



# Deep Learning Models for Lung Nodule Segmentation: A Comparative Study

Aliya Orazalina<sup>2</sup> · Heechul Yoon<sup>1</sup> · Sang-II Choi<sup>2</sup> · Seokhyun Yoon<sup>1</sup>

Received: 13 February 2024 / Revised: 7 August 2024 / Accepted: 21 August 2024  
© The Author(s) under exclusive licence to The Korean Institute of Electrical Engineers 2024

## Abstract

Lung nodule detection is clinically crucial but challenging and time-consuming. The development of automated segmentation approaches could be helpful. To assess the capability of deep learning methods for lung diagnosis, this paper compares recent deep learning models and evaluates their performance. We implemented several preprocessing steps, including windowing, thresholding, and resizing, to improve the image quality, adjust the image dimension suitable for the network, and focus on specific areas of interest within an image. The evaluation was conducted on the Lung Image Database Consortium (LIDC) dataset using Dice similarity coefficient (DSC) and Hausdorff distance (HD) metrics with model complexity parameters for a multifaceted comparison of the models. The experiments showed that the highest accuracy among the five chosen models (97.80% DSC and 1.29 HD) was reached by the Connected-UNets model, which also has the highest computational complexity. In this paper, we quantitatively evaluated and compared 5 deep learning models namely Salient Attention UNet, Connected-UNets, DDANet, UTNet, and EdgeNeXt. The evidence-based overview of current deep learning achievements for the clinical community investigated in this study can be useful to the research community in developing a new model and, thus, designing computer-aided detection and diagnosis (CAD) systems.

**Keywords** Nodule detection · Lung cancer · Deep learning · Neural networks · CAD

## 1 Introduction

Respiratory diseases such as lung cancer, tuberculosis, bronchitis, emphysema, pneumonia, and cystic fibrosis are still emerging, prevalent, and life-threatening [1]. One of the most recent detrimental diseases is the well-known COVID-19, which is caused by the SARS-CoV-2 virus [2]. As the symptoms of COVID-19 are similar to pneumonia, this disease is often misdiagnosed. Early diagnosis is essential for

the prevention of COVID-19 spreading as well as for effective treatment [3]. In addition to COVID-19, according to World Health Organization (WHO) statistics for 2020, 10 million deaths from various cancer types were reported, and 1.8 million are associated with lung cancer worldwide [4].

One of the ways to reduce lung cancer mortality is by using low-dose computed tomographic screening, which can be an effective life-saving tool if the screening results are correctly interpreted and analyzed [5]. However, manual analysis of medical images is laborious, operator-dependent, and subjective, which can often lead to misdiagnosis as pathological information can be misinterpreted or even neglected. Like many other diseases, detecting and interpreting lung nodules is challenging. First of all, lung nodules are often patient specific and heterogeneous, which is not ideal for their identification [6]. Secondly, lung nodules are geometrically diverse, with various sizes. Finally, there is a complication in distinguishing nodules from the surrounding normal tissues, as they are often visually similar [7].

Recent advances in deep learning methods and their application in medical image analysis have shown promising potential in computer-aided detection and diagnosis, referred

✉ Heechul Yoon  
heechul.yoon@dankook.ac.kr

Aliya Orazalina  
orazalina.aliya@dankook.ac.kr

Sang-II Choi  
choisi@dankook.ac.kr

Seokhyun Yoon  
wireless@dankook.ac.kr

<sup>1</sup> Department of Electronic and Electrical Engineering, Dankook University, Yongin-Si, Korea

<sup>2</sup> Department of Computer Engineering, Dankook University, Yongin-Si, Korea

**Table 1** Summary of common single networks and hybrid networks

Method type	Description	Advantages	Disadvantages	Application	Papers
Auto encoder	Learning using dimensionality reduction	operating without the need for labeled data and falls within the realm of unsupervised learning	predicting only on similar data (specific data that it was trained on)	dimensionality reduction, image compression, image denoising, image generation, recommendation system	[30–32]
CNN	Using convolution operations for extracting features from data	Learning local connections, has sharing weights among groups of connections, reducing the number of parameters using the pooling layer	Having limitations in learning long range dependencies, requiring a large amount of labeled data	Image segmentation, image classification, image recognition	[33–35]
RNN	Modeling sequence of data due to the ability of memorizing temporal dependencies	processing any length data, model size is not dependent to input size	having high computational complexity, has problems such as exploding and gradient vanishing	natural language processing, image translation, image detection, image recognition	[36–38]
ViT	Learning from considering image patches as a sequence of data. But comparing with RNN does not require sequential data be in order before processing	learning global and local dependencies effectively	having quadratic computational complexity	image segmentation, image classification, image recognition	[39–41]
GAN	Generating new data instances using CNN	generating different versions of data that can be used for training	training complexity due to various data requiring	data generation	[42–44]
Hybrid Network	Consisting two or more deep learning methods	Having limitation compensation of the included deep learning methods	Having computational complexity	depending on the included architectures	[45–47]

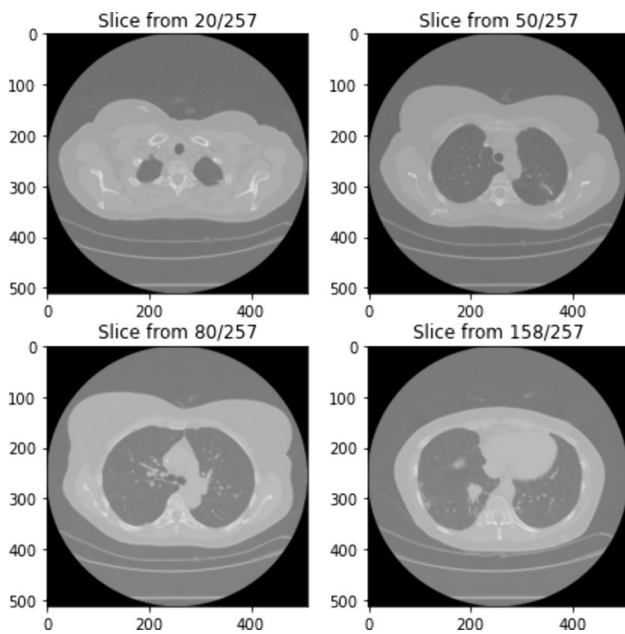


Fig. 1 Sample images from LIDC dataset

to as CAD. The capability of the deep learning based CAD techniques has almost reached the level of experienced radiologists [8]. CAD systems are generally divided into four stages: pre-processing, segmentation, feature extraction, and classification. Pre-processing is the stage of preparing data for a deep learning model where denoising and adjusting image contrast are performed [9]. In the segmentation stage, targeted objects such as organs, tissue, mass, and tumors are segmented. Features including size, location, texture, and patterns are then extracted from segmented lesions of interest. At the last classification stage, the selected features

are analyzed, and a segmented area is assigned to one of the given groups. In the case of the lung cancer determination task, the target area is the tumor, the features are the characteristics of the tumor, and the classification groups are benign or malignant [10].

This comparative study focuses on state-of-the-art deep learning-based lung nodule segmentation and annotation systems. The majority of earlier research was conducted on small datasets [11–14], which might not be a fair comparison given that deep learning-based techniques require a larger number of training images to achieve optimal performance. There was also a lack of validation by other groups because many of these studies were performed on private datasets [15–18]. Additionally, the primary focus of most papers lies in reliability metrics, such as overall accuracy, with little attention paid to the complexity of the model. However, considering the complexity of the model is vital as it provides valuable insights into the computational resources required and sheds light on the trade-offs between model intricacy and performance. Comparing the latest trends in the field gives a clearer landscape in the direction of current studies and the horizons of future research. Most of the recently proposed models are based on transformer technology, which was originally presented in [19] (2017). Initially, the transformer was developed for application in the natural language processing (NLP) task [20]. Nowadays, transformers have a great deal of applications in computer vision. Similar to the transformers in NLP, the main component of the vision transformer is the self-attention layer, which captures long-term dependencies and thus increases performance in modeling global features [21].

Deep learning based medical image segmentation methods can be categorized in many ways, such as network architecture, training method, input data type, etc. [22].

