Traitement du Signal
Vol. 41, No. 4, August, 2024, pp. 2081-2091
Journal homepage: http://iieta.org/journals/ts

An Improved Deep Network Model to Isolate Lung Nodules from Histopathological Images Using an Orchestrated and Shifted Window Vision Transformer

Ponnan Sabitha[1], Ramalingam Aroul Canessane[2], Manickarasi Sivathanu Pillai Minu[1], Vinayagamoorthy Gowri[1], Maria Soosai Antony Vigil[1*]

[1] Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai 600089, India
[2] Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai 600119, India

Corresponding Author Email: antonym@srmist.edu.in

https://doi.org/10.18280/ts.410436

Received: 12 December 2023
Revised: 8 April 2024
Accepted: 15 June 2024
Available online: 31 August 2024

Keywords:
attention units, lung nodules, isolation, shifted window transformer, Vision Transformer (ViT)

ABSTRACT

Cancer is a major health issue worldwide. Classification of pulmonary (lung) nodules into benign and malicious is one of the stimulating exploration domain as it is the second most serious malignancy and the crucial source of universal deaths. Accurate identification of lung cancer from Computed Tomography (CT) scans achieves an important role in cancer diagnostics system. Besides, the accuracy of the manual isolation framework for lung cancer is dependent on the severity of the malignancy and the efficiency of the radiologist, which frequently cause inappropriate decisions. Thus, the segmentation of the affected area from the CT images is a very challenging task since the morphological features of pulmonary nodules are very complex. Recently, Machine Learning (ML) approaches, particularly Deep Learning (DL) methods enable medical industry to analyse huge data at remarkable speeds without debasing the accuracy of tumour segmentation algorithms. However, due to minute inter-class variances between the affected area and its adjacent tissues and the huge diversity of isolation targets, the deep models often fail to segment lung nodules accurately. To solve these issues, we develop an Orchestrated and Shifted Window Transformer (OSWT) with Multi-head self-attention (MSA) units to isolate the abnormal (diseased) area from pulmonary CT images precisely. We assess OSWT on a CT lung image dataset, called The Cancer Genome Atlas (TCGA or Atlas), and relate the performance of the proposed OSWT against 7 innovative classification models in terms of performance measures. The segment or using an OSWT delivers 98.4% dice similarity index (DSI), 96.5% of Jaccard similarity measure (JSM), 0.73% of volume error (VE), and 0.99s average computational cost. The extensive experimental results demonstrate that the OSWT model realizes improved performance and is more suitable for isolating abnormal cancer area from CT scans.

1. INTRODUCTION

The diagnostic imaging in current medical industry is considered as a leading discipline for deriving useful insights about medical abnormalities. Digital medical imaging is used to make photographic depictions of the internal tissues of the human body to observe, examine, or treat therapeutic disorders. Indeed, several pioneering modalities and techniques have been employed in medical industry to observe, examine, and understand cancer-like diseases. These techniques are used to obtain swift and more accurate results for making accurate verdicts in disease classification and management. According to the statistics obtain from the International Agency for Research on Cancer, there were 17 million newly identified malignant patients and 9.5 million cancer-related deaths globally in 2018 [1]. The overall cancer burden is projected to rise to 27.5 million new patients and 16.3 million fatalities in 2040. More precisely, one in 18 males and 1 in 46 females are anticipated to suffering from lung malignancy over a lifetime [2]. This burden will certainly be even more due to the increasing omnipresence of reasons that increase risk such as smoking, unnatural diet, fewer childbirths, physical idleness, etc.

Generally, cancer is characterized by the abandoned partition of anomalous cells in tissues of any body parts. Occasionally, these cells also extent to other body parts by the process of metastasis, therefore leading to comorbidities [3]. There are different malignancies such as breast, brain, ovarian, lung, and cervical cancers. Among them, pulmonary malignancy is a critical cancer which is becoming the major source of 1.69 million mortalities annually [4]. Lung cancer is an abnormal growth of cells in the lung. The most significant lung cancer is carcinoma. This type of cancer is classified into two types: (i) small-cell lung carcinoma (SCLC) which is related to any kind of smoking; and (ii) non-small cell lung carcinoma (NSCLC) like squamous and adenocarcinoma tumors. About 10-15% of lung nodules are recognized as SCLC but 85-90% of patients with pulmonary cancer are affected by NSCLC [5].

Generating right information about lung nodules can be applied for indexing more number of medical images with enormous lung cancer databases. This can be employed for analyzing and training the proposed segmentors and classifiers using ML/DL methods. Eventually, this data may support physicians to diagnose and treat patients by demonstrating the effectiveness of applied treatments and courses with related features of the cancer. Earlier research works have demonstrated that deep networks can effectively increase the efficiency of segmentation algorithm, especially for isolating the anomalous regions from medical scans.

Traditionally, the radiologist performs nodule segmentation physically [6]. Although the physical nodule isolation realized by a medical professional is deemed the reference (i.e., gold standard) it is an arduous and very laborious task. Besides, it comprises of complex procedures, and the results are dependent on the expertise of radiologists. Furthermore, the decisions vary from one professional to another professional and they are not reproducible by the same radiologist. Thus, automatic segmentation and reproducible approaches are vital for managing and treating pulmonary nodules.

From a comprehensive survey, it is observed that DL approaches can inevitably isolate lung tumors from images precisely [7, 8]. However, these segmentors have some restrictions such as overfitting problems, morphological variations, poor accuracy, lower sensitivity, and higher computational cost. Overfitting befalls when the system fits too well into the training databases. It then becomes challenging for the segmentors to perform isolation from new image that were not in the learning set. Many DL methods are suffered from this issue when using high-dimensional sparse database. Designing segmentation models from high-dimensional and low-sample-size databases is becoming ever more vital. Therefore, a fast, cost-effective, and very sensitive DL-based segmentation model for isolating lung nodules is a key research domain. With more medical images being collected from treatments the queries arise of just how the care quality can be sustained or possibly even enhanced more. The processing time in isolating affected area from diagnostic images is a continuous challenge in the domain of clinical image processing.

Deep networks, especially Convolutional Neural Networks (CNNs), has drawn huge attention from several researchers and is effectively used to segment Region of Interest (RoI) from lung images with improved performance. The convolution networks usually get an RGB (red, green, and blue) image as an input and perform a sequence of convolutional, regularization, and pooling functions. The convolution unit defines the inherent correlation among features (e.g., size, shape, and edge statistics) in the input image. CNNs encode biases, including translation equivariance and spatial relationships. These features support in creating standard and competent analytical models. Conversely, the local receptive module in a CNN confines the range of the distant relations in a medical scan.

The convolution operation is content-free since the kernel biases are assigned with the identical weights used for all inputs, irrespective of their form. Since CNNs evolve, their depth also increases, and accordingly, the problem of gradient burst also increased. Hence, integrating extra layers leads to complex training inaccuracies. The CNN models often fail to acquire morphological and edge information from CT images. To resolve this problem of insufficient receptor areas and to simulate the complete representation, a self-attention method was proposed [9]. The attention-based models include an image-grid-based gating module that contains distinguished skip connections to permit signals to navigate and gather the localized data gradient from the encoding unit before it integrates with features of the decoding unit. This approach allows the model to control itself to a particular object segmentation procedure. Moreover, CNNs typically based on the learning process using huge datasets. Hence, an effective and reliable model for isolating lung tumours from medical scans is indispensable, mostly in the initial stage of the cancer.

