Received: December 6, 2023. Revised: January 18, 2024.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Semantic Location Aware Swin Transformer Based U-Net Model for Improving Lung Disease Prediction

Nagaraj Dhivya[1]* **Palaniappan Sharmila[2]**

[1]*Department of Computer Science, Navarasam Arts and Science College for Women,*
Affiliated to Bharathiar University, Arachalur, Erode- 638101, Tamilnadu, India
[2]*School of Computing Science, KPR College of Arts Science and Research,*
Affiliated to Bharathiar University, Arasur, Coimbatore- 641048, Tamilnadu, India
*Corresponding author's Email: dhivyarubiphd@gmail.com

Abstract: Lung diseases have been a significant concern throughout history, necessitating early disease prediction using high-level knowledge. Deep Learning models have proven effective in diagnosing lung disorders using clinical imaging modalities like Computerized Tomography (CT) and Chest X-Ray (CXR) images. The Ensemble Deep Lung Disease Predictor (EDEPLDP) framework has been proposed for the rapid detection of various diseases using CT and CXR images. However, the U-Net model used for segmentation tasks lacks sufficient low-level localization abilities. To address this, a Semantic Location enhanced Swin Transformer-based U-Net (SLST-U-Net)+EDEPLDP model is proposed in this article. This model leverages Location Attention (LA) and De Mejora Progresiva (DMP) to enhance feature discrimination at the level of spatial information and semantic position. The Contextual Guidance Attention (CGA) method combines spatial and semantic information. The DMP enhances feature discrimination by increasing edge data inference and providing a richer depiction of the target position. The CGA reduces the semantic gap and effectively fuses spatial texture information and semantic information. The LA mechanism improves computational capacity for semantic features and precision of semantic position data, enabling retrieval of long-range contextual data in channel and geographic contexts. Additionally, Swin Transformer (ST) is added in the encoder and decoder section of U-net to increase the finer details of spatial and semantic information. Finally, the features extraction and classification part of EDEPLDP is employed to detect and classify the lung diseases. Experimental results revealed that the proposed SLST-U-Net+EDEPLDP model outperforms the CNN, E2E-DNN LungNet22, EfficientNet-SE, LDDNet and EDepLDP models with an accuracy of 94.94% and 95.42% on CXR and CT images, respectively.

Keywords: Lung diseases, Location attention, De mejora progresiva contextual guidance attention, Swin transformer.

1. Introduction

Lung diseases or pulmonary disorder negatively influences the lung conditions and their connective tissues in humans [1]. The examples include Coronavirus Disease 2019 (COVID-19), pneumonia, Chronic Obstructive Pulmonary Disease (COPD), etc. Millions of people die annually due to lung illness, making it a major cause of mortality and disability. Early identification is crucial for improved survival and diagnosis [2].

Traditionally, medical practitioners have used diagnostic imaging techniques such as CT, CXR, Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI) to diagnose lung illnesses for the early lung disease prediction [3]. But, these imaging techniques require skill and multiple models to accurately interpret the images. DL is an emerging field that is applied to diagnose a variety of lung diseases and provides valuable assistance for healthcare providers in making accurate medical decisions [4]. This method enables immediate disease detection for analyzing complex medical images. DL approaches are often sorted as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short Memory Term (LSTM), and others [5]. CNN is a common DL approach that trains clinical image structures and features to identify all disease types using improved clinical images [6].

For example, an automatic DL-based lung ailment detection model [7] has been developed to categorize healthy and infected CXR scans. The model utilizes manual lung masks to segment the lung area and a new CNN architecture with extra layers and tweaked hyper-parameters. However, this model was plagued by epistemic uncertainty, which significantly impacts the efficacy of DL models used to identify lung diseases. To solve this, the EDepLDP model has been proposed [8]. The model uses U-Net architecture to segment CXR and CT images, then uses InceptionResNetV2 and Xception to identify informative and discriminative features. These features are then used in conGRU-LSTM to categorize lung disorders. However, the U-Net model applied for the segmentation applications results in poor low-level localization abilities.

In this paper, SLST-U-Net+EDepLDP model is developed to overcome the inadequacies of U-net's localisation capabilities in EDepLDP when applied to CT and CXR images for lung disease diagnosis. SLST-U-Net increases feature discrimination at the granularity of geographic detail and semantic location by using DMP and LA. The CGA technique integrates spatial and semantic data, improving feature discrimination and enriching data with a comprehensive target representation. It bridges the gap between these two types of data, allowing seamless integration of spatial texture and semantic data. The LA mechanism enhances representational capacity for semantic features and location information, efficiently collecting long-range contextual information in channel and geographic settings. Both encoder and decoder in U-net use ST for extracting coarse and fine-grained information. Finally, the EDepLDP feature extraction and classification segment is employed to accurately classify the lung diseases types.

The following portions are prepared as follows: Section II examines related studies. Section III explains the SLST-U-Net+EDepLDP model for lung disease classification. Section IV illustrates the performance effectiveness of the proposed model. Section V summarizes the whole work and suggests future enhancement.

2. Literature survey

A hybrid technique using DL networks was developed [9] for categorizing Interstitial Lung Disease (ILD). CT images were segmented using a conditional GAN, followed by a multiscale feature retrieval module. Pre-trained ResNet50 classifier extracted features and Support Vector Machine (SVM) was employed for ILD classification. However, this results with lowest accuracy results.

A multi-stage approach called SPFKMC method was developed [10] for segmenting and categorizing ILD patterns by utilizing superpixel analysis and k-means cluster fusion. But, this model faces substantial low-order localization concerns.

The EfficientNet v2-M deep learning model [11] was built for classifying lung disorders on CXR scans. This model enhances the detection task accuracy by utilizing pre-trained weights from ImageNet and augmented data to enhance sample diversity. But, the hyper-parameters were not fine-tuned well resulting in lower precision rate.

An End-to-End Deep Neural Network (E2E-DNN) [12] was developed to classify lung disorders from CT image patches. The data was cleaned and a new structure was developed to better categorize lung tissue images, trained using categorical cross-entropy and optimized using Adam. However, this model resulted with lower F1-Score.

The LungNet22 model was developed [13] for lung disease types, using raw CXR data from multiple sources. It uses CLAHE techniques and Green Fire Blue filtering, upsampling classes with less data, and CNN models for categorization. But, the model doesn’t considered the negative cases equally resulting to lower recall values.

A multichannel EfficientNet-based stacking ensemble method was developed [14] using CXR images for lung disease identification. The dataset combines retrieved traits into a model partitioned into dynamic layers. This is fed into a stacked ensemble learning classifier for lung disease diagnosis. But, this model provides lower accuracy on larger datasets.

A CNN framework for lung segmentation was constructed [15] using CT and CXR images, capturing relevant features with concatenate blocks and a transpose layer for enhanced spatial resolution. However, this model necessitates advanced segmentation model to improvise the accuracy result.

An optimized DenseNet201 layer called LDDNet was constructed [16] for detecting infectious lung diseases was developed using CT and CXR images. It incorporates dropout, batch normalization, global average pooling layer, and

Received: December 6, 2023. Revised: January 18, 2024.

