Computers in Biology and Medicine 150 (2022) 106083

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Attention UW-Net: A fully connected model for automatic segmentation and annotation of chest X-ray

Debojyoti Pal [a], Pailla Balakrishna Reddy [b], Sudipta Roy [a,*]

[a] Artificial Intelligence and Data Science, Jio Institute, Navi Mumbai, 410206, India
[b] Reliance Jio - Artificial Intelligence Centre of Excellence (AICoE), Hyderabad, India

ARTICLE INFO

Keywords:
Deep learning
U-net
Automatic annotations
Chest X-rays
Attention gates
Segmentation
Medical image

ABSTRACT

Background and objective: Automatic segmentation and annotation of medical image plays a critical role in scientific research and the medical care community. Automatic segmentation and annotation not only increase the efficiency of clinical workflow, but also prevent overburdening of radiologists. The objective of this work is to improve the accuracy and give a probabilistic map for automatic annotation from small data set to reduce the use of tedious and prone to error manual annotations from chest X-rays.

Method: In this paper, we have proposed an attention UW-Net, which introduces an intermediate layer acting as a bridge between the encoder and decoder pathways. The intermediate layer is a series of fully connected convolutional layers generated from the upsampling of the final encoder layer connected to the corresponding up sampled and down sampled blocks via skip-connections. The intermediate layer is further connected to the decoder pathway using a downsampling layer.

Results: The proposed attention UW-Net is giving a very good performance, achieving an average F1-score of 95.7%, 80.9%, 81.0% and 77.6% for lung (large), heart (medium), trachea (small), and collarbone (small) object segmentations, respectively. The attention UW-Net outperforms not only in comparison to U-Net and its variations but also with respect to other standard recent automatic and semi-automatic segmentation/annotation models. An ablation study was also performed to find the best suited high-performing architecture.

Conclusion: The uniformity in prediction accuracy of segmentation masks for all kinds of segmentation masks (large, medium, and small lesions) makes this model best for automatic annotation of organs.

1. Introduction

The increasing volume of clinical data in medical imaging needs a fast identification and analysis of specific features in X-ray images. But analysis of many X-rays for a long time reduces the speed and accuracy of radiologists' annotation and diagnosis of diseases, thereby slowing down imaging technicians in capturing, screening, and diagnosing patient data. With the advent of U-Net [1], there has been a resurgence in the field of computer vision in medical imaging using deep learning architectures, especially for semantic segmentation. The state-of-the-art techniques for semantic segmentation include variants of U-Net [2] such as Residual and Attention U-Nets along with the characteristic encoder-decoder architectures such as ResNet [3] and Dense Net [4]. These architectures share a key similarity, the skip connections. These skip connections are followed by copy and concatenation blocks. The skip connections along with attention blocks have enabled the model to emphasize key semantic features and dependencies which in turn helps in the detection of finer details of target objects.

One major difficulty in chest X-ray (CXR) segmentation and annotation is due to its variation in shape, size (due to age), gender, and the overlapping of clavicles and rib cage [5]. Moreover, the positional relationship of the heart with respect to the left and right lung is crucial for the generation of an accurate segmentation mask of the heart. However, when medical experts annotate the lung fields, they look for certain consistent structures surrounding the lung fields. Most of the recent works on medical imaging have been focused on Magnetic resonance imaging (MRI) [6], positron emission tomography (PET) [7] and classification of X-ray scan images [8] of brain, breast, liver, chest and prostate regions for an anomaly, tumor, and cancer detections. The U-Net has outperformed almost all previously known deep-learning models for target lesion segmentation and is recognized as a go-to model for segmentation tasks. However, the previously mentioned

* Corresponding author.
E-mail addresses: Debojyoti.Pal@jioinstitute.edu.in (D. Pal), balakrishna.pailla@ril.com (P.B. Reddy), sudipta1.roy@jioinstitute.edu.in (S. Roy).

https://doi.org/10.1016/j.compbiomed.2022.106083
Received 7 June 2022; Received in revised form 14 August 2022; Accepted 3 September 2022
Available online 10 September 2022

models along with U-Net have several limitations that have not been addressed. One of the limitations is their inability to accurately predict smaller objects and showcase the anatomical features intrinsic to the segmented organs/regions of interest (ROIs) in the predicted masks. These features include the physical distance of separation between multiple organs. One of the limitations the proposed model aims at solving is the identification of the objects of interest with a nonstandard shape when no other image-derived features are involved.

Given the limitations of U-Net architecture and the lack of a generalized model in the domain of X-ray scans that work equally well for small and large objects of interest using a small image data set, this paper focuses on the 2D CXR images and proposes a modified version of the popular U-Net model which we call UW net. Finally, we implement attention gates to further improve the prediction accuracy of attention UW-Nets for small lesion segmentation. We introduced an intermediate layer which is a series of fully connected convolution layers connected to the corresponding encoder and decoder layers by skip connections. It helps in reducing the semantic gap between feature maps of encoder and decoder layers and makes the model focus on intrinsic details by re-introducing the features lost during downsampling actions. We also implement attention layers in the UW-Net architecture to enable localization of essential features and improve accuracy for small lesion segmentation.

The rest of the section is described as follows: Section 2 describes the literature review, and Section 3 discusses the model, the dataset used and the implementation details. Section 4 showcases the results obtained after performing a comparative study of the proposed attention UW-Net via comparison with other models of the U-Net family and with other segmentation models. Section 5 is used to discuss the ablation study performed on the attention UW-Net. Finally, we conclude our paper in Section 6.

2. Literature review

X-ray is a widely used manual screening technique used by medical practitioners to detect abnormalities in the chest region. CXR images provide information about the size, shape and location of the heart, lungs, bones and bronchi using X-rays. The implementation of several deep learning architectures trained on Graphics processing unit (GPU) based platforms have elevated the accuracy of medical image segmentations, due to their capability of handling large number of image datasets. However, this requires many time-consuming annotated data not only for the medical practitioners but also for the training of a model.

Early works on lung segmentation from CXR image included knowledge-based method approach [9] to make a clear distinction between low-level and high-level processing applied on raw data by mapping image edges associated with the anatomical model of lung boundary using parametric features. However, such techniques are noise-sensitive, thereby failing to generate a proper segmentation mask. To overcome this problem, various edge detection filters such as canny edge-detection and morphological operations [10] like dilation and erosion were used to produce segmented masks similar to the ground truth masks. However, these techniques can neither handle complex images nor segment smaller regions of interest.

With the advent of fully convolutional networks (FCNs) like Visual Geometry Group (VGG) Net [11] and ResNet were used to segment bio-medical images. The ability to extract high-level features using a fully connected down-sampling path followed by an up-sampling path led to the development of the Structural Correcting Adversarial Network (SCAN) [12] to segment lung fields and heart in CXR images by exploiting the generative power of graph variational encoders. The critic network guides the fully convolutional segmentation model to achieve precise segmentation masks. However, the fixed receptive size of FCNs and the huge class imbalance between foreground and background results in inaccurate segmentations of small organs.

The FCN approach needed a large amount of training data. Training such datasets incurred an increased runtime for the successful implementation of these models. To address the issue of increased time frame, stochastic computing was interwoven into Deep Convolutional Neural Networks (DCNNs) [13]. After the development of DCNN, Ronneberger et al. developed a novel state-of-the-art architecture, U-Net, as an extension of the FCN approach to address the problem of data availability. As shown in Fig. 1. U-Net consists of two main parts: the convolutional encoding and decoding units. The basic convolution operations are performed, followed by RELU activation in both network parts. For downsampling in the encoding unit, 2×2 max-pooling operations are performed. In the decoding phase, the convolution transpose (representing up-convolution or de-convolution) operations are performed to up-sample the feature maps. One main drawback of this architecture is that the skip connections impose an unnecessarily restrictive fusion scheme, forcing aggregation only at the same scale feature maps of the encoder and decoder sub-networks.

The architecture is for an input image of size ($128 \times 128 \times 3$). Each orange box corresponds to a multi-channel feature map. White boxes denote the multi-channel feature maps which are concatenated to the up-sampled feature maps in orange. Arrows denote operations.

The inability to solve the problem of small lesion segmentation and limitation to detect the organs of interest with a non-standard shape resulted in the development of several modified versions of the aforementioned architecture, which included the addition of attention gates, previously limited to natural language processing to be used in medical image segmentation tasks. The recursive usage of attention gates [14] was used to increase the receptive fields of convolutional filters and consider the relationship between tissues at a global level. However, the difficulty in reducing false positive predictions for small objects posed a hurdle for the effective generation of segmentation masks for smaller objects of interest. Deep learning models like bi-directional LSTMs (Long Short-Term Memory) [15] are used in NLP (Natural Language Process) to further enhance the outcomes of predicted masks by incorporating both spatial and temporal information. Most of the recent works on medical imaging, such as COVID-Net [16] have focused on anomaly detection and localization of lungs from CXR images [17,18] which in turn helps in the detection and classification of pulmonary diseases such as pneumonia [19] and COVID-19 [20] to name a few. However, COVID-Net uses a high number of images for training. In contrast, fully convolutional neural network for automatic lung segmentation [19] uses a number of post-post-processing techniques such as hole-filling algorithm to obtain desirable results. Other mentioned models incorporate pre-processing techniques such as image cropping to enlarge the aspect ratio of the targeted anomalous regions. The mentioned drawbacks have paved the way for the introduction of meta-heuristic approaches [21] to deep learning models. Their ability lies in solving complex optimization problems and multi-objective problems. As a result, learning-based optimization approaches such as Monarch Butterfly Optimization (MBO) [22], Earthworm Optimization Algorithm (EWA) [23], Moth Search (MS) algorithm [24], Slime Mould algorithm (SMA) [25], Hunger Games search (HGS) [26], Runge Kutta optimizer (RUN) [27], Colony Predation Algorithm (CPA) [28], and Harris Hawks Optimization (HHO) [29], Particle Swarm Optimization (PSO) [30], Dynamic Learning Evolution Algorithm [31] and Learning based Elephant Herding Optimization (EHA) [32] in deep-learning approaches. These algorithms are highly scalable and robust in terms of handling problems involving huge data points. However, the solutions obtained by swarm intelligent algorithms such as PSO, HHA, MBO, EHA, SMA, and CPA converge prematurely and have poor local optimization ability resulting in poor performance in complex optimization problems. Evolutionary algorithms such as GA and algorithms such as RUN are complex and have a higher runtime to find a convergent solution. The trade-off in terms of complexity versus accuracy is not good to replace classical/traditional optimizers. In addition, these models not only suffer from the problem of unbalanced exploitation but also fail to account