At present, transformer-based models have infiltrated into the domain of healthcare diagnostic imaging, where the self-attention module is used as a supernumerary to the representative convolutional function to define reserved relations in an image. Presently, the Vision Transformer (ViT) network improves the performance of the attention mechanism using convolutional and recurrence functions. ViT employs a distinct configuration to get high-resolution data from features and coded global correlation from encoder unit [10]. Though ViT provides more accurate outcomes, most of the ViT-based segmentors hampered by the layered architecture of CNNs. Being inspired by the vast utilization of ViT in several clinical image processing applications, this research proposes a ViT-based lung nodule segmentor to circumvent the shortcomings of conventional convolutional networks.

This work targets to enhance the amalgamation mechanism used by the rudimentary ViT. It uses attention weights computed in the encoding layer to show up valuable tokens. The proposed OSWT model with MSA unit employs a cross-contextual attention method to re-calculate the set of features. In contrast to the model proposed [11], which uses a transformer to produce the attention, our OSWT model uses the previously estimated attention vector from the encoder which does not include any additional computational and storage costs. Moreover, our model utilizes the attention approach in the scale of the encoding/decoding unit to form a multi-resolution feature vector. The main contributions of this research are given below:

1) This work develop an effective segmentor using an OSWT for segmenting the pulmonary nodules from CT images. The application of the OSWT with MSA unit upsurges the capacity of the system to excerpt features associated with size, shape, edge, and flat region of tumors from medical images.
2) This work proposes a low-frequency extraction unit with a MSA unit to compute the distant reliance feature from the clinical scan.
3) To process distant reliance scans, we employed an OSWT to overcome the restriction of dividing the input image into patches with constant size in traditional ViT model.
4) This study assesses the performance of the proposed OSWT with MSA unit in isolating lung tumours from the TCGA database.

The other sections of this article are organized as follows. This study analyses some relevant studies on lung cancer detection models in Section II. Section III explains the architecture and workings of the OSWT transformer in detail. Section IV discusses the application of the OSWT in the isolation of lung lesions. Section V and Section VI describe the experimental analysis and performance measures used in our study. Section VII summarize this research.

2. LITERATURE SURVEY

The revolution of DL approaches has revitalized the field of medical data analysis, forming a base for radical improvements and novel insights. Several deep networks have been developed in the literature to isolate diseased or cancerous cells from the histopathological scans. In this section, we have analyzed some state-of-the-art deep networks for addressing lung cancer segmentation problems from CT scans. Badrinarayanan et al. [12] developed a novel deep CNN model for semantic pixel-wise segmentation, known as SegNet. The proposed deep network comprises an encoder, a decoder module, and a pixel-wise classifier. The decoding module relates the feature map of the encoding module with lower resolution to the feature map of input with higher resolution to perform pixel-wise analysis. The decoding module employs pooling factors measured in the max-pooling layer of the encoding module to realize non-linear unpooling. This removes the training inevitability of unpool. The unpooled vectors are not dense. Hence, this model performs convolution operations with learnable kernels to create dense attribute vectors. Alam et al. [13] proposed a semantic segmentation CNN (SegNet) with a conventional computer vision method, called level sets. The level set provides more precise isolation results but it is sensitive to parameter initialization, which is often executed manually.

A new improved convolutional model is proposed by Chen et al. [14], called 3D LungNet for lung cancer isolation. This model uses the 3D information present in the large volume of image. Initially, a binary classifier chooses image parts that may contain shares of a lump. To segment the nodules, the selected scans are given to the isolation model which selects feature matrix from each 2D image using dilated convolutions and then integrates the pooled maps through 3D convolution functions and combines the 3D structural attributes in the CT scan into the output. Li et al. [15] proposed a multi-view convolution model to isolate pulmonary tumors by applying coronal, axial, and sagittal information about any voxel of the malignant tumor.

U-Net is a commonly used convolution network for segmenting pulmonary cancers [16]. U-Net structure encompasses two paths; a symmetric increasing path to realize accurate localization and a decreasing path to collect context. The decreasing path includes consecutive convolutional and up sampling modules. It is used to choose attributes by restraining the dimension of the attribute map. The increasing path comprises a convolutional module and completes up-conversion to find the size of the feature vector associated with the loss of structural attributes. Moreover, the positional information is improved from the decreasing path to the increasing path using dropout links. These links are worked independently and allow data to be transferred from one system unit to another without adding further computational cost. Nair and Hinton [17] developed an improved U-Net in which the encoder is replaced by a pre-trained ResNet-34 unit. This model employs a bidirectional convolution of long short-term memory to combine the designated feature map of the corresponding decreasing path into the previous intensification of the up-convolutional unit. Now, a densely connected convolutional unit is utilized for the decreasing path.

Huang et al. [18] proposed a 3D U-Net and contextual CNN to segment and classify lung malignancies automatically and help radiologists understand CT images. The skip connections in classical U-Net distort the clinical image attributes. The high-level features designated by this network frequently do not include adequate high-resolution edge data of the clinical scan, instigating higher indecision in which higher resolution boundaries disrupt the system results considerably. To tackle these problems, Szegedy et al. [19] introduced a Modified U-Net (mU-Net) for isolating pulmonary malignancy from CT images. This network uses a residual module with deconvolution and activation tasks by applying drop out connections. This technique handles the problems in basic U-Net related to features with low-resolution structures.

Ioffe et al. [20] developed a novel 3D DL network for pulmonary cancer segmentation from CT images, named multiple-attention U-Net (MAU-Net). This network initially exploits a twofold attention module at the restricted access of the U-Net that describes the key correlation between spatial dimensions and layers. The multiple attention module is then employed to adaptively re-calculate and syndicate multi-scaling features from the twofold attention modules, the preceding feature matrix of the decoding unit, and the equivalent attributes from the encoder. ResNet uses diverse residual convolution units to effectively excerpt the important features of the CT scans. All attributes from different layers of the ResNet were pooled into a distinct result. This model achieved an amalgamation of superficial features with high-level semantic attributes to generate dense outputs.

Szegedy et al. [21] proposed a transformer-based segmentation model using U-Net, called TransUNet. This research demonstrated that ViT and its hybrid architectures provide improved results than CNN-based self-attention networks. This network uses a fused CNN-Transformer model to derive comprehensive high-resolution dimensional data from attributes. Szegedy et al. [22] modified the structure of TransUNet by assimilating the attention technique into the Transformer dropout links for the skin lesion isolation task. Though TransUNet presents better outputs, this approach suffers from being reliant on CNNs layered attribute selection. To cope with this issue, Zhu et al. [23] designed a new network model using only a transformer called the Swin U-Net model. This network exploits the concept of a Swin transformer to construct the U-Net structure without any convolution module. Several aforementioned deep networks have met their target competently. Conversely, their isolation performance with respect to dice similarity index, volume error, Jaccard similarity measure, and average time for isolation is frequently not superior. Bearing the above issues in mind, this work aims to develop a new effective lung nodule segmentation model using a Vision Transformer. We develop an OSWT with MAM to segment the diseased or abnormal area from lung CT scans accurately.

