Malignancy-Based Classification of CT Scan Images for Lung Cancer Patients

A Deep Learning Approach for Malignancy Prediction

(The Assignment's Report of Applied AI in Biomedicine Course)

Christian Ferrareis[1], Bahram Hedayati[2], and Kristina Tas[3]

[1]Student of Computer Science, christian.ferrareis@mail.polimi.it, 10725804

[2]Student of Telecommunication Engineering, bahram.hedayati@mail.polimi.it, 10870276

[3]Student of Biomedical Engineering, kristina.tas@mail.polimi.it, 10968804

1 Introduction

Cancer is a major social, public health, and economic problem in the 21st century, responsible for almost one in six deaths (16.8%) and one in four deaths (22. 8%) from non-communicable diseases (NCDs) worldwide [1]. The late-stage diagnosis of lung cancer significantly affects its prognosis, contributing to a low 5-year survival rate of less than 20% in most cases [2]. Traditional cancer detection relies on radiologists analyzing CT scans, which can be time-consuming and prone to human error. Early-stage lung cancer is difficult to detect because small tumors often look like benign nodules which are small masses of tissue that may indicate malignancy. Implementing lung cancer screening programs has been a crucial step toward early diagnosis and intervention. However, the effectiveness of such programs heavily relies on the accurate detection and classification of nodules [3]. Automated computer-aided detection (CAD) systems and deep learning-based approaches have shown great potential in assisting radiologists in identifying malignant nodules with high accuracy as they can detect subtle abnormalities invisible to the human eye [4].

This study proposes a deep learning-based malignancy classification model to enhance the detection of lung nodules in CT scan images. The goal is to improve early diagnosis accuracy, thereby increasing the chances of effective treatment and reducing lung cancer mortality rates. To do this, we did some standardization tasks, performed a number of pre-processing techniques, and implemented four different classifiers based on the images' type and the sort of classification including multi-class and binary classifications.

2 Materials and Methods

2.1 Machine Learning Framework

In order to achieve the goal mentioned in the introduction, we followed a standard workflow implemented in [5] and modeled four different classifiers based on the images' type and the sort of classification including multi-class and binary classifications as we were asked to do in the assignment. Figure 1 shows the workflow of the proposed framework for the detection of malignancy scores with respect to each classifier.

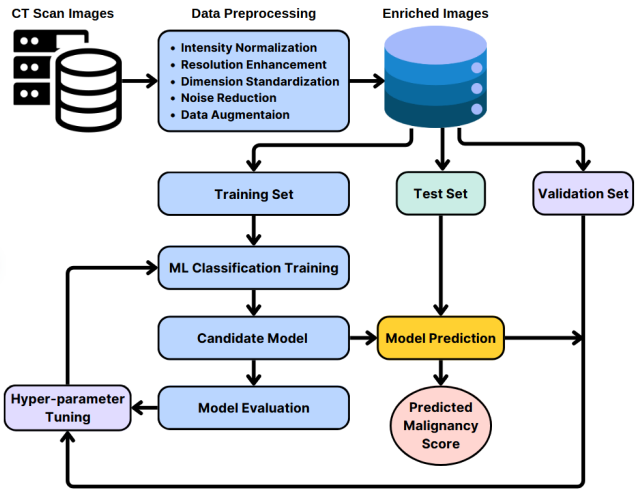

Figure 1: Work flow diagram of ML-based framework for detection of malignancy score of CT scan images

2.2 Dataset Exploration

Data exploration is crucial for understanding patterns, detecting anomalies, and ensuring data quality before analysis or modeling. The given dataset includes two categories of CT scan images per patient: full view and nodule view. Each view consists of 2363 slices of a CT scan. In general,

the output of a performed CT scan includes many slices covering a volume. The slices in the dataset are those with the largest nodule area. Each patient recognized a specific level of malignancy falling between 1 and 5. The dataset is highly unbalanced in terms of malignancy score for both multiple and binary classes as you can see in Figure 3 which will add bias to the AI model. Bias in AI systems can arise if the training data overrepresents one group and underrepresents others, leading to inaccurate predictions. AI models in healthcare must be trained on diverse datasets to ensure they perform well across different patient populations. Ensuring balanced datasets is crucial to avoid misdiagnosis, ineffective treatments, and healthcare disparities [6]. Therefore, it is necessary to exploit the methods in order to balance the dataset before training the model.