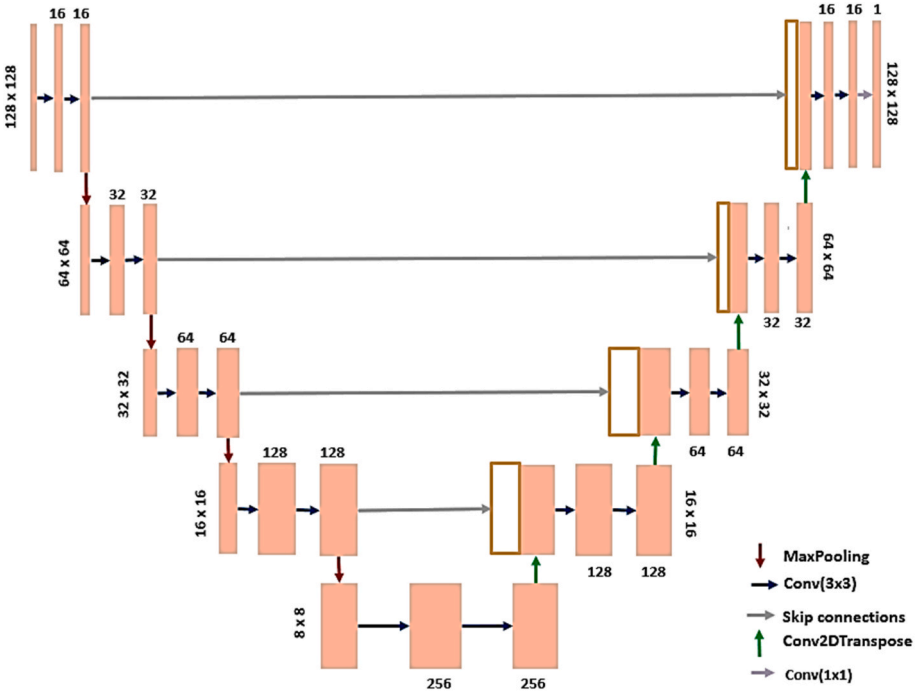

Fig. 1. The basic U-Net architecture.

for the intrinsic qualities such as texture which are inherent to the segmented organs. Recent developments take care of the drawback of ignoring the spatial distribution of the regions of interest in a higher dimensional space by including the addition of image-to-graph localized skip connections to give anatomically plausible segmentation results. The output graphs are directly sampled from a 2-D distribution learnt during training from the bio-medical images and are then projected on a 2-D latent space [33]. This modification is further implemented in 3-D images by a model architecture known as Voxel2Mesh [34]. It is built on the idea of deforming an elliptical mesh template topology.

The main challenge of the heterogeneous appearance of the target organ segmentation and annotation is still an open problem in medical image segmentation. The target organ or lesion may vary hugely in size, shape, and location from patient to patient. The size of the convolution kernel in the encoding and decoding layers of U-Net is fixed. Therefore, the diversity of features is lost due to the fixed receptive field of the convolutional layers. The sliding window approach was implemented in a few research to solve this problem. However, in the case of smaller targets, the sliding window approach fails to efficiently detect essential features required for accurate segmentation due to the receptive field of the convolution kernel being too large. Another limitation lies in detecting the physical distance relationship for multiple annotated segments/regions of interest (ROIs) or when the border of the organs of interest is unclear during manual annotations. All existing methods have failed to address the abovementioned problem, thereby failing to give accurate segmentation masks for small, medium, and big lesion segmentation on the same scale. To address the existing problem, we proposed an attention UW net with the following major contributions:

- The encoder-decoder architecture of U-Net has a sequence of convolutional layers which are fed forward to the decoder layer. This leads to the network learning redundant feature. The proposed network addresses the problem by adding a novel intermediate layer in between blocks B-5 and B-6 of the original U-Net architecture. This intermediate layer acts as a densely connected layer which helps the network learn a diverse set of features instead of redundant features.
- The intermediate layer aims at reinforcing the bottleneck of U-Net to enable a better exchange of intrinsic feature vectors. An added skip connection makes the network learn the details lost in the prior max-pooling steps and improves the representational power by reusing the features and enables the flow of information and combining the extracted features between the up sampled and down sampled layers.
- A modified attention gate was implemented in the proposed architecture to further improve the prediction accuracy for segmentation of small lesions. Modifications were made, precisely in the resampling of attention vectors, by replicating the vector space generated in the channel axis. This modification improves the performance of the network for small lesion segmentations by reducing information loss.
- Changing the way attention vectors are generated reduces the number of redundant pixels predicted as a segmentation mask (false positive predictions) not only improves the accuracy but also provides precise segmented masks of small lesions.
- Generating accurate segmentations for small and large lesions with a limited number of images (125 images) as training data. The proposed model achieves these results without using any pre- or post-processing techniques and without any pre-trained weights.

3. Proposed methodology

3.1. Dataset

Since the availability of large number of annotated samples is a challenge, we use the NIH Chest X-ray Dataset [35] which consists of 112,120 X-Ray images with disease labels of 14 different disease classes with an additional class for “no findings”. However, the unavailability of ground truth was a big challenge. Since manual annotations of small lesions takes considerable time a subset of 200 CXR was used as a dataset. In order to maintain generality and uniformity of the available data, the images were randomly sampled from the 14 classes.

3.2. Data annotations and splitting

Manual annotation of the dataset was done using VGG annotator [36]. Each image had an original dimension of 1024 × 1024 which was down sampled to a fixed dimension of 128 × 128 for significantly

accelerating the implementation of segmentation without compromising the accuracy. The n-gon was filled with 1 using fill function and augmented to an array of dimension 128 × 128 comprising of 0s to create the binary masks, representing the ground truth for validation of the data. The dataset was split into 120 images for training, 30 for validation and 50 for testing.

3.3. Attention UW-net architecture

The overall architecture of Attention UW net is shown in Fig. 2. The orange boxes represent feature vectors; the white boxes represent copied feature vectors to be concatenated. The yellow boxes represent the modifications done to the existing U-Net architecture. The grey arrows represent skip connections. Attention gates are represented by coloured circles, and dotted arrows represent the gating signals. The dashed rectangle represents the modified attention gate. The novel architecture is named as attention UW-Net since the architecture has a resemblance to the letter 'W,' uses a modified version of attention gates and has U-Net as a backbone (see Fig. 2). The proposed architecture has an encoding path followed by the intermediate layer and the decoding path. The attention layers followed by the skip connection act as a connection between different layers.

Encoding Path: The encoding path consists of five fully connected steps. Each step comprises of a series of block. The operation performed on each block of any step of the encoding path can be represented as a series of convolutional layer followed by a max-pooling layer. Each convolutional layer consists of a convolutional filter having a filter size of 3 × 3 followed by a RELU activation function, repeated twice. It is subsequently followed by a max-pooling layer of filter size 2 × 2. The purpose of the encoding path is to progressively extract features for image representations, increasing the dimensional representation of image. However, the proposed attention UW net architecture deviates from the U-Net architecture in the final step. Unlike the final step of U-Net, which comprises of a series of convolutional layers, the final step in UW-Net is considered as an extension of the previous encoding path and that has the same number of operations performed as in the other layers.

Intermediate Layer: The encoding path is connected to the intermediate layer (represented as a light blue box in Fig. 2), which is represented as an orange box in Fig. 2. The intermediate layer consists of a series of densely connected convolutional layers, which is responsible for improving the model performance by allowing the model to learn better intrinsic features. Since the last block of encoding layer has the largest filter size and the convolutional operations take place over the smallest input vector, it is responsible for the generation of the richest feature vectors. This layer acts as a bridge between the encoded and decoded paths of the network. However, unlike the U-Net, the intermediate layer combines feature maps of the fourth step of the encoder path with the last block of the encoder path using the horizontal skip connections, thus re-injecting the details lost during the max-pooling of the final encoder block. Inspired by the ability of attention gates to localize intrinsic feature vectors, as showcased in attention UW-Net, the model uses attention gates before the skip connections to enable better transfer of feature vectors to the corresponding steps of the encoder and decoder paths. The intermediate layer then goes through a series of consecutive convolutional layers before being down-sampled via max-pooling function. The series of operation is formulated as:

$$I_A = Conv\left(\left[A\left(x^{i+1}, x^i\right)\right]\right) \tag{1}$$

where I_A represents the output feature maps of the intermediate layer of the attention UW-Net, $Conv(.)$ represents the convolutional operation, $A(y, x)$ represents a bivariate function which takes two feature maps as an input and returns an attention guided feature map as an output, U(.) represents the Up-sampling operation, [] represents the concatenation operation, x represents the output feature map of ith layer of the encoder pathway, where 'i' represents the layer number from the top to the bottom of the attention UW-Net 'varying from 1 to 5. The feature vectors are represented as $x^i \in R^{H \times W \times C}$ where H, W and C represents the dimensions of height, width and channel of vector. The intermediate layer I_A takes two feature vectors from the last two encoder blocks and feeds them into the attention gate. The output goes through a series of convolutional filters followed by activation layers. The intermediate layer is down sampled and is then connected to the decoding path. The repeated usage of attention gates over the re-distributed feature maps suppresses feature responses from irrelevant background regions. This effectively acts as a cropping operation of a denoised image based on the ROIs without using morphological operations. The repeated distribution followed by accumulation of the feature maps highlights the essential features and prune redundant feature responses, thereby preserving

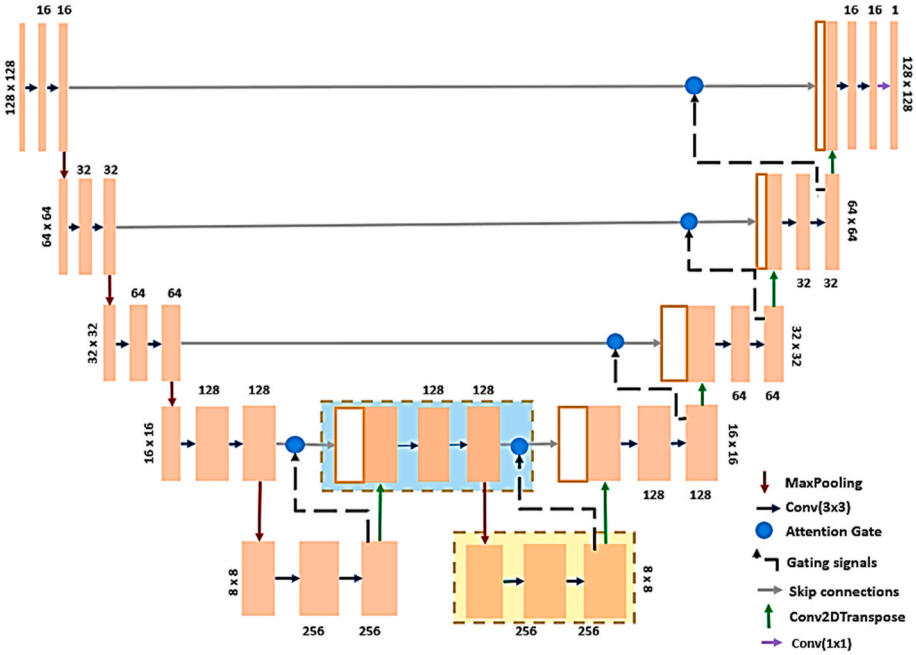

Fig. 2. The proposed attention UW-Net architecture.

features relevant to the segmented task.

