Pneumothorax Segmentation Using Feature Pyramid Network and MobileNet Encoder Through Radiography Images

Ayush Singh, Gaurav Srivastava, and Nitesh Pradhan

1 Introduction

Medical imaging incorporates methods and techniques to better understand medical images with aid of algorithms. It plays a crucial role in the classification and treatment of diseases [1, 2]. One of the major techniques used in medical imaging is image segmentation which involves breaking down an image into smaller segments based on different parameters ranging from textures to shapes. It has been an area of active research with use cases in autonomous flight/navigation, satellite, and medical imaging. With advancements in deep learning, image segmentation has been proven to be a boon in the diagnosis of a vast variety of diseases ranging from brain tumors to cancers.

Pneumothorax is one such lungs disease that results in sudden breathlessness because of some underlying symptoms or with no symptoms at all [3]. Pneumothorax can be diagnosed by clinical diagnosis using a stethoscope. In contrast, in some cases where the pneumothorax is smaller, X-ray scans are used to determine the location of the pneumothorax. The diagnosis of which is done by a radiologist by looking at the chest X-ray and recognizing the region of the pneumothorax. If pneumothorax is small, then it might heal on its own or else treatment includes the insertion of a needle between the ribs to remove excess air. Image segmentation can aid this process by segmenting the region of the pneumothorax thus providing a confident diagnosis.

Several sophisticated methods, especially convolutional neural networks [4], have introduced a new paradigm in image segmentation. The way of segmenting an image

A. Singh · G. Srivastava · N. Pradhan)
Department of Computer Science and Engineering, Manipal University Jaipur, Jaipur, Rajasthan, India
e-mail: nitesh.pradhan@jaipur.manipal.edu

G. Srivastava
e-mail: gaurav2001@gmail.com

V. S. Reddy et al. eds.), *Soft Computing and Signal Processing*,
Smart Innovation, Systems and Technologies 313,
https://doi.org/10.1007/978-981-19-8669-7_10

can thus be divided into two: Semantic segmentation and instance segmentation where the former relates to tasks of classifying an image at pixel level and the latter refers to the more complex task of classifying instances of the same pixels in the image. One of the major roadblocks in medical imaging is the availability of labeled data in large quantities, but these are tackled by a state-of-the-art convolutional neural network in medical image segmentation called U-Net [5]. This network is based on an encoder-decoder model where the encoder and decoder can be replaced by more efficient CNN architecture like VGG-16 [6], ResNet50 [7], MobileNet [8], and EfficientNet [9]. Following a research hierarchy, these networks improve one upon another. This paper investigates the training of U-Net with different backbone networks along with different architectures like FPN [10] and LinkNet [11] on the pneumothorax X-ray dataset provided by the Society for Imaging Informatics in Medicine SIIM).

We start presenting our findings by introducing earlier work in image segmentation in the next section followed by a description of the dataset and networks used thus explaining different network architectures. We share our experimental results in Section 4.

2 Related Works

Image segmentation has been an area of active research as its applications are boundless. It covers a wide range of applications in the engineering and medical fields. Today's state-of-the-art algorithms make use of decades of research progress in this field. Earlier work in image segmentation made use of image thresholding [12], clustering [13], edge detection, and graph-based methods.

Image thresholding is a simple idea in which the image is first converted into gray scale, and then, thresholding is applied to segment an image. Clustering starts by considering each pixel as an individual cluster and then merging these individual clusters that have the least inter-cluster distance. K-means [14] clustering is a commonly known clustering algorithm. Edge detection algorithms work on major disruptions in the image, i.e., detecting boundaries of different objects in the image by making use of 2D filters. Graph-based segmentation is the most famous of all the methods and is still used in many state-of-the-art image segmentation algorithms. The most used algorithm in graph-based segmentation is Felzenszwalb et al. [15] which first creates an undirected graph where every pixel in the image is a node, and the difference between intensities of two nodes is the weight of the vertex connecting them. This algorithm is still widely used in current state-of-the-art deep learning networks to provide region proposals.

Deep learning methods, especially convolutional networks, have made significant improvements in image segmentation. Since their resurrection at ILSVRC 2010 where AlexNet [16], the largest CNN at that time was introduced. Since then, CNNs have made significant contributions to signal processing tasks. Abedella et al. [17]

trained B-U-Nets which comprised of four networks ResNet50, DenseNet, Efficient-netb4, and SE-ResNet50) as the backbone. They combined BCE and dice coefficient to make the loss function of the network which achieved 0.8608 on the test set which is among the highest dice score on the SIIM-pneumothorax dataset. Jarakar et al. [18] trained the same dataset on U-Net with ResNet as backbone network achieving a dice score of 0.84. Noticeable work of Malhotra et al. [19] which incorporates Mask R-CNN with ResNet101 as the backbone FPN. This model has a lower loss than ResNet50 as the backbone.