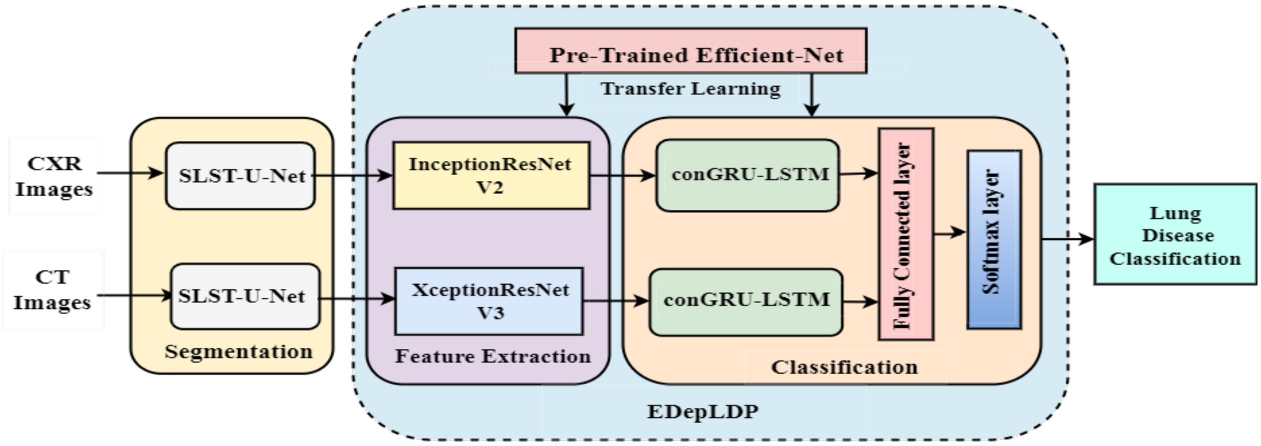


Figure. 1 Overall Pipeline of SLST-U-Net+EDepLDP

Table 1. Lists of notations

Notations	Description
c	Number of Channels
$h * w$	spatial resolution of the image
P_{uv}	Position Relevance Attention Map
H_{high}	Large receptive field feature
H_{low}	Small receptive field feature
g	final output of GC
w_a and w_b	Embedding matrices of different convolution projections
$Gate$	Attention Map
σ	Sigmoid Function maps all the values between 0 and 1
H	Encoded Feature
ϕ	ReLU activation
η_u	Weighted sum of a linearly transformed input elements
w^y, w^x and w^j	Parameter matrices of each layer and each attention head
Υ_{uv}	Each weight coefficient calculated by using softmax function
$\varkappa_{u,v}$	Correlation among two input elements intended by scaled dot product
i_u and i_v	Definite position embedding between input variable
$z_{u,v}$	Matrix form for i_u and i_v
C_{map}	Context-dependent attention map
y_{uv}^j	Trainable position parameter matrix.
H_c	Feature of the c^{th} route of feature H
C_w	Tasks correlation of all channel feature
$\bar{\bar{H}}$	Key Channel Feature
n	Total number of pixels in each image
T	Decoded Feature
s_u	Ground-truth value of the u^{th} pixel
z_u	Confidence score of the u^{th} pixel
S	Patch Size
$\widehat{H}_{out}^u$	Final output of small scale branch in TCF
h^a	Primary feature Branch in TCF
g^u	Secondary feature Branch in TCF

dense layer. However, this accuracy would be degraded if the model trains on larger datasets.

3. Proprsed Methodology

This section briefly outlines the complete framework of SLST-U-Net+ EDepLDP, as shown in Fig. 1. Table 1 lists the notations used in this study.

3.1 Semantic Localization for Semantic Feature Enhancement

Swim transformers are used in SLST-U-Net for medical image segmentation improving the performance of an ST-based model. The SLST-U-Net is built using U-Net for segmenting clinical images, following previous work [8].

The SLST-U-Net system consists of three steps: encoding, decoding, and improving semantic features. ResNet50 is a model used for characterizing features in medical images with four encoding blocks as depicted in 2(a) that down-sample feature maps by a factor of two.

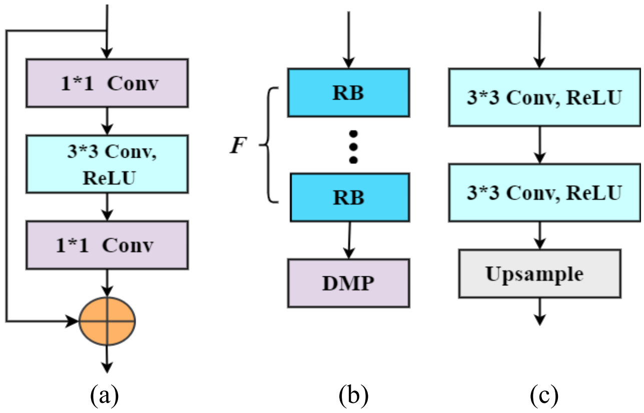


Figure. 2 Design of the encoding and decoding: (a) Residual Block (RB) in encoding, (b) Position of DMP, and (c) Design of Decoding.

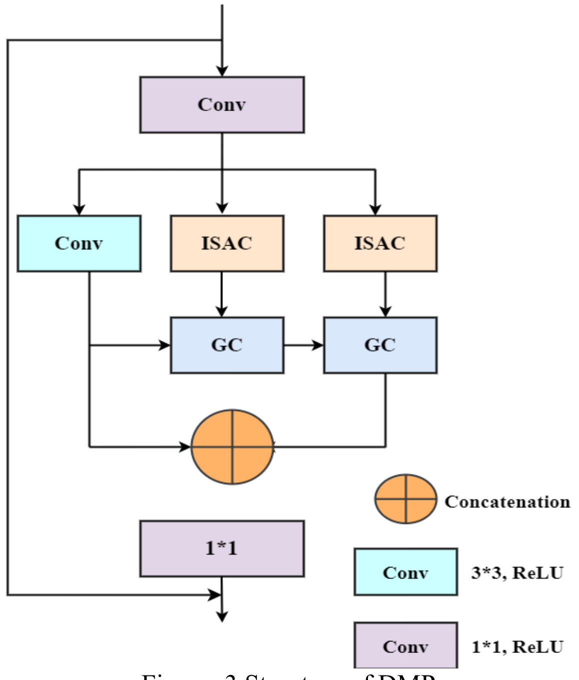


Figure. 3 Structure of DMP

Fig. 2(b) and 2(c) depicts the DMP location and encoding structure respectively. The decoding process is compatible with traditional U-Net, involving two convolutions and one up-sampling for segmentation. The semantic feature enhancement, referencing LA, improves representation capacity whereas DMP and CGA are included in the third and second encoded to reduce processing costs.

3.1.1 Framework of DMP unit

The Intensive Self-Attention Convolution (ISAC) and the Gated Convolution (GC) are the two components of the gradual improvement mechanism. The structure of DMP model is depicted in Fig. 3.

The convolution function uses the $3 * 3$ convolution module and two dilated SAC blocks to collect features from various receptive fields. The GC receives features from both operations, and the larger receptive field feature is used for differential extraction. The discriminative features are captured again before the original 3*3 convolution output and the results of the two GCs are merged.

