THE IMPACT OF USING DIFFERENT COLOR SPACES IN HISTOLOGICAL IMAGE CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS

Ronny Velastegui and Marius Pedersen

Department of Computer Science, Norwegian University of Science and Technology

ABSTRACT

Classification is an important aspect of medical image analysis. Nowadays, Convolutional Neural Networks (CNNs) are extensively used in the field of medical image classification. There are several kinds of research on medical image classification combining different CNN architectures and data sets. When using color image data sets, most of those research works use RGB as the standard color space to train and test the models. While RGB is a standard color space to represent images on multimedia devices, RGB might not be the best color space to train CNN models for medical image classification applications. We implement an AlexNet CNN to classify colon tissue images to detect tumors. We perform this task using several color spaces, such as RGB, XYZ, CIELAB, HSV, and YCbCr. We analyze the results and indicate which color spaces give the best accuracy in performing this medical image classification task.

***Index Terms*—** Medical Image Classification, AlexNet, Convolutional Neural Networks, Color Spaces

1. INTRODUCTION

Image classification has been a crucial factor in many medical imaging applications [1]. This is because, based on medical images such as lung x-rays, colon tissue images, skin photos, or skull x-ray, an experienced physician can evaluate and detect abnormalities and potential diseases with precision [2]. For example, based on colon tissue images, a gastroenterologist could detect if a patient has a tumor or not [3].

In most cases, due to a large number of patient images to be analyzed added to the lack of well-experienced physicians, this task becomes very difficult to complete efficiently [4]. One solution would be to hire more doctors to analyze the images, but this would be expensive for some hospitals or care centers. Even if we talk about pandemic situations such as COVID-19, where the lung x-rays of infected patients reached thousands per day [5], it is practically impossible for any care center to reach a sufficient number of doctors to supply the analysis of this huge number of medical images [6].

It is for this reason that ways to automate the medical image classification process have been sought. Thus, this search led to the implementation of traditional machine learning methods to complete this task, but the precision and efficiency achieved with those methods were very inferior compared to an expert physician [7]. In the last few years, it is with the implementation of deep learning, specifically with Convolutional Neural Networks (CNNs), that advances in the classification of medical images are achieved [8]. This is due to the effectiveness of these networks to extract characteristics from the analyzed images in a similar way to the functioning of the visual cortex of humans [9].

One of the most popular CNN architectures implemented for image classification is Alexnet [10]. The advantages that Alexnet brought versus other machine learning methods were its high precision and good efficiency when being trained using GPUs [11]. Because of this, in 2012, Alexnet showed excellent results in the ImageNet Large Scale Visual Recognition Challenge [12]. The original architecture of Alexnet is formed by eight layers, five of them are convolutional layers combined with pooling operations, and the last three are fully connected layers [13]. Although today, there are several deeper and more complex modern CNN architectures which, through a long training time, are capable of surpassing the precision of Alexnet, Alexnet is still widely used in many image classification types of research due to its simple architecture and quick training [14].

Although there is research on medical image classification using CNN, they mostly focus on different data sets and architectures [15]. Most of the time, when using color image data sets, they share a common point which is the use of images represented in the RGB color space to train and test the model [16]. RGB is a standard color space to manipulate images on displays, cameras, projectors, etc [17]. But, we cannot assure that RGB is the best color space to train CNN models for image classification tasks, especially in the medical field. There is research on image classification using CNNs using other color spaces, such as HSI or CIELAB. These show better results compared with the use of RGB [18]. These have focused on different image classification areas, such as face classification [19], fruit classification [20], scene classification [18], etc. As far as we know, currently, there is no detailed research about the influence of different color spaces in medical image classification using CNNs.

In this work, we implement an AlexNet CNN to classify colon tissue images to detect tumors. We perform this task using several color spaces, such as RGB, XYZ, CIELAB, HSV, and YCbCr. We analyze the results and indicate which color spaces give the best accuracy in performing this medical image classification task.

2. RELATED WORKS

Currently, as far as we know, there is little research on the influence of using different color spaces to perform CNN-based medical image classification. We will mention some related works in which similar ideas are implemented for different image classification applications.