Tolkachev et al. [20] examined U-Net with different backbone networks. They used ResNet34, SE-ResNext50, SE-ResNext101, and DenseNet121. They also put their system to test against experienced radiologists thus examining how confident diagnosis can affect the treatment processes. Their system achieved a dice coefficient of 0.8574.

In this paper, we present a comparative study of the three most widely used network architectures, i.e., U-Net, LinkNet, and FPN with each architecture trained with four different backbone networks.

3 Materials and Methods

Data Description

The dataset is provided by Society for Imaging Informatics in Medicine SIIM) on Kaggle. It consists of 12,000 DICOM files which consist of metadata about the image and the X-ray image in .jpg or .png format. In general, Digital Imaging and Communications in Medicine DICOM) file consists of a header file and an image file. The header file contains information about the patient and information containing a description of the image such as pixel intensity, dimensions of the image. The image can be from any medical scan such as MRI, X-ray, and ultrasound. The annotations are provided in the form of run-length encoding RLE) along with image IDs. The X-ray scans which don't have pneumothorax are marked -1.

Data Preprocessing

Since the dataset is in DICOM .dcm), it must be preprocessed to be used for model training. These Dicom files are read using the pydicom library in Python. It may seem that some of the images have lower contrast, so as a preprocessing step, the contrast of the images is increased using histogram equalization [21]. The images are also resized to 128x128 and converted into a NumPy array of $128 \times 128 \times 3$. The rest of the content from the Dicom file is read and converted into Pandas DataFrame. The annotations are RLE-encoded, so they are read and converted to Numpy arrays of dimension $128 \times 128 \times 1$. The dataset was then splitted into training and validation set in the ratio of 8:2. A sample from the preprocessed dataset is shown in Figure 1.