[a] **Intensified Self-Attention Convolution (ISAC):** The ISAC uses Multi-Head Self-Attention (MHSA) transformer method which focuses on local and global data using dilated convolution embedding. This method optimizes feature representations by selectively aggregating global context and incorporating broader contextual positioning information into local features. At first, three distinct dilated convolution functions are conducted to the encoder feature a $\in ¡^{c*h*w}$ to yield

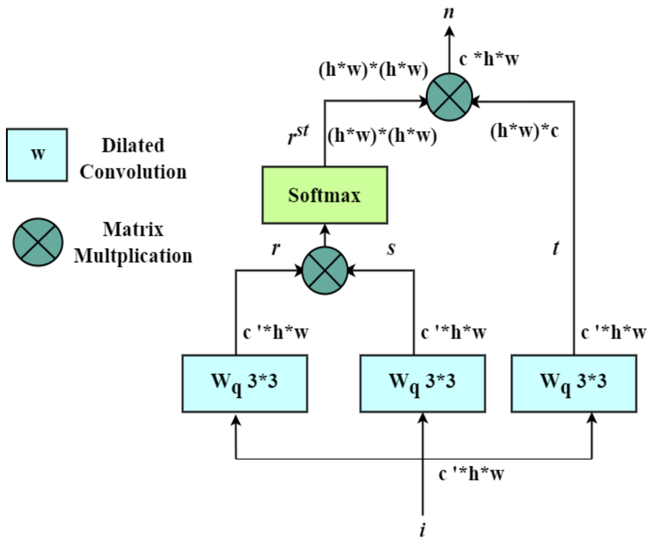


Figure. 4 Structure of ISCA

the feature maps $r \in ¡^{c'*h*w}$, $s \in ¡^{c'*h*w}$ and $t \in ¡^{c'*h*w}$. In order to decrease the computation and the amount of model variables where c is specified as $c' = {}^{c}/_{4}$. Following that, r and s are re-formed as the feature maps $E \in ¡^{(h*w)*c'}$, $F \in ¡^{(h*w)*c'}$ and t into $K \in ¡^{(h*w)*c}$, respectively. The fig. 4 depicts the ISAC model.

The positional relevant attention is generated through matrix multiplication and softmax normalization using feature maps E and F, as described in Eq. (1).

$$P_{uv} = \frac{\exp(E_u \,.\, F_v)}{\sum_{x=1}^{n} \exp(E_x \,.\, F_v)} \qquad (1)$$

In Eq. (1), P_{uv} represents the influence of the u^{th} position on the v^{th} position in Eq. (1), and $n = h * w$ is the pixel numbers. Following that, Q is multiplied by P and the resultant feature at each point may be expressed in Eq. (2),

$$ISAC(P, Q) = P.Q \qquad (2)$$

Ultimately, the optimised feature maps are re-modified in order to keep the ISAC output, i.e., $H_{isac} \in ¡^{c*h*w}$.

Gated Convolution (GC): The GC module has two inputs, $H_{high} \in ¡^{c*h*w}$ and $H_{low} \in ¡^{c*h*w}$ which represent major andminor adaptive field features respectively, as depicted in Fig. 3. The feature maps $Gate \in ¡^{c*h*w}$ and $H \in ¡^{c*h*w}$ are then generated by applying two distinct convolutional methods to the features as H_{high} and H_{low} . After that, a multiplication function is accomplished instantly. To calculate the absolute result g of GC, i.e., $H_{GC} \in$

¡^{c*h*w}. The computing procedure is expressed as following

$$Gate = w_a . H_{high} \tag{3}$$

$$H = w_b . H_{low} \tag{4}$$

$$g = \phi(H) * \sigma(Gate) \tag{5}$$

In the preceding equations (3), (4), and (5), w_a and w_b are the embedding vectors of various convolution mappings. The attention map and sigmoid operations is defined as $Gate$ and σ which maps each value to the range from 0 to 1. H defines the feature integration, and represents ReLU stimulation.

3.1.2 Local Attention (LA)

The LA model employs Routed MHSA (RMHSA) and Dimensional MHSA (DMHSA) mechanisms to accurately represent lesion regions, with RMHSA utilizing SA for improved route and dimensional data, and DMHSA using relative position embedding for spatial feature connections. The task of definite position embedding is defined by

$$\varkappa_{u,v} = \frac{(i_u w^y . (i_v w^x)^Q}{\sqrt{D_\eta}} \tag{6}$$

$$z_{u,v} = \left(i_u w^j\right) . (y^j_{uv})^Q \tag{7}$$

$$\Upsilon_{uv} = \frac{exp(\varkappa_{u,v} + \varrho_{u,v})}{\sqrt{D_\eta}} \tag{8}$$

$$\eta_u = \textstyle\sum_{v=1}^{n} \Upsilon_{uv} . i_v \tag{9}$$

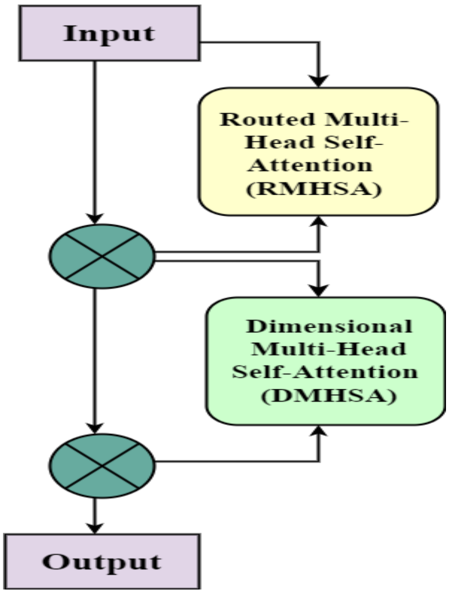

Input
Routed Multi-Head Self-Attention (RMHSA)
Dimensional Multi-Head Self-Attention (DMHSA)
Output

Figure. 5 Depiction of LA

The output component η_u is calculated by calculating the weighted sum of linearly changed input components. Variable vectors w^y, w^x and w^j are varied for all layers and attention heads. e $z_{u,v}$ matrix defines the definitive positional integration of input variables i_u and i_v which is improved through backward propagation and can be trained using y^j_{uv}. The fig. 5 depicts the LA mechanism.

3.1.3 Channel Guide Attention (CGA)

The CGA effectively reduces superfluous textual data, narrows the semantic gap and enhances feature concatenation efficiency in skip connection. The ability of MHSA is to determine the associative between feature and semantic positions, as depicted in Fig. 6. The lower level input from encoder feature $H_{low} \in$ ¡^{c*h*w} is initially moulded into $X_{low} \in$ ¡$^{heads*(h*w)*dim}$ and $J_{low} \in$ ¡$^{heads*(h*w)*dim}$ respectively. After that, channel selection (CS) to extract essential K channels.

The significant channels of T are selected using CS after reformulating the high-level input from the decoder feature $H_{High} \in$ ¡^{c*h*w} . A context-dependent attention map $C_{map} \in$ ¡$^{heads*(h*w)*(h*w)}$ is generated by performing a multi-head scaled dot-product function with softmax normalization among T and the transformed variant of J, indicating the global factors with respect to spatial attributes.