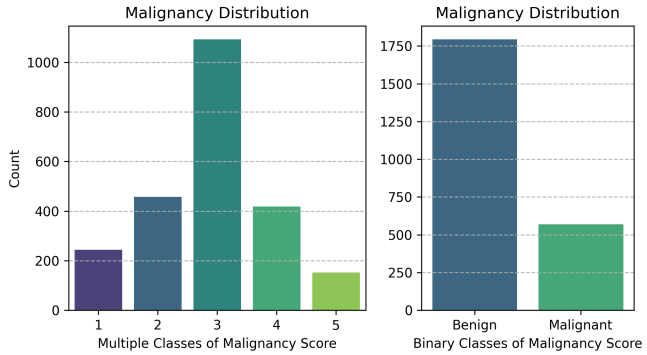

Figure 2: Unbalanced distribution of class (malignancy score)

When it comes to the brightness, the shape of the lungs, and the contrast of images, we observed that they are quite different as some samples are illustrated in Figures 4 and 5. Furthermore, the shapes of nodule slice images vary from (44×45) to (124×108) which adds the size standardization constraint to the pre-processing step. The presence of random noise, haze-like effect, and darkness are the other issues regarding the quality of images that should be addressed effectively.

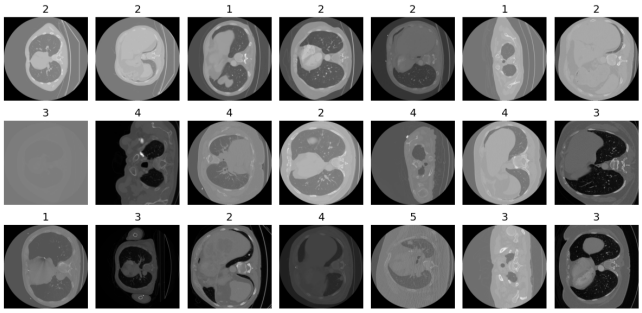

Figure 3: Some samples of full slice images. The label of each image is specified on the top.

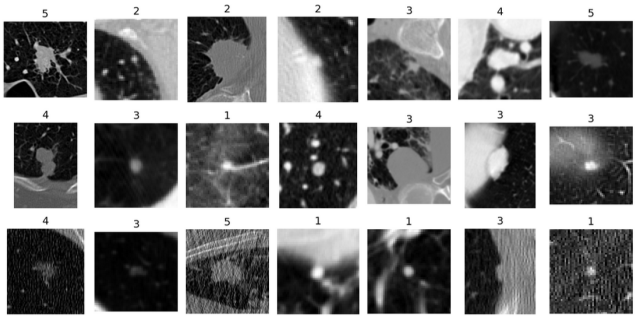

Figure 4: Some samples of nodule slice images. The label of each image is specified on the top.

In order to tackle the aforementioned issues, we exploited the weight class approach to balance the dataset, used a variety of image processing methods to improve the quality of images, and tested lots of standardization in terms of image dimensions which are described in the Data Pre-processing section.

2.3 Data Pre-processing

Raw healthcare data is often noisy, incomplete, and inconsistent. Preprocessing ensures that the dataset is clean, balanced, and standardized before feeding it into AI models [6]. Intensity normalization, resolution enhancement, dimension standardization, noise reduction, and data augmentation are necessary when working with medical images to ensure consistency across datasets. In the following sections, you can find out how we did these data preprocessing steps.

2.3.1 Intensity Normalization

The file format of the images in the dataset is the NRRD (Nearly Raw Raster Data). It is a flexible and efficient format for storing n-dimensional raster data, commonly used in medical imaging and scientific visualization to store CT, MRI, and other volumetric data along with metadata in a separate or embedded header [7]. The values in each NRRD file are scaled in the Hounsfield scale (HU), which is a quantitative scale used in CT imaging to measure tissue density. On the HU scale, air is -1000 HU, water is 0 HU, and dense bone is around +1000 HU, helping to differentiate tissues based on their X-ray attenuation properties [8]. Since the variety of intensities is extremely high, we clipped the intensities to remove unnecessary intensities for analyzing the lung and related tissues.

2.3.2 Resolution Enhancement

After the intensity normalization step, we performed the Gamma transformation, which is a nonlinear transformation used in image processing to adjust the brightness and enhance the contrast of an image [9]. The optimal Gamma value found for

the majority of images is 1.35 which improves the quality of images while preserving the details.