Osadebey et al. [21] investigated the discrimination power of five different color spaces, and a total of 16 color channels, for two unsupervised approaches and a deep learning approach on the segmentation of skin lesions in dermatoscopy images. The results indicated that depending on the task, no single color space could be found for optimal results.

Castro et al. [20] implemented and compared different methods to classify fruit images according to their ripeness level. They compare four different machine learning approaches combined with three different color spaces. In the results, they found that the models that obtain the best accuracy performing this task were: artificial neural networks and support vector machines. Also, they realized that both models are prone to their precision being affected according to the color space used. This is why both models obtain a high accuracy using CIELAB color space, but exactly the same models obtain a low accuracy using the other color spaces.

Rajan et al. [18] used a pre-trained CNN to classify a scene image data set, using different color spaces and intensity planes. The objective of this work was to find out which color space or intensity plane allows obtaining the highest accuracy. In the results, they found that there is no unique color space or intensity plane which obtains the best accuracy classifying all the classes in this data set. Each of the eight classes is classified better than others in particular color spaces or intensity planes. They conclude that the impact of choosing one color space or other is very important for image classification.

Cely et al. [22] explored the impact of color spaces in the transfer learning of CNNs. They analyzed the behavior of some pre-trained CNN at the time of classifying images in a new data set that contained images of cats and dogs. With the difference that this new data set to be classified will be previously transformed into other color spaces, and not only in the RGB space. They found each one of the three CNN architectures reaches different accuracy in different color spaces.

Gowda and Yuan [23] investigated the influence of color spaces on CNNs to perform image multi-class classification of the CIFAR-10 data set. They also proposed a CNN architecture for image classification that combined seven different color spaces at the same time. They found that the proposed architecture obtained slightly better accuracy than other CNN architectures that use just the RGB color space.

3. METHODOLOGY

First, we define the medical image data used. Then, we define the color spaces used to represent that data set. Finally, we define the CNN architecture and configuration that we use for image classification.

3.1. Data Set

The data set that will be used in this work is formed by a collection of 7180 histological images obtained from 50 patients diagnosed with Colorectal Adenocarcinoma [24]. This data set will be divided into two parts, 60% of the set will be used for training and the remaining 40% for testing. The images have a size of 224x224 pixels and are represented in the RGB color space. While there are many medical image data sets that can be used to perform our experiment, we chose this one because the Colorectal Adenocarcinoma is one of the most common types of cancer nowadays [25]. The general procedure of our experiment would be very similar to other types of medical image data sets.

The data set contains images of colorectal cancer and healthy tissue divided into nine categories; Adipose (ADI), Background (BACK), Debris (DEB), Lymphocytes (LYM), Mucus (MUC), Smooth Muscle (MUS), Normal Colon Mucosa (NORM), Cancer-associated Stroma (STR), and Colorectal Adenocarcinoma Epithelium (TUM). Each image was assigned to one category by doctors. The last category can be seen as the most important because it shows the presence of colorectal cancer. Figure 1 shows samples of each category.

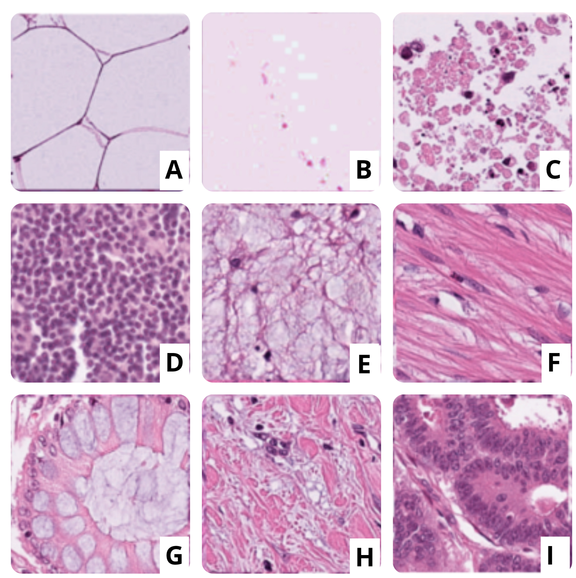


Fig. 1: Examples of each category in the data set. A) Adipose, B) Background, C) Debris, D) Lymphocytes, E) Mucus, F) Smooth Muscle, G) Normal Colon Mucosa, H) Cancer-associated Stroma, I) Colorectal Adenocarcinoma.