Multiplying the map C by J would provide an aggregate of values with weights determined by J. The last step in determining the CGA output is to

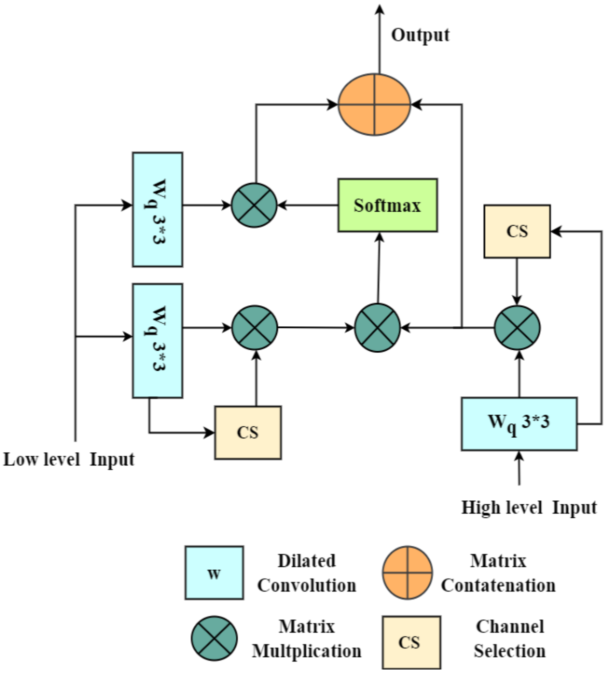


Figure. 6 Illustration of CGA

Received: December 6, 2023. Revised: January 18, 2024.

merged together to adjusted the low and high level features, i.e., $H_{cga} \in ¡^{(2*c)*h*w}$. The CS can be formulated as follows,

$$\mathcal{Z}_c = \frac{1}{h*w}\sum_{u=1}^{h}\sum_{v=1}^{w}\left(H_c(u,v)\right), \mathcal{Z} \in ¡^{c}, F \in ¡^{c*h*w} \quad (10)$$

$$C_w = sigmoid\ (w.\mathcal{Z}\ ), w\ \in ¡^{c*c}, C_w \in ¡^{c} \quad (11)$$

$$\bar{\bar{H}} = C_w > H, H \in ¡^{c*h*w} \quad (12)$$

The task-correlated channel-wide feature vector C_w. At last, the key channel feature $\bar{\bar{H}}$ is calculated by multiplying the input feature H by C_w . By continuously optimising the weight w in the model training tasks in which the most important channel feature is identified with high precision.

3.1.4 Loss Function

The End-to-end training is used by SLST-U-Net throughout the training stage. The binary cross entropy loss (L_{bce}) and Dice Loss (L_{dice}) are examples of entropy losses. The following is the expression for determining L_{bce} and L_{dice}.

$$\ell_{bce} = -\sum_{v=1}^{n}(s_u \log(z_u)) + (1-s_u)\log(1 - s_u)) \quad (13)$$

$$\ell_{dice} = 1 - \frac{\sum_{v=1}^{n} s_u z_u + \mu}{\sum_{v=1}^{n}(s_u z_u) + \mu} \quad (14)$$

$$\ell_{total} = \alpha\ .\ L_{bce} + \beta\ .\ L_{bce} \quad (15)$$

Where, z_u is the confidence score for pixel u depending on the estimations, s_u is the real value for pixel u and n is the cumulative pixel numbers in the images. In this system, α and β have values of 0.6 and $\mu = 10^{-5}$ respectively.

3.2 Swin Transformer block in Semantic Feature module for Image Segmentation

In this model, ST block is used in the U-net encoder and decoder part, which can be effective for medical imagine segmentation tasks when combined with U-Net for finer spatial and semantic data details. The ST encoder consists of l identical units, each composed of Multi-Layer Percepton (MLP) and MHSA, with a residual association and a LayerNorm (ln) layer is executed before and every MSA component and every MLP. As a result, the Transformer encoder's l −layer output q_l may be written as follows:

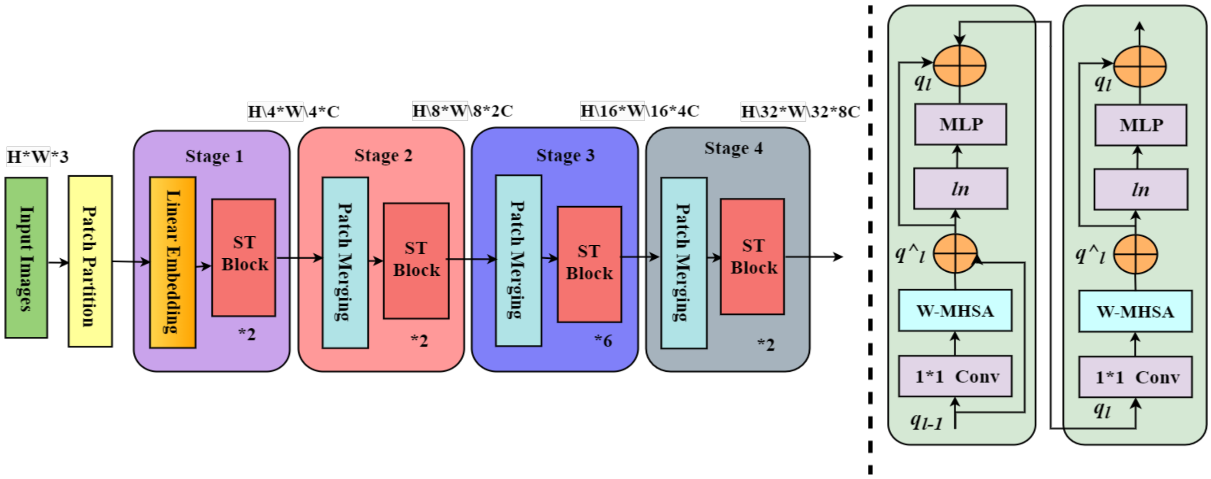
$$\hat{q_l} = MSA(\ln(q_{l-1}) + q_{l-1}) \quad (16)$$

$$q_l = MLP(ln(\hat{q}_l) + \hat{q}_l\ ) \quad (17)$$

The traditional transformer structure is deemed unsuitable for complex estimations and high-quality imaging applications due to its exponential computational cost. ST recommends Window-based MHSA (W-MHSA) or Shifted Window MHSA (SW-MHSA) for effective simulation. W-MHSA constrains $m * m$ patches and partitions input features into windows, with self-reflection only during operational periods. The l^{th} layer of W-MHSA and MLP provide the following outputs denoted by $\hat{q_l}$ and q_l respectively,

$$\hat{q_l} = W - MHSA(ln(q_{l-1})) + q_{l-1} \quad (18)$$

$$q_l = MLP(ln(\hat{q}_l)) + \hat{q}_l\ ) \quad (19)$$



(a) (b)

Figure. 7: (a) Architecture of ST model and (b) Two successive ST i.e., W-MHSA and SW-MHSA