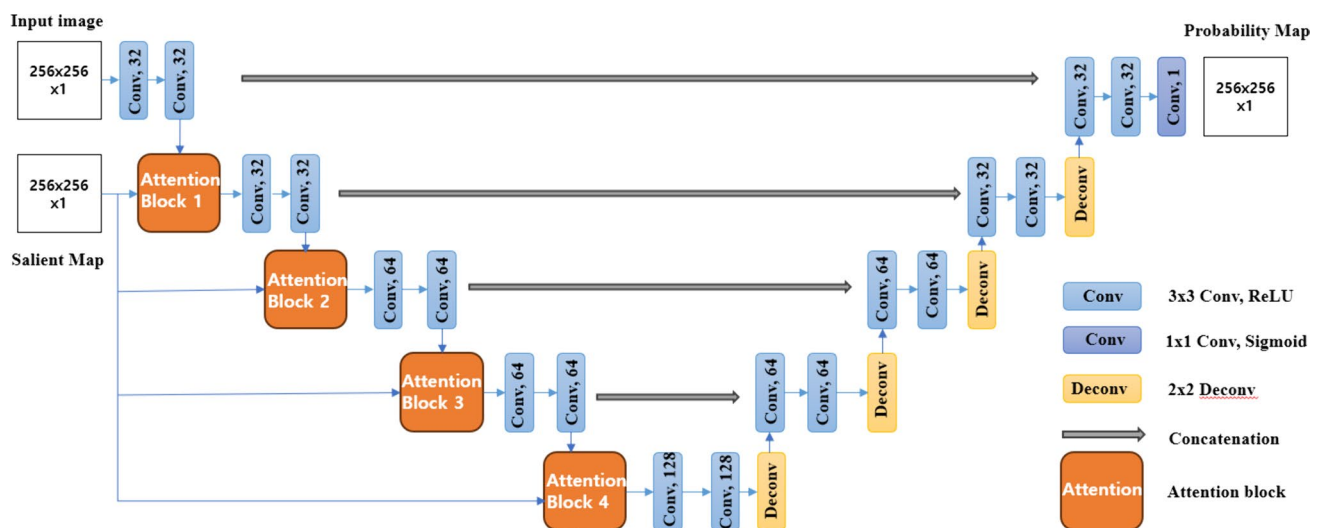


Fig. 2 Architecture of the salient attention UNet

Examples of network architecture kinds might be RNN [23], LSTM [24], CNN [25], ViT [26], etc. The most common of them were summarized in Table 1. The references in this table are implementation examples of these architectures in medical image segmentation. In training methods, deep learning models are broadly divided into supervised learning [27], unsupervised learning [28], and reinforcement learning [29]. The first two of them are distinguished by input data size and labels (annotated data). Supervised learning generally uses a large amount of data and labels, whereas no annotated data is required in unsupervised learning. Reinforcement learning is a special one whose learning is held by interactions between agent and environment. It is more focused on goal-directed learning than other types of machine learning.

We organized the rest of the paper to include the following: (i) A review of the used dataset and methods in terms of key features (Sect. 2); (ii) A performance comparison of the methods that were presented in (i) (Sect. 3); and (iii) Discussion and conclusions (Sect. 4).

## 2 Materials and Methods

### 2.1 CT Database for Model Quantification

The LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) is the largest publicly available dataset with annotations of lung nodules. In Fig. 1, sample CT image slices from the dataset are

presented. The vastness and accuracy of this dataset could allow us to produce a reliable benchmark for evaluating the performance of different segmentation models [48]. The dataset includes 1010 cases of helical thoracic CT images, annotated masks with lung nodules, and XML files with subjective characteristics of the nodules such as spiculation, lobulation, subtlety, internal structure, shape (sphericity), margin, solidity, and likelihood of malignancy. Images and masks are provided as DICOM files, which means that there is not only pixel data but also metadata about device information, image acquisition parameters, and anonymized patient identification data [49].

Annotations were made by 12 expert radiologists in a two-phase process. The first phase was the blinded-read phase, where radiologists separately reviewed each CT slice and annotated target lesions, classifying them into one of three categories based on the nodule size: nodules with a size in the 3–30 mm range, nodules with a size less than 3 mm, and non-nodules with a size greater than or equal to 3 mm. The main focus of the database is nodules with sizes in the 3–30 mm range, as they are most likely to be malignant [50]. During the second phase, known as the unblinded-read phase, each radiologist conducted an independent review of their own annotations, in addition to examining the anonymized annotations made by the three other radiologists, in order to formulate a final opinion. The objective of this process was to detect, to the fullest extent possible, all lung nodules present in each CT scan without the necessity of enforcing unanimous consensus.

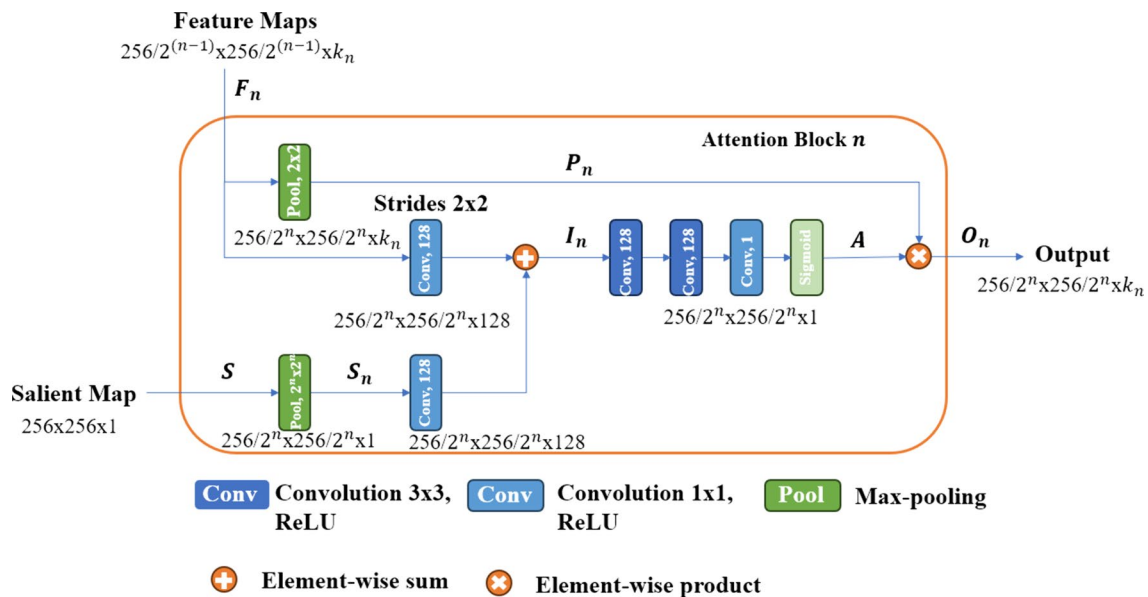


Fig. 3 The structure of the attention block

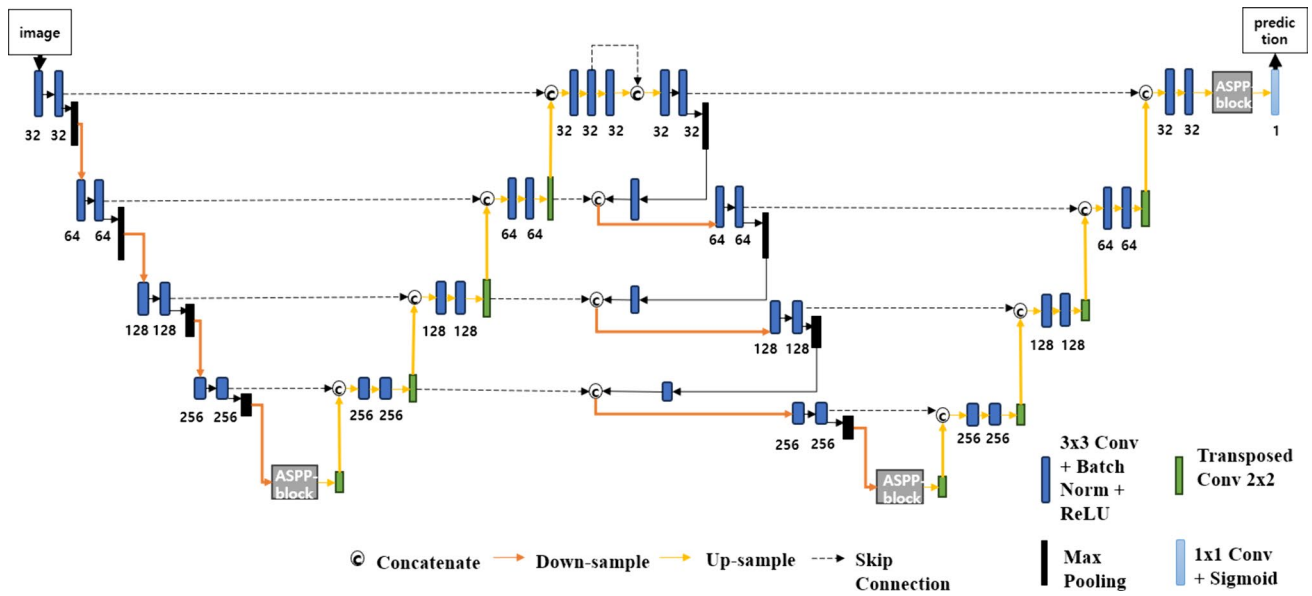


Fig. 4 Architecture of the connected-UNets

## 2.2 Deep Learning Models for Lung Nodule Segmentation

Recent advances in deep learning have substantially impacted the computer vision field [51]. This section will provide the major features, such as structure, functionality, and pros and cons, of deep learning models that were used in this comparison study. We have chosen five deep learning models for further comparison and evaluation based on the following criteria:

Firstly, all models consider local and global connections, which are crucial in the medical image segmentation field

[52]. These connections are vital in the context of medical image segmentation because they allow the model to understand both the fine details and the broader context of the image. Local connections focus on specific features and their relationships in proximity, while global connections provide a wider context, capturing overall structures and relationships between different regions. This combined approach ensures that the model can effectively discern intricate details while maintaining a comprehensive understanding of the entire image.