3.2. Color Spaces

By default, like most image classification data sets, our data set is represented using the RGB color space. For this experiment, it will be converted to seven different color spaces. The first four color spaces are CIEXYZ, HSV, YCbCr, and CIELAB. In addition, we include duplicates for the L, A, and B channels of CIELAB, being LLL, AAA, and BBB. Thus, we will also be able to explore the influence of each CIELAB component on the overall performance of our image classification model. In this way, a total of eight different versions of the original data set will be used and compared during the training and testing of the CNN classification model.

3.3. CNN Architecture

The CNN architecture we use is Alexnet [9]. Because the objective of this work encompasses the analysis of CNN architectures in a general way, the use of a simple and general CNN, such as Alexnet, is a good option. It is composed of eight principal layers. The first five layers are convolutional layers combined with pooling operations. The last three layers are fully connected layers. The activation function used is ReLU. We use the Stochastic Gradient Descent with Momentum (SGDM) optimizer and back-propagation learning algorithm to train our model. Both the implementation of the CNN architecture, as well as the manipulation of the color spaces, have been carried out using MATLAB R2020a.

To train our classification model we use two different approaches. The difference between them is the presence of transfer learning. In the first approach, we will use transfer learning to initialize the weights before the training phase. While in the second approach, all the weights of the network will be initialized with random values before the training phase. For this work, it is important to implement both training approaches. Since, in most cases, when implementing transfer learning, the new model receives knowledge from an architecture previously trained using an image data set represented in the RGB color space. This situation could affect negatively our experiment since we will not only use the RGB color space but also different ones. To compare all the results, we have set a maximum number of training iterations equal to 650. This is approximately the number of iterations in which the original RGB version converges to good prediction accuracy, so it will serve as a basis for comparison.

4. RESULTS

First, we will show all the results obtained using the transfer learning. Then, we will show all the results obtained training from scratch, that is without transfer learning.

4.1. With Transfer Learning

Figure 2 shows the model training process using transfer learning and RGB color space. In the blue graph, we can see how the model reaches a high validation accuracy of 92%. Note that the training graphs obtained with the other color spaces have been omitted due to the similarity with RGB.

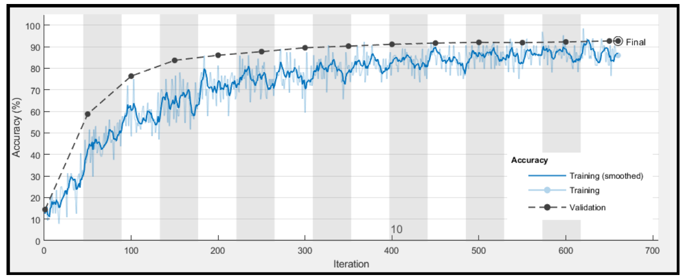


Fig. 2: Classification accuracy evolution during the training of the model based on transfer learning and RGB color space.

After the training stage, the final accuracy of the model was evaluated using the test data, and the results were summarized in a confusion matrix. Figure 3 shows the confusion matrix for each of the nine classes for the RGB color space. The test accuracy using RGB is 92.5% for the overall classification case and 90.7% for the tumor class classification.

