Editorial Team. (2020). Note from the editors: World Health Organization declares novel coronavirus (2019-nCoV) sixth public health emergency of international concern. *Eurosurveillance*, *25*(5), 200131e.

[28] Subramanian, R. R., Achuth, D., Kumar, P. S., kumar Reddy, K. N., Amara, S., & Chowdary, A. S. (2021, January). Skin cancer classification using Convolutional neural networks. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 13-19). IEEE.

[29] Wolff, K., Johnson, R. A., Saavedra, A. P., & Roh, E. K. (2017). Fitzpatrick's color atlas and synopsis of clinical dermatology (; K. G. Edmonson, R. Pancotti, & C. Yoo, Eds.). *USA: McGrawHill Companies Inc*.

[30] Celebi, M. E., Kingravi, H. A., Uddin, B., Iyatomi, H., Aslandogan, Y. A., Stoecker, W. V., & Moss, R. H. (2007). A methodological approach to the classification of dermoscopy images. *Computerized Medical imaging and graphics*, *31*(6), 362-373.

[31] Oliveira, R. B., Mercedes Filho, E., Ma, Z., Papa, J. P., Pereira, A. S., & Tavares, J. M. R. (2016). Computational methods for the image segmentation of pigmented skin lesions: a review. *Computer methods and programs in biomedicine*, *131*, 127-141.