Secondly, models have high generalization ability, which means that they can be used for a variety of medical image

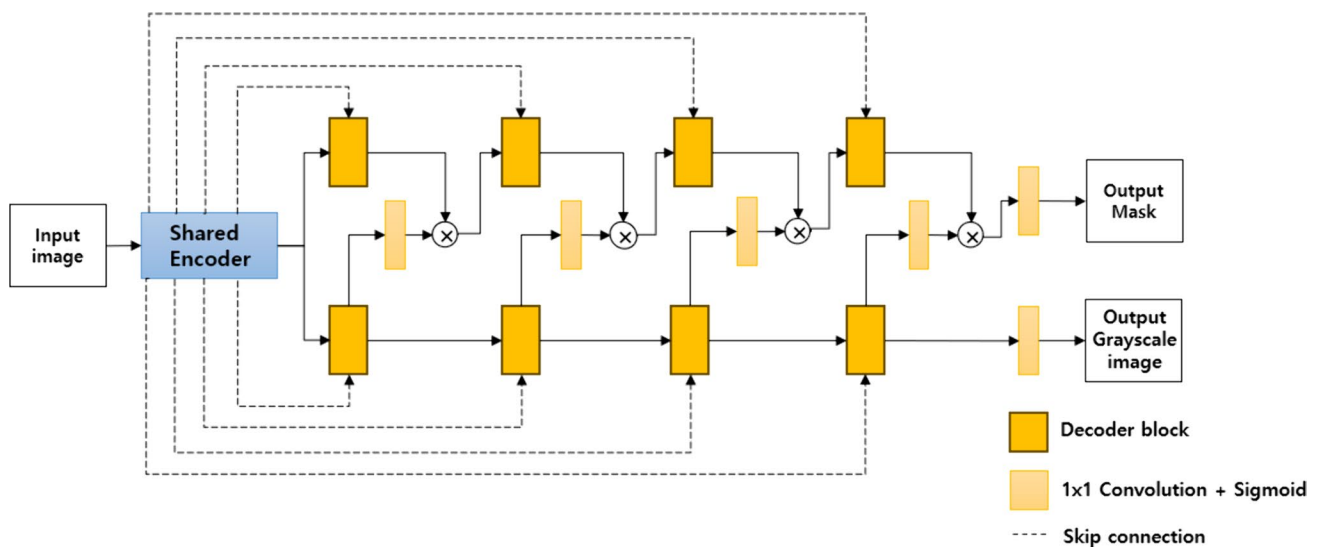


Fig. 5 Architecture of the DDANet

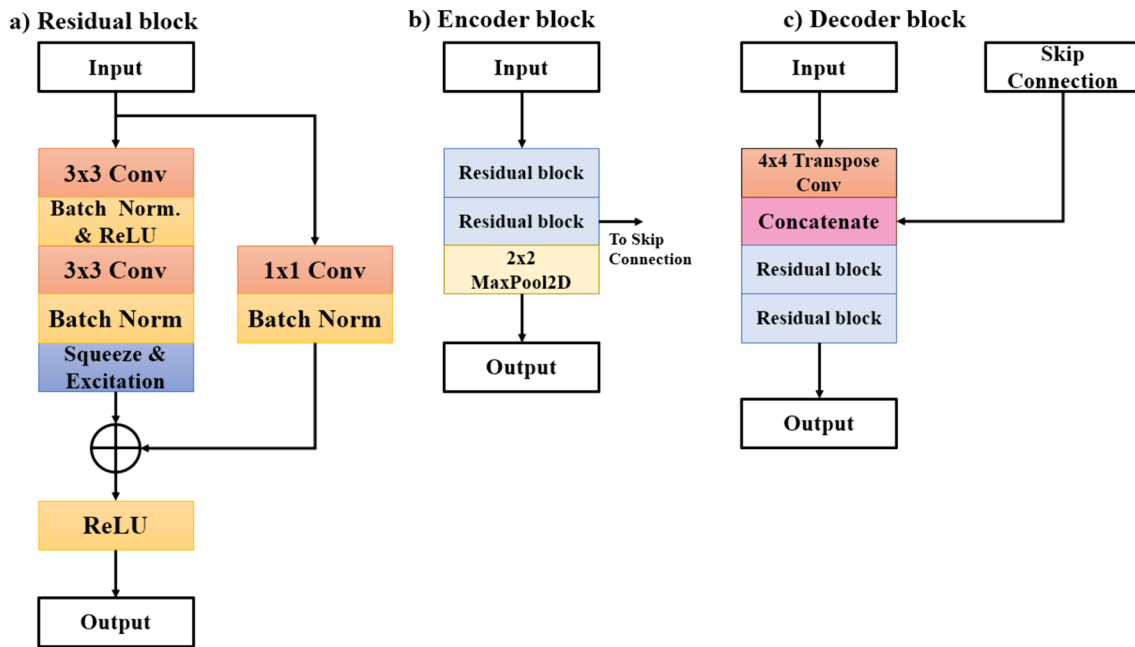


Fig. 6 a residual block, b encoder block, c decoder block of the DDANet

tasks [53]. This attribute is crucial as it implies that the models can effectively adapt and perform well on various types of medical images, even those they have not specifically been trained on. This adaptability is especially valuable in the dynamic and diverse field of medical imaging, where

a single model may be required to handle a wide array of image modalities, anatomical regions, or clinical scenarios.

Finally, the models have been developed and published over the past five years and show the latest trending technologies in the computer-aided detection and diagnosis

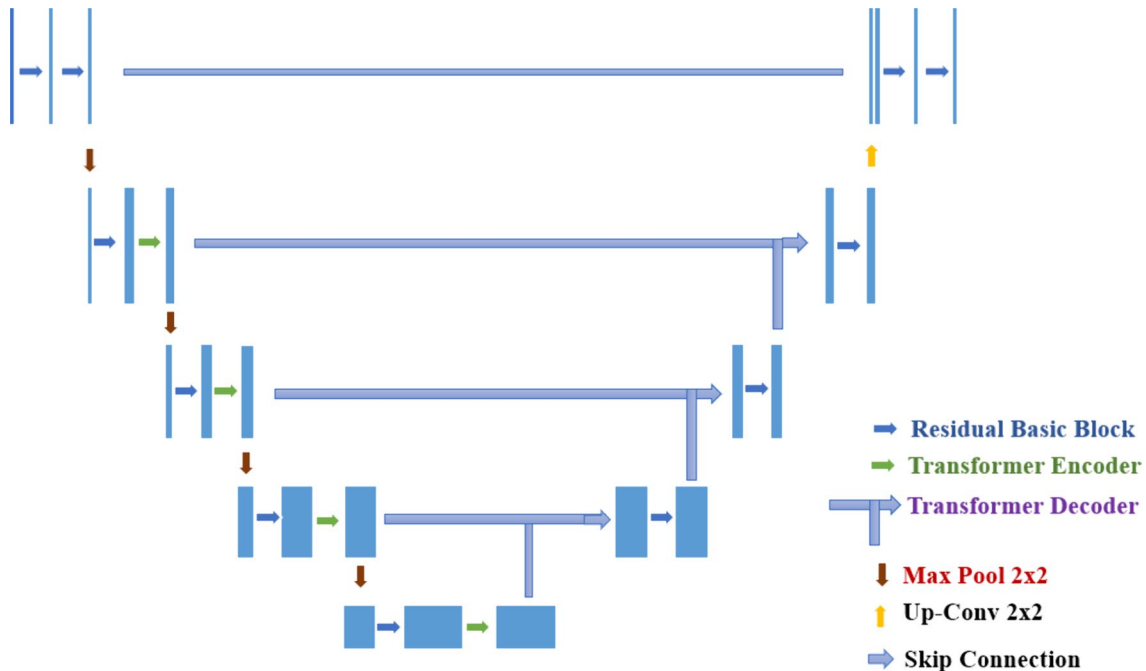


Fig. 7 Architecture of the UTNet

industry, which can be implemented as a base and improved in future model development.