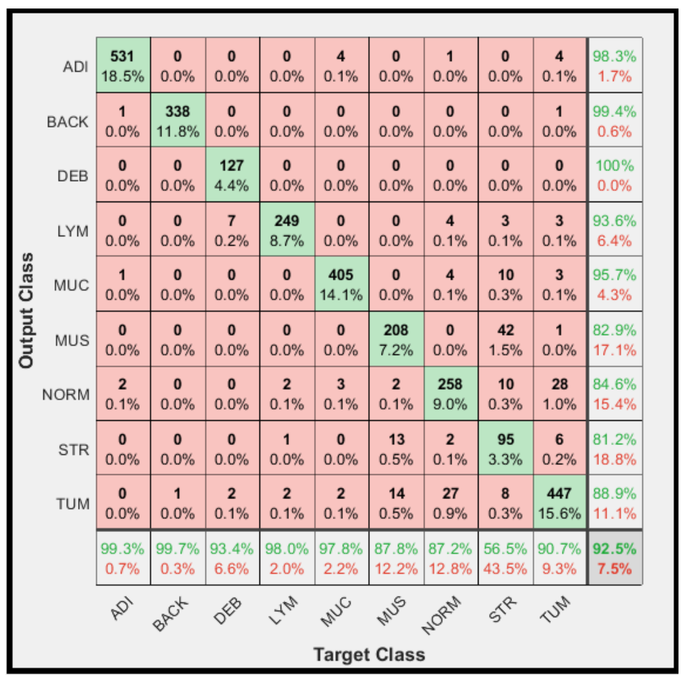


Fig. 3: Confusion matrix obtained after the test process of the model based on transfer learning and RGB color space.

The results of the training and testing process with all the other seven color spaces and color channels are shown in Table 1. In the last column, we can see the overall test accuracy of the eight color spaces. We can see that the model that obtains the highest overall accuracy of 93.0% is CIELAB, followed by RGB and XYZ models with 92.5%. The lowest overall accuracy of 90.3% was obtained using the BBB plane. Although the difference in the overall classification accuracy may seem small, in the case of accuracy per class, the difference is more noticeable.

Table 1: Classification accuracy of each of the eight spaces based on transfer learning.

COLOR SPACE	Classification accuracy per class (%)									OVERALL ACCURACY
	ADI	BACK	DEB	LYM	MUC	MUS	NORM	STR	TUM	
RGB	99.3	99.7	93.4	98.0	97.8	87.8	87.2	56.5	90.7	92.5
CIELAB	99.4	100.0	92.6	98.0	98.1	92.0	81.8	62.5	92.3	93.0
HSV	99.4	99.4	94.1	98.4	96.9	91.6	83.1	48.2	90.5	91.9
YCbCr	99.4	99.7	90.4	98.4	95.9	94.5	82.8	47.0	89.7	91.6
XYZ	99.4	99.1	91.9	98.4	96.6	93.2	86.1	57.1	89.9	92.5
LLL	99.4	98.8	90.4	98.4	96.9	92.0	83.1	60.7	90.7	92.4
AAA	98.9	99.1	92.6	98.8	96.6	91.1	76.4	59.5	92.3	91.9
BBB	99.8	99.7	87.5	98.8	98.1	87.8	80.4	37.5	88.2	90.3

In the DEB (Debris) class, the most accurate space was HSV with 94.1%, while the least accurate model was BBB with 87.5%. For the MUC (Mucus) class, the most accurate models were LAB and BBB with 98.1%, while the least accurate model was YCbCr with 95.9%. For the MUS (Smooth Muscle) class, the most accurate model was YCbCr with 94.5%, while the least accurate models were RGB and BBB with 87.8%. For the NORM (Normal) class, the most accurate model was RGB with 87.2%, while the least accurate model was AAA with 76.4%. For the STR (Cancer-associated Stroma) class, the most accurate model was LAB with 62.5%, while the least accurate model was BBB with 37.5%. For the TUM (Colorectal Adenocarcinoma Epithelium) class, the most accurate models were LAB and AAA with 92.3%, while the least accurate model was BBB with 88.2%. In the case of the ADI (Adipose), BACK (Background), and LYM (Lymphocytes) classes, the accuracy obtained with all the methods is similar, with a difference of less than 1% approximately.

Some classes are classified in a more accurate way when they are represented in certain color spaces. The original RGB color space obtained the highest accuracy only in the case of the NORM class, while in the other classes, the precision when using the RGB space was lower than other color spaces. Also, we can see that the CIELAB color space led to the highest classification accuracy in some classes, including the TUM class. Finally, as we already mentioned before, the CIELAB color space led to a slightly better overall classification accuracy. This comparison can be better seen in Figure 4.