3. THE PROPOSED OSWT MODEL

Attention in image processing is either realized with CNN or used to adapt some layers of CNN while conserving its general structure. On the other hand, some studies demonstrate that the application of convolutional networks is not mandatory and a transformer hoard image patches directly. The transformer models attain promising results in image classification [24]. This model contains the encoding/decoding units which hoards multiple image blocks concomitantly without demanding any recurrent model. This type of parallel computing is unfeasible in traditional convolutional networks. The implementation of the

transformer has primarily employed the notion of self-attention to control remote relations among the image patches.

In this work, the ViT is recommended as an effort to upturn the utilization of the original transformer for vision. By optimizing a ViT model, we can attain better enactment on a medical image database. It outdoes the basic convolutional network-based classification algorithm by around 4 × regarding computational performance and accurateness. In recent times, ViT has proposed as a feasible supernumerary to convolutional networks that typically use local receptor areas with suitable kernels, the attention module in ViT allows it to consider every pixel of the input scan and assimilate data across the whole picture. ViT employs the transformer encoding module to perform disease identification by relating a patch of scan stacks to the corresponding tag through the attention mechanism as shown in Figure 1. An attention unit provides relationships between a stack of inputs and outputs to be modeled regardless of the distance between them. It has become a most important part of persuasive stack modeling and transformation models for different tasks, and self-attention is an attention mechanism that investigates the statistics of a single stack by mapping different RoI in the image stack.

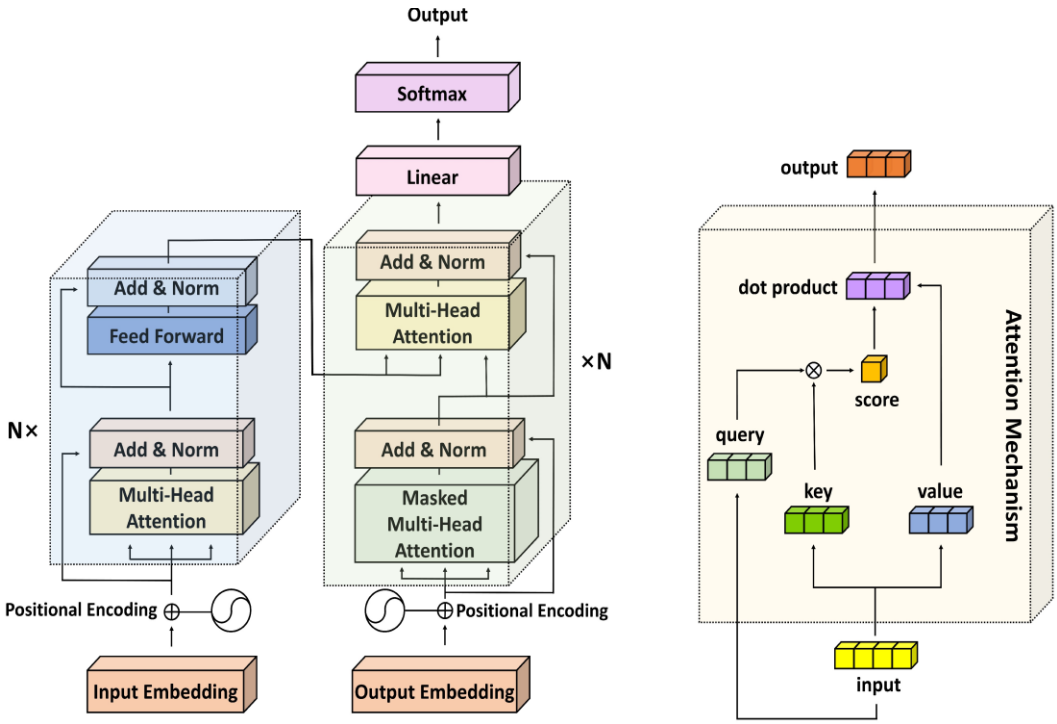

Figure 1. Architecture of ViT with attention module

Consider $K = \{M_i, t_i\}_{i=1}^{s}$ is s set of CT scans, in which M_i is an individual scan and t_i represents its equivalent tag $t_i \in \{1, 2\}$. Mostly, the rudimentary ViT design encompasses three layers such as a linear embedding layer, an encoding layer, and a head classifier. Primarily, an input image M is split into a set of non-overlapping patches. Let a medical image M with the size of $w \times d \times c$, in which w is the width, d is the depth, and c is the number of channels in the image. To process a 2D image, ViT divides each image into patches of length and width β (i.e., the scan is converted into square patches). Consequently, we obtain sub-image patches m_i with $\beta \times \beta \times c$ size from these pictures. It converts the medical scan $M \in \mathbb{R}^{d \times w \times c}$ into a flattened stack of 2D patches $m \in \mathbb{R}^{m \times \beta^2 c}$, where, $(d \times w)$ is the resolution of the input image, $(\beta \times \beta)$ is the resolution of each image block, and $n = \frac{wd}{\beta^2}$ is the number of blocks in the input image. This creates a sequence of patches $(m_1, m_2, m_3, \ldots m_n)$ of length n. Generally, the block size β is selected as 16×16 or 32×32 in which a reduced dimension of patch causes an extended block and vice versa. The transformer processes each patch as a separate token. Thus, input pictures are managed as a stack of sub-image blocks in which each block is flattened into a single vector by integrating the channels of all the pixels in a patch and then linearly projecting it to the selected input size. ViT transforms the flattened image blocks into s size through a learnable linear projection unit as discussed in the following section.

3.1 Embedding layer

The heap of sub-image blocks are directly related to a matrix of the size s using a trained embedding vector a. These embedding (i.e., data representation technique in n-dimensional space to cluster similar data points together) representations are then combined with a trainable class token. These tokens are very important in this study to perform the disease classification. Then, the transformer considers the embedded patches as a heap regardless of their sequence. To preserve the spatial organization of the patches as in the input scan, the location data a_{PI} is calculated and attached to the patch. The output-embedded image blocks with the token T_0 is defined by Eq. (1):

$$T_0 = [T_c; x_1 a; x_2 a; \ldots . x_n a] + a_{PI}, a \in \mathbb{R}^{\beta^2 c \times s}, a_{PI} \in \mathbb{R}^{(n+1) \times s} \quad (1)$$

The position (location) of an object in a heap of image patches is calculated by the positional encoder to facilitate each position is given a unique depiction. From a comprehensive survey, it is witnessed that 1D and 2D

positional encodings produce almost the same outputs [25]. Therefore, a simple 1D encoding module is used in this work to collect the location information of the flattened image patches.