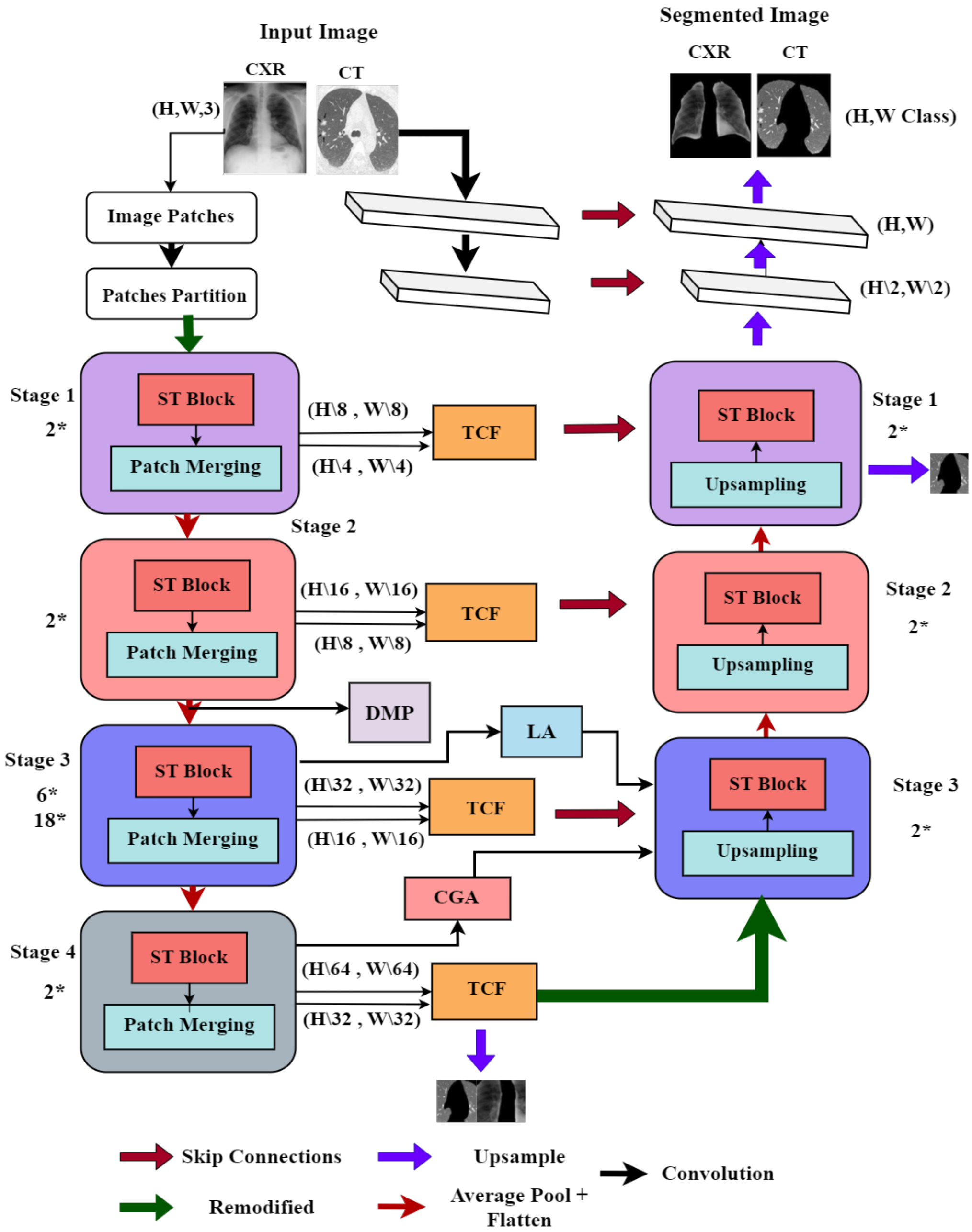


Figure. 8 Structure of Proposed SLST-U-Net Model

SW-MSA is a batch-processing method that maintains identical number of batch windows in a fixed partitioning scheme, while both W-MSA and SW-MSA are used for relative location bias. The results of the SW-MSA and the MLP blocks might be expressed in the form of

$$\hat{q}_{l-1} = SW - MHSA(ln(q_{l-1})) + q_{l-1} \quad (20)$$

$$q_{l+1} = MLP(ln(\hat{q}_l)) + \hat{q}_l \quad (21)$$

The complete structure of SLST-U-Net a medical image segmentation model is shown in Fig. 8. The ST blocks in subsequent stages function in the same way, with the exception that their input is the output of the previous stage. Fig. 7(a) depicts the structure of ST model and Fig. 7(b) depicts the two successive models of ST, W-MHSA and SW-MHSA with regular and shifted windowing configurations, respectively.

[a] Encoder: The model segmented images using U-Net architecture and extracted features using ST encoder. A linear embedding layer sliced the input medical images into $H/_S * H/_S$ non-overlapping patches, tokenized, and projected to dimension k. Tokens were sent into ST, a four-phase system with a fixed block count, which decreased as the network expanded, with the first three steps involving a patch merging layer. This technique combines features of all sets of $2 * 2$ neighbouring patches and performs a linear layer on the channel-dimensional combined features, lowering many tokens $2 * 2 = 4,2 *$ down-sampling and raising the output dimension by 2. The resulted resolutions of 4 levels are $^{H}/_{S} * {}^{H}/_{S}$, $^{H}/_{2S} * {}^{H}/_{2S}$, $^{H}/_{4S} * {}^{H}/_{4S}$ and $^{H}/_{8S} * {}^{H}/_{8S}$ with dimensions of $k, 2k, 4k, and\ 8k$, correspondingly.

[b] Decoder: The proposed decoded method uses upsampling, skip connection, and a ST block to improve decoding speed. The input is level 4 encoder output, with features up-sampled and combined with disconnected feature maps. The ST block develops long-term relationships and global context connection. Low-level attributes with ranges of $h * w$ and $^{H}/_{2} * {}^{H}/_{2}$ are used to generate the $^{H}/_{4} * {}^{H}/_{4}$ output with three layers like $3 * 3$ convolutional layer, group normalization layer, and ReLU layer. The skip connection is used to obtain ultimate mask predictions.

[c] Multi-Scale Feature Representations: The model proposes a multi-scale ST for feature extraction, enhancing segmentation efficiency and strengthening patch relationships. It employs a main extend with a patch size of 4 and an additional branch with a patch size of 8. As a consequence, the encoding branch produces results with resolutions of $^{H}/_{4} * {}^{H}/_{4}$, $^{H}/_{8} * {}^{H}/_{8}$, $^{H}/_{16} * {}^{H}/_{16}$ and $^{H}/_{32} * {}^{H}/_{32}$ and H⁄32*H⁄32, while the decoding side produces output with values of $^{H}/_{8} * {}^{H}/_{8}$, $^{H}/_{16} * {}^{H}/_{16}$, $^{H}/_{32} * {}^{H}/_{32}$ and $^{H}/_{64} * {}^{H}/_{64}$.

[d] Transformer Conjoint Integrative (TCF) Module: The MHSA method enhances image segmentation by creating a TCF module after obtaining encoder features, generating a token based on one branch's feature map, and focusing TCF as the primary focus of the ST model. Assume, for results of 2 branches from an equal level u ($u = 1,2,3,4$) represented by $h^{a} = [H_1^u, H_2^u, \dots, H_{h*w}^u] \in \mathbb{R}^{c*(h*w)}$ (primary branch) and $g^u = \left[G_1^u, G_2^u, \dots, G_{\frac{h}{2}*\frac{h}{2}}^u\right] \in \mathbb{R}^{c*(\frac{h}{2}*\frac{h}{2})}$ (complementary branch), correspondingly. Afterward, the transformation result of g^u is obtained as follows,

$$\widehat{G^u} = Flatten\,(Avgpool(g^u)) \tag{22}$$

$\widehat{G^u} \in \mathbb{R}^{c*1}$ is a 1D average pooling (Avgpool) layer which is preceded by a flatten procedure in Eq. (22), In order to engage with Hu at the pixel level, the token $\widehat{G^u}$ stands for the global conceptual data of $\widehat{g^u}$. For the computation of global SA, H^u is combined with $\widehat{G^u}$ to produce a stream of $1 + h * w$ tokens.

$$\widehat{H^u} = ST\,([\widehat{G^u},]h_1^u, h_2^u, \dots, h_{h*w}^u) \tag{23}$$

$$= ST\,(h_1^u, h_2^u, \dots, h_{h*w}^u) \in \mathbb{R}^{c*(1*h*w)} \tag{24}$$

$$\widehat{H}_{out}^u = (h_1^u, h_2^u, \dots, h_{h*w}^u) \in \mathbb{R}^{c*(1*h*w)} \tag{25}$$

Where, ST is analogous to Eq. (16) and (17) and $\widehat{H}_{out}^u$ is the small-scale branch's ultimate output in TCF. This method establishes relations among each token in $H^u = [h_1^u, h_2^u, \dots, h_{h*w}^u] \in \mathbb{R}^{c*(1*h*w)}$ and the complete g^u allowing for the optimized features to access coarse-grained data from the massive branch. In order to improve segmentation efficiency, TCF in ST module might deliver efficient feature fusion of multi-scale branch.