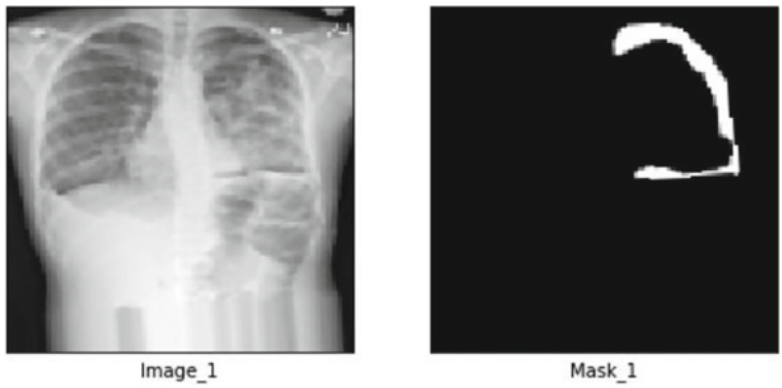

Fig. 1 Dataset visualization with both CXR image and after applying mask

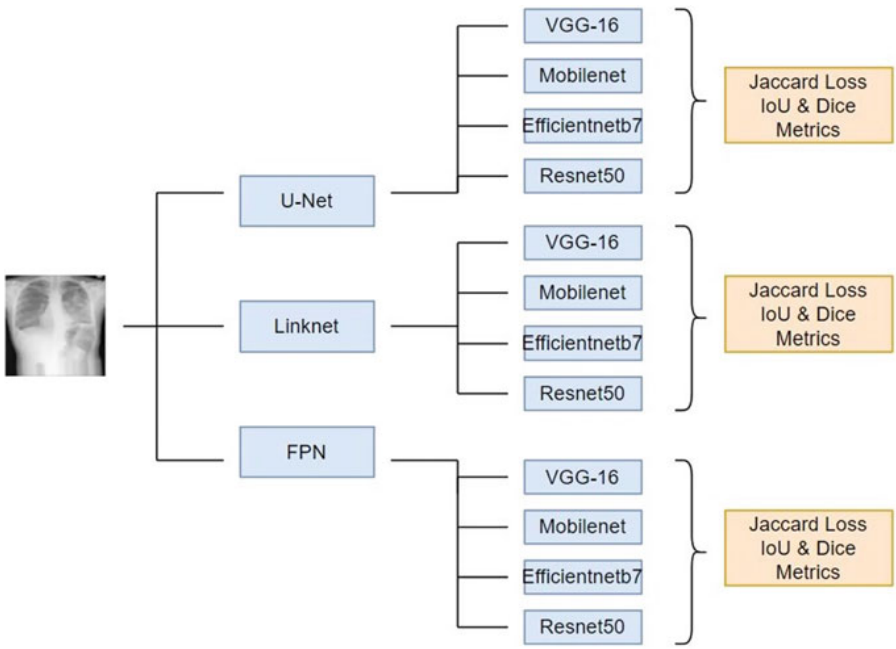

Fig. 2 Graphical abstract of the proposed work

Image Segmentation

Image segmentation at its crux is classifying an image at the per-pixel level. It means assigning each pixel in the image with some class labels. To divide an image into segments, we first extract features from the image. These features are captured using convolutional neural networks. The early part of CNN's captures low-level pixels features while later layers capture high-level features in the image. The task of the

CNN is to produce a per-pixel prediction of every object identified in the image. This task of dividing an image at the pixel level can then be classified into two ways—semantic segmentation and instance segmentation. Semantic segmentation classifies every object belonging to the same class as one whereas instance segmentation classifies every object belonging to the same class as distinct. Depending upon the number of distinct classes we want to classify in an image, segmenting images can further be classified as single-class segmentation and multiclass segmentation. In our project, since we are dealing with only one type of class, i.e., small regions of pneumothorax, so it is a single-class semantic segmentation.

Network Architectures

The popularity of CNNs rose after Alex Krizhevsky [22] trained a deep convolutional neural network AlexNet) that achieved the highest accuracy in ILSVRC 2010. Since then, the research interest in CNN has grown which has led to the development of better, efficient, and more robust CNN networks. After the development of AlexNet, the inception network [23] was developed by Google. Bigger networks like AlexNet are more prone to over-fitting and the larger the network the more difficult it becomes to transfer gradients throughout the network.

The idea behind inception was to instead of using fixed-size kernels for convolution operation to use multiple, as it will help capture sizes of different sizes in the image. The object distributed throughout the image will be captured by large kernels while the objects distributed locally will be captured by smaller kernels. This makes the network wider instead of longer which helps in the gradient flow without being computationally expensive.

VGG-16: VGG is another widely used network. Instead of using larger kernels for convolution, it uses smaller 3x3 kernels. This significantly reduced the computational cost of the network and made the network training easier. One of the major problems in AlexNet and VGG is computational cost.

ResNet50: ResNet showed that networks can be made to go deeper and deeper without being too computationally expensive. It introduced the concept of "skip connection" which made gradient flow easier throughout the network. MobileNet is a lightweight convolutional neural network designed to be run on mobiles and microcomputers like Arduino, Raspberry Pi. It achieves this using depthwise separable convolutions which significantly reduces the network's computational cost.

EfficientNetB7: EfficientNet highlighted the use of scaling for achieving higher accuracy. The scaling can be done by adjusting the width of the network which means scaling the number of filters in the network. Another scaling factor is the depth which means adjusting the length of the CNN while scaling the resolution of the input image also helps in improving accuracy. The scaling of the network is done using compound scaling introduced in the EfficientNet paper.

As shown in , our experiment uses VGG-16, MobileNet, EfficientNetb7, and ResNet50 networks as backbone networks for encoder-decoder segmentation models, i.e., U-Net, LinkNet, and FPN networks.

Network Architectures for Segmentation

Earlier contribution using CNN for image segmentation was made by proposing full-convolutional networks FCNs) [24]. FCNs are networks without fully connected networks at the end, so this network uses feature maps from the last convolution to make predictions. Since the last layers produce coarse feature maps, dense output in the final prediction is obtained by deconvolution operation on previous layers and adding them to the output. Deconvolution network is another well-known image segmentation architecture that unlike FCN learns the deconvolution parameters. SegNet [25] based on encoder-decoder architecture was introduced along the lines of a deconvolution network, but instead of deconvolution, it uses upsampling for producing dense feature maps as output.

U-Net: U-Net [5] was essentially introduced for biomedical imaging and is known as the state-of-the-art encoder-decoder network in the medical field which can learn from a few-labeled dataset which makes it suitable for biomedical segmentation. Like SegNet, U-Net also uses upsampling with no pooling layers which helps in improving the resolution of the output layer. It does not have any fully connected layers like the FCN but is a drastic improvement over FCN. Skip connections are introduced in the network which helps in carrying semantic information to later layers while also providing indirect pathways for smooth gradient flow.