3.2 Encoding layer

The output heap of embedded image patches T_0 is transferred to the dynamic encoding unit of ViT. The main architecture of the encoder in a transformer encompasses L identical layers as shown in Figure 2. Each layer contains two core components: (i) a MSA unit for generating attention maps from a specific embedded visual token. This process enables the model to emphasis on the most vital zones in the image (e.g., cancer cells); and (ii) a multi-layer perceptron (MLP) unit which is a classifier and encompasses two dense layers with a GeLU (Gaussian Error Linear Unit) activation unit. The MSA unit is the most important component in this layer. The sovereign attention outcomes are then combined and linearly transformed into the projected size. This study employs 12 MSA unit modules in the encoder unit. The encoding module implements residual dropout links and is controlled by a normalization module. The layer norm (LN) keeps the training procedure on target and allows the model to adjust the differences in the training database samples. The mathematical functions of MSA unit and MLP are defined by Eqs. (2) and (3).

$$T_l' = MAU(LN(T_l - 1) + T_l - 1, \quad l = 1,2 \dots L) \tag{2}$$

$$T_l = MLP(LN(T_l') + T_l', \quad l = 1,2 \dots L \tag{3}$$

In the final encoding layer, we accept the first part in the heap T_L^0 and transfer it to an exterior MLP head classification unit for calculating the tag t as defined in Eq. (4).

$$t = LN(T_L^0) \tag{4}$$

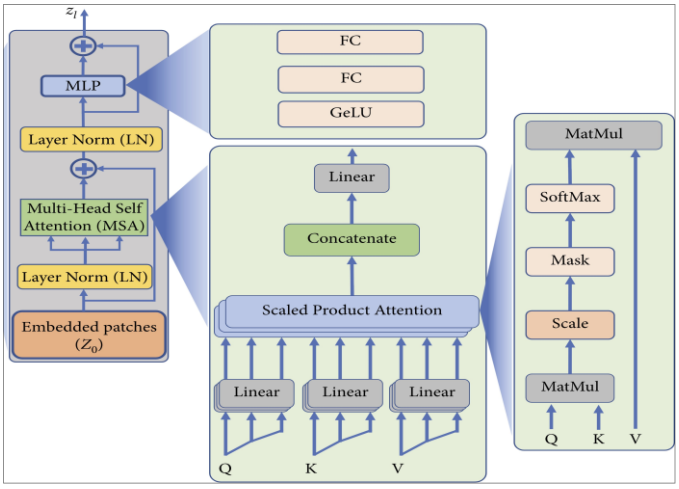

Figure 2. Encoding module in ViT

3.3 Multiple-head self-attention unit

The multi-head attention allows the model to extract local and global relationship in the input scan. It calculates the relative importance of an embedding of a block as compared to the other embeddings in the heap (i.e., determines the image blocks with maximum and minimum significance level) and excludes the input image stack with minimum significance level. This module contains 4 layers such as the linear layer to match the expected output size, the scaled dot-product attention layer where the dot products of queries, keys, and values are scaled down, the concatenation layer to connect a trainable (class) embedding with the other block predictions, and a final linear layer to obtain a linear block projection.

In general, attention is defined by its weight at a high level and calculated from the weighted sum of entire values of the patch order. The self-attention module computes the weights of attention through calculating and scaling down the dot-product of the query (Q), key (K), and values (V). Figure 3 illustrates the details of the calculation that is performed by the self-attention unit. For each pixel in the heap, Q, K, and V are calculated by multiplying the pixel against the trained vector χ_{QKV} as given in Eq. (5).

$$[Q, K, V] = E\chi_{QKV}, \quad \chi_{QKV} \in \mathbb{R}^{s \times 3S_{key}} \tag{5}$$

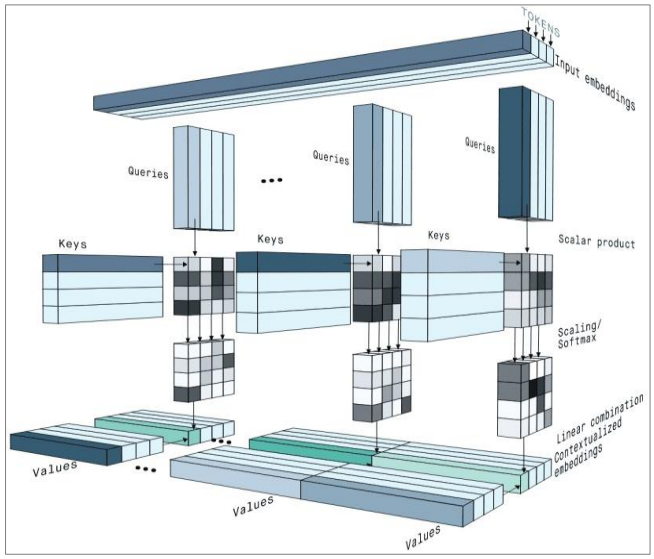

Figure 3. Illustration of attention mechanism

In order to determine the relative significance of a specific pixel as compared to other pixels in the picture sequences, the dot-product is computed between the query vectors of this pixel with the key of other pixels. The output defines the comparative significance of blocks in the stack. The outcome of the dot-product is then scaled down and given to a SoftMax unit. The softmax function is defined by Eq. (6).

$$A = softmax\left(\frac{QK^T}{\sqrt{S_{key}}}\right), \quad A \in \mathbb{R}^{n \times n} \tag{6}$$

Similar to usual dot product, the self-attention module performs the scaling function. However, it employs the size of the key (S_{key}) as a scaling element. To end, the value of every matrix of sub-image patch embeddings is multiplied by the result of the softmax layer to find out the area with the maximum attention values as defined by in Eq. (6). The comprehensive self-attention (η) function is defined by Eq. (7).

$$\eta(E) = A.V \tag{7}$$

MSAU calculates the scaled attention for h heads by

applying a range of values for Q, K, and V instead of applying a specific value. The result of each head is integrated and calculated using a feed-forward function with trainable weights w to the selected size. This function is defined by Eq. (8).

$$MSAU(T) = Concat\big(\eta_1(T);\ \eta_2(T);\ \ldots.\eta_h(T)\big)w, w \in \mathbb{R}^{hS_{key} \times S} \quad (8)$$

Inspired by the present encroachments of the Swin U-Net model [26], this study proposes an orchestrated Swin transformer model for RoI isolation in lung nodule classification. The intended model provides a twofold attention mechanism where the initial fold integrates the attention weight calculated from the encoding units to reveal the substantial tokens based on spatial attention. But the successive attention unit considers pair-wise tokens for updating significant features of the medical image.

3.4 OSWT model

In this research, we propose the orchestrated and shifted window approach to integrate the data with diverse scales. It is used as the backbone system to extract the features from the medical scan. Shifted window transformers are used as the filters in the transformer-based models that successfully implement the attention mechanism. It creates hierarchical feature vectors by integrating image blocks in deeper layers. The computational cost of this model is directly proportional to the size of the image. This is because of the calculation of self-attention only within every native window. Hence, it provides a flexible structure for identification and diagnosis of medical images. Conversely, the basic transformers produce a specific low resolution feature vectors and have quadratic computational cost to the input size due to the computation of global attention. Swin modules are often organized in a heap to excerpt more radical and deeper features. Within a Swin module, a shifted window is employed to calculate both global and local self-attention. The shifted windows are non-overlapping windows that divide the input scans into blocks. To reduce the quadratic computational cost in computing attention, two serial Swin units can realize attention mechanism with reduced computational cost.

Conventional self-attention mechanisms require determining correlations among all pixels, resulting in high processing cost. However, the Swin Transformer introduces a Window-based Self Multi-head Attention (W-SMA) mechanism shown in Figure 4. Initially, it splits the input scan into fixed-size blocks and then employs attention mechanism separately to every block, considerably minimizing processing overhead. When applying the W-SMA mechanism, despite the drop in processing overhead realized using the split function, attention calculation remains limited within individual windows, which thwart the data communication between various windows. To handle this problem, the Swin Transformer employs the Shifted Window Self Multi-Head Attention (SW-SMA) mechanism. This mechanism aims to accurately define both global and local attributes, different from the traditional MSA model normally used in classic ViT models. The standard Transformer structure for vision applications use a global attention approach that involves identifying correlations between a token and all other tokens. This global attention generates quadratic cost in terms of the number of tokens, making it inappropriate for several applications that need huge amount of tokens for computation or for presenting high-dimensional images.