The segmented CXR and CT images from SLST-U-Net are inputted into InceptionResNetV2 and Xception of EDEPLDP to identify informative and discriminative features respectively. Then, the extracted deep features are fed into the softmax layer of conGRU-LSTM for lung disease classification. [8]. Hence, the proposed SLST-U-Net+EDEPLDP model effectively resolves the low-level localization issues in U-Net based segmentation for CT and CXR images for the efficient prediction of lung diseases and its types.

4. Author name(s) and affiliation(s)

4.1 Dataset description: In this study, two benchmark databases are considered which is briefly illustrated below.

CXR data: The CXR dataset [17], which includes 112,120 frontal-view X-ray images of 30,805 patients with 14 thoracic pathologies, includes bacterial or viral diseases, chronic obstructive lung disease, and COVID-19 and non-Covid chest X-ray instances [18]. For experimental purposes, only five pathologies, atelectasis, infiltration, pneumonia, along with COVID-19 and non-Covid, are considered.

CT data: As like CXR data, CT images were also categorized into five diseases: atelectasis, infiltration, pneumonia, COVID-19, and non-COVID. This study analyzed various open public portals to collect CT data for experimental purposes. The lung atelectasis images were obtained from [19], COVID and non-COVID (Normal) images from [20], viral pneumonia from [21], and infiltration from [22].

Table 2. Observed CXR and CT scans

Diseases\ Images observed	CXR Images			CT Images		
	Number of CXR images Considered	Training Images (70 %)	Testing Images (30 %)	Number of CXR images Considered	Training Images (70 %)	Testing Images (30 %)
Covid-19	1345	942	403	1002	702	300
Normal	1345	942	403	984	689	295
Pneumonia	1443	1011	432	1762	1234	528
Atelectasis	290	203	87	310	217	93
Infiltrate	270	189	81	260	182	78

Table 3. Parameter settings for existing and proposed model

Framework	Parameters	Range	Framework	Parameters	Range
CNN [7]	No. of convolutional layer	3	LDDNet [16]	No. of convolutional layer	3
	Max Pooling	2		Learning rate	0.001
	Stride	2		Batch size	96
	No. of. Epochs	200		Epochs	30
	Dropout	0.5		Optimizer	SGD
	Optimizer	Adam		Loss Function	Cross Entropy (CE)
	Learning rate	0.0001		Dropout	0.5
	Batch Size	32		Activation Function	ReLU
E2E-DNN [12]	No. of LSTM units	64	EDepLDP [8]	No. of convolutional layer	3
	Activation function	tanh		Learning Rate	0.001
	Stride size	2		Batch Size	128
	Optimizer	Adam		Optimizer	Adam
	Batch size	64		Epochs	25
	No. of epochs	18		Activation Function	ReLU
	Learning rate	0.001		Stride	2
	Dropout	0.4		Dropout	0.7
LungNet22 [13]	No. of convolutional layer	3		Loss Function	BCE
	Pooling size	2	Proposed SLST-U-Net + EDEPLDP	Layers	3 Encoding; 3 Decoding
	Learning rate	0.000001		No. of convolutional layer	3
	Batch size	128		Kernel Size	3
	Optimizer	Adam		Learning Rate	0.01
	Activation Function	ReLU		Dimensions	8
	No. of epochs	100		Optimizer	SGD
	Dropout	0.9		Momentum	0.9
EfficientNet-SE [14]	No. of. Dimensions	1408		Weight Decay	0.0005
	Learning rate	0.0001		Stride	2
	Batch size	64		Epochs	72
	Optimizer	Adam		Batch Size	256
	Epochs	15		Dropout	0.7
	Dropout	0.001		Loss Function	BCE
	Loss Function	Binary Cross Entropy (BCE)			

Table 2 lists the total number of CXR and CT scans observed in the dataset used for training and testing.

Performance evaluation: This study investigates the effectiveness of the SLST-U-Net+EDEPLDP model using CXR and CT images in MATLAB 2019a. In both CXR and CT datasets, 70% of the images are used for training, while 30% are used for evaluation in each category of lung disorders.

A quantitative performance is conducted to evaluate the efficiency of SLST-U-Net+EDEPLDP model by comparing with as CNN [7], E2E-DNN [12], LungNet22 [13], EfficientNet-SE [14], LDDNet [16] and EDepLDP [8] which are also implemented and tested using the above-considered datasets. Table 3 lists parameter settings for the proposed SLST-U-Net + EDEPLDP and existing models.

The assessment measures used to evaluate the effectiveness of the proposed and current models in the both CXR and CT are briefly discussed.

4.2.1 Accuracy: It is the proportion of correctly classified instances in each class of lung disorders over the total number of samples examined.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{26}$$