### 2.2.1 Salient Attention UNet

The Salient Attention UNet model is based on the UNet model and Salient Attention blocks [54]. The architecture of the model presented in Fig. 2. The U-shaped architecture in the UNet implies three main parts, which are the encoder, skip-connections, and decoder [55]. In the encoder part, learning of global contextual representations takes place with gradual downsampling layers of CNN. In the decoder section, deep features are extracted, and missing spatial information is restored through skip-connections. These features are then combined and upsampled to match the input resolution, enabling precise pixel/voxel-wise semantic

prediction. In the Salient Attention UNet model, in order to improve attentional control of the network, attention blocks are incorporated with the encoder layers, which forces it to learn feature representations that focus attention on high-priority target regions. Figure 3 presents the structure of the attention block. Input feature maps, denoted as  $F_n$ , are processed through two sequential blocks. The first block performs max-pooling to generate  $P_n$  features, while the second block utilizes a  $1 \times 1$  convolutional layer to expand the channel dimension to 128. The saliency map, S, undergoes downsampling via max-pooling and upsampling through the convolutional layer to match the feature map's channel dimension. Subsequently, the upsampled  $F_n$  and S are added to produce intermediate maps  $I_n$ . This result then passes through several convolutional layers followed by a sigmoid function to normalize it within the [0,1] range. Finally, the product of the attention map A and max-pooled feature map  $P_n$  yields the output of the attention block, denoted as  $O_n$ .

In the case of tumor segmentation, the focus of the network attention is a crucial part of accurate semantic prediction. In the original paper, the model was evaluated on the breast ultrasound dataset collected from three hospitals.

The strength of the model is the utilization of prior knowledge about the organ and tumor, which resulted in the generation of salient attention image maps during training. However, it is important to note that the model comes with a quadratic computational cost.

### 2.2.2 Connected-UNets

The Connected-UNets model is based on two basic UNet models that are connected using skip-connection [56]. Figure 4 details the architecture of this model. Atrous Spatial Pyramid Pooling (ASPP) blocks play the role of bottleneck in the network [57]. ASPP extracts multi-scale contextual information and employs it to address the issue of losing resolution in the case of small sized tumors. There are a few variants on this architecture-e.g. Connected-UNets, Connected-AUNets, and Connected-ResUNets. The last two models are Connected-UNets application on Attention

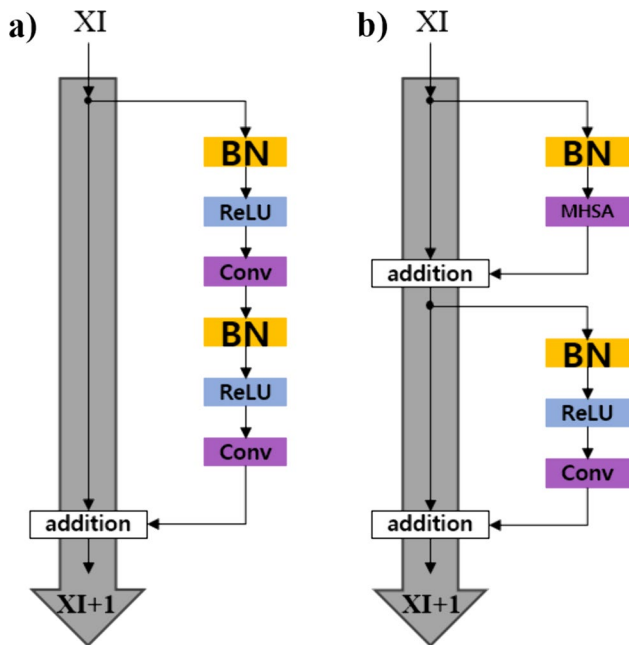


Fig. 8 a residual basic block of the UNet, b transformer encoder block of the UNet

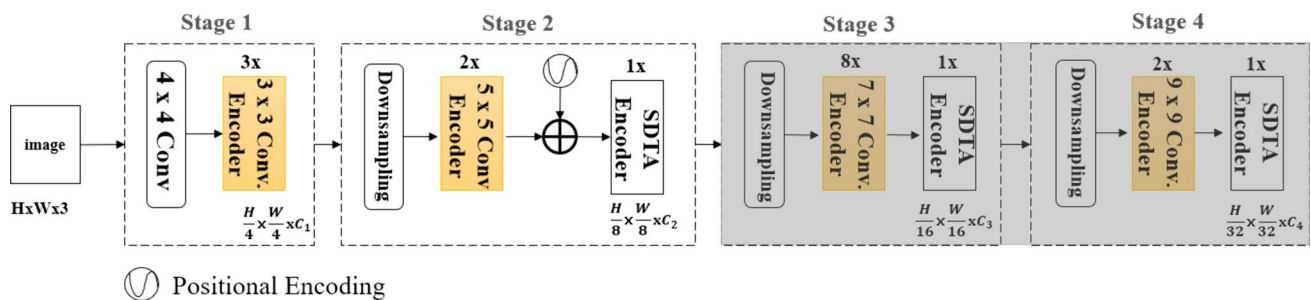


Fig. 9 Architecture of the EdgeNeXt

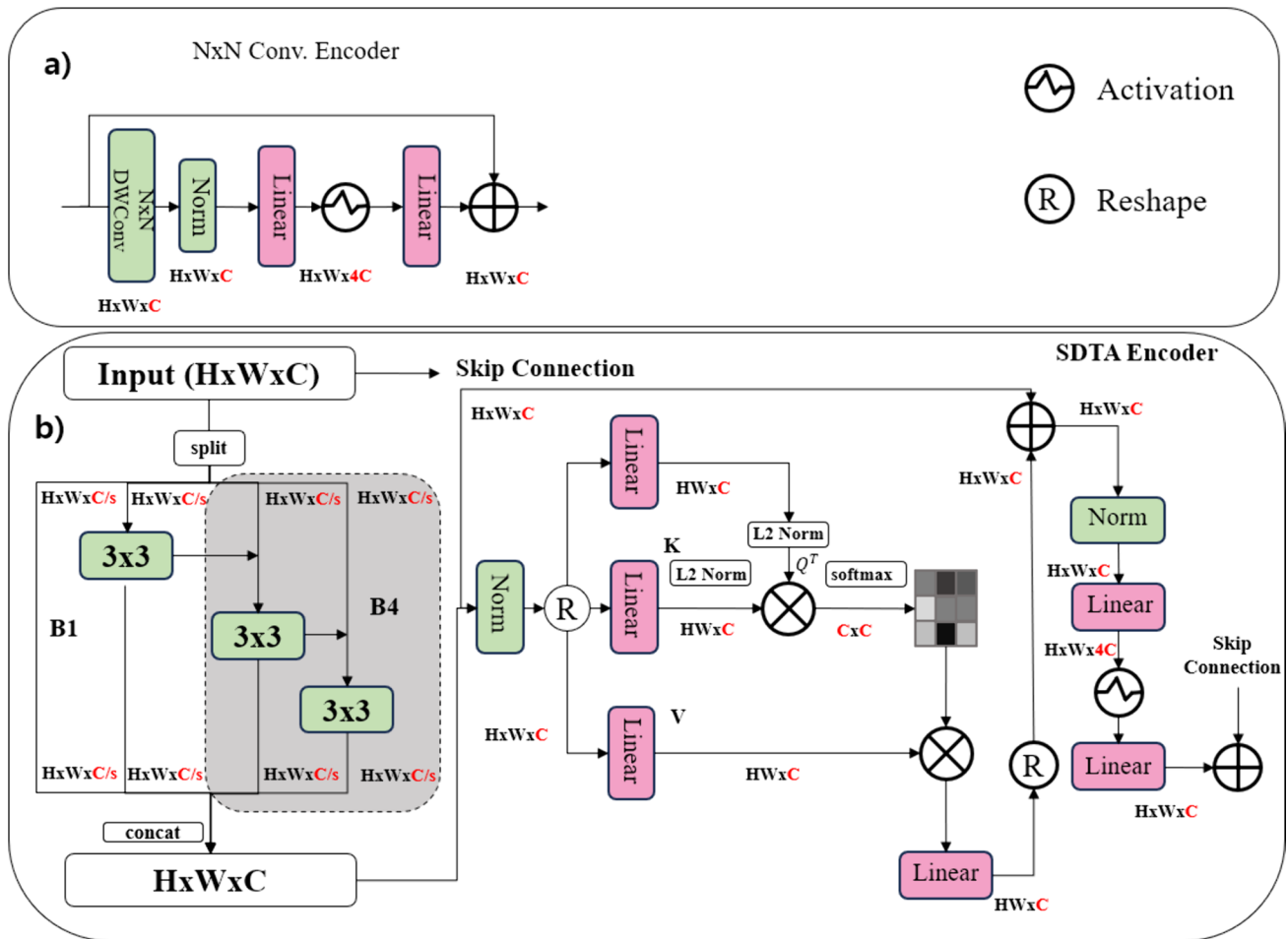


Fig. 10 a the convolutional encoder, b the SDTA encoder

UNet [58] and Residual UNet [59], respectively. In their original paper for model evaluation, three mammography datasets were used, which are the Curated Breast Imaging Subset of Digital Database for Screening Mammography [60], INbreast [61], and a private dataset.