The key purpose of the shifted window is to perform self-attention within localized windows. Each window consists of non-overlapping blocks, and self-attention is performed within this window. Consequently, there is a drop in processing overhead; while the original multi-head self-attention shows quadratic cost about the block number, the window-based MSA establishes linear overhead. The Swin Transformer assimilates a shifted window splitting approach, alternating between two structures across successive blocks to effectively model window connections. The early unit uses a standard window structure, allowing for local self-attention calculation from equally spaced windows, starting from the top-left pixel. Then, the successive Swin Transformer module implements a shifted window mechanism. This intended shift enables the model to calculate different spatial correlations efficiently.

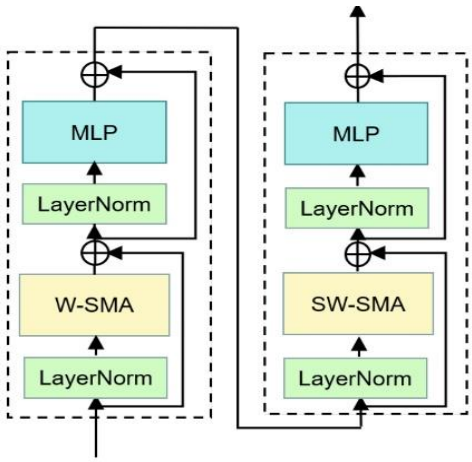

Figure 4. Structure of Swin block

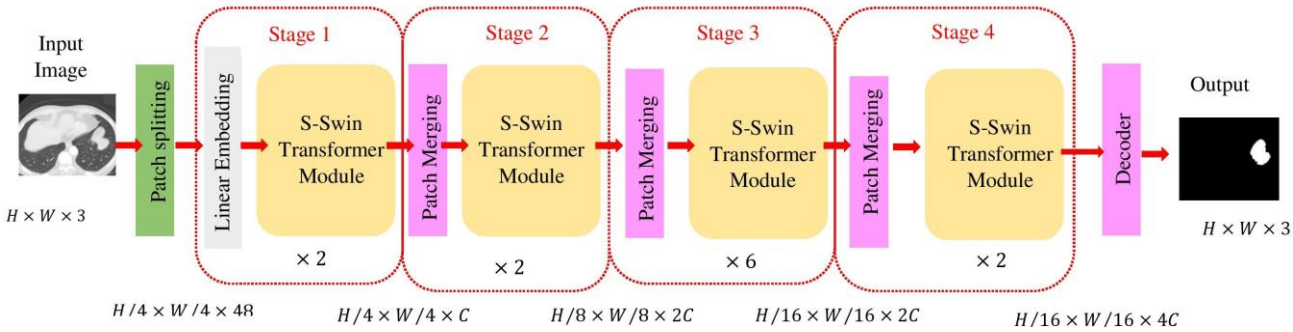

Figure 5. Architecture of the OSWT and image-merging process

The intended OSWT comprises of two sequential transformers. Conversely, all the MSA units are replaced by a window-based attention module. The first unit encompasses a window-based MSA unit that estimates the attention within the window, and the subsequent unit encompasses a multi-attention unit using shifted-window. This unit calculates intra attention through the windows by interchanging two splitting configurations in following Swin transformer units as given in Figure 5. Consequently, two sequential Swin Transformer modules can estimate the intra attention concurrently across the entire scan, and consumes reduced time for processing. It increases the efficiency of the model in multi-scale nodule segmentation without including more hyperparameters. OSWT splits the CT scan into many windows and computes the relationship among the features within the window using the multi-attention unit, which not only increases the receptor region of the shallow network but also ensures the isolation ability of very small nodules in the lung image.

4. OSWT-BASED SEGMENTATION MODEL

Isolation of lung tumor on CT scans is particularly very important task for cancer disease management processes such as diagnosis, treatment, and response evaluation. In this work, we develop a completely automated OSWT lung tumor isolation method that can handle a large variety of CT scans. We assess our segmentation model using the OSWT on the TCGA dataset by relating its performance with similar advanced segmentation models regarding VE, DSI, JSM, and speed. Since the attributes directly extracted from CT images presented different gray scales and strengths, it is indispensable to apply a preprocessing technique before such images are given to the isolation model. We use normalization and standardization methods for preprocessing. This study adopts a min-max strategy for normalization using Eq. (9).

$$\bar{\mu} = \frac{\mu - \mu_{min}}{\mu_{max} - \mu_{min}} \tag{9}$$

where, $\bar{\mu}$ are normalized features retrieved from feature space; the terms μ and μ_{min} are minimum feature values; and μ_{max} is maximum value of feature. The CT images are preprocessed in this work by excluding noise and artifacts in the input images. Preprocessing comprises the following steps: (i) the input images are of different dimensions and intensity. Therefore, all the images have been converted into a constant dimension of 224×224 before implementing isolation technique; (ii) A filter with values ([-1, 0, -1], [0, 5, 0], [-1, 0, -1]) is applied for finding edges of the nodules [27]; (iii) the values of each pixel are calculated by converting the red-green-blue (RGB) color to the luma and chroma (YUV) color space. Luminance is more authoritative than color for isolation. Therefore, the resolution of V (red projection) and U (blue projection) are reduced but Y is conserved at full resolution (iv) then, the intensity values of each pixel are stabilized by converting the YUV color space to RGB color by smoothing edges and balancing histogram.

5. IMPLEMENTATION OF PROPOSED MODEL

The proposed OSWT model is implemented and its performance is analyzed through experimentation. A comprehensive empirical analysis is performed on an Intel Core i7-4790 CPU with 3.6GHz base speed, 16GB storage, and Windows 10 operating system. The performance of the OSWT approach is evaluated by comparing the empirical outcomes with 7 relevant segmentation models, viz. SegNet [13], 3D LungNet [14], U-Net [16], mU-Net [19], MAU-Net [20], TransUNet [21], Swin U-Net [23]. All of these networks including our OSWT use DL algorithms for isolating the lung nodule and are trained through the similar learning setup. The SegNet network structure solve the semantic segmentation problem of lung nodule. It contains an encoder-decoder module, and a pixel-wise classifier. The decoding module relates the feature map of the encoding module with lower resolution to the feature map of input with higher resolution to perform pixel-wise analysis. The 3D LungNet uses the 3D information present in the large volume of image. To segment the nodules, the selected scans are given to the isolation model which selects feature matrix from each 2D image using dilated convolutions and then integrates the pooled maps through 3D convolution functions and combines the 3D structural attributes in the CT scan into the output.