In addition to Gamma transformation, we performed CLAHE (Contrast Limited Adaptive Histogram Equalization) which is an image enhancement technique that improves contrast by applying local histogram equalization to small regions of an image instead of the whole image, preventing over-enhancement and noise amplification. It limits the contrast in each region to avoid excessive brightness differences, making it useful for medical images, low-light conditions, and X-rays. This method ensures balanced contrast enhancement while preserving important details [10]. Fine-tuning CLAHE's parameters is challenging and can be specified after doing a considerable number of experiments. However, we could find clipLimit=2.5 and tileGridSize = (8, 8) after studying some papers like [11] and [12]. You can see the results of performing these operations on some of the full-slice images in Figure 6.

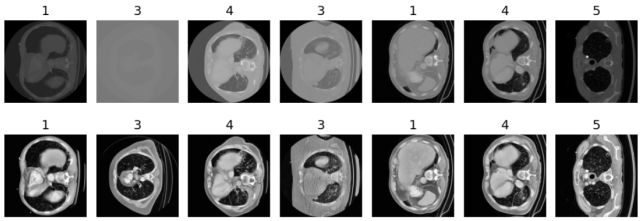

Figure 5: Some of pre-processed full slice images. The bottom images are preprocessed editions of the upper ones after intensity standardization and performing Gamma transformation.

2.3.3 Dimension Standardization

The dimension of full-view images is fixed (512×512), whereas nodule-view images have different dimensions. Thus, nodule slice images need resizing as they are too much different in terms of dimension. Figure 7 demonstrates a general view of the dimension distribution of nodule slices while it is simplified to show only the shapes that there are more than 25 images with those shapes due to the space limitation.

We consider a mean of the dimension majority as the standard dimension for nodule slices. Some samples of the resulting images of nodule slices after the pre-processing phase are shown in Figure 8.

2.3.4 Noise Reduction

Detection of tumors accurately necessitates high-quality images with minimal noise interference. Given the limited number of images in the dataset, it's crucial to preserve as many images as possible, filtering out only those with excessive noise that

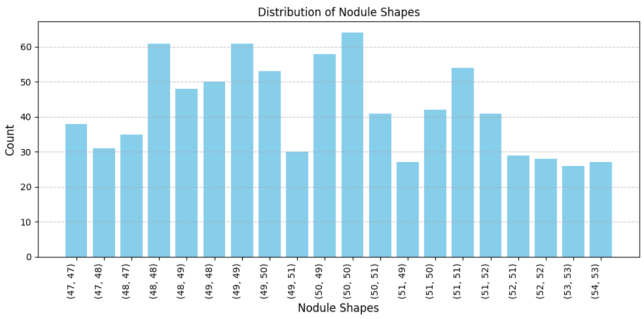

Figure 6: Shape distribution of nodule slice images. For limited space reasons, only shapes that have a frequency greater than 25 are shown.

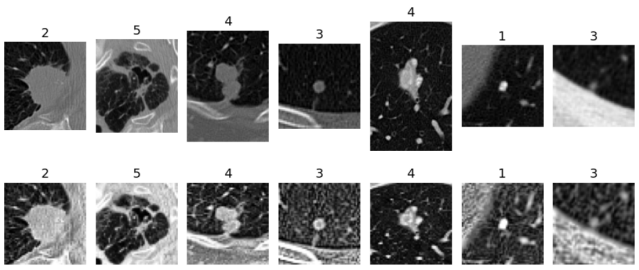

Figure 7: Some of pre-processed nodule slice images. The bottom images are preprocessed editions of the upper ones after intensity standardization, resizing, and performing Gamma transformation.

could impede analysis. A practical approach involves computing the Laplacian variance of each image to quantify the level of detail and noise. By analyzing the distribution of these variances across the dataset as shown in Figure 9, we can establish a threshold to identify and exclude only the most super-noisy images and preserve the majority of the data for tumor detection tasks. This method aligns with the techniques discussed in [13].