LinkNet: LinkNet is like U-Net and was developed to be efficient with fewer parameters and FLOPs. LinkNet uses only 11.5 million parameters and 21.2 GFLOPs. The efficiency of LinkNet allows it to be used in segmenting live videos as well. It was developed by keeping efficiency along with better performance in mind. It provided state-of-the-art results on the CamVid dataset.

Feature Pyramid Network: Unlike U-Net and LinkNet, Feature Pyramid Network FPN) is a pyramid feature network that makes use of the bottom-up and top-down approaches for making predictions. High-level features are extracted from a bottom-up approach which increases the semantic value of the feature maps at later layers. Then, top-down approach is used to construct high-resolution layers from the semantic-rich output of the bottom-up approach. Features from the bottom-up approach are added to top-down layers for better detection of the objects in the image while also acting as skip connections for the easy flow of gradients.

4 Experiment and Results

In this section, we present a detailed overview of our experimental setup and the metric we used for evaluating different network architectures.

Experimental Setup

All 12 networks were trained using Kaggle kernels which are equipped with NVIDIA P100 GPU with GPU memory of 16GB capable of performing 9.3 TFLOPS. Each network required 4–5 h of training time. Each network is trained for 50 epochs.

Intersection over Union

Intersection over Union [26] IOU) is a metric to calculate the performance of the semantic model. It is calculated by calculating the intersection between the predicted mask and the ground truth and dividing it by the total number of pixels in both the predicted mask and ground truth mask. If IoU is 0, then it indicates that our segmentation model has poorly performed, and if IoU is 1, then it means that it has performed nicely. It is calculated as shown in Equation 1.

$$J \quad , B) = \frac{| \quad B|}{| \ \cup B|} = \frac{| \quad B|}{| \ |+|B|-| \quad B|} \quad 1)$$

$J =$ Jaccard distance

$=$ Set 1

$B =$ Set 1

In multiclass segmentation, the IoU is calculated for each class separately and then averaged over all calculated IoUs which predicts the total IoU for the semantic model.

Dice Coefficient

It is the harmonic coefficient of precision and recall [27]. The dice coefficient is calculated by multiplying 2 by the total of true positives TP) divided by 2 times the number of TP + false negatives FN) + false positives FP). It is defined as shown in Equation 2.

$$\text{Dice Coefficient} = \frac{2|X \quad Y|}{|X|+|Y|} \quad 2)$$

Results

We trained U-Net, FPN, and LinkNet on the pneumothorax dataset. We experiment with them using different backbone networks as encoders and decoders. These backbone networks include VGG-16, MobileNet, EfficientNet, and ResNet50. The main objective was to derive the best network combination for the pneumothorax dataset. We used Adam optimizer for training all 12 networks. We start by examining the loss values of U-Net as presented in Table 1 followed by LinkNet with a detailed description of evaluation metric values in Table 2 and at last presenting the performance description of FPN in Table 3. For training the networks, we used Jaccard loss with Intersection over Union IoU) and dice coefficient as the metrics for evaluating the networks.

Below we present our findings in tabular form for each network architecture with different backbone networks. They are compared based on the Jaccard loss, IoU, and dice coefficient obtained on the training and validation set, followed by the graphs for the same.

Table 1 Experimental results of different pre-trained encoders with U-Net architecture

Backbone networks	Training Jaccard loss	Training IoU	Training dice coefficient	Validation Jaccard loss	Validation IoU	Validation dice coefficient
VGG-16	0.34537	0.65460	0.78623	0.29936	0.70061	0.81916
MobileNet	0.46473	0.53533	0.68999	0.43130	0.56882	0.71826
EfficientNetB7	0.43055	0.56955	0.71935	0.37270	0.62732	0.76591
ResNet50	0.34410	0.65590	0.78797	0.32202	0.67802	0.80348

Table 2 Experimental results of different pre-trained encoders with LinkNet architecture

Backbone networks	Training Jaccard loss	Training IoU	Training dice coefficient	Validation Jaccard loss	Validation IoU	Validation dice coefficient
VGG-16	0.35843	0.66159	0.79254	0.31247	0.70429	0.82304
MobileNet	0.39980	0.62218	0.76251	0.36521	0.65366	0.78657
EfficientNetB7	0.41013	0.61216	0.75510	0.35961	0.65936	0.79099
ResNet50	0.38842	0.61157	0.75317	0.36906	0.63097	0.76852