Connected-UNets variations are capable of small sized tumor predictions and over perform the standard architectures like UNet, Attention UNet, and Residual UNet. The limitation of the model is using ASPP blocks, which easily discard local detail characteristics in large amounts [62].

### 2.2.3 DDANet

The DDANet model is based on an encoder-decoder architecture with applied residual blocks and a squeeze and excitation layer [63]. Architecture of the model presented in Fig. 5. Due to the fact that the training error tends to rise as neural networks become more complex and a particular layer's activation tends to zero deeper in the network, the residual blocks were applied to form an identity function

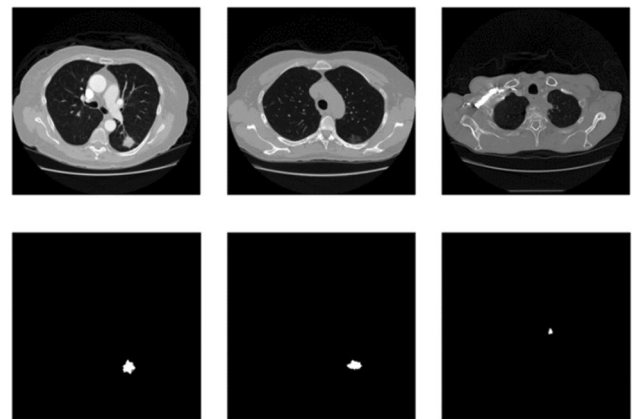


Fig. 11 Images from LIDC dataset after windowing and masks (segmented lung nodules)

that addresses this problem. The squeeze and excitation layer plays a role as a channel-wise attention mechanism for addressing the problem of CNN, where every feature



channel is equally important. The detailed structure of the residual block, encoder block, and decoder block is provided in a), b), and c) parts of Fig. 6, respectively.

There are two parallel decoders after the shared encoder. Each of the decoders give two outputs; the first output is a mask and the second output is a grayscale image. The second decoder generates an attention map, which helps improve the semantic representation of the feature maps [64]. Attention mechanisms applied in the network increase the performance of the model by capturing essential local features, filtering irrelevant information, and improving long-range dependencies.

The model can be used for real-time prediction. However, the real-time predictions require increased computational resources due to the necessity for parallel processing of data, which can be considered a potential limitation of the model.

### 2.2.4 UTRNet

The UTRNet model is based on Transformer technology combined with CNN in a U-shape encoder-decoder architecture [65]. The architecture of the model is provided in Fig. 7. The hybrid architecture offers a beneficial impact on capturing local and global dependencies. The attention block of the transformer is efficient self-attention, which was proposed in the original paper. The key distinction between pair-wise attention and efficient self-attention lies in the latter's ability to capture feature maps from all regions, including boundary regions. This is facilitated by employing distinct projections to transform a key and a value into the embedding space. The residual block and transformer encoder block are shown in a), b) parts of Fig. 8, respectively. The model was evaluated on a cardiac MRI dataset that consisted of 150 annotated images of segmented left ventricle (LV), right ventricle (RV), and left ventricular myocardium (MYO).

Due to efficient self-attention, the model has linear computational complexity. Nevertheless, the model exhibits outcomes that suffer from either over-partitioning or under-partitioning issues in cases of large amounts of small object detection.

### 2.2.5 EdgeNeXt

The EdgeNeXt model was developed as a general purpose model with light weights that can be implemented on edge devices like cameras, sensors, embedded systems, and personal devices [66]. The EdgeNeXt has a hybrid architecture that combines CNN and transformer. Figure 9 presents the architecture of the model. The model uses split depth-wise transpose attention (STDA) as an encoder, which increases receptive field and encoded multi-scale features by applying depth-wise convolution along with self-attention across channel dimensions of multichannel groups from splitted input tensors. Figure 10 shows Convolutional encoder and SDTA encoder in the a) and b) parts, respectively. Previous models like ViT used multi-headed self-attention (MHA), which has a high computational cost. The EdgeNeXt model contains efficient multi-head self-attention (MSHA).

EdgeNeXt, known for its remarkable computational efficiency and versatility across tasks, particularly excels in real-time predictions owing to its lightweight architecture. However, its original testing on non-medical images necessitates careful evaluation before applying it to medical imagery for an accurate prognosis.

## 2.3 Data Preprocessing

Prior to feeding the dataset into the models, a data preprocessing stage was conducted [67]. To enhance image quality and optimize input for the network, we applied a series of preprocessing steps such as windowing, thresholding, and resizing. These preprocessing steps are fundamental in preparing medical images for analysis. Windowing helps to highlight certain ranges of pixel values, enhancing the visibility of particular structures or pathologies. Thresholding aids in isolating regions of interest based on pixel intensity. Resizing ensures that the image is compatible with the input requirements of the network. This diligent preparation ultimately leads to more accurate and meaningful analysis, benefiting medical professionals in their diagnostic and treatment decisions. Original images are CT images in DICOM format with a size

**Table 2** Models implementation details

Model	Dataset size	Learning rate	Batch size	Optimizer	Loss function
Salient attention UNet	510 images	0.0001	4	Adam optimizer	dice loss
Connected-UNets	After augmentation: 13,302 images	0.0001	16	Adam optimizer	dice loss
DDANet	1000 images	0.0001	1	Adam optimizer	binary cross-entropy and Dice loss
EdgeNeXt	ImageNet: 1.28 M training and 50 K validation images for 1000 categories	0.006	4096	AdamW	cross-entropy loss
UTNet	150 images	0.05	16	SGD optimizer	dice loss and cross-entropy loss

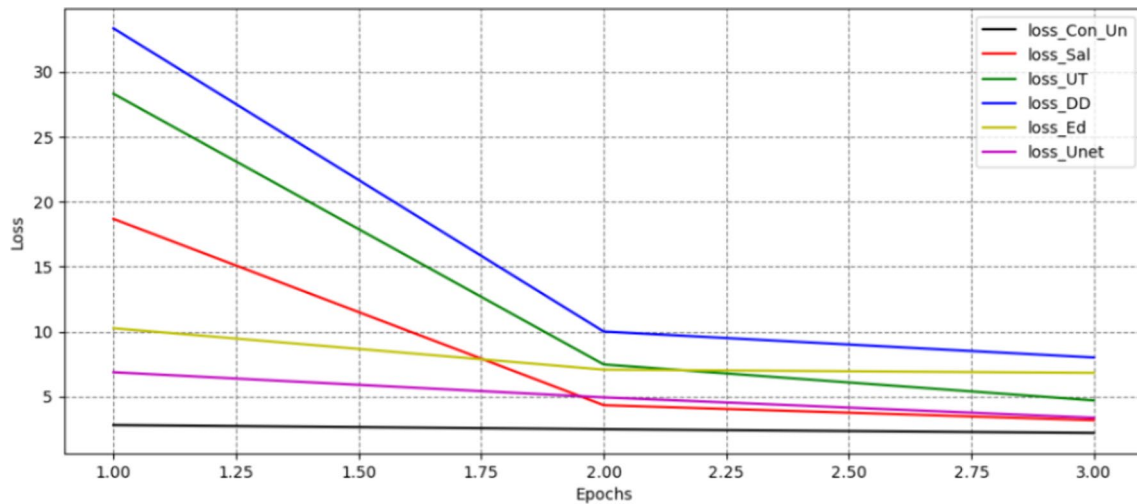


Fig. 12 Loss and dice score through epochs of the models

Table 3 Models performance evaluation

#	Reference	Model name	Avg DSC/performance evaluation (%)	Avg HD, Hausdorff distance
1	[54]	Salient Attention UNet	95.66	1.40
2	[56]	Connected-UNets	<b>97.80</b>	<b>1.29</b>
3	[63]	DDANet	78.33	1.60
4	[65]	UTNet	93.66	1.60
5	[66]	EdgeNeXt	93.33	1.90
6	[70]	U-Net	72.50	1.62

of 512\*512 pixels. We scaled all images to the size of 224\*224 pixels for models, and then we windowed them to the pixel range [0,255]. After the windowing substage, thresholding was applied. Figure 11 displays images from the LIDC dataset post-windowing, alongside corresponding masks highlighting segmented lung nodules.