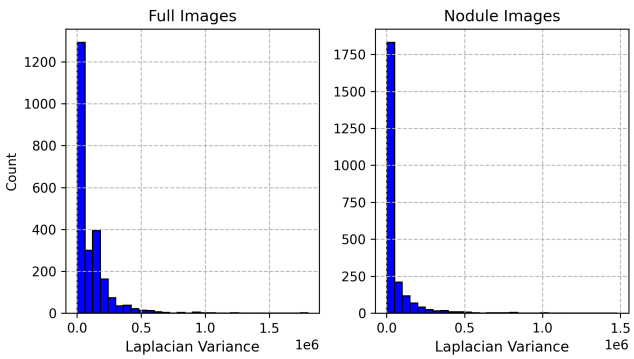

Figure 8: Laplacian Variance of Full and Nodule Images.

Specifying a certain threshold for excluding noisy images from the dataset can be done by doing a statistical analysis of the Laplacian variance values to figure out how many slices will be removed with respect to the percentile of the Laplacian variance as illustrated in Table 1.

Furthermore, doing some experiments to train

Percentile	Noisy Full Slices	Noisy Nodule Slices
99	23	24
98	48	48
97	71	71
96	95	95
95	119	119
94	142	142
93	166	166
92	189	189
91	213	213
90	237	237

Table 1: The number of noisy images with respect to the percentile of the Laplacian variance values.

the model and check the performance metrics can be helpful for making an optimal decision on noisy image removal procedure. Looking at Table 1, it is clearly seen that there is a linear correlation between the percentile of the Laplacian variance distribution and the number of noisy images for both full slices and nodule slices. According to the fact that we do not want to exclude so many images from the dataset and retain as many images as possible, we decided to exclude the most 2% noisy slices which led to the ignoring of 71 images from each full and nodule slices. You can see some examples of the most noisy images that exist in the dataset even after resolution enhancement in Figure 10.

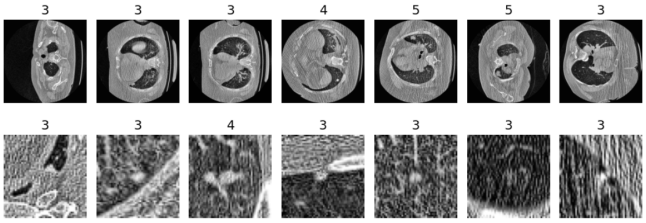

Figure 9: Example of noisy full images (first row) and noisy nodule images (second row)

2.3.5 Data Augmentation

Data augmentation refers to a group of techniques whose goal is to battle limited amount of available data to improve model generalization and push sample distribution toward the true distribution [14]. There are different augmentation strategies and performing a certain combination of them depends on the dataset. Data augmentation techniques that we used in this assignment are:

- *Rotation:* This involves rotating images by a certain degree, aiding models in recognizing objects from various orientations.
- *Flipping:* Horizontal and vertical flips create mirror images, enhancing the model's ability to generalize across different spatial orientations.
- *Zooming:* It involves scaling images in or out, allowing models to become invariant to size changes.
- *Shearing:* It skews the image along the x or y-axis, providing a perspective shift that helps models learn from distorted versions of the original images.

2.4 Deep Learning Models

Deep Learning (DL) is a subset of machine learning (ML) and AI that extracts a complex hierarchy of features from images by its self-learning ability. It involves neural networks with many layers that extract a hierarchy of features from raw input images. Several types of DL approaches have been developed for different purposes, such as object detection and segmentation in images. In this section we discuss some of these DL models we found them useful and effective for the given dataset in terms of malignancy score prediction.

2.4.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are the algorithms most commonly applied to images. CNN architectures are increasingly complex, with some systems having more than 100 layers, which means millions of weights and billions of connections among neurons. A typical CNN architecture contains multiple convolution, max-pooling, and activation layers.

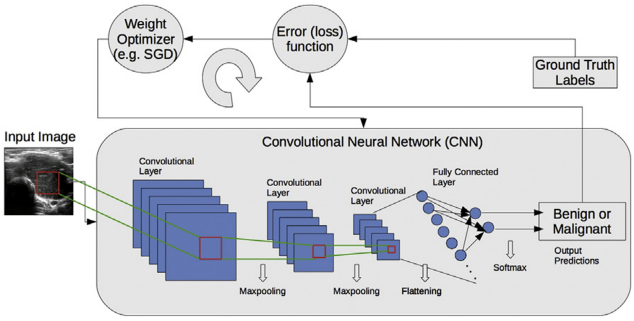

Figure 10: A schematic illustration of a deep learning process

As you can see in Figure 10, to perform a prediction from input data, the output scores of the final CNN layer are connected to a softmax nonlinearity function that normalizes scores into a multinomial distribution over labels. Also, an optimizer that minimizes the error between prediction and ground-truth labels through a loss function and a gradient backpropagation method that updates

weights at each iteration is used to train CNN architectures until they converge to a steady state [5]. We implemented a structure like this as a base model or a starting point to develop more robust models. The other models that gave us a better performance are explained in the following sections.