Decoding Path: The decoding path consists of five steps. Each step starts with an up-sampling of the feature maps followed by the concatenation of the attention guided feature maps. These concatenated feature maps are passed through two consecutive convolution operations followed by RELU activations. The convoluted and transformed feature vectors serve as a gating signal for the superior attention gate in terms of the hierarchy of individual block. The final step of the decoder path comprises of three blocks. The first two blocks comprise of two convolutional layers. The output of the last convolutional layer is fed to a 1 × 1 convolutional filter to map each 16-component feature vector to the required number of classes. In total this architecture has two paths each containing five steps, one intermediate layer, 5 attention gates and 5 skip connections to concatenate the required feature maps from encoded path to the decoded path. The entire architecture with respect to the feature maps with respect to the feature maps of the intermediate layer defined in equation (2) is represented as follows:

$$y^i = \begin{cases} Conv\left(\left[A\left(y^{i+1}, x^i\right), U\left(y^{i+1}\right)\right]\right) & i \neq 4 \\ Conv\left(\left[A\left(I_A, y^{i+1}\right), U\left(y^{i+1}\right)\right]\right) & i = 4 \end{cases} \forall \ \mathrm{i} \ \in [1,4] \tag{2}$$

where y^i represents the output feature map obtained from the ith step of the decoding path.

Modified Attention gate: The modified attention gates, represented in Fig. 3, enable the model to generate effective segmentation masks for segmentation of smaller regions of interest. The attention UW-Net makes the full utilization of the attention gates and their ability to simplify the localization of intrinsic features required for better representation of images in a higher feature space. The implementation of attention gates is mathematically formulated as:

$$A\left(y^{i+1}, x^i\right) = G\left(g\left(y^{i+1}\right), x^i\right) \tag{3}$$

where G(.) represents the attention gate and g(.) represents the gating signal and a represents the output feature map of attention gate (AG).

Fig. 3 showcases the attention gate architecture where x^i and $g(y^{i+1})$ represent the input signal and gating signal respectively. These signals are fed to the attention gate to obtain an output signal α_i comprising of a series of attention vectors ranging between 0 and 1.

The attention gates (AGs) used in the proposed model use soft attention to address the issue of detection of extra pixels as predicted output for small objects that show huge variations in shape. Soft attention therefore reduces the false positive predictions for small objects with ambiguous shape. These AGs produce attention coefficients $\alpha_i \in [0,1]$ for each pixel which helps in scaling the input feature maps denoted as x^i to produce relevant features as output, depicted as $\widehat{x^i}$. The attention gates take two feature vectors namely as $F_G \in R^{H_G \times W_G \times C_G}$ and $F_I \in R^{H_I \times W_I \times C_I}$ as input vectors. F_G represents the features from the gating signal represented as g(.) in equation (3) and F_I represents the input signal from convolutional blocks connected via skip connections. The vectors F_G and F_I are represented in equation (3) as $g(y^{i+1})$ and x^i. In the attention gate represented in Fig. 3, F_I is passed through a Convolution layer with (2 × 2) filter having stride = 2 to generate a vector $F_I^{Conv} \in R^{H_G \times W_G \times C_I}$ and F_G is passed through a Conv2DTranspose layer with a (3 × 3) filter to generate a vector $F_G^{C2T} \in R^{H_G \times W_G \times C_I}$.

$$x^i = F_I^{Conv} = W_x^T x^i \tag{4}$$

$$g\left(y^{i+1}\right) = F_G^{C2T} = W_g^T g\left(y^{i+1}\right) \tag{5}$$

where W_x and W_g are the linear transformation vectors and b_g is the bias term.

These feature vectors are then added and passed onto a RELU activation function (σ_1) to produce intermediate activation maps α_i with layer-wise attention co-efficient represented as q_A. The layer-wise attention co-efficient passed through a sigmoid activation function (σ_2) and duplicated along the channel axis 'C' to obtain an attention vector $F_A \in R^{H_I \times W_I \times C_I}$.

$$q_A = Conv\left(\sigma_1\left(F_I^{Conv} + F_G^{C2T}\right)\right) \tag{6}$$

Normally attention gates obtain feature vectors by resampling them using bilinear interpolation. Since linear interpolation techniques uses local tendency to make guesses aimed at finding intermediate values, it might result in distortion of feature vectors with respect to the intermediate vector space. The linearization of the resampled data also results in the generation of fewer numbers of extreme points (global maxima). The duplication along the channel axis avoids resampling feature vectors by converting the 2-dimensional vector space to 3-D, thereby matching the dimensions of the input vector F_I. The attention co-efficient is computed by an element-wise product of the input vector, F_I with the attention vector:

$$\alpha_i = \sigma_2(U(q_A) \times C_I).F_I \tag{7}$$

where $\times C_I$ represents the duplication of feature vector operation and '.'

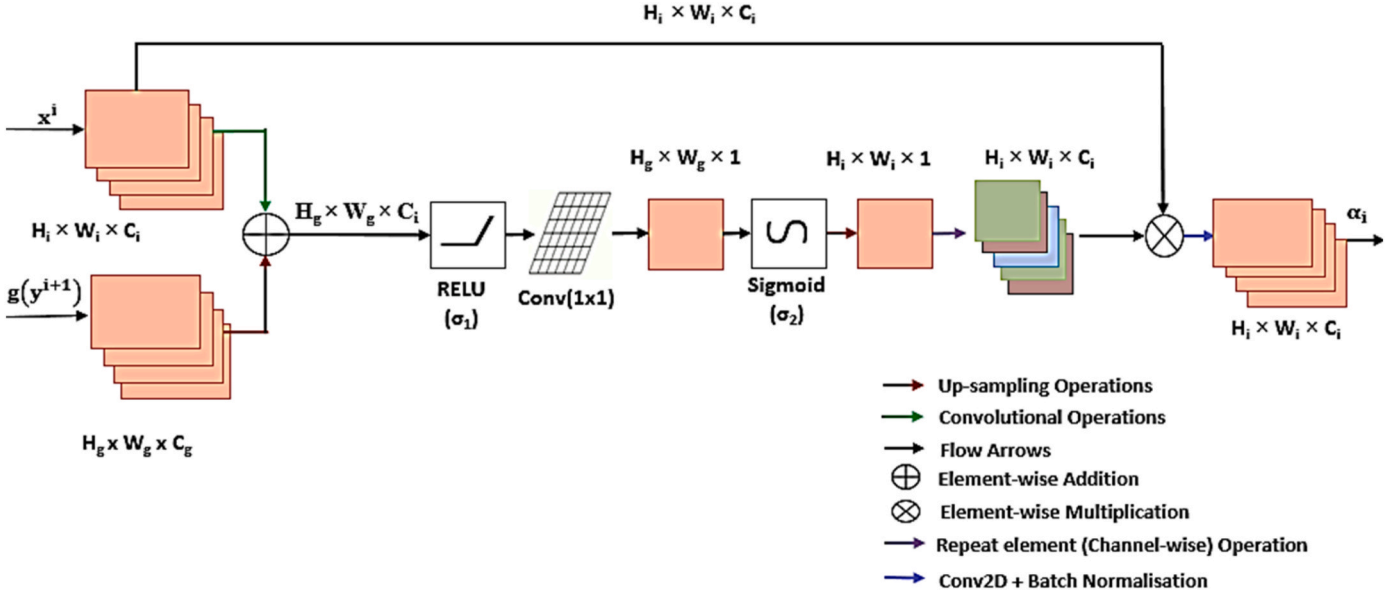

Fig. 3. The schematic diagram of the modified attention gate.

represents the element-wise multiplication operation. The attention coefficient scales the low-level signals and retain the intrinsic features essential for accurate segmentations. These features are then passed through a convolution layer followed by a batch-normalisation layer.

3.4. System and environment

The proposed semantic segmentation model is developed in Tensor Flow 2.0. The network is trained on an Intel RTX A4000 chip and Intel i9 processor, using Adam as an optimizer for 200 epochs with a learning rate of 0.001 and binary cross-entropy is used as loss function. The batch size is set to 8 and the convolutional layer filter size is set to 16, 32, 64, 128 and 256 from the top to the bottom of the attention UW-Net.

4. Results

One output of the proposed Attention UW-Net for individual segmentation and annotation tasks of the lungs, heart, trachea and collarbone is shown in Fig. 4(A). Fig. 4(B) and (C) comprises the probability maps of the final augmented and segmented image. The final images (as referred to in Fig. 4(B) and (C)) give complete information regarding the outlines, positions and areas covered by individual organs, collectively, in terms of the input CXR image. The accuracy of the individual segmented and annotated outputs has been visualized from the individual ground truth masks in the second column. The augmented output is the final output of a 2D CXR image generated by the proposed model.