Table 3 Experimental results of different pre-trained encoders with FPN architecture

Backbone Networks	Training Jaccard Loss	Training IoU	Training Dice Coefficient	Validation Jaccard Loss	Validation IoU	Validation Dice Coefficient
VGG-16	0.47703	0.54882	0.70226	0.44197	0.58129	0.73027
MobileNet	0.28901	0.72649	0.83946	0.23795	0.77301	0.87055
EfficientNetB7	0.40222	0.62012	0.76136	0.33454	0.68251	0.80799
ResNet50	0.35407	0.64591	0.77967	0.32804	0.67201	0.79928

U-Net In our experiment, we found that for all the backbone networks we got similar Jaccard loss for U-Net architecture. The loss varied by very little amount among these backbones. ResNet50 and VGG-16 backbones produced the lowest loss with VGG-16 producing slightly better results on the validation set.

LinkNet LinkNet produced similar results for VGG-16, MobileNet, and Efficient-Netb7 while ResNet50 produced the lowest dice score. Among the three, the best performing backbone was VGG-16 followed by EfficientNetb7.

Feature Pyramid Network FPN) The best combination was with MobileNet as the backbone network followed by EfficientNetb7. VGG-16 produced the lowest dice score on the validation set. FPN with MobileNet combination performed the best among all 11 networks.

Next, we present the learning curves which comprise Jaccard loss, Intersection over Union, and dice coefficient. Each of these curves shows improvement in the

networks when trained for 50 epochs. We share 3 different graph plots, i.e., Figs. 3 and 4 displaying Jaccard loss, IoU, and dice coefficient of all the 12 networks, respectively.

Jaccard Loss graph The following graphs plot the learning curve for all 11 networks when trained for 50 epochs. Inference from the graph shows MobileNet FPN with the lowest loss while VGG-16 FPN with the highest loss value.

Intersection Over Union graph We can conclude from the graph that the highest IoU is achieved by MobileNet FPN, and the least is produced by MobileNet U-Net. We observe a steep increase in the MobileNet FPN curve while VGG-16 FPN and MobileNet U-Net follow a similar curve.

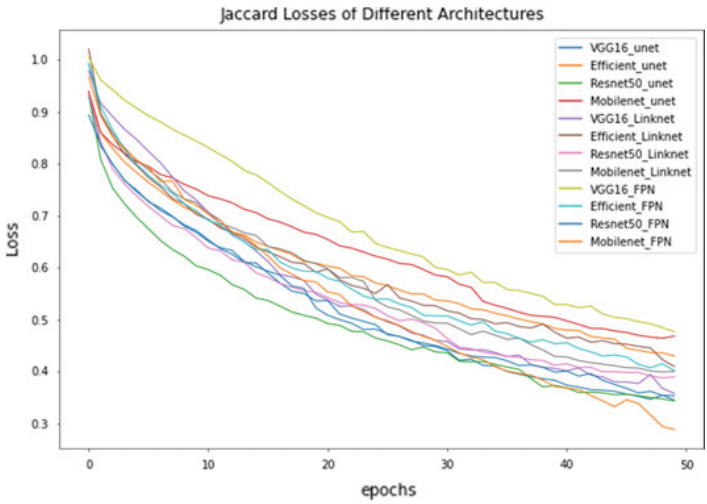

Fig. 3 Loss curve during training of different pre-trained encoders combined with U-Net, FPN, and LinkNet architecture

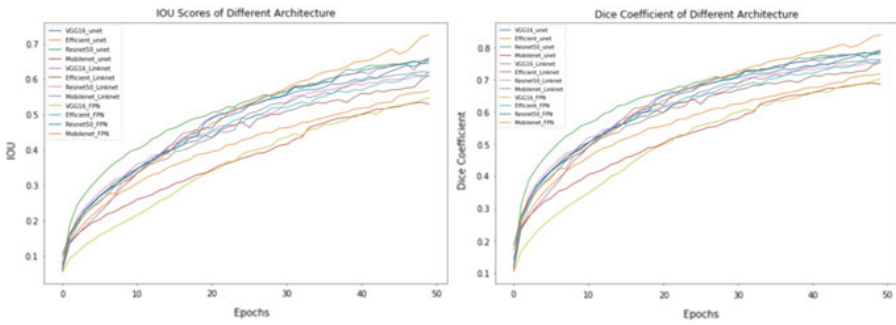

Fig. 4 Dice coefficient and Jaccard index curve during training of different pre-trained encoders combined with U-Net, FPN, and LinkNet architecture

Dice Coefficient We can easily infer that MobileNet FPN has the best dice score hence the best network combination among the 12 networks while VGG-16 FPN and MobileNet U-Net are the poor performing combinations. U-Net with ResNet50 backbone is the second-best backbone network.