U-Net structure encompasses a symmetric increasing path to realize accurate localization and a decreasing path to collect context. Moreover, the positional information is improved from the decreasing path to the increasing path using dropout links. These links are worked independently and allow data to be transferred from one system unit to another without adding further computational cost. The 3D U-Net segments and classifies lung malignancies automatically. The skip connections in classical U-Net distort the clinical image attributes. The high-level features designated by this network frequently do not include adequate high-resolution edge data of the clinical scan, instigating higher indecision in which higher resolution boundaries disrupt the system results considerably. To tackle these problems, mU-Net uses a residual module with deconvolution and activation tasks by applying drop out connections. This technique handles the problems in basic U-Net related to features with low-resolution structures. MAU-Net network initially exploits a twofold attention module at the restricted access of the U-Net that describes the key correlation between spatial dimensions and layers. The multiple attention module is then employed to adaptively re-calculate and syndicate multi-scaling features from the twofold attention modules, the preceding feature matrix of the decoding unit, and the equivalent attributes from the encoder. TransUNet uses a fused CNN-Transformer model to derive comprehensive high-resolution dimensional data from attributes. Though TransUNet presents better outputs, this approach suffers from being reliant on CNNs layered attribute selection. To cope with this issue, Swin U-Net model exploits the concept of a Swin transformer to construct the U-Net structure without any convolution module. Several aforementioned deep networks have met their target competently.

5.1 Dataset preparation

To evaluate the performance of any DL algorithm, we need a large dataset that provides an improved solution. In this work, we employ numerous labeled lung CT images from the TCGA dataset. This dataset was collected from the National Cancer Institute Lung Cohort Consortium [25]. The collected information is associated with proteomic, genomic, and medical scans. The assembled images are stored in the form of

DICOM (digital imaging and communications in medicine) with certain tags such as gender, birth date, study dates, etc. This database comprises of 251,135 de-identified CT images from lung malignancy patients. Radiologist annotations on the cancer localities were also given for each scan. 5 academic thoracic radiologists performed the annotations: the bounding box was drawn by one radiologist and then verified by the other four. For our analysis, we only considered CT scans with a resolution of 1 mm. CT images with resolutions other than 1 mm were omitted for the analysis. We made this choice since CT scans of diverse intervals may present changes in the Radiomics attributes that confuse the understanding of the results. A resolution of 1 mm is the most generally assimilated slice thickness in hospitals, and such CT scans were represented in this database efficiently. Hence, considering 1 mm thick CT scans was the most appropriate selection for upcoming medical utilization.

In this research, we use 5043 input scans for the experimentation and 70% of samples (i.e., 3530 scans) is used for learning and 30% (i.e., 1513 scans) of samples for testing. Figure 6 displays some example TCIA images. Vision Transformers are prone to overfitting, especially when the target dataset is small or different from the pre-training dataset. To prevent overfitting problem, we employed 10-fold cross-validation (10-FCV). In this approach, the complete database is divided into ten parts (each of 10% of the entire dataset). Then, one part (10%) is used for validation, whereas the remaining data (90% of the entire dataset) are employed for testing and training. The application of 10-FCV guarantees that all the images in the database get to be in a trial just once.

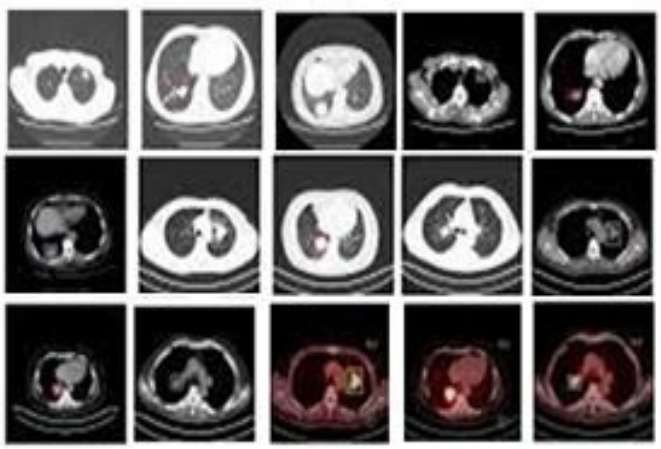

Figure 6. Sample lung cancer CT images

5.2 Evaluation metrics

The effectiveness of the OSWT network is numerically evaluated by some significant performance measures such as the VE, DSI, JSM, and average computational time. These measures are estimated by computing the difference between the isolation results and a physically labelled reference. VE is defined by Eq. (10).

$$VE = \frac{2 \times (S - G)}{(S + G)} \tag{10}$$

where, G is the gold standard image (i.e., ground truth) and S is the isolation output gained by the OSWT. For any healthcare application, $VE < 5\%$ is more likely acceptable [26]. DSI is normally used to calculate the performance of the isolation task. It is defined as a similarity measure between two picture elements. Also, it reflects the fitness level between the input image and the isolated image. The DSC value is always in [0, 1] and it is calculated by Eq. (11).

$$\text{DSI} = \frac{2 \times |G \cap S|}{|G| + |S|} \tag{11}$$

JSM is a performance indicator used to evaluate the efficiency of any segmentation model. For a particular database, the JSM presents the resemblance between the output image and the reference image. It is calculated using Eq. (12).

$$\text{JSM} = \frac{|G \cap S|}{|G \cup S|} \tag{12}$$

This work also takes the average computational time into account as the evaluation metric.

6. RESULT ANALYSIS

The proposed OSWT segmentation model is realized through MATLAB R2018b/deep learning toolbox software package. Table 1 shows the complete results achieved from the OSWT isolation network. The 10-FCV method is employed to achieve superior results. Therefore, the entire dataset is split into ten fragments. For every test, one fragment is used for evaluation, and the other fragments are employed for learning the model. Then, the mean value of all ten trials is calculated for assessment. An inclusive investigation of our outcomes reveals the fortes and weaknesses of our OSWT segmentation model. In most cases, despite the size of the reference image, our model isolates the nodules very well. Figure 7 displays sample input scans employed for testing, isolated RoI achieved by the OSWT model, and their equivalent gold standard. Although the cancer-affected regions are in different arbitrary vicinities within the lung image and appear in numerous dimensions, the isolated RoI appear to overlap impeccably.

Table 1. The segmentation results gained by OSWT

Algorithm	Criteria	VE (%)	DSI	JSM	Average Processing Time (s)
SegNet [13]	Mean	4.821	0.649	0.680	3.422
	SD	1.703	0.152	0.113	0.002
3D LungNet [14]	Mean	3.634	0.778	0.653	3.052
	SD	0.041	0.014	0.012	0.001
U-Net [16]	Mean	2.302	0.835	0.704	3.128
	SD	0.058	0.177	0.011	0.001
mU-Net [19]	Mean	1.334	0.845	0.750	3.146
	SD	0.049	0.181	0.008	0.002
MAU-Net [20]	Mean	1.134	0.857	0.796	1.187
	SD	0.015	0.132	0.010	0.003
TransUNet [21]	Mean	1.332	0.943	0.806	1.156
	SD	0.016	0.014	0.009	0.004
Swin U-Net [23]	Mean	0.952	0.961	0.847	1.035
	SD	0.023	0.011	0.011	0.005
OSWT	Mean	0.733	0.984	0.965	0.992
	SD	0.003	0.009	0.009	0.008