2.4.2 VGG 16

VGG 16 is characterized by its simplicity and depth, utilizing 16 layers composed of small 3×3 convolutional filters. This design choice allows the network to capture intricate features while maintaining manageable computational complexity. As we learned during the practical lessons, its architecture's uniformity and depth have made it a benchmark in image recognition tasks.

2.4.3 ResNet

ResNet addresses the degradation problem in deep networks by incorporating residual learning. This is achieved through shortcut connections that enable the network to learn identity mappings, facilitating the training of much deeper networks without performance degradation [15]. We saw a simple architecture of ResNet during the practical lessons which is shown in Figure 11.

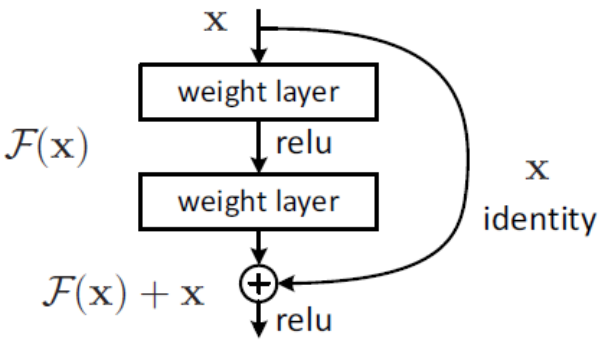

Figure 11: The ResNet architecture

2.4.4 EfficientNet

EfficientNet introduces a compound scaling method that uniformly scales network depth, width, and resolution using a set of fixed scaling coefficients. This approach leads to a family of models that achieve state-of-the-art accuracy while being computationally efficient. EfficientNet-B0, the baseline model, serves as the foundation for this scalable architecture [15]. These architectures are often employed in transfer learning scenarios.

2.4.5 Transfer Learning

Transfer learning is a technique in ML where a model developed for a particular task is reused as the starting point for a model on a different but related task. This approach is especially beneficial when dealing with limited data in the target domain, like the given dataset in this assignment, as it allows leveraging knowledge from a source domain where ample data is available. VGG16, ResNet, and EfficientNet models are pre-trained on large datasets such as ImageNet to capture a wide range of features that can be fine-tuned for specific tasks. The goal is to improve performance and reduce training time. Utilizing these pre-trained models and fine-tuning them can significantly enhance image analysis and classification tasks in medical images. The practical advantages of transfer learning in scenarios with limited labeled data [16], like the given dataset.

2.4.6 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of deep learning models that belong to the unsupervised learning family. They are primarily used for generating synthetic data when there is insufficient real data. A GAN consists of two neural networks that compete against each other in a game-theoretic framework:

- *Generator (G):* Creates fake samples from random noise.
- *Discriminator (D):* Evaluates whether a sample is real or fake.

The goal of the Generator is to fool the Discriminator, while the goal of the Discriminator is to correctly distinguish between real and fake samples. In other words, a generative model G captures the data distribution, and a discriminative model D estimates the probability that a sample came from the training data rather than G. The training procedure involves G attempting to maximize the probability of D making a mistake, leading to a minimax two-player game [17]. In each iteration, if the Discriminator correctly detects fake data, the Generator is penalized. If the Generator successfully fools the Discriminator, it is rewarded. Both networks continuously update using gradient Descent and back-propagation. Over multiple iterations, the Generator improves and starts producing highly realistic data.

Conditional GAN (cGAN) is an extension of the traditional GAN framework that incorporates additional information into both the generator and discriminator models. This conditioning information can be in the form of class labels, attributes, or any other data that provides context, enabling the generation of outputs with specific desired characteristics. By integrating this auxiliary data, cGANs offer enhanced control over the data generation pro-

cess, allowing for the creation of more targeted and relevant samples [18].