The proposed attention UW-Net achieved the best effect on the aforementioned training set with the highest average F1 score [37] of 95.7, 80.9, 81.0 and 77.6 for lung, heart, trachea and collarbone segmentations, respectively (see Table 1). The average, maximum, minimum and standard deviation scores of performance metrics obtained using other evaluation metrics specificity, recall, precision, and F1 are shown in Table 1.

Table 1
Performance metrics of proposed attention UW net in terms of evaluation metrics such as specificity, recall, precision and F1 score.

Organs	Format	Specificity	Recall	Precision	F1
Lungs	Mean ± SD	98.8 ± 0.8	94.9 ± 2.5	96.6 ± 1.8	95.7 ± 1.4
	Max- Min	99.7–94.3	98.8–89.0	98.9–92.7	97.8–88.7
Heart	Mean ± SD	97.8 ± 1.1	95.8 ± 7.8	77.7 ± 9.6	80.9 ± 7.3
	Max- Min	99.9–94.9	99.5–70.7	99.8–48.7	91.6–58.3
Trachea	Mean ± SD	99.4 ± 0.2	79.4 ± 7.4	83.1 ± 6.0	81.0 ± 5.2
	Max- Min	99.8–98.7	99.9–62.7	95.4–69.7	89.9–70.8
Collarbone	Mean ± SD	99.4 ± 0.2	81.7 ± 8.9	74.7 ± 8.5	77.6 ± 6.3
	Max- Min	99.8–98.9	92.2–57.4	86.8–54.4	85.7–62.4
Average of (Mean ± SD)		**98.8 ± 0.5**	**87.9 ± 6.6**	**83.0 ± 6.5**	**83.8 ± 5.0**

* Avg = Average (Mean values), SD = Standard Deviation, Max = Maximum and Min = Minimum F1.

The metrics in Table 1 showcase the accuracy metrics, namely specificity, recall of 94, precision and F1 of the attention UW-Net. The proposed model gives an average specificity score of 98.8. The specificity score gives an estimate of how well the model can detect false positive pixels. The average precision score of 83.0 by the model signifies its ability to classify the relevant number of positive pixels as positive, whereas the average recall score of 87.9 means that the model

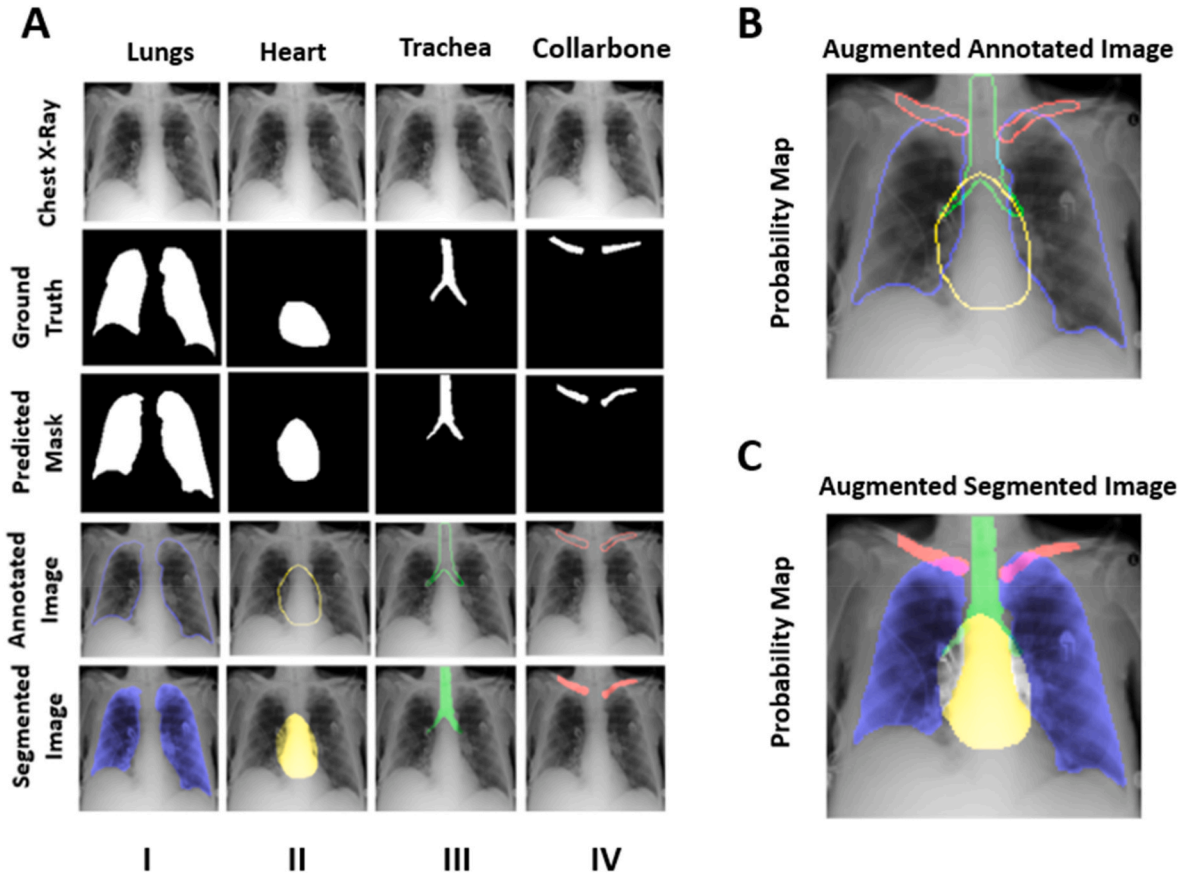

Fig. 4. A) The predicted annotation and segmentation images for individual segmentation tasks, namely lungs (I), heart (II), trachea (III), and the collarbone (IV), using attention UW-Net with respect to the ground truth mask. B) The probability map of the annotated CXR image obtained as final output. C) The probability map of the segmented CXR image obtained as final output.

can classify negative pixels as negative with a very high success rate. Table 1 displays the model's average F1 score for the four different segmentation tasks to be 83.8. The F1 score is the harmonic mean of precision and recall scores account for such a high score. This high precision score results in the model providing near similar predicted masks as the ground truth mask, which is validated in the comparative studies performed against other models and the ablation study on the intermediate late of the attention UW-Net.

To establish the robustness and effectiveness of the proposed Attention UW-Net model, we also compared U-Net, and its different variations and other widely used automatic and semi-automatic segmentation models.

4.1. Comparison with respect to U-net families

The segmented mask for annotation of attention UW-Net with other U-Net families is shown in Fig. 5. The visualization shows attention UW-Net performing better than all other U-Net [2], U-Net with ResNet backbone [37], Attention U-Net [13], Trans U-Net [38], and Swin U-Net [39] with U-Net family.

The Swin U-Net and U-Net perform better than Trans U-Net, as shown in Fig. 5(G), (H) and 5(I), respectively. The inability of the Trans U-Net to provide a segmented output mask for trachea segmentation and an abrupt-shaped heart mask is clearly portrayed in Fig. 5(I). The U-Net was unable to provide a detailed and complete collarbone segmentation mask in Fig. 5(H), thereby showcasing its limitations in segmenting small lesions. Swin U-Net, on the other hand, not only failed to provide well-defined boundaries for the segmented organs but also was providing an abrupt shape for the prediction of a heart mask (as seen in the predicted heart mask in Fig. 5(G)). Attention U-Net, shown in Fig. 5 (E), gives the best output segmentation masks among the U-Net variations, followed by the U-Net model having a ResNet backbone, as seen in Fig. 5(F). The attention U-Net performed surprisingly well in terms of small lesion segmentations, whereas the ResNet backbone variation failed to predict the collarbone mask with an acceptable shape and distance of separation between the left and right collarbones.

Attention U-Net maintains the shape of the segmented organs but fails to capture the intrinsic details of the segmented organs, especially in the case of trachea and heart segmentation, showcased in Fig. 5(E), where the latter output mask fails to showcase the proper size to height ratio. The U-Net with ResNet as a backbone also fails to maintain the size ratio of the trachea with the branched bronchi, as is visible in Fig. 5(F). The proposed architecture outperforms all the compared U-Net and its variations, as seen in the predicted segmented masks showcased in Fig. 5 (C). The output segmentation masks maintain both the shape and details of the target organs except for heart segmentation which fails to showcase the perfect height to width ratio. The details showcased in the lung and collarbone segmentation mask clearly establish the need for attention gates in the proposed architecture. The UW-Net without attention gates, as a result, fails to showcase the intrinsic details such as the left and right bronchi for the trachea mask in Fig. 5(D) and the proper size for the heart segmentation. The high imbalance of classes is visible, especially for the collarbone segmentation mask as showcased in the ground truth mask for collarbone segmentation as is showcased in Fig. 5 (B), making it difficult for automatic segmentation.

Five different models were implemented on the same dataset for evaluation and comparison purpose for quantitative analysis, which includes: U-Net [2], U-Net with ResNet backbone [39], Attention U-Net, Trans U-Net [40], Swin U-Net [41], the proposed attention UW-Net and UW-Net. To maintain uniformity in all the models in terms of architecture, the same set of convolutional filter sizes is maintained. Since small lesion segmentation masks contain an overwhelming number of negative class pixels, F1 score was used as an accuracy metrics for comparing the proposed model with the aforementioned models. In Table 2, we compared the proposed attention UW-Net for collarbone, trachea, heart and lung segmentation on the average, maximum, minimum and standard deviation (SD) of F1 metrics with respect to other variations of U-Net models.