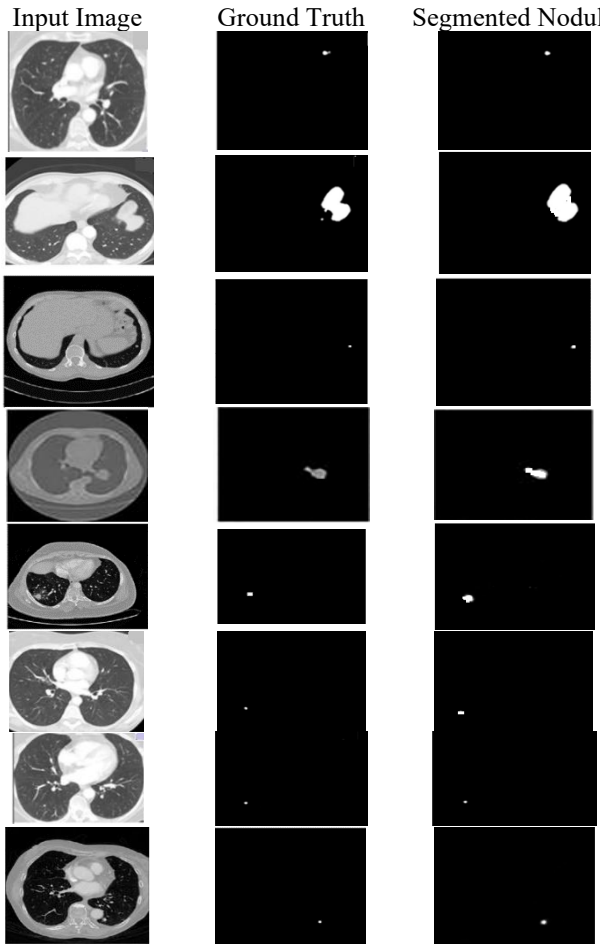

Figure 7. Segmentation results

The segmentation results obtained from different isolation approaches with respect to average and standard deviation (SD) are listed in Table 1. By applying max-pooling coefficients of the feature matrix and decoder network, the conventional SegNet model provides nominal segmentation results such as 4.821±1.703% volume error, 0.649±0.125 of DSI, 0.680±0.113 of JSM and 3.422±0.002 of average processing time. This model utilizes the 3D information existing in a CT image efficiently. Hence, this network can gain better segmentation fallouts than SegNet. This network achieves 3.634±0.041% of VE, 0.778±0.014 of DSI, and 0.653±0.012 of JSM. The average running time per trial of the 3D LungNet is 3.052±0.001s.

The U-Net model achieves better results as compared to SegNet and 3D LungNet using the idea of global location and contextual information concurrently. Besides, it ensures the conservation of the whole texture of the input images. Thus, it realizes comparatively reduced VE (2.302±0.058%), higher DSI (0.835±0.177), and improved JSM (0.704±0.011). Besides, it consumes more time per case (3.128±0.001s). On the other hand, higher-level features designated by this network do not include adequate resolution to edge information of the input always, causing augmented uncertainty in which edges with higher-resolution mostly affect the results of pulmonary tumor isolation and classification.

The modified U-Net (mU-Net) includes a residual path with deconvolution and triggering operations to the skip link of the U-Net to circumvent the repetition of low-resolution statistics of features. For minor object scans, attributes in the drop-out link are not assimilated with features in the residual module. Besides, the proposed network has additional convolution units in the dropout link to choose higher-level features of small object in scans and high-resolution features of high-resolution edge information of huge objects. Therefore, this model delivers improved enactment for lung noodle isolation regarding VE (1.334±0.049%), DSI (0.845±0.181), and JSM (0.750±0.008). For efficient segmentation of anticipated RoI, it takes 3.146 ± 0.002s for every slice.

Using a 3D encoding/decoding modules in CNN architecture, MAU-Net achieves better results in segmenting RoI from volumetric CT scans. This network gains 1.134± 0.015% VE, 0.857±0.132 DSI, 0.796±0.010 JSM, and 1.187±0.003s computational time. By implementing the idea of ViT with attention mechanism TransUNet provides 1.332±0.016% of VE, 0.943±0.014 of DSI, 0.806±0.009 of JSM, and 1.156 ±0.004s of average computational time. The Swin U-Net deep network exploits the Swin Transformer modules with the U-Net structure without using any convolution module. Hence it provides better isolation performance with respect to VE (0.952±0.023%), DSI (0.961±0.011), JSM (0.847±0.011), and average isolation time (1.035±0.005s).

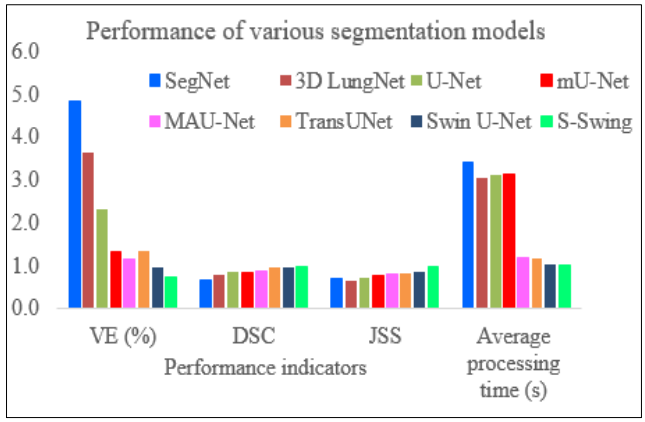

Figure 8. The results obtained from different segmentation networks regarding mean values

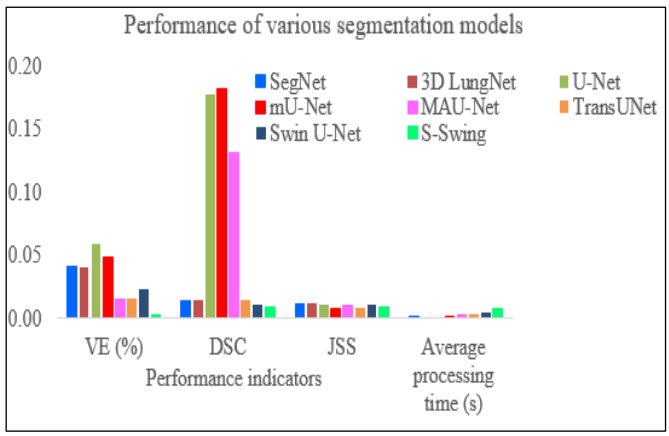

Figure 9. The results obtained from different segmentation networks regarding SD values

The proposed OSWT outdoes all other approaches regarding the evaluation measures. We also emphasize the concept that the attention principle of OSWT ensures an efficient yet less processing cost as compared with other RoI

isolation approaches. It realizes results of 0.733±0.003%, 0.984±0.009, 0.965±0.009, and 0.992±0.008s in VE, DSI, JSM, and average computational time, correspondingly. Figures 8 and 9 demonstrate the superiority of the proposed OSWT -based segmentation model in terms of mean and SD values of the evaluation metrics, correspondingly. The higher mean value and lower SD values reveal the dependability and sturdiness of the proposed model.

The proposed OSWT-based segmentation model delivers 98.4% dice similarity index (DSI), 96.5% of Jaccard similarity measure (JSM), 0.73% of volume error (VE), and 0.99s average computational cost. These results indicates the effectiveness of the proposed model. The reduced mean volume error (0.73%) indicates that it does not differ significantly from the expert's decision. The increased value of DSI (98.4%) indicates improved performance of the segmentation process. Also, it reflects the fitness level between the input image and the isolated image. This model provides 96.5% JSM which indicates the better performance of the segmentation model. For a particular database, the JSM presents the resemblance between the output image and the reference image. Moreover, this model is quite faster than existing segmentation models since it consumes 0.99s for realizing segmentation.