2.5 Model Training

2.5.1 Class Weighting

As mentioned before, the dataset is highly imbalanced, meaning some classes have far more images than others. This imbalance can cause the deep learning model to favor majority classes, leading to poor performance in minority classes. There are several approaches to handle the class imbalance problem like oversampling and undersampling. The method we used is class weighting which assigns higher weights to underrepresented classes and lower weights to overrepresented ones. The weights are computed for multi-label and binary classifications separately using the formula (1).

$$w_i = \frac{n_{\text{samples}}}{n_{\text{classes}} \times n_i} \tag{1}$$

where:

- w_i: The weight for class i
- $n_{samples}$: The total number of images in the dataset
- $n_{classes}$: The total number of classes in the dataset
- n_i: The number of samples with class i in the dataset

The corresponding computed weights for each class using the scikit-learn library are shown in Table 1. The class weighting approach helps the model penalize misclassifications of minority classes more heavily.

Class	Weight
1	1.937
2	1.034
3	0.433
4	1.131
5	3.109
Benign	0.659
Malignant	2.073

Table 2: Weights of different classes of the dataset

3 Experimental Results

4 Discussion and Conclusion

5 References

[1] Sung, H., Ferlay, J., Siegel, R. L., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3), 209-249.

[2] Siegel, R. L., Miller, K. D., & Jemal, A. (2022). Cancer statistics, 2022. CA: A Cancer Journal for Clinicians, 72(1), 7-33.

[3] Aberle, D. R., Adams, A. M., Berg, C. D., et al. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine, 365(5), 395-409.

[4] Ardila, D., Kiraly, A. P., Bharadwaj, S., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25(6), 954-961.

[5] Akkus, Z., Cai, J., Boonrod, A., Zeinoddini, A., Weston, A. D., Philbrick, K. A., & Erickson, B. J. (2019). A survey of deep-learning applications in ultrasound: Artificial intelligence–powered ultrasound for improving clinical workflow. Journal of the American College of Radiology, 16(9), 1318–1328.

[6] Bohr, A., & Memarzadeh, K. (Eds.). (2020). Artificial Intelligence in Healthcare. Academic Press.

[7] Bourke, P. (2004). NRRD (Nearly Raw Raster Data) File Format Specification. Retrieved from www.paulbourke.net.

[8] Ramphal, R., & Raniga, S. B. (2020). Hounsfield Unit. Published by StatPearls. Retrieved from www.ncbi.nlm.nih.gov.

[9] Gonzalez, R. C., & Woods, R. E. (2008). Digital Image Processing (3rd ed.). Prentice Hall.

[10] Tawfik, N., Emara, H., El-Shafai, W., Soliman, N., Alarni, A. & Abd El-Samie, F. (2024). Enhancing Early Detection of Lung Cancer through Advanced Image Processing Techniques and Deep Learning Architectures for CT Scans. Computers, Materials, and Continua. 81. 271-307.

[11] Fawzi, A., Achuthan, A., & Belaton, B. (2021). Adaptive Clip Limit Tile Size Histogram Equalization for Non-Homogenized Intensity Images. IEEE Access. PP. 1-1.

[12] Khomduean, P., Phuaudomcharoen, P., Boonchu, T. et al. Segmentation of lung lobes and lesions in chest CT for the classification of COVID-19 severity. Sci Rep 13, 20899 (2023).

[13] Ranjbaran A, Hassan AH, Jafarpour M, Ranjbaran B. A Laplacian based image filtering using switching noise detector. Springerplus. 2015 Mar 8;4:119.

[14] Hao R, Namdar K, Liu L, Haider MA, Khalvati F. A Comprehensive Study of Data Augmentation Strategies for Prostate Cancer Detection in Diffusion-Weighted MRI Using Convolutional Neural Networks. J Digit Imaging. 2021

Aug;34(4):862-876.
[15] Yang Y, Zhang L, Du M, Bo J, Liu H, Ren L, Li X, Deen MJ. A comparative analysis of eleven neural networks architectures for small datasets of lung images of COVID-19 patients toward improved clinical decisions. Comput Biol Med. 2021 Dec;139:104887.
[16] Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., Siddiqui, T., & Mellit, A. (2023). A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope. Sustainability, 15(7), 5930.
[17] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). "Generative Adversarial Nets." Advances in Neural Information Processing Systems (NeurIPS).
[18] Coursera Website