4.2. Without Transfer Learning

For this case, we completed the training and testing process similar to the previous section. It is important to note that the final precision of this model is lower than that obtained in the previous section with transfer learning. The absence of transfer learning leads to the need for a much larger number of iterations to achieve a classification accuracy similar to the previous one. However, since our objective is to evaluate the performance of both models, we kept the maximum limit of iterations the same as in the previous section, which is 650.

Table 2 summarizes the classification accuracy of each class, as well as the overall test accuracy, using different color space models. We can see that the model that obtains the highest overall accuracy of 79% is CIELAB, followed by the HSV with 78.3%. The lowest overall accuracy of 76.4% was obtained using BBB. Although the difference in the overall classification accuracy may seem small, in the case of accuracy per class, the difference is more noticeable.

Table 2: Classification accuracy of each of the eight spaces trained without transfer learning.

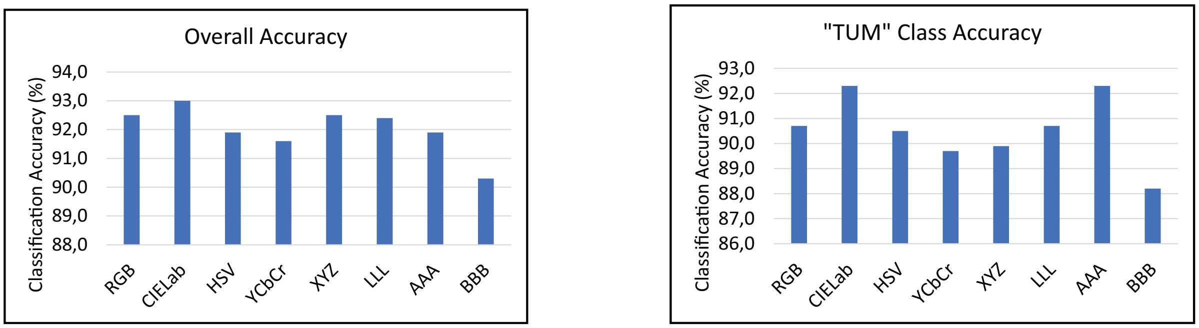
COLOR SPACE	Classification accuracy per class (%)									OVERALL ACCURACY
	ADI	BACK	DEB	LYM	MUC	MUS	NORM	STR	TUM	
RGB	98.5	100.0	50.0	83.9	79.7	90.7	33.1	10.1	79.3	76.5
CIELAB	99.4	100.0	50.0	83.5	82.9	87.3	54.1	5.4	80.9	79.0
HSV	98.9	100.0	50.0	87.4	80.2	92.8	45.6	5.4	79.9	78.3
YCbCr	98.5	100.0	50.0	81.1	83.6	89.0	45.6	6.0	78.9	77.7
XYZ	99.1	100.0	50.7	89.0	80.7	90.3	37.5	0.6	77.3	76.8
LLL	98.5	100.0	50.0	83.9	79.7	90.7	33.1	10.1	79.3	76.5
AAA	98.5	100.0	50.0	83.9	79.7	90.7	33.1	10.1	79.3	76.5
BBB	99.1	100.0	50.0	82.3	85.3	84.4	31.1	5.4	80.1	76.4

In the LYM class, the most accurate was XYZ with 89.0%, while the least accurate was YCbCr with 81.1%. For the MUC class, the most accurate model was BBB with, 85.3%, while the least accurate models were RGB, LLL, and AAA with 79.7%. For the MUS class, the most accurate model was HSV with 92.8%, while the least accurate model was BBB with 84.4%. For the TUM class, the most accurate model was LAB with 80.9%, while the least accurate model was XYZ with 77.3%. In the case of the ADI, BACK, and DEB classes, the accuracy obtained by all the methods is similar, with a difference of less than 1%. Finally, in the case of NORM and STR classes, a low precision was achieved by all the models. This may mean that both classes are possibly the most difficult to classify correctly and therefore require a very large number of iterations during training.