As shown in Table 2, the attention UW-Net outperforms all the existing models. U-Net and its variations, on the other hand, fail to produce accurate outputs for small lesions. The Trans U-Net, which uses transformers as the backbone, fails to give an output mask for smaller regions of interest, such as the collarbone and trachea. Swin U-Net, on the other hand, outperforms U-Net and its ResNet backbone variant for collarbone and heart segmentation but falls behind, although ever so slightly in lungs and trachea segmentation. By comparing the results on lungs, heart, trachea and collar bone segmentation, the UW-Net achieves a performance gain of 3.2%, 2.9%, 7.1% and 11.8% and a rise in

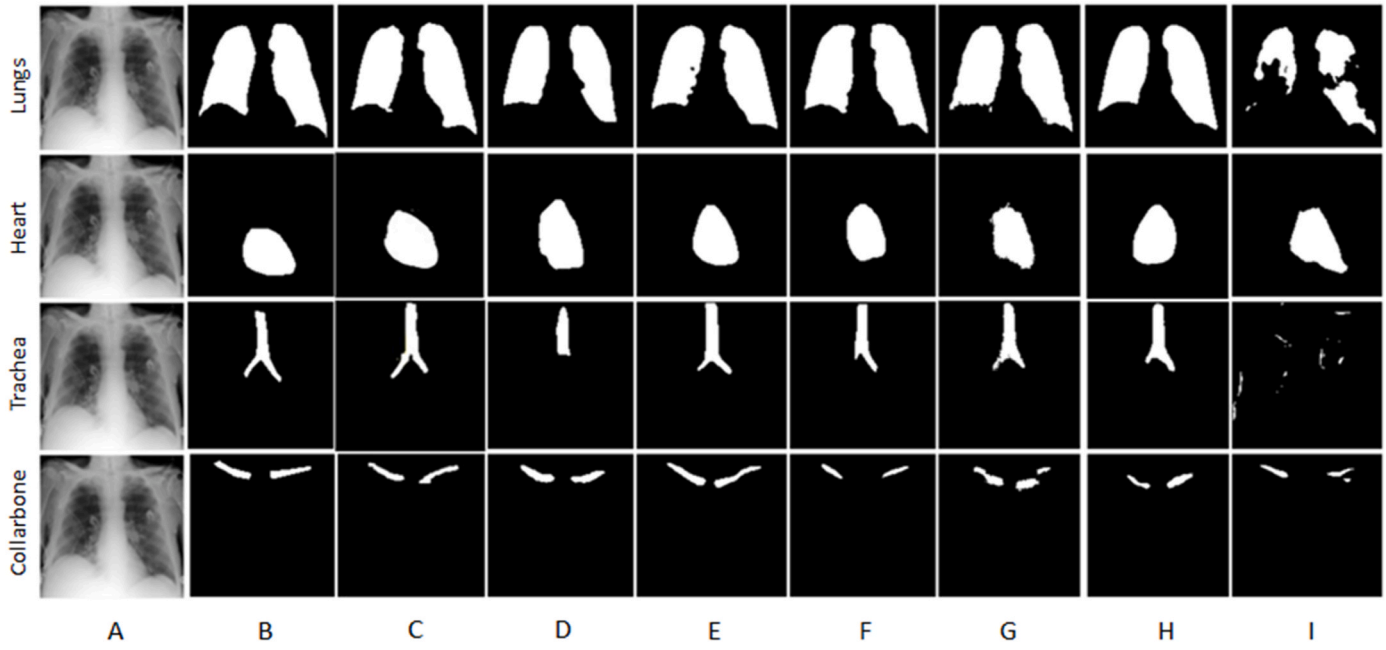

Fig. 5. Comparison of the predicted masks obtained from a 2D CXR image (A) as an input with respect to the ground truth masks (B) between the proposed Attention UW-Net (C), UW-Net (D), Attention U-Net (E), U-Net with ResNet backbone (F), Swin U-Net (G), U-Net (H) and Trans U-Net (I) (from left to right).

Table 2
The F1 scores of the attention UW-Net, base U-Net and its other variations in segmentation of lungs, heart, trachea and collarbone.

Model		Lungs	Heart	Trachea	Collarbone
U-net with ResNet backbone	Avg ±SD	94.6 ± 10.8	72.7 ± 10.4	67.2 ± 9.2	56.2 ± 16.1
	Max - Min	94.6–19.9	91.6–47.0	84.3–48.8	79.3–24.7
U-Net	Avg ±SD	94.1 ± 1.7	77.0 ± 9.0	69.0 ± 14.6	34.2 ± 19.9
	Max-Min	96.4–88.1	89.3–49.5	85.3–7.2	71.1–0.0
UW-Net	Avg ±SD	93.9 ± 4.1	68.9 ± 13.6	79.0 ± 6.4	54.0 ± 15.9
	Max-Min	96.8–69.1	90.6–35.1	87.8–65.8	79.4–26.5
Attention U-Net	Avg ±SD	93.4 ± 6.4	74.6 ± 10.8	76.8 ± 6.8	72.2 ± 7.4
	Max-Min	97.2–54.4	92.8–37.4	84.8–62.2	83.1–57.5
Proposed Attention UW-Net	**Avg ±SD**	**95.7 ± 1.3**	**80.9 ± 7.3**	**81.0 ± 5.2**	**77.5 ± 6.3**
	Max-Min	**97.8–88.7**	**91.6–58.3**	**89.9–70.8**	**85.6–62.4**
Swin U-Net	Avg ±SD	93.7 ± 3.2	77.7 ± 8.2	71.9 ± 7.3	58.5 ± 12.0
	Max-Min	97.2–85.6	92.6–58.1	82.2–54.1	74.3–32.8
Trans U-Net	Avg ±SD	79.2 ± 11.6	78.2 ± 13.4	9.2 ± 2.4	21.1 ± 14.5
	Max-Min	93.9–41.9	19.5–13.4	10.0–0.1	47.0–0.0

* Avg = Average (Mean values), SD = Standard Deviation, Max = Maximum, and Min = Minimum F1.

performance by 1.9%, 3.9%, 0.9% and 5.0% over U-Net, Attention U-Net, Swin U-Net and U-Net with ResNet backbone on the trachea and heart segmentation respectively in terms of F1 score. However, the improvement vanishes for lung and collarbone segmentations. This is because the shape, size and orientation of the collarbone vary from patient to patient, and there is a limited contrast with the neighbouring pixels (cells). The UW-Net also fails to effectively leverage global structural information to resolve the local details, which leads to the deterioration in the performance of lung segmentation. This drawback is solved by the usage of attention layers in the proposed model. The attention mode of the UW-Net outperforms significantly in the collarbone segmentation with an average improvement of 43.3%, 21.4%, 19.0% and 5.4% over U-Net, U-Net with ResNet backbone, Swin U-Net and attention U-Net with respect to the F1 score. The attention UW-net outperforms the U-Net by 1.7%, the Swin U-Net by 2.6% and the attention U-Net by 4.2% but falls behind the U-Net with ResNet backbone in lung segmentation by 1.4%. However, the latter has more layers in comparison to the attention UW-Net. Therefore, more parameters and more time are required for training.

The boxplots represent the distribution of F1 scores for the compared architectures are shown in Fig. 6. The median and mean of the distribution of F1 scores for each corresponding method are represented by orange and green lines, respectively. The box represents the range of the central 50% of the F1-scores and gives an idea of the data distribution. The whiskers or the lines extending from the boxes represent the range of the remaining data points. The black dots past the whiskers represent the outliers.

From Fig. 6, it is clear that the proposed attention UW-Net leads all the existing architectures that are compared in this paper. The proposed attention UW-Net architecture not only has less deviation (spread) in the F1-score but also has a higher mean and median value. The number of outliers and the distance of the outliers from the lower whisker are lower than the compared architectures. The proposed attention UW-Net has the bulk of scores in the range of 92–95 for lung segmentation, 77 to 86 for heart segmentation, 76 to 82 for trachea and 74 to 81 for collarbone segmentation. The proposed architecture also has a lesser number of outliers in comparison to the other architectures. From the graphical representation of the F1 scores across the boxplot in Fig. 6, attention UW-Net outperforms other U-Net variations in all aspects.

4.2. Comparison with other models

The model has been compared with four segmentation models apart from the U-Net architecture to further our claim of the proposed attention UW-Net is the best among the other prevalent segmentation models. The results are tabulated in Table 3 and visualized as predicted segmentation masks in Fig. 7. These models include LinkNet [41], FPN [42], PSP Net [43], region growth algorithm [44], RU-Net [45] and 2ST-UNet [46].

The proposed attention UW-Net provides the best output segmentation by a long shot, as seen in Fig. 7(C). The predicted segmentation masks of lungs for LinkNet in Fig. 7(D), though lacks in showcasing the details, come close to the lung segmented mask of the proposed model. The failure to give proper segmentation masks for smaller regions is showcased in the heart and trachea segmentation in Fig. 7(D). The trachea masks fail to show the splitting into bronchi, and the shape of the segmented heart doesn't match with the ground truth (Fig. 7(B)). Since the region growing algorithms depend on the input seed, four different input seeds were allocated in correspondence with the supposed position of the located organs. Its inability to pin-point the precise locations and features is evident in Fig. 7(E). As is evident, the region-growing algorithm is better off giving generalized output rather than a specified output mask as required in certain cases. RU-Net and 2ST UNet perform a good job in generating lung and trachea masks, as is showcased in Fig. 7(F). However, the predicted heart mask is not showing a proper shape. The predicted collarbone mask is also not showcasing the proper position in terms of the ground truth mask, as is shown in Fig. 7(G). Other segmentation models, such as COVID-Net [16] were unable to give any output mask due to the absence of pre-processing steps which were incorporated while training the model. COVIDX-Net [20], which is designed for classification tasks, was also unable to give any significant results for the mentioned training data. As PSP Net and FPN were unable to generate any segmentation masks, there were no true positive pixels. Therefore, their entries were not included in Table 3. As a result, Fig. 4 is limited to 5 columns only. Their inability might be due to the lack of training data and the size of the convolutional filters being too large to capture receptive features essential for the pixel-level classification of input CXR images.

The proposed attention UW-Net outperforms all the other segmentation models in different segmentation tasks and performance is shown in Table 3. FPN and PSP Net fail to generate a segmentation mask for the given segmentation task due to the class imbalance of segmented masks for small lesions such as collarbone, heart and trachea segmentation. The small region of interest results in a huge difference between the true-positive and false-positive pixels. The automatic segmentation method implemented by using the region growth algorithm also fares bad in the segmentation tasks at hand. However, from Table 2, proposed model in lung segmentation task, leading the attention UW-Net by an average F1-score of 1.5%. Apart from this, the proposed model has a significant upper hand over the other compared segmentation models. The proposed model leads the LinkNet by an average F1 score of 4.4%, 5.3% and 4.7%, the RU-Net by an average F1 score of 3.4%, 1.5%, 3.9% and 2.6% and the 2ST-UNet by an average F1 score of 0.8%, 2.3%, 4.4% and 7.0% for heart, trachea and collarbone segmentation (see Table 3). The consistency of the proposed model in generating segmentation masks irrespective of the size and shape of the region of interest is clearly established from the experiment performed with respect to the state-of-the-art RGB models as well as segmentation models developed specifically for CXR images as is showcased in Fig. 7 and Table 3.