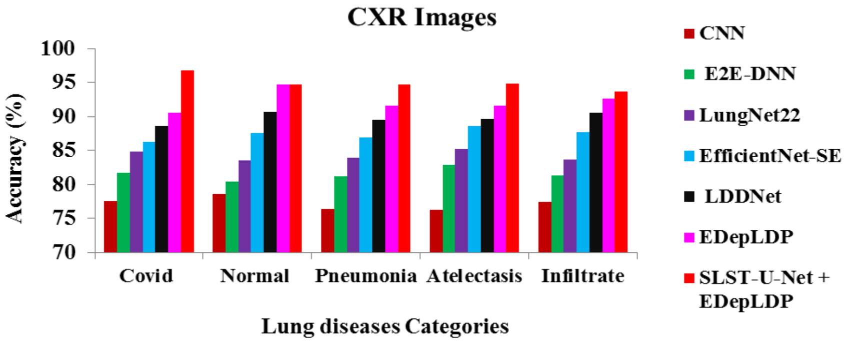


Figure. 9 Accuracy Comparison for lung disease category using CXR images

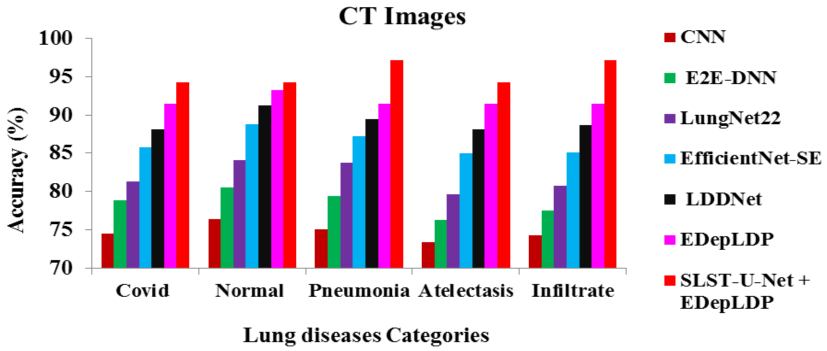


Figure. 10 Accuracy Comparison for lung disease category using CT images

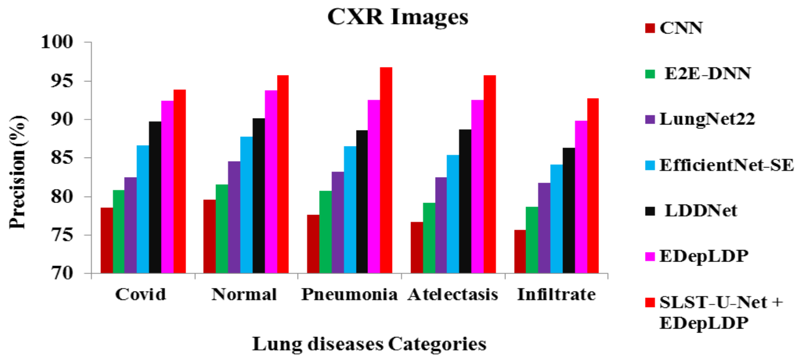


Figure. 11 Precision Comparison for lung disease category using CXR images

Received: December 6, 2023. Revised: January 18, 2024.

In above Eq. (26), TP (True Positive) is the result in which the model correctly labels the lung disease categories as themselves, for example, pneumonia is categorised as pneumonia. The result indicates that the classifier successfully identifies the Covid-19 as Covid-19 is defined as TN (True Negative). The term FP (False Positive) refers to the result in which the model incorrectly identifies lung disorders (AtelectasisCovid-19InfiltrateNormalPneumonia) as Atelectasis. The result FN (False Negative) demonstrates that the model incorrectly labels the Normal as Infiltrate.

4.2.2 Precision: It is the proportion of correctly classified instances of lung disease types at the TP and FP rates.

$$Precision = Precision = \frac{TP}{TP+FP} \tag{27}$$

4.2.3 Recall: It is the ratio of instances of accurately defined lung illnesses at TP and FN rates.

$$Recall = \frac{TP}{TP\ + FN} \tag{28}$$

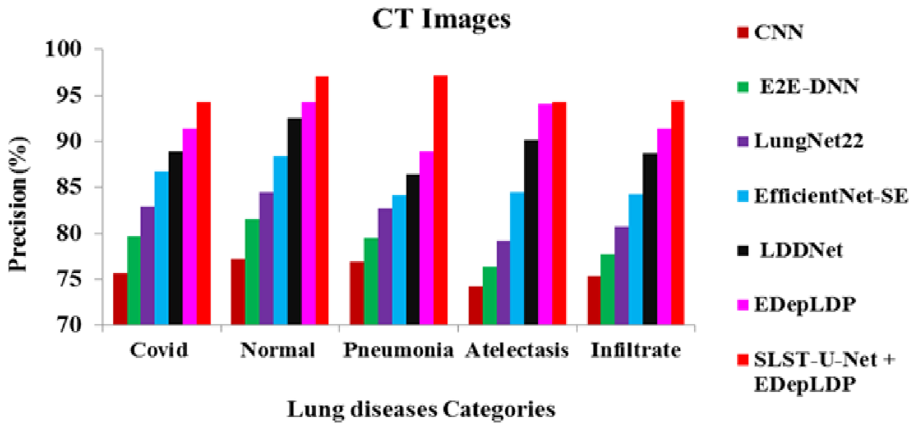


Figure. 12 Precision Comparison for lung disease category using CT images

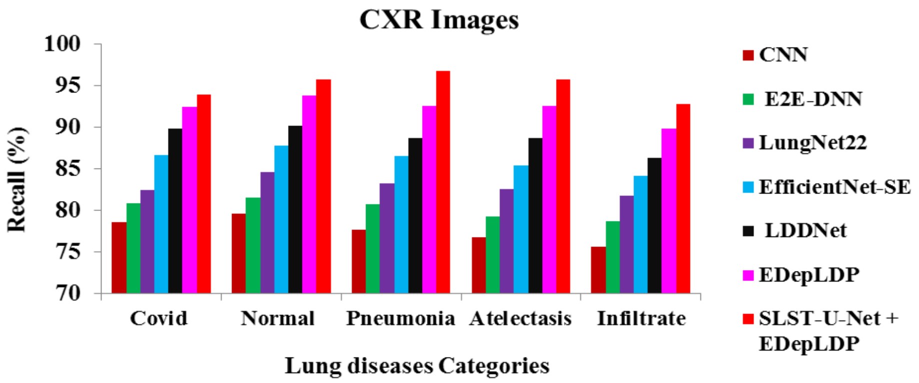


Figure. 13 Recall Comparison for lung disease category using CXR images

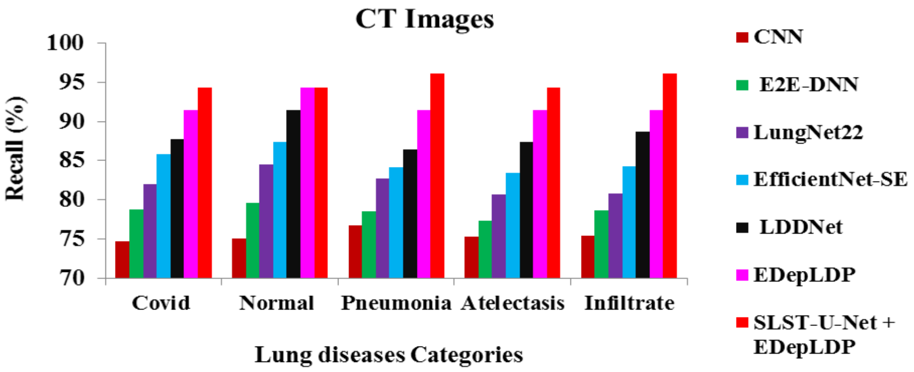


Figure. 14 Recall Comparison for lung disease category using CT images

Figs. 9 and 10 show the accuracy of various models in diagnosing Covid-19, Normal, Pneumonia, Atelectasis, and Infiltrate using CT and CXR. The SLST-U-Net+EDEPLDP model outperforms other models in lung condition classification due to its increased learning instances from CXR and CT images. For example, in the classification of atelectasis, the accuracy of SLST-U-Net+EDEPLDP is 24.28% and 28.51% higher than CNN, 14.29% and 23.69% higher than E2E-DNN, 11.27% and 18.44% higher than LungNet22, 6.9% and 10.99% higher than EfficientNet-SE, 5.7% and 6.96% higher than LDDNet, and 3.51% and 3.13% higher than EDepLDP for the CXR and CT images respectively.