In a similar way to the previous method based on transfer learning, we can also see that some classes are classified in a more accurate way when they are represented in certain color spaces. For example, the LYM, MUC, and MUS classes are classified with greater precision when they are represented in the XYZ, BBB, and HSV models respectively. Finally, the CIELAB color space led to the highest classification accuracy in the TUM class, as well as the general classification accuracy. We can see this comparison in Figure 5.

5. CONCLUSION

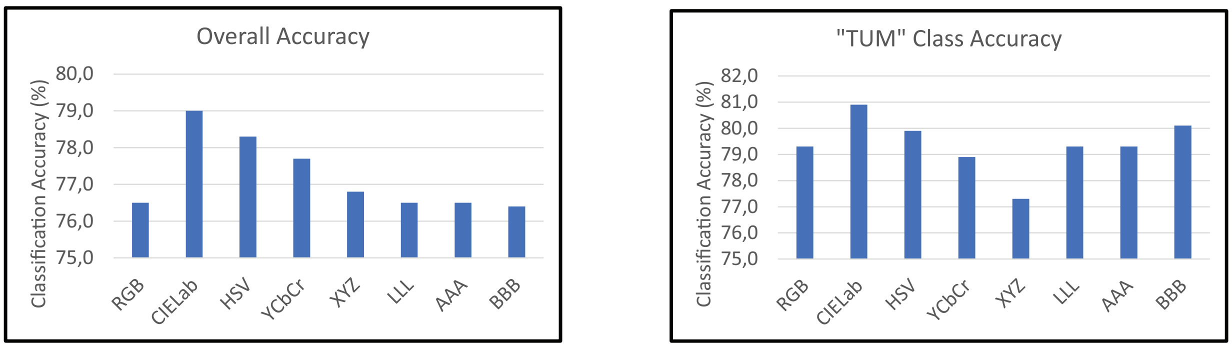
We have investigated the influence of different color spaces in a medical image classification task using CNN. We have implemented AlexNet to classify colon tissue images to detect tumors. We have trained our model using a data set composed of 7180 histological images, divided into nine different



(a) Overall accuracy

(b) TUM class accuracy

Fig. 4: Classification accuracy of the eight color spaces based on transfer learning.



(a) Overall accuracy

(b) TUM class accuracy

Fig. 5: Classification accuracy of the eight color spaces without transfer learning.

classes, obtained from 50 patients diagnosed with Colorectal Adenocarcinoma which is the most common type of colon and rectal cancer.

We have performed this task using five different color spaces and three color channels. The color spaces evaluated were: RGB, XYZ, CIELAB, HSV, and YCbCr. The channels were: LLL, AAA, and BBB, which represent each component of the CIELAB color space. For the training stage, two different CNN training approaches were taken into account. The first approach used transfer learning, while the second approach used training from scratch. Both approaches were implemented to determine the influence of color spaces in both cases independently.

The results showed that there is no color space that classifies all the classes with the highest accuracy, but each class is classified with higher accuracy in certain color spaces. Also, it was shown that the original RGB space, in which the image data set was represented by default, did not lead to the highest classification accuracy. Instead, it was observed that the highest classification accuracy, both in the overall case and in the tumor case, was obtained using the CIELAB color space.

As future work, one should evaluate other CNN architectures, as well as other medical image data sets. In this way, the results obtained could be implemented to improve other medical image classification applications.

6. REFERENCES

[1] Atam P Dhawan, *Medical image analysis*, vol. 31, John Wiley & Sons, 2011.

[2] Hiam Alquran, Isam Abu Qasmieh, Ali Mohammad Alqudah, Sajidah Alhammouri, Esraa Alawneh, Ammar Abughazaleh, and Firas Hasayen, "The melanoma skin cancer detection and classification using support vector machine," in *2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*. IEEE, 2017, pp. 1–5.

[3] Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Medical image analysis*, vol. 18, no. 3, pp. 591–604, 2014.