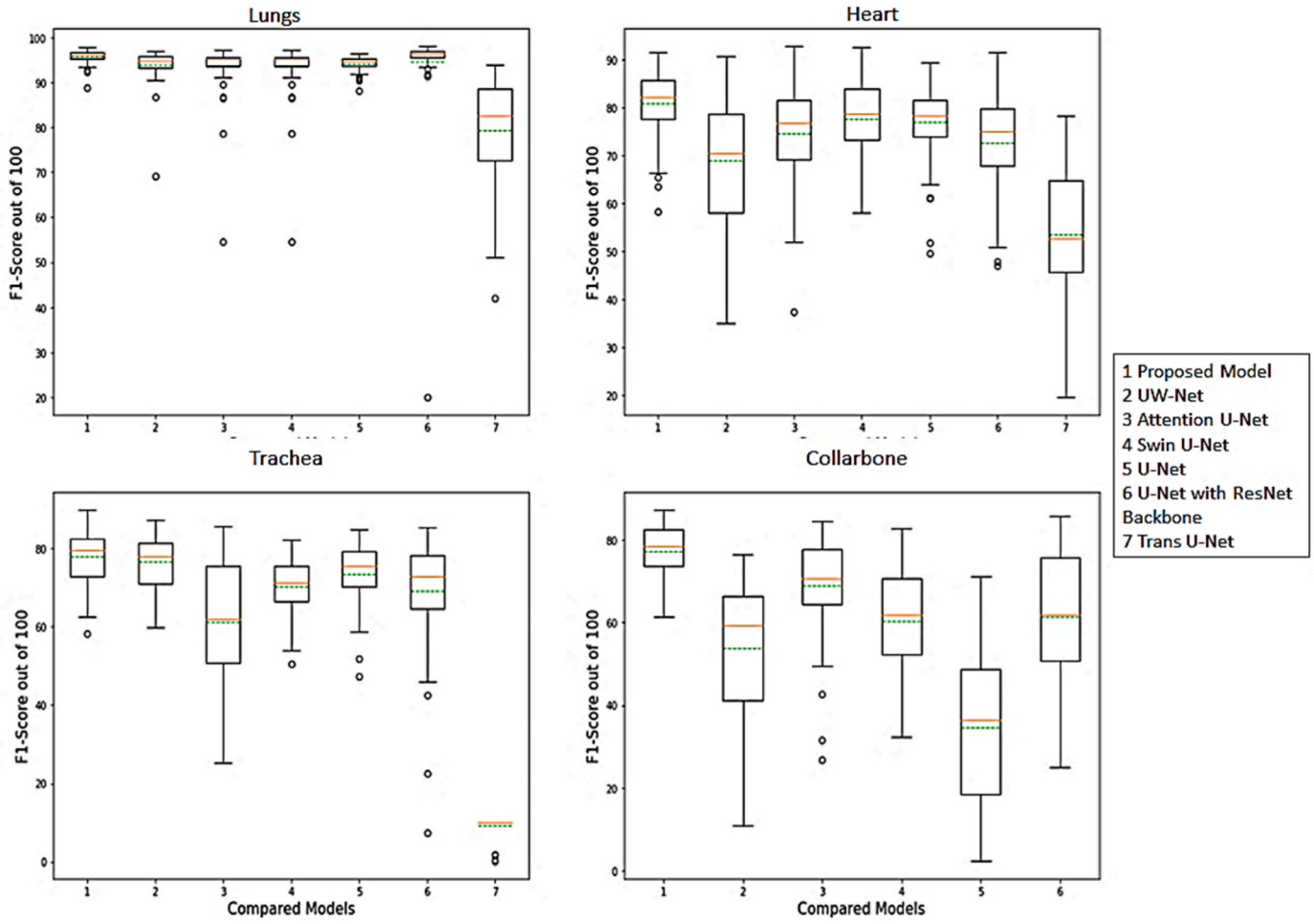

Fig. 6. A series of boxplots representing the distribution of F1 scores on the four different segmentation tasks, namely lungs, heart, trachea and collarbone.

Table 3
F1 scores comparison of attention UW-Net with other segmentation models.

Model	Format	Lungs	Heart	Trachea	Collarbone
LinkNet	Avg ±SD	94.1 ± 1.1	76.5 ± 9.8	75.7 ± 6.8	72.9 ± 8.9
	Max - Min	97.4–92.1	93.9–48.1	87.4–62.1	86.1–44.9
Region Growing Algorithm	Avg ±SD	39.6 ± 5.5	1.2 ± 10.4	13.7 ± 2.3	6.4 ± 3.9
	Max-Min	47.3–28.8	35.0–0	31.5–0.1	27.7–0.1
Proposed Attention UW-Net	**Avg ±SD**	**95.7 ± 1.3**	**80.9 ± 7.3**	**81.0 ± 5.2**	**77.5 ± 6.3**
	Max-Min	**97.8–88.7**	**91.6–58.3**	**89.9–70.8**	**85.6–62.4**
RU-Net	Avg ±SD	92.3 ± 4.8	79.4 ± 7.0	77.1 ± 7.7	74.9 ± 8.2
	Max - Min	97.1–74.6	90.3–53.1	90.3–56.6	85.5–53.8
2ST-UNet	Avg ±SD	94.9 ± 2.2	78.6 ± 7.9	76.6 ± 8.4	70.5 ± 12.8
	Max - Min	97.5–87.6	90.1–54.9	89.1–52.6	86.5–14.7

* Avg = Average (Mean values), SD = Standard Deviation, Max = Maximum, and Min = Minimum F1.

4.3. Discussion

The proposed model has showcased a commendable performance in terms of small and large region segmentation and annotation task without using any pre- or post-processing techniques as shown in Tables 2 and 3 The limited training data to obtain these results without using any pre-trained weights should also be kept into consideration. Since the proposed model can effectively annotate large as well as small regions of interests, this model has a wide range of use cases in medical vision especially in predicting and precisely annotating fractures in bones, detecting anomalies in any 2D scan images as inputs. A 3D version of this model can also be implemented for the detection of tumors in breast and brain MRI scans.

4.4. Potential limitations

The proposed attention UW-Net outperforms the aforementioned models for organ segmentation and annotations for small (collarbone, trachea and heart) and large (lungs) organs. However, the mentioned model has certain limitations in terms of lung field annotations. The proposed model adds extraneous regions of colonic and gastric air present below the diaphragm (as shown in Fig. 8(C)) as lung masks in some cases. This may be due to the addition of an extra skip connection which connects the intermediate layer with the decoder layer. This results in an additional positional embedding of the feature vector for the first decoder layer, which results in capturing extraneous noises surrounding the regions of interest whose pixel correlation is almost the same as the pixels of the regions which are to be segmented. However, the proposed model shows the lesser area for the misclassified region, unlike other good segmentation models such as Attention U-Net (shown in Fig. 8(D)) and LinkNet (shown in Fig. 8(E)).

The proposed attention UW-Net includes the adjacent regions in the

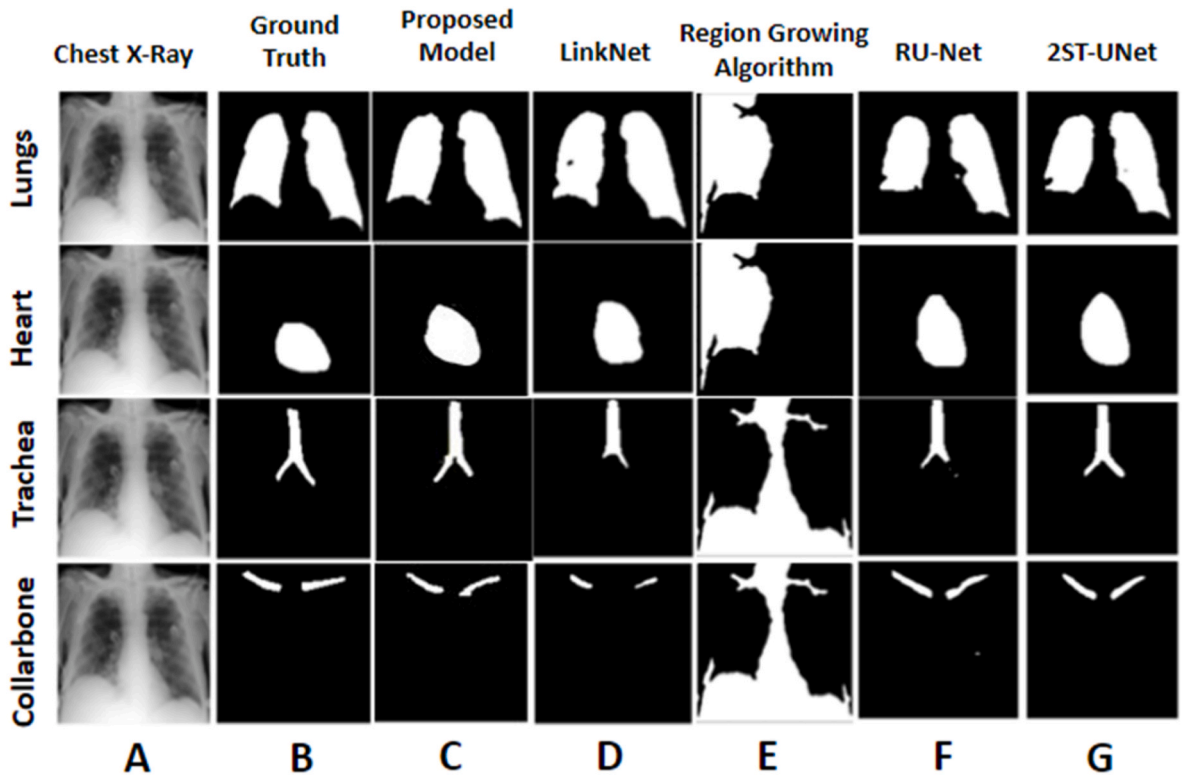

Fig. 7. Comparison on the predicted masks obtained from a 2D CXR image (A) as a input with respect to the ground truth masks (B) between the proposed attention UW-Net (C) with other segmentation models namely LinkNet (D), the region growing algorithm (E), RU-Net (F) and 2ST-UNet (G) (from left to right).