Figs. 11 and 12 indicates the precision (in%) achieved by CNN, E2E-DNN, LungNet22, EfficientNet-SE, LDDNet, EDepLDP, and SLST-U-Net+EDEPLDP models in identifying Covid-19, Normal, Pneumonia, Atelectasis, and Infiltrate using CT and CXR images, accordingly. It is found that the precision of SLST-U-Net+EDEPLDP for every category of lung illness is superior to that of other categorisation models from CXR and CT images for each diseases classes. In the case of Covid categorization, the precision of SLST-U-Net+EDEPLDP is 19.53% and 24.56% higher than CNN, 16.15% and 18.34% higher than E2E-DNN, 13.85% and 13.71% higher than LungNet22, 8.37% and 8.67% higher than EfficientNet-SE, 4.60% and 6.09% higher than LDDNet, 1.52% and 3.13% higher than EDepLDP for CXR and CT images.

Figs. 13 and 14 show the recall (%) achieved by existing and proposed models when utilising CT and CXR images to diagnose different lung disease categories such as Covid-19, Normal, Pneumonia, Atelectasis and Infiltrate. It is established that the recall of SLST-U-Net+EDEPLDP for each category of lung illnesses is superior to that of other classification models from CXR and CT images for each category of diseases. In the case of pneumonia classification, the recall of SLST-U-Net+EDEPLDP is 24.59% and 25.29% higher than CNN, 19.83% and 22.38% higher than E2E-DNN, 16.34% and 16.29% higher than LungNet22, 11.89% and 14.24% higher than EfficientNet-SE, 9.17% and 11.19% higher than LDDNet, and 4.56% and 5.15% higher than EDepLDP for CXR and CT images.

5. Conclusion

The paper proposes SLST-U-Net+EDEPLDP to enhance U-net's low localisation abilities for CT and CXR images to identify lung illnesses. This model utilizes DMP and LA to distinguish spatial and semantic features, while CGA combines spatial and semantic data. DMP optimizes edge data interpretation and increases objective position depiction, while CGA reduces semantic gaps by merging spatial texture and semantic data. LA improves semantic feature representation and location information accuracy, enabling long-range contextual information in channel and geographic situations. ST efficiently extracts coarse and fine-grained characteristics from geographical and semantic information. The SLST-U-Net+EDEPLDP model outperforms traditional segmentation methods on CXR and CT images with an accuracy of 94.94% and 95.42% for efficient lung diseases prediction.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, methodology, software, validation, Dhivya; formal analysis, investigation, Sharmila; resources, data curation, writing—original draft preparation, Dhivya; writing—review and editing, Dhivya; visualization, supervision, Sharmila.

References

[1] M. Avalos-Fernandez, T. Alin, C. Métayer, R. Thiébaut, R. Enaud and L. Delhaes, “The respiratory microbiota alpha-diversity in chronic lung diseases: first systematic review and meta-analysis”, *Respiratory Research*, Vol. 23, No. 1, pp. 214, 2022.

[2] H. M. Emara, M. R. Shoaib, W. El-Shafai, M. Elwekeil, E. E. D. Hemdan, M. M. Fouda, ... & F. E. A. El-Samie, “Simultaneous Super-Resolution and Classification of Lung Disease Scans”, *Diagnostics*, Vol. 13, No. 7, pp. 1319, 2023.

[3] M. O. Wielpütz, C. P. Heußel, F. J. Herth and H. U. Kauczor, “Radiological diagnosis in lung disease: factoring treatment options into the choice of diagnostic modality”, *Deutsches Ärzteblatt International*, Vol. 111, No. 11, pp. 181, 2014.

[4] S. T. H. Kieu, A. Bade, M. H. A. Hijazi and H. Kolivand, “A survey of deep learning for lung disease detection on medical images: state-of-the-art, taxonomy, issues and future directions”, *Journal of imaging*, Vol. 6, No. 12, pp. 131, 2020.

[5] J. Ma, Y. Song, X. Tian, Y. Hua, R. Zhang and J. Wu, “Survey on deep learning for pulmonary

medical imaging", *Frontiers of medicine*, Vol. 14, pp. 450-469, 2020.

[6] V. Agarwal, M. C. Lohani, A. S. Bist, U. Rahardja, A. Khoirunisa and R. D. Octavyra, "Analysis of Emerging Preprocessing Techniques Combined with Deep CNN for Lung Disease Detection", In: *Proc. of 2022 1st International Conference on Technology Innovation and Its Applications (ICTIIA) IEEE*, pp. 1-6, 2022.

[7] S. Z. Y. Zaidi, M. U. Akram, A. Jameel and N. S. Alghamdi, "Lung segmentation-based pulmonary disease classification using deep neural networks", *IEEE Access*, Vol. 9, pp. 125202-125214, 2021.

[8] N. Dhivya and P. Sharmila, "Multimodal feature and transfer learning in deep ensemble model for lung disease prediction", *International Journal of Data Acquisition and Processing*, Vol. 38, No. 2, pp. 271, 2023.

[9] A. U. Gupta and S. Singh Bhadauria, "Multi-Level Approach for Segmentation of Interstitial Lung Disease (ILD) Patterns Classification Based on Superpixel Processing and Fusion of KMeans Clusters: SPFKMC", *Computational Intelligence and Neuroscience*, 2022.

[10] M. S. Junayed, A. A. Jeny, M. B. Islam, I. Ahmed and A. S. Shah, "An efficient end-to-end deep neural network for interstitial lung diseaserecognition and classification", *Turkish Journal of Electrical Engineering and Computer Sciences*, Vol. 30, No. 4, pp. 1235-1250, 2022.

[11] S. P. Pawar and S. N. Talbar, "Two-stage hybrid approach of deep learning networks for interstitial lung disease classification", *BioMed Research International*, 2022.

[12] S. Kim, B. Rim, S. Choi, A. Lee, S. Min and M. Hong, "Deep learning in multi-class lung diseases' classification on chest X-ray images", *Diagnostics*, Vol. 12, No. 4, pp. 915, 2022.

[13] F. J. M. Shamrat, S. Azam, A. Karim, R. Islam, Z. Tasnim, P. Ghosh and F. De Boer, "LungNet22: A Fine-Tuned Model for Multiclass Classification and Prediction of Lung Disease Using X-ray Images", *International Journal of Personalized Medicine*, Vol. 12, No. 5, pp. 680, 2022.

[14] V. Ravi, V. Acharya and M. Alazab, "A multichannel EfficientNet deep learning-based stacking ensemble approach for lung disease detection using chest X-ray images", *Cluster Computing*, Vol. 26, No. 2, pp. 1181-1203, 2023.

[15] A. Sulaiman, V. Anand, S. Gupta, Y. Asiri, M. A. Elmagzoub, M. S. A. Reshan and A. Shaikh, "A Convolutional Neural Network Architecture for Segmentation of Lung Diseases Using Chest X-ray Images", *Diagnostics*, Vol. 13, No. 9, pp. 1651, 2023.

[16] P. Podder, S. R. Das, M. R. H. Mondal, S. Bharati, A. Maliha, M. J. Hasan and F. Piltan, "LDDNet: A Deep Learning Framework for the Diagnosis of Infectious Lung Diseases", *Sensors*, Vol. 23, No. 1, pp. 480, 2023.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases", In: *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 2097-2106, 2017.

[18] https://www.kaggle.com/paultimo/thymooney/chestxray-pneumonia

[19] https://radiopaedia.org/articles/lung-atelectasis?lang=us

[20] https://www.kaggle.com/datasets/mehradaria/covid19-lung-ct-scans

[21] https://radiopaedia.org/articles/viral-respiratory-tract-infection.

[22] https://radiopaedia.org/playlists/41156?lang=us