With the exception of the Connected-UNets model, most models underwent only the resizing preprocessing step. In contrast to Connected-UNets, we opted for windowing instead of histogram equalization and Otsu's thresholding instead of normalization within the [0, 1] range for the dataset. While histogram equalization enhances global contrast, it may not preserve local contrast as effectively as windowing, which is specifically tailored for targeted enhancements. Moreover, windowing offers users the ability to manipulate the window center and width interactively, enabling adjustments to optimize the visualization of specific structures.

In terms of pixel value scaling, Otsu's thresholding operates based on intensity levels, while normalization to the range [0, 1] standardizes pixel values for easier comparison

or image processing. Notably, Otsu's thresholding is particularly well-suited for image segmentation, as it identifies an optimal threshold, effectively separating the image into distinct classes. This adaptive approach is advantageous in scenarios where automatic and accurate determination of the threshold is essential for reliable segmentation results.

## 2.4 Evaluation Metrics

In order to quantitatively assess the segmented results from all five models of the experiment, we used the Dice similarity coefficient (DSC) and Hausdorff distance (HD). The DSC metric is defined as follows:

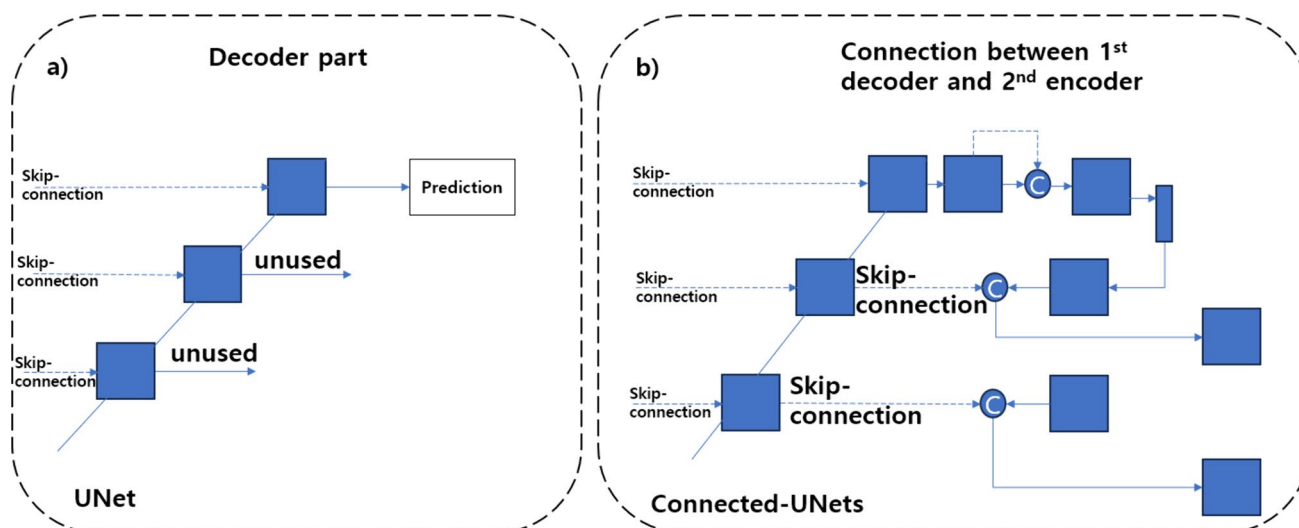
$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \quad (1)$$

where  $|X|$  and  $|Y|$  represent the number of elements in each set.

The main purpose of HD is to obtain all the locations in the image that match the model. This metric for A to B is defined as follows:

$$\delta_H^{\sim}(A, B) = \max_{a \in A} \min_{b \in B} ||a - b|| \quad (2)$$

where  $A = \{a_1, a_2, \dots, a_n\}$  and  $B = \{b_1, b_2, \dots, b_n\}$  are two point sets in used in  $E^2$ .



**Fig. 13** The difference between the basic UNet and the connection of two basic UNets in connected-UNets **a** basic UNet decoder **b** modified skip-connection part of the connected-UNets

**Table 4** Models complexity evaluation

#	Model name	Trainable parameters	Depth of the model (number of layers)
1	Salient attention UNet	1.912.517	77
2	Connected-UNets	<b>20.052.673</b>	128
3	DDANet	6.835.638	<b>398</b>
4	UTNet	28.865	278
5	EdgeNeXt	1.158.049	192

### 3 Results

#### 3.1 Performance Comparison

The comparison of five models' performance was presented in this section. The models were trained using a NVIDIA GeForce RTX 3090 GPU, and they were trained on 16 batch sizes on the PyTorch [68] and TensorFlow [46] frameworks. We modified the EdgeNeXt model, originally designed for multi-class classification, to suit our specific task by converting it into a binary classification model. Binary cross entropy was used as a loss function for all models. For model optimization, the adaptive moment estimation optimizer (Adam) [69] was applied.

We changed two crucial hyperparameters, specifically the learning rate and batch size, from their original values in

the model. The learning rate, which dictates the magnitude of steps taken during the optimization process, plays a significant role in shaping the training dynamics of the model. The initial values for the learning rate and batch size of the models can be found in Table 2.

In our experimental setup, we adjusted the learning rate and batch size to 0.001 and 16, respectively. A larger learning rate is beneficial for navigating smooth and flat optimization landscapes, helping the model escape flat regions and converge to the optimal solution more rapidly, particularly when dealing with large datasets. It's worth noting that, when comparing dataset sizes across models (excluding EdgeNeXt, as outlined in Table 2), LIDC-IDRI stands out with a dataset containing 244, 527 images. In the case of EdgeNeXt, our study indicated that such high learning rates or batch sizes were not necessary.

In contrast to the Salient Attention UNet and DDANet, we employed larger batch sizes, offering more stable gradient estimates that could result in a smoother convergence. This approach reduces noise in updates, potentially enhancing the stability of the optimization process. On the other hand, when compared to EdgeNext, the utilization of a smaller batch size impacts the model by infrequent updates to the parameters during each iteration. This might result in a slower progression of the optimization process while also demanding less memory.

The evaluated models' performance is listed in Table 2. The training was stopped after 10 epochs when the loss of the validation set did not improve. Figure 12 illustrates the progression of loss and DSC score across epochs for the models.

Table 3 includes the evaluated results on model complexity. There are two different values that were used for model

**Table 5** The conclusion on the results obtained from the comparative study

#	Method	Advantages	Limitation
1	Salient attention UNet	Learning ability of salient features across multiple levels of feature abstraction	Middle computational complexity
2	Connected-UNets	Recovering ability of the fine-graded features and consideration of both global and local features	High computational complexity
3	DDANet	High generalization ability as well as high speed of segmentation	High computational complexity
4	UTNet	Ability of capturing long-range dependencies and linear computational complexity	Middle level segmentation ability
5	EdgeNeXt	A resource-efficient network with the ability to learn multi-scale features	Low segmentation ability

description-e.g. the number of trainable parameters and the model depth, which equal the number of layers.

## 4 Discussion and Conclusions

In this paper, a comparative study is presented for lung nodule detection using computed tomography images. Preprocessing was carried out to obtain better performance results by enhancing image quality, modifying the input size for the network, and fine-tuning specific areas of interest within the image. We used two metrics such as Dice similarity coefficient (DSC) and Hausdorff distance (HD) for model evaluation. Additionally, Table 3 provides the evaluated results on the model complexity, such as the number of trainable parameters and the model depth. Our experiments show that Connected-UNets reached the best results with 97.80% Dice score and HD equal to 1.29 Table 2 due to connection between two UNets which contributed in saving fine-grained features from the layers of the first decoder. The difference of using single UNet and Connected-UNets is detailed in Fig. 13.

However, Connected-UNets is the most complex model with the highest number of trainable parameters and layers (Table 3) over other models. Consequently, it has a high computational cost. Due to the presence of a large number of trainable parameters in the model, it can potentially lead to overfitting which occurs when a model acquires an excessively detailed understanding of the training data, capturing noise or random fluctuations rather than the underlying patterns.

The second best performance results reached by Salient Attention UNet. This is due to usage of Salient maps and attention blocks which highlighted the most visually important regions in an image. The internal structure of the salient attention block is also important. The cascade of convolutional layers Conv  $3 \times 3$ , Conv  $3 \times 3$ , and Conv  $1 \times 1$  (Fig. 3) allowed for the capture of more complex patterns in a non-linear way. The receptive field of such a sequence might be larger than the usage of individual convolution  $7 \times 7$ . Additionally, the application of salient maps addresses the lack

of interpretability in deep learning models by highlighting regions of input data that are deemed most relevant for a given prediction.

The UTNet achieved the third best performance by adapting an efficient self-attention mechanism which is a light version of standard multi-head attention. Similarly, with Connected-UNets, UTNet also tried to save fine-grained details from decoder layers by applying a Transformer mechanism on top of the skip connections. However, compared with Connected-UNets, the UTNet cannot achieve the same results (Table 4).