[4] Dinggang Shen, Guorong Wu, and Heung-Il Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[5] Huwen Wang, Zezhou Wang, Yinqiao Dong, Ruijie Chang, Chen Xu, Xiaoyue Yu, Shuxian Zhang, Lhakpa Tsamlag, Meili Shang, Jinyan Huang, et al., "Phase-adjusted estimation of the number of coronavirus dis-

ease 2019 cases in wuhan, china," *Cell discovery*, vol. 6, no. 1, pp. 1–8, 2020.

[6] Ioannis D Apostolopoulos and Tzani A Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, vol. 43, no. 2, pp. 635–640, 2020.

[7] Michael W Kattan, Dennis A Adams, and Michael S Parks, "A comparison of machine learning with human judgment," *Journal of Management Information Systems*, vol. 9, no. 4, pp. 37–57, 1993.

[8] Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen, "Medical image classification with convolutional neural network," in *2014 13th International Conference on Control Automation Robotics & Vision*. IEEE, 2014, pp. 844–848.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[10] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari, "The history began from alexnet: A comprehensive survey on deep learning approaches," *arXiv preprint arXiv:1803.01164*, 2018.

[11] Pedro Ballester and Ricardo Matsumura Araujo, "On the performance of googlenet and alexnet applied to sketches," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 1124–1128.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[13] Yang You, Zhao Zhang, C Hsieh, James Demmel, and Kurt Keutzer, "100-epoch imagenet training with alexnet in 24 minutes," *arXiv preprint arXiv:1709.05011*, 2017.

[14] Zheng-Wu Yuan and Jun Zhang, "Feature extraction and image retrieval based on alexnet," in *Eighth International Conference on Digital Image Processing (ICDIP 2016)*. International Society for Optics and Photonics, 2016, vol. 10033, p. 100330E.

[15] Kenji Suzuki, "Overview of deep learning in medical imaging," *Radiological physics and technology*, vol. 10, no. 3, pp. 257–273, 2017.

[16] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, no. 11, pp. 226, 2018.

[17] Sabine Süsstrunk, Robert Buckley, and Steve Swen, "Standard rgb color spaces," in *Color and Imaging Conference*. Society for Imaging Science and Technology, 1999, vol. 1999, pp. 127–134.

[18] Rajan Sachin, V Sowmya, D Govind, and KP Soman, "Dependency of various color and intensity planes on cnn based image classification," in *International symposium on signal processing and intelligent recognition systems*. Springer, 2017, pp. 167–177.

[19] Vanessa Buhrmester, David Münch, Dimitri Bulatov, and Michael Arens, "Evaluating the impact of color information in deep neural networks," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2019, pp. 302–316.

[20] Wilson Castro, Jimy Oblitas, Miguel De-La-Torre, Carlos Cotrina, Karen Bazán, and Himer Avila-George, "Classification of cape gooseberry fruit according to its level of ripeness using machine learning techniques and different color spaces," *IEEE Access*, vol. 7, pp. 27389–27400, 2019.

[21] Michael Osadebey, Marius Pedersen, and Dag Waaler, "Evaluation of color spaces for unsupervised and deep learning skin lesion segmentation," in *International Conferences Computer Graphics, Visualization, Computer Vision and Image Processing 2020*, 2020, p. 8 pages.

[22] Javier Diaz-Cely, Carlos Arce-Lopera, Juan Cardona Mena, and Lina Quintero, "The effect of color channel representations on the transferability of convolutional neural networks," in *Science and Information Conference*. Springer, 2019, pp. 27–38.

[23] Shreyank N Gowda and Chun Yuan, "Colornet: Investigating the importance of color spaces for image classification," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 581–596.

[24] Jakob Nikolas Kather, Niels Halama, and Alexander Marx, "100,000 histological images of human colorectal cancer and healthy tissue," https://doi.org/10.5281/zenodo.1214456, apr 2018, Zenodo v0.1.

[25] Sri G Thrumurthy, Sasha SD Thrumurthy, Catherine E Gilbert, Paul Ross, and Amyn Haji, "Colorectal adenocarcinoma: risks, prevention and diagnosis," *bmj*, vol. 354, 2016.