Fig. 8. Comparison on CXR image (A) in terms of predicted lung field segmentation masks produced by the proposed model (C), Attention U-Net (D) and LinkNet (E) with respect to the ground truth mask (B). The extraneous regions denoting the colonic and gastric air are marked as a yellow circle (showcased in sub-figure A, C, D and E).

predicted lung field masks, as showcased in Fig. 8. However, the extraneous region included in the predicted mask is less as compared to the predicted mask of Attention U-Net (as shown in Fig. 8(D)). The LinkNet not only fails to give an accurate segmentation mask for the lungs but also adds a greater number of false positive pixels in the predicted lung mask. Post-processing techniques such as denoising and erosion can be considered as a possible solutions to this limitation. However, since the paper focuses on introducing a novel model with little to no pre- or post-processing techniques, the applicability of the mentioned techniques has not been covered in this paper.

5. Ablation study

The ablation study is performed in order to demonstrate the effectiveness of the proposed attention UW-Net architecture based on (i) duplicating the novel intermediate layer and stacking them one level on top of another, and (ii) stacking the replicated intermediate layer side by side as an extension of second last decoder layer in the decoder path as shown in Fig. 8. The mentioned models are trained and tested on the same dataset with the same system configuration. The results of the experiment are tabulated in Tables 3 and 4. The study is conducted to prove that stacking similar intermediate layer level-wise or side-ways does not, in any way, improve the segmentation accuracy of the proposed model. The level-wise addition of the novel intermediate layer on the proposed UW-Net architecture is shown in Fig. 9 (Level-wise). The stacking of the intermediate layer over the UW-Net is shown in Fig. 9 (Stack-wise).

The F1 score of the level-wise and stack-wise addition of the intermediate layer, respectively, is shown in Table 4. The "Model – 1" layer is the attention U-Net architecture. The model with +1, +2 and +3-layers refers to the UW-Net with 2, 3 and 4 intermediate layers, respectively, stacked in the previously mentioned orientation. The proposed UW-Net model is the most consistent among the modified models that are compared in Tables 3 and 4 The proposed UW-Net model significantly outperforms the models mentioned in the ablation study for small lesion segmentation such as heart, trachea and collarbone. However, the average F1 scores are almost the same for large organ segmentation, precisely lung segmentation. In terms of the level-wise addition of intermediate layers, the proposed model leads the other models by an average F1 score of 5.4, 4.9, 5.1 and 1.1 for collarbone segmentation, by

Table 4
The F1 scores of the ablation study on the level-wise as well as stack-wise duplication of the intermediate layer.

Type	Model		Lungs	Heart	Trachea	Collarbone
Level-wise	Model – 1 Layer	Avg ±SD	93.4 ± 6.4	74.6 ± 10.8	76.8 ± 6.8	72.2 ± 7.3
		Max - Min	97.2–54.4	92.8–37.4	84.8–62.2	83.1–57.4
	Proposed Model	**Avg ±SD**	**95.7 ± 1.4**	**80.9 ± 7.3**	**81.0 ± 5.2**	**77.6 ± 6.3**
		Max-Min	**97.8–88.7**	**91.6–58.3**	**89.9–70.8**	**85.7–62.4**
	Model +1 Layer	Avg ±SD	95.7 ± 2.0	79.5 ± 7.6	76.2 ± 8.4	72.5 ± 7.9
		Max-Min	98.1–87.1	90.2–54.8	89.8–70.8	85.4–41.1
	Model +2 Layers	Avg ±SD	95.4 ± 2.2	79.8 ± 6.8	69.7 ± 9.6	74.1 ± 7.9
		Max-Min	97.7–87.8	91.1–57.8	87.0–47.4	85.8–53.1
	Model +3 Layers	Avg ±SD	95.6 ± 1.8	79.4 ± 6.9	78.6 ± 7.7	76.5 ± 6.8
		Max-Min	97.8–90.8	88.8–57.1	91.9–54.3	85.8–57.4
Stack-wise	Model – 1 Layer	Avg ±SD	93.4 ± 6.4	74.6 ± 10.8	76.8 ± 6.8	72.2 ± 7.3
		Max-Min	97.2–54.4	92.8–37.4	84.8–62.2	83.1–57.4
	Proposed Model	**Avg ±SD**	**95.7 ± 1.4**	**80.9 ± 7.3**	**81.0 ± 5.2**	**77.6 ± 6.3**
		Max-Min	**97.8–88.7**	**91.6–58.3**	**89.9–70.8**	**85.7–62.4**
	Model +1 Layer	Avg ±SD	95.6 ± 1.7	80.7 ± 6.7	77.4 ± 6.8	74.9 ± 7.9
		Max-Min	97.8–89.9	92.6–62.1	89.7–62.1	87.7–49.9
	Model +2 Layers	Avg ±SD	95.6 ± 1.8	80.0 ± 7.5	77.3 ± 7.7	77.0 ± 6.8
		Max-Min	97.7–89.6	92.9–62.4	90.4–54.8	86.0–58.3
	Model +3 Layers	Avg ±SD	95.8 ± 1.4	80.8 ± 7.4	77.2 ± 7.1	75.7 ± 9.3
		Max-Min	97.8–91.7	93.8–61.6	89.2–52.4	88.9–45.5

* Avg = Average (Mean values), SD = Standard Deviation, Max = Maximum value and Min = Minimum F1.

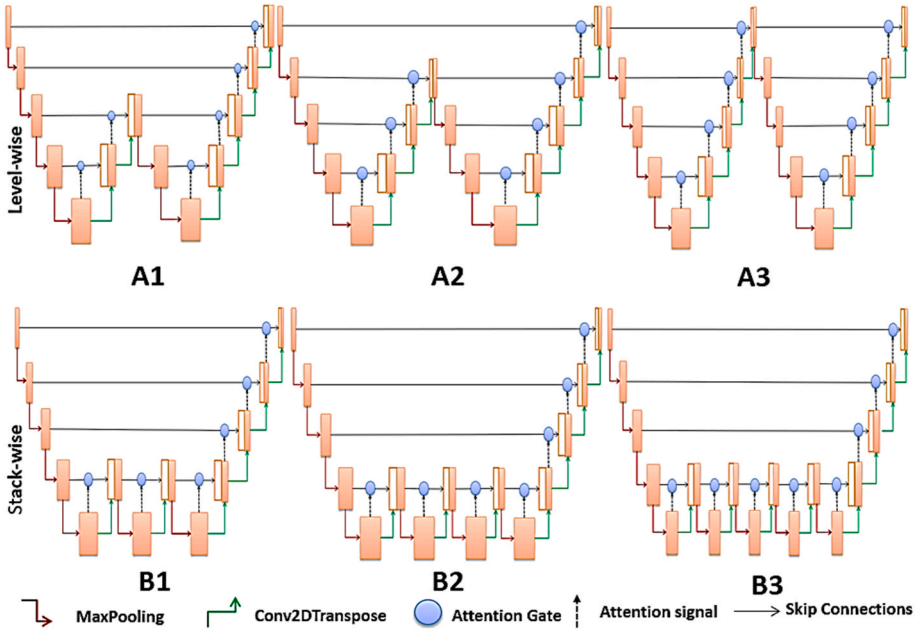

Fig. 9. The feature vectors of the different ablation techniques used: Level-wise (A1-A3) and Stack-wise (B1–B3). As we move from left to right, the number of intermediate layers added to the UW-Net increases from once in (1) to thrice in (3).

4.8, 3.8, 11.3 and 2.4 for trachea segmentation and by 5.3, 1.4, 1.1 and 1.5 for heart segmentation in terms of level-wise increment of the intermediate layer.

The proposed model is the best performing variation in terms of the intermediate layer, though the other models have almost similar performance. In terms of the stack-wise addition of intermediate layers, the proposed model leads the other models by an average F1 score of 5.4, 2.7, 0.6 and 1.9 for collarbone segmentation, by 4.2, 3.6, 3.7 and 3.8 for trachea segmentation and by 6.3, 0.2, 0.9 and 0.1 for heart segmentation in terms of level-wise increment of the intermediate layer. However, the model with four intermediate layers, i.e., the intermediate layer in between the four skip connections, has the highest average F1 score, outperforming the proposed model by a mere 0.1 in F1-score. The increased number of skip connections, intermediate and attention layers should be taken into consideration for the increase in average F1-score by a mere 0.1. The comparison in terms of F1-score as evaluation has established the fact that the proposed UW-Net is the best possible method in terms of addition or subtraction of the novel intermediate layer.

The ablation study shows how well the proposed attention UW-Net performs over small and large region segmentation tasks, not only in terms of average F1-score but also the standard deviation. The deviation from the mean score is represented as black vertical lines on the top of each bar in an individual unstacked plot. The consistent performance is reflected by the small vertical lines on top of the bars representing the attention UW-Net model. The variation in F1 scores for the proposed model is significantly less in comparison to other models compared in the study.

6. Conclusion

In this paper, we have presented a novel attention UW-Net to address

the need for a more generalized model that not only deals with small lesion segmentation but also considers the anatomical features and the positional relationship between the lung field and heart. The experimental results show that the novel attention UW-net architecture, along with the modified attention gates, is highly efficient for the identification and segmentation of small lesions such as collarbone and trachea as well as large region segmentation such as lungs. Furthermore, the attention UW-Net outperforms not only the existing U-Net models and their variations but also other existing state-of-the-art segmentation models with respect to both consistency and accuracy obtained via accuracy metrics. The results of the ablation study prove that the model is the best variation in terms of the position and stacking of intermediate layers.

The future work will be to incorporate computational intelligent algorithms, especially swarm-based algorithms such as MBO, EWA and EHA, into the proposed model. Combating the problem of generating false positive pixels in certain specific cases (such as lung field annotation) as mentioned in the potential limitations section would be another aspect which would be focused on in future.

Source code

The complete source code of this work is available in GitHub repository at: https://github.com/Dynamo13/Attention_UWNet.

Declaration of competing interest

Authors have no conflict of interest to declare.

Acknowledgement

This research work was supported by the Jio Institute "CVMI-Computer Vision in Medical Imaging" research project fund by RFIER-Jio Institute, grant # 2022/33185004 under the "AI for ALL" research center. We would like to thank Dr Kalyan Tadepalli, H N Reliance Foundation Hospital for his advice in this CVMI project.