Overall, this comparative study results revealed a dichotomy in the field of medical image processing. The summary of five models regarding their advantages and limitations evaluated in the study can be found in Table 5 in the Appendix. On one hand, there are sophisticated, computationally expensive models, and on the other hand, there are lighter models that, unfortunately, do not exhibit significant accuracy. This underscores the imperative for continued research and development in medical image processing to strike a balance between computational efficiency and model accuracy. Further exploration and innovation are needed to bridge the existing gaps and advance the capabilities of models in this critical domain.

## Appendix

See Table 5.

**Acknowledgements** This research was supported by the National Research Foundation of Korea grant funded by the Korea government (MSIT) (NRF-2022R1C1C1012107). This research was partially supported by the Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220101). This research was also partially supported by the Technology Innovation Program (No. 20022442 and No. 20024893) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

1. Early-life origins of respiratory diseases: a key to prevention. (2020) *Lancet respiratory medicine*, The 8 (10): 935–935. <https://pubmed.ncbi.nlm.nih.gov/33007283/>
2. Ochani RK, Asad A, Yasmin F, Shaikh S, Khalid H, Batra S, Sohail MR, Mahmood SF, Ochani R, Arshad MH, Kumar A, Surani S (2021) COVID-19 pandemic: from origins to outcomes. A comprehensive review of viral pathogenesis, clinical manifestations, diagnostic evaluation, and management. *Infez Med* 29(1):20–36
3. Yüce M, Filiztekin E, Özkaya KG (2021) COVID-19 diagnosis — A review of current methods. *Biosens Bioelectron* 172:112752
4. World Health Organization (WHO) (2022) Cancer statistics worldwide. <https://www.who.int/news-room/fact-sheets/detail/cancer>.
5. National Lung Screening Trial Research Team, Aberle DR, Adams AM et al (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 365:395–409
6. Aresta G, Jacobs C, Araújo T, Cunha A, Ramos I, Ginneken B, Campilho A (2019) iW-Net: an automatic and minimalistic interactive lung nodule segmentation deep network. *Sci Rep* 9:11591
7. Wang S, Zhou M, Liu Z, Liu Z, Gu D, Zang Y (2017) Central focused convolutional neural networks: developing a data-driven model for lung nodule segmentation. *Med Image Anal* 40:172–183
8. Erickson BJ, Bartholmai B (2002) Computer-aided detection and diagnosis at the start of the third millennium. *J Digit Imaging* 15(2):59–68
9. Raad KB de, Garderen KA van, Smits M, Voort SR van der, Incekara Oei EHG, Hirvasniemi J, Klein S, Starmans MPA (2021) The effect of preprocessing on convolutional neural networks for medical image segmentation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
10. Balkenende L, Teuwen J, Mann RM (2022) Application of deep learning in breast cancer imaging. *Semin Nucl Med* 52(5):584–596
11. Sari S, Soesanti I, Setiawan NA (2021) Best performance comparative analysis of architecture deep learning on ct images for lung nodules classification. In: *Proceedings-2021 IEEE 5th international conference on information technology information systems and electrical engineering: applying data science and artificial intelligence technologies for global challenges during pandemic era ICITISEE 2021*, 138–143
12. Cui X, Zheng S, Heuvelmans MA, Yihui Du, Sidorenkov G, Fan S, Li Y, Xie Y, Zhu Z, Dorrius MD, Zhao Y, Veldhuis RNJ, de Bock GH, Oudkerk M, van Ooijen PMA, Vliegenthart R, Ye Z (2022) Performance of a deep learning-based lung nodule detection system as an alternative reader in a Chinese lung cancer screening program. *European J Radiol* 146:110068
13. Traoré A, Ly AO, Akhloufi MA (2020) Evaluating deep learning algorithms in pulmonary nodule detection. In: *2020 42nd annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. Montreal, QC, Canada, pp 1335–1338
14. Zhou L, Li Li, Tianran Li, Douqiang L, Xiaoliang W, Dehong L (2020) Does a deep learning-based computer-assisted diagnosis system outperform conventional double reading by radiologists in distinguishing benign and malignant lung nodules? *Front Oncol*. <https://doi.org/10.3389/fonc.2020.545862>
15. Yang K, Liu J, Tang W, Zhang H, Zhang R, Jun G, Zhu R, Xiong J, Xiaoshuang R, Jianlin W (2020) Identification of benign and malignant pulmonary nodules on chest CT using improved 3D U-Net deep learning framework. *European J Radiol* 129:109013
16. Shi J, Ye Y, Zhu D, Lianta S, Huang Y, Huang J (2021) Comparative analysis of pulmonary nodules segmentation using multiscale residual U-Net and fuzzy C-means clustering. *Comput Method Programs Biomed* 209:106332. <https://doi.org/10.1016/j.cmpb.2021.106332>
17. Murchison JT, Ritchie G, Senyszak D, Nijwening JH, van Veenendaal G, Wakkie J et al (2022) Validation of a deep learning computer aided system for CT based lung nodule detection, classification, and growth rate estimation in a routine clinical population. *PLoS ONE* 17(5):e0266799
18. Ciompi F, Chung K, van Riel S et al (2017) Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Sci Rep* 7:46479. <https://doi.org/10.1038/srep46479>
19. Vaswani, A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. <https://arxiv.org/abs/1706.03762>
20. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M, Botsis T (2017) Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J Biomed Inform* 73:14–297
21. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M (2022) Transformers in vision: a survey. *ACM Comput Surv* 54(10s):1–41
22. Hao S, Zhou Y, Guo Y (2020) A brief survey on semantic segmentation with deep learning. *Neurocomputing* 406:302–321
23. Rumelhart D, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature*. <https://doi.org/10.1038/323533a0>
24. Hochreiter S, Schmidhuber J (1997) Long Short-term Memory. *Neural Comput* 9:1735–1780
25. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Part of *advances in neural information processing systems 25 (NIPS 2012)*
26. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Thomas U (2020) An image is worth 16x16 words: transformers for image recognition at scale. <https://arxiv.org/abs/2010.11929>
27. Hastie T, Tibshirani, R, Friedman J (2008) Overview of supervised learning. *The elements of statistical learning*, 9–41
28. Khalid K, Nripendra KS (2021) A tour of unsupervised deep learning for medical image analysis. *Curr Med Imag* 17(19):1059–1077
29. Li Y (2018) Deep reinforcement learning: an overview. <https://arxiv.org/abs/1701.07274>
30. Ma S, Li X, Tang J, Guo F (2022) EAA-Net: rethinking the autoencoder architecture with intra-class features for medical image segmentation. <https://arxiv.org/abs/2208.09197>
31. Huang J, Li H, Li G, Wan X (2022) Attentive symmetric autoencoder for brain MRI segmentation. *MICCAI 2022: medical image computing and computer assisted intervention–MICCAI 2022*. 203–213
32. Subramaniam S, Jayanthi KB, Rajasekaran C, Kuchelar R (2020) Deep learning architectures for medical image segmentation. In: *Annual IEEE symposium on computer-based medical systems*
33. Kayalibay B, Grady J, Smagt P (2017) CNN-based segmentation of medical imaging data. <https://arxiv.org/abs/1701.03056>
34. Tseng K, Zhang R, Chen C, Hassan M (2021) DNetUnet: a semi-supervised CNN of medical image segmentation for super-computing AI service. *J Supercomput* 77:3594–3615

35. Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D (2018) DRINet for medical image segmentation. *IEEE Trans Med Imaging* 37(11):2453–2462
36. Bai W, Suzuki H, Qin C, Tarroni G, Oktay O, Matthews PM, Rueckert D (2018) Recurrent neural networks for aortic image sequence segmentation with sparse annotations. *MICCAI 2018: medical image computing and computer assisted intervention–MICCAI 2018* 586–594
37. Kim S, An S, Chikontwe P, Park S (2021) Bidirectional RNN-based few shot learning for 3D medical image segmentation. In: *Proceedings of the AAAI conference on artificial intelligence: AAAI-21 technical tracks vol 35* 3 1808–1816
38. Monteiro M, Figueiredo MAT, Oliveira AL (2018) Conditional random fields as recurrent neural networks for 3D medical imaging segmentation. <https://arxiv.org/abs/1807.07464>
39. Huang X, Deng Z, Li D, Yuan X (2021) MISSFormer: an effective medical image segmentation transformer. <https://arxiv.org/abs/2109.07162>
40. Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, Fu Y, Feng J, Xiang T, Philip HST, Li Z (2020) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. <https://arxiv.org/abs/2012.15840>
41. Karimi D, Vasylychko SD, Gholipour A (2021) Convolution-free medical image segmentation using transformers. *MICCAI 2021: medical image computing and computer assisted intervention–MICCAI 2021*
42. Xun S, Li D, Zhu H, Chen M, Wang J, Li J, Chen M, Wu B, Zhang H, Chai X, Jiang Z, Zhang Y, Huang P (2022) Generative adversarial networks in medical image segmentation: a review. *Comput Biol Med* 140:105063
43. Sun Y, Yuan P, Sun Y (2020) MM-GAN: 3D MRI data augmentation for medical image segmentation via generative adversarial networks. In: *2020 IEEE international conference*
44. Yan W, Wang Y, Gu S, Huang L, Yan F, Xia L, Tao Q (2019) The domain shift problem of medical image segmentation and vendor-adaptation by Unet-GAN. *Medical image computing and computer assisted intervention–MICCAI*
45. Xie Y, Zhang J, Shen C, Xia Y (2021) CoTr: efficiently bridging CNN and transformer for 3D medical image segmentation. *Medical image computing and computer assisted intervention–MICCAI*
46. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth H, Xu D (2021) UNETR: transformers for 3D medical image segmentation. <https://arxiv.org/abs/2103.10504>
47. Tang Y, Yang D, Li W, Roth H, Landman B, Xu D, Nath V, Hatamizadeh A (2022) Self-supervised pre-training of swin transformers for 3D medical image analysis. <https://arxiv.org/abs/2111.14791v2>
48. Armato SG III, Roberts RY, McNitt-Gray MF, Meyer CR, Reeves AP, McLennan G, Engelmann RM, Bland PH, Aberle DR, Kazerooni EA, MacMahon H, van Edwin JRB, Yankelevitz D, Croft BY, Clarke LP (2007) The lung image database consortium (LIDC): ensuring the integrity of expert-defined ‘truth.’ *Acad Radiol* 14:1455–1463
49. Mustra M, Delac K, Grgic M (2008) Overview of the DICOM standard. In: *50th international symposium ELMAR*
50. McNitt-Gray MF, Armato SG III, Meyer CR, Reeves AP, McLennan G, Pais RC, Freymann J, Brown MS, Engelmann RM, Bland PH, Laderach GE, Piker C, Guo J, Towfic Z, Qing DPY, Yankelevitz DF, Aberle DR, van Beek DJR, MacMahon H, Kazerooni EA, Croft BY, Clarke LP (2007) The lung image database consortium (LIDC) data collection process for nodule detection and annotation. *Acad Radiol* 14:1464–1474
51. Doulamis N, Doulamis A, Protopapadakis E (2018) Deep learning for computer vision: a brief review. *Comput Intell Neurosci* 2018:7068349
52. Ahamed MA, Imran AAZ (2022) Joint learning with local and global consistency for improved medical image segmentation. *MIUA 2022: medical image understanding and analysis*. 298–312
53. Neyshabur B, Bhojanapalli S, McAllester D, Srebro N (2017) Exploring generalization in deep learning. *Advances in neural information processing systems* 30
54. Vakanski A, Xian M, Freer PE (2020) Attention enriched deep learning model for breast tumor segmentation in ultrasound images. *Ultrasound Med Biol* 46(10):2819–2833
55. Liu L, Cheng J, Quan Q, Wu F, Wang Y, Wang J (2020) A survey on U-shaped networks in medical image segmentations. *Neurocomputing* 409:244–258
56. Baccouche A, Garcia-Zapirain B, OleaCastillo C, Elmaghraby AS (2021) Connected-UNets: a deep learning architecture for breast mass segmentation. *NPJ Breast Cancer* 7:151
57. Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2016) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. <https://arxiv.org/abs/1606.00915>
58. Oktay O, Schlemper J, Folgoc L, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B, Rueckert D (2018) Attention U-Net: learning where to look for the pancreas. <https://arxiv.org/abs/1804.03999>
59. Zhang Z, Liu Q, Wang Y (2017) Road extraction by deep residual U-Net. <https://arxiv.org/abs/1711.10684>
60. Smith K, Rutherford M (2017) Cancer imaging archive. Curated breast imaging subset of digital database for screening mammography (CBIS-DDSM). <https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM>
61. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS (2020) INbreast: toward a full-field digital mammographic database. *Acad Radiol* 19:236–248
62. Lian X, Pang Y, Han J, Pan J (2021) Cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation. *Pattern Recogn* 110:107622
63. Tomar NK, Jha D, Ali S, Johansen HD, Johansen D, Riegler AM, Halvorsen P (2021) DDANet: dual decoder attention network for automatic polyp segmentation. *ICPR international workshop and challenges*. <https://arxiv.org/abs/2012.15245>
64. Lu Y, Zhang W, Jin C, Xue X (2012) Learning attention map from images. In: *2012 IEEE conference on computer vision and pattern recognition*
65. Gao Y, Zhou M, Metaxas D (2021) UTNet: A hybrid transformer architecture for medical image segmentation. *MICCAI 2021*. <https://arxiv.org/abs/2107.00781>
66. Maaz M, Shaker A, Cholakkal H, Khan S, Zamir SW, Anwer RM, Khan FS (2022) EdgeNeXt: efficiently amalgamated CNN-transformer architecture for mobile vision applications. *ECCVW 2022 (oral, CADL: computational aspects of deep learning)*. <https://arxiv.org/abs/2206.10589>
67. Sonka M, Hlavac V, Boyle DR (2014) *Image pre-processing. Image processing, analysis and machine vision*. 56–111
68. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G (2019) PyTorch: an imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32
69. Kingma DP, Ba J (2017) Adam: a method for stochastic optimization. <https://arxiv.org/abs/1412.6980>
70. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. *MICCAI 2015: medical image computing and computer-assisted intervention–MICCAI 2015*, 234–241

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



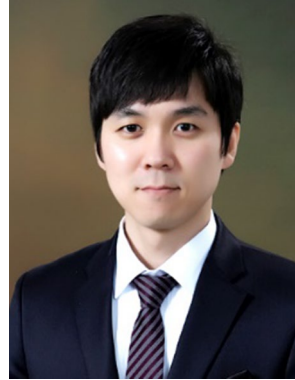
**Aliya Orazalina** completed her M.S. degree in Computer Science and Engineering from Dankook University, specializing in medical image processing. She received her B.S. degree in Computer Science and Engineering from Omsk State Technical University. During her graduate studies, she conducted research on comparing deep learning algorithms for lung cancer detection using medical imaging data, with a particular emphasis on improving diagnostic accuracy and efficiency.



**Heechul Yoon** received his B.S. and M.S. degrees in electrical engineering from Sogang University, Seoul, South Korea, in 2008 and 2010, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2018. From 2010 to 2014, he worked at Samsung Electronics, Suwon, South Korea, as a Researcher of ultrasound signal and image processing. In 2014, he began his Ph.D. studies at The University of Texas at Austin, Austin, TX,

USA, where he conducted studies at the Ultrasound Imaging and Therapeutics Laboratory. In 2015, the Laboratory moved to the Georgia Institute of Technology. He conducted his research at Georgia Tech and Emory University School of Medicine, Atlanta. From 2019 to 2021, he worked at Siemens Healthineers, Issaquah, WA, USA, as a Staff Engineer. He is currently an Assistant Professor with the Department of Electronics and Electrical Engineering, Dankook university. His

research interests include the development and clinical translation of medical imaging technologies including shear-wave elasticity imaging and ultrasound-guided photoacoustic imaging.



**Sang-Il Choi** received the B.S. degree from the Division of Electronic Engineering, Sogang University, South Korea, in 2005, and the Ph.D. degree from the School of Electrical Engineering and Computer Science, Seoul National University, South Korea, in 2010. He was a Post-Doctoral Researcher with BK21 Information Technology, Seoul National University, in 2010, and the Computer Science Department, Institute for Robotics and Intelligent Systems, University of Southern California, Los

Angeles, CA, USA, until 2011. He is currently a Professor with the Department of Computer Engineering, Dankook University, South Korea. His research interests include pattern recognition, machine/deep learning, computer vision, and their applications.



**Seokhyun Yoon** received his B.S. and M.S. degrees in electronics engineering from Sung Kyun Kwan University, Suwon, Korea, in 1992 and 1996, respectively, and his Ph.D. degree in electrical and computer engineering from the New Jersey Institute of Technology, Newark, in 2003. In 1999, he was with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea. During 2003–2005, he was with the Telecomm. R&D Center, Samsung Electronics Co., Ltd., Suwon, Korea,

where he worked on technologies for wireless/mobile air interfaces. Currently, he is a professor in the Department of electronics engineering, Dankook University, Yongin-si, Kyunggi-do, Korea. His research activities are focused on wireless communications, signal processing and bioinformatics. Dr. Yoon was awarded the Hashimoto Prize in 2003 for his research achievements through the Ph.D. course.