5 Conclusion

In this paper, we investigated three well-known network architectures with four different backbone networks each hence training 12 networks altogether to determine the best network combination for the pneumothorax segmentation dataset. We trained the network with Jaccard loss and used IoU and dice coefficient as metrics. We conclude that the best network combination was FPN with MobileNet backbone while U-Net with MobileNet and FPN with VGG-16 as the worst-performing architectures. We communicated our findings with the aid of tables and graphs. Future works include using other more sophisticated architectures like Mask R-CNN for segmentation and comparing it with our existing results.

References

1. Himabindu, G., Ramakrishna Murty, M.: Classification of kidney lesions using bee swarm optimization. Int. J. Eng. Technol. **7** 2.33): 1046–1052 2018).
2. Himabindu, G., Ramakrishna Murty.M.: Extraction of texture features and classification of renal masses from kidney images. Int. J. Eng. Technol **7** 2.33): 1057–1063 2018)
3. MacDuff, A., Arnold, A., Harvey, J.: Management of spontaneous pneumothorax: British thoracic society pleural disease guideline 2010. Thorax **65**, no. Suppl 2 2010): ii18-ii31.
4. LeCun, Y., Haffner, P., Bottou, L., Bengio, Y.: Object recognition with gradient-based learning. In: Shape, Contour and Grouping in Computer Vision, pp. 319–345. Springer, Berlin, Heidelberg 1999)
5. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer, Cham 2015)
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition 2014). arXiv:1409.1556
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 2016)
8. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H.: Mobilenets: efficient convolutional neural networks for mobile vision applications.arXiv:1704.04861 2017)
9. Mingxing, T., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR 2019)
10. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Featurepyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 2017)

11. Chaurasia, A., Culurciello, E.: Linknet: exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing VCIP), pp. 1–4. IEEE 2017)
12. Gurung, A., Tamang, S.L.: Image segmentation using multi-threshold technique by histogram sampling 2019). arXiv:1909.05084
13. Naous, T., Sarkar, S., Abid, A. and Zou, J.: Clustering plotted data by image segmentation 2021).arXiv:2110.05187.
14. Andrecut, M.: K-Means Kernel Classifier 2020). arXiv:2012.13021
15. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. Int. J. Comput. Vision **59** 2), 167–181 2004)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 2012)
17. Abedalla, A., Abdullah, M., Al-Ayyoub, M., Benkhelifa, E.: Chest X-ray pneumothorax segmentation using U-Net with Efficient-Net and ResNet architectures. PeerJ Comput Sci. **29** 7), e607 2021). https://doi.org/10.7717/peerj-cs.607.PMID:34307860;PMCID:PMC8279140
18. Pneumothorax segmentation: deep learning image segmentation to predict pneumothorax by karan Jarkhar [arXiv:1912.07329]
19. Malhotra, P., Gupta, S., Koundal, D., Zaguia, A., Kaur, M. and Lee, H.N.: Deep learning-based computer-aided pneumothorax detection using chest X-ray images. Sensors **22** 6): 2278 2022). https://doi.org/10.3390/s22062278
20. Tolkachev, A., Sirazitdinov, I., Kholiavchenko, M., Mustafaev, T., Ibragimov, B.: Deep learning for diagnosis and segmentation of pneumothorax: the results on the kaggle competition and validation against radiologists. IEEE J. Biomed. Health Inform. **25** 5), 1660–1672 2020)
21. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B. and Zuiderveld, K.: Adaptive histogram equalization and its variations. Comput. Vis. Graphics Image Process. **39** 3): 355–368 1987)
22. Hong, S., Noh, H., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1520–1528. 2015.
23. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Goingdeeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 2015)
24. Long, J., Shelhamer, E., Darrell, T.: Fullyconvolutional networks for semantic segmentation." In Pro- ceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440. 2015.
25. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39** 12), 2481–2495 2017)
26. Generalized intersection over union: a metric and a loss for bounding box regression. arXiv:1902.09630
27. Continuous dice coefficient: a method for evaluating probabilistic segmentations. arXiv:1906.11031