References

[1] N. Siddique, S. Paheding, C.P. Elkin, V. Devabhaktuni, U-net and its variants for medical image segmentation: a review of theory and applications, IEEE Access 9 (2021) 82031–82057.
[2] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, medical image computing and computer-assisted intervention – MICCAI 2015. MICCAI, Lect. Notes Comput. Sci. 9351. (2015).
[3] M. Farooq, A. Hafeez, Covid-ResNet: A Deep Learning Framework for Screening of COVID19 from Radiographs, 2020 arXiv preprint arXiv:2003.14395.
[4] X. Gao, T. Chen, R. Niu, A. Plaza, Recognition and mapping of landslide using a fully convolutional DenseNet and influencing factors, IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 14 (2021) 7881–7894.
[5] M.S. Brown, L.S. Wilson, B.D. Doust, R.W. Gill, C. Sun, Knowledge-based method for segmentation and analysis of lung boundaries in chest X-ray images, Comput. Med. Imag. Graph. 22 (6) (1998) 463–477.
[6] S. Roy, K.I. Shoghi, Computer-Aided Tumor Segmentation from T2-Weighted MR Images of Patient-Derived Tumor Xenografts, International Conference on Image Analysis and Recognition, Springer, Cham, 2019, pp. 159–171.
[7] S. Roy, T.D. Whitehead, S. Li, F.O. Ademuyiwa, R.L. Wahl, F. Dehdashti, K. I. Shoghi, Co-clinical FDG-PET radiomic signature in predicting response to neoadjuvant chemotherapy in triple-negative breast cancer, Eur. J. Nucl. Med. Mol. Imag. 49 (2) (2022) 550–562.
[8] A. Mittal, D. Kumar, M. Mittal, T. Saba, I. Abunadi, A. Rehman, S. Roy, Detecting pneumonia using convolutions and dynamic capsule routing for chest X-ray images, Sensors 4 (2020) 1068.
[9] N.C. Mithun, S. Das, S.A. Fattah, Automated detection of optic disc and blood vessel in retinal image using morphological, edge detection and feature extraction technique, in: 16th Int'l Conf. Computer and Information Technology, 2014, pp. 98–102.
[10] L. Wang, S. Guo, W. Huang, Y. Qiao, Places205-vggnet Models for Scene Recognition, 2015 arXiv preprint arXiv:1508.01667.
[11] W. Dai, N. Dong, Z. Wang, X. Liang, H. Zhang, E.P. Xing, Scan: structure correcting adversarial network for organ segmentation in chest x-rays, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, Cham, 2018, pp. 263–273.
[12] A. Ren, Z. Li, C. Ding, Q. Qiu, Y. Wang, J. Li, X. Qian, B. Yuan, Sc-dcnn: highly-scalable deep convolutional neural network using stochastic computing, ACM SIGPLAN Not. 52 (4) (2017) 405–418.
[13] Ozan Oktay, et al., Attention U-Net: Learning where to Look for the Pancreas, 2018 arXiv preprint arXiv:1804.03999.
[14] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, S. Escalera, Bi-directional ConvLSTM U-Net with densley connected convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, 0-0.
[15] N. Gaggion, L. Mansilla, C. Mosquera, D.H. Milone, E. Ferrante, Improving Anatomical Plausibility in Medical Image Segmentation via Hybrid Graph Neural Networks: Applications to Chest X-Ray Analysis, 2022 arXiv preprint arXiv: 2203.10977.
[16] L. Wang, Z.Q. Lin, A. Wong, Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images, Sci. Rep. 10 (1) (2020) 1–12.
[17] M. Sirshar, T. Hassan, M.U. Akram, S.A. Khan, An incremental learning approach to automatically recognize pulmonary diseases from the multi-vendor chest radiographs, Comput. Biol. Med. 134 (2021), 104435.
[18] R. Rashid, M.U. Akram, T. Hassan, Fully convolutional neural network for lungs segmentation from chest X-rays, in: International Conference Image Analysis and Recognition, Springer, Cham, 2018, pp. 71–80.
[19] A.M. Khan, T. Hassan, M.U. Akram, N.S. Alghamdi, N. Werghi, Continual learning objective for analyzing complex knowledge representations, Sensors 22 (4) (2022) 1667.
[20] E.E.D. Hemdan, M.A. Shouman, M.E. Karar, Covidx-net: A Framework of Deep Learning Classifiers to Diagnose Covid-19 in X-Ray Images, 2020 *arXiv preprint arXiv:2003.11055*.
[21] W. Li, G.-G. Wang, A.H. Gandomi, A survey of learning-based intelligent optimization algorithms, Arch. Comput. Methods Eng. 28 (5) (2021) 3781–3799, https://doi.org/10.1007/s11831-021-09562-1.
[22] Gai-Ge Wang, Suash Deb, Zhihua Cui, Monarch butterfly optimization, Neural Comput. Appl. 31 (7) (2019) 1995–2014.
[23] Gai-Ge Wang, Suash Deb, Dos Santos Coelho Leandro, Earthworm optimisation algorithm: a bio-inspired metaheuristic algorithm for global optimisation problems, Int. J. Bio-Inspired Comput. 12 (1) (2018) 1–22.
[24] Gai-Ge Wang, Moth search algorithm: a bio-inspired metaheuristic algorithm for global optimization problems, Memetic Computing 10 (2) (2018) 151–164.
[25] Mohamed Abdel-Basset, Victor Chang, Reda Mohamed, HSMA_WOA: a hybrid novel Slime mould algorithm with whale optimization algorithm for tackling the image segmentation problem of chest X-ray images, Appl. Soft Comput. 95 (2020), 106642.
[26] W.S. AbuShanab, M. Abd Elaziz, E.I. Ghandourah, E.B. Moustafa, A.H. Elsheikh, A new fine-tuned random vector functional link model using Hunger games search optimizer for modeling friction stir welding process of polymeric materials, J. Mater. Res. Technol. 14 (2021) 1482–1493.
[27] H. Shaban, E.H. Houssein, M. Pérez-Cisneros, D. Oliva, A.Y. Hassan, A.A. Ismaeel, M. Said, Identification of parameters in photovoltaic models through a Runge Kutta optimizer, Mathematics 9 (18) (2021) 2313.
[28] B. Shi, H. Ye, L. Zheng, J. Lyu, C. Chen, A.A. Heidari, P. Wu, Evolutionary warning system for COVID-19 severity: Colony predation algorithm enhanced extreme learning machine, Comput. Biol. Med. 136 (2021), 104698.
[29] E.H. Houssein, M.E. Hosney, D. Oliva, W.M. Mohamed, M. Hassaballah, A novel hybrid Harris hawks optimization and support vector machines for drug design and discovery, Comput. Chem. Eng. 133 (2020), 106656.
[30] T.M. Shami, A.A. El-Saleh, M. Alswaitti, Q. Al-Tashi, M.A. Summakieh, S. Mirjalili, Particle Swarm Optimization: A Comprehensive Survey, IEEE Access, 2022.
[31] G. Li, G.G. Wang, J. Dong, W.C. Yeh, K. Li, DLEA: a dynamic learning evolution algorithm for many-objective optimization, Inf. Sci. 574 (2021) 567–589.
[32] W. Li, G.-G. Wang, A. H Alavi, Learning-based Elephant Herding Optimization Algorithm for Solving Numerical Optimization Problems, Knowledge-Based Systems, 2020, 105675, https://doi.org/10.1016/j.knosys.2020.105675.
[33] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, Y.G. Jiang, Pixel2mesh: generating 3d mesh models from single rgb images, in: Proceedings of the European Conference on Computer Vision, ECCV), 2018, pp. 52–67.
[34] U. Wickramasinghe, E. Remelli, G. Knott, P. Fua, Voxel2mesh: 3d mesh model generation from volumetric data, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, Cham, 2020, pp. 299–308.
[35] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, Ronald Summers, ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, IEEE CVPR (2017) 3462–3471.
[36] A. Dutta, A. Zisserman, The VIA annotation software for images, audio and video, in: Proceedings of the 27th ACM International Conference on Multimedia (MM '19) vol. 4, Nice, France. ACM, New York, NY, USA, 2019, https://doi.org/10.1145/3343031.3350535. October 21–25.
[37] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. 21 (1) (2020) 1–13.
[38] A. Abedalla, M. Abdullah, M. Al-Ayyoub, E. Benkhelifa, Chest X-ray pneumothorax segmentation using U-Net with EfficientNet and ResNet architectures, PeerJ Computer Science 7 (2021) e607.
[39] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers Make Strong Encoders for Medical Image Segmentation, 2021 arXiv preprint arXiv:2102.04306.

[40] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like Pure Transformer for Medical Image Segmentation, 2021 arXiv preprint arXiv: 2105.05537.
[41] Cihan Akyel, Nursal Arıcı, LinkNet-B7: noise removal and lesion segmentation in images of skin cancer, Mathematics 10 (5) (2022) 736.
[42] Syazwany, Nur Suriza, Ju-Hyeon Nam, and Sang-Chul Lee. "MM-BiFPN: multi-modality fusion network with Bi-FPN for MRI brain tumor segmentation." IEEE Access 9 (2021): 160708-160720.
[43] L. Yan, D. Liu, Q. Xiang, Y. Luo, T. Wang, D. Wu, H. Chen, Y. Zhang, Q. Li, PSP net-based automatic segmentation network model for prostate magnetic resonance imaging, Comput. Methods Progr. Biomed. 207 (2021), 106211.
[44] E.S. Biratu, F. Schwenker, T.G. Debelee, S.R. Kebede, W.G. Negera, H.T. Molla, Enhanced region growing for brain tumor mr image segmentation, J. Imag. 7 (2) (2021) 22.
[45] Leclerc, Sarah, Erik Smistad, Thomas Grenier, Carole Lartizien, Andreas Ostvik, Frédéric Cervenansky, Florian Espinosa et al. "RU-Net: a refining segmentation network for 2D echocardiography." In 2019 IEEE International Ultrasonics Symposium (IUS), pp. 1160-1163. IEEE.
[46] Ayat Abedalla, et al., The 2ST-UNet for Pneumothorax Segmentation in Chest X-Rays Using ResNet34 as a Backbone for U-Net, 2020 arXiv preprint arXiv: 2009.02805.