7. CONCLUSION

Lung cancer is the most detrimental type of malignancy. Designing an automated and reliable model to segment pulmonary nodules from a CT image is a very expedient tool in the medical industry. This study develops a novel OSWT network for isolating the affected regions from CT images. This network utilizes the concept of the simultaneous shifted window to assimilate the data with various scales and is used as the mainstay model to excerpt the attributes. Shifted window transformers are filters in the ViT-based models that effectively apply the attention mechanism. It creates hierarchical attribute vectors by combining image blocks in deeper modules and has direct computational cost to the dimension of scans owing to the processing of intra attention only depending on the local window. Thus, it acts as an adaptable configuration for both dense recognition and image classification endeavors. We evaluate OSWT on an open-access lung image database, known as the TCGA, and relate the performance of OSWT against seven advanced segmentation models regarding performance indicators. The proposed segmentation model using an OSWT transformer provides 98.4% dice similarity coefficient DSI, 96.5% of JSM, 0.73% of volume error (VE), and 0.99s average processing time. The extensive experiments prove that the OSWT model achieves a better performance and is appropriate for diseased area segmentation from CT scans.

REFERENCES

[1] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3): 209-249. https://doi.org/10.3322/caac.21660

[2] Llovet, J.M., Zucman-Rossi, J., Pikarsky, E., et al. (2021). Hepatocellular carcinoma. Nature Reviews Disease Primers, 2: 16018. https://doi.org/10.1038/nrdp.2016.18

[3] Suresh, D., Srinivas, A.N., Kumar, D.P. (2020). Etiology of hepatocellular carcinoma: Special focus on fatty liver disease. Frontiers in Oncology, 10: 601710. https://doi.org/10.3389/fonc.2020.601710

[4] Aatresh, A.A., Alabhya, K., Lal, S., Kini, J., Saxena, P.P. (2021). LiverNet: Efficient and robust deep learning model for automatic diagnosis of sub-types of liver hepatocellular carcinoma cancer from H&E stained liver histopathology images. International Journal of Computer Assisted Radiology and Surgery, 16: 1549-1563. https://doi.org/10.1007/s11548-021-02410-4

[5] Yao, C., Jin, S., Liu, M., Ban, X. (2022). Dense residual transformer for image denoising. Electronics, 11(3): 418. https://doi.org/10.3390/electronics11030418

[6] Simona Răboacă, M., Dumitrescu, C., Filote, C., Manta, I. (2020). A new adaptive spatial filtering method in the wavelet domain for medical images. Applied Sciences, 10(16): 5693. https://doi.org/10.3390/app10165693

[7] Hoque, M.Z., Keskinarkaus, A., Nyberg, P., Seppänen, T. (2021). Retinex model-based stain normalization technique for whole slide image analysis. Computerized Medical Imaging and Graphics, 90: 101901. https://doi.org/10.1016/j.compmedimag.2021.101901

[8] Zhou, S., Li, X. (2020). Feature engineering vs. deep learning for paper section identification: Toward applications in Chinese medical literature. Information Processing & Management, 57(3): 102206. https://doi.org/10.1016/j.ipm.2020.102206

[9] Gu, D., Su, K., Zhao, H. (2020). A case-based ensemble learning system for explainable breast cancer recurrence prediction. Artificial Intelligence in Medicine, 107: 101858. https://doi.org/10.1016/j.artmed.2020.101858

[10] Jussupow, E., Spohrer, K., Heinzl, A., Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. Information Systems Research, 32(3): 713-735. https://doi.org/10.1287/isre.2020.0980

[11] Chai, Y., Bian, Y., Liu, H., Li, J., Xu, J. (2021). Glaucoma diagnosis in the Chinese context: An uncertainty information-centric Bayesian deep learning model. Information Processing & Management, 58(2): 102454. https://doi.org/10.1016/j.ipm.2020.102454

[12] Badrinarayanan, V., Kendall, A., Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12): 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[13] Alam, M., Wang, J.F., Cong, G.P., Lv, Y.R., Chen, Y.F. (2021). Convolutional neural network for the semantic segmentation of remote sensing images. Mobile Networks and Applications, 26: 200-215. https://doi.org/10.1007/s11036-020-01703-3

[14] Chen, X., Duan, Q., Wu, R., Yang, Z. (2021). Segmentation of lung computed tomography images based on SegNet in the diagnosis of lung cancer. Journal of Radiation Research and Applied Sciences, 14(1): 396-403. https://doi.org/10.1080/16878507.2021.1981753

[15] Li, D.D., Chi, Z.Q., Wang, B.L., Wang, Z., Yang, H., Du, W.L. (2021). Entropy-based hybrid sampling ensemble learning for imbalanced data. International Journal of

Intelligent Systems, 36(7): 3039-3067. https://doi.org/10.1002/int.22388

[16] LeCun, Y., Bottou, L., Orr, G.B., Müller, K.R. (2002). Efficient backprop. In Neural networks: Tricks of the trade. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 9-50. https://doi.org/10.1007/3-540-49430-8_2

[17] Nair, V., Hinton, G.E. (2010). Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807-814.

[18] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, pp. 4700-4708. https://doi.org/10.1109/CVPR.2017.243

[19] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, pp. 1-9. https://doi.org/10.1109/CVPR.2015.7298594

[20] Ioffe, S., Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167. https://doi.org/10.48550/arXiv.1502.03167

[21] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[22] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. (2017). Inception-v4, inception - ResNet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 31(1): 4278-4284. https://doi.org/10.1609/aaai.v31i1.11231

[23] Zhu, C., Chai, X., Xiao, Y., Liu, X., Zhang, R., Yang, Z., Wang, Z. (2024). Swin-Net: A Swin-transformer-based network combing with multi-scale features for segmentation of breast tumor ultrasound images. Diagnostics (Basel), 14(3): 269. https://doi.org/10.3390/diagnostics14030269

[24] Ferreira, C.A., Melo, T., Sousa, P., Meyer, M.I., Shakibapour, E., Costa, P., Campilho, A. (2018). Classification of breast cancer histology images through transfer learning using a pre-trained inception ResNet v2. In International Conference Image Analysis and Recognition. Cham: Springer International Publishing, pp. 763-770. https://doi.org/10.1007/978-3-319-93000-8_86.

[25] Doğantekin, A., Özyurt, F., Avcı, E., Koc, M. (2019). A novel approach for liver image classification: PH-C-ELM. Measurement, 137: 332-338. https://doi.org/10.1016/j.measurement.2019.01.060

[26] Alom, M.Z., Yakopcic, C., Nasrin, M.S., Taha, T.M., Asari, V.K. (2019). Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network. Journal of Digital Imaging, 32: 605-617. https://doi.org/10.1007/s10278-019-00182-7

[27] Toğaçar, M., Özkurt, K.B., Ergen, B., Cömert, Z. (2020). BreastNet: A novel convolutional neural network model through histopathological images for the diagnosis of breast cancer. Physica A: Statistical Mechanics and its Applications, 545: 123592. https://doi.org/10.1016/j.physa.2019.123592