FABNet: Fusion Attention Block and Transfer Learning for Laryngeal cancer Tumor Grading in P63 IHC Histopathology Images

Pan Huang, Xiaoheng Tan, Xiaoli Zhou, Shuxian Liu, Francesco Mercaldo and Antonella Santone

***Abstract*—Laryngeal cancer tumor (LCT) grading is a challenging task in P63 Immunohistochemical (IHC) histopathology images due to small differences between LCT levels in pathology images, the lack of precision in lesion regions of interest (LROIs) and the paucity of LCT pathology image samples. The key to solving the LCT grading problem is to transfer knowledge from other images and to identify more accurate LROIs, but the following problems occur: 1) transferring knowledge without a priori experience often causes negative transfer and creates a heavy workload due to the abundance of image types, and 2) convolutional neural networks (CNNs) constructing deep models by stacking cannot sufficiently identify LROIs, often deviate significantly from the LROIs focused on by experienced pathologists, and are prone to providing misleading second opinions. So we propose a novel fusion attention block network (FABNet) to address these problems. First, we propose a model transfer method based on clinical a priori experience and sample analysis (CPESA) that analyzes the transfer ability by integrating clinical a priori experience using indicators such as the relationship between the cancer onset location and morphology and the texture and staining degree of cell nuclei in histopathology images; our method further validates these indicators by the probability distribution of cancer image samples. Then, we propose a fusion attention block (FAB) structure, which can both provide an advanced non-uniform sparse representation of images and extract spatial relationship information between nuclei; consequently, the LROI can be more accurate and more relevant to pathologists. We conducted extensive experiments, compared with the best Baseline model, the classification accuracy is improved 25% ,and It is demonstrated that FABNet performs better on different cancer pathology image datasets and outperforms other state of the art (SOTA) models.**

***Index Terms*—Laryngeal cancer Tumor (LCT) Grading, Attention Mechanism, Transfer Learning, Fusion Attention Block (FAB), Lesion Region of Interest (LROI).**

Pan Huang, College of Optoelectronic Engineering, Chongqing University , Xiaoheng Tan and Xiaoli Zhou, School of Microelectronics and Communication Engineering, Chongqing University, 400044, Chongqing, China. Shuxian Liu, School of Information Science and Engineering, Xinjiang University, 830046, Urumqi, China. Antonella Santone and Francesco Mercaldo, Department of Medicine and Health Sciences "Vincenzo Tiberio", University of Molise, 86100, Campobasso, Italy.

Corresponding author: Xiaoheng Tan (txh@cqu.edu.cn)

I. Introduction

Laryngeal cancer is one of the most common cancers in otolaryngology. Approximately 3% of new laryngeal cancer patients have a poor prognosis each year. The prognosis of laryngeal cancer patients is determined mainly by the stage of the cancer. Accurate diagnosis of the laryngeal cancer tumor (LCT) grade at the first diagnosis will greatly improve the 5-year survival rate of patients. Patients with low-grade LCTs especially have a higher chance of being cured if they undergo appropriate treatment surgery and practice good lifestyle habits. Therefore, accurate grading is very important [1]. P63 can be used as a specific marker for laryngeal squamous cell carcinoma, which is of great significance for the early diagnosis and differential diagnosis of laryngeal squamous cell carcinoma. P63 may actively participate in the regulation of the occurrence of laryngeal squamous cell carcinoma, but in laryngeal squamous cell carcinoma. It is in an activated state during the development of morphogenetic cell carcinoma, and may play a negative feedback role in the development process [2-3]. In addition, many methods for precise grading of cancer tumors have been proposed; these methods include Gleason grading for prostate cancer [4-5] and Nottingham grading for breast cancer [6-7]. In contrast, LCTs are a type of squamous cell carcinoma, and the grade of an LCT is distinguished mainly by the degree of staining, number, morphology and interrelationship of atypical squamous cells [1,8,9]. LCTs are divided into three main grades(as shown in Figure 3).

Tumor grading by using pathological tissue images of laryngeal cancer is difficult due to small intergrade differences (as shown in Figure 3a) and a paucity of pathological data. Transfer learning can solve such problems well, but numerous image datasets can be transferred, blind empirical transfer is often ineffective due to differences in data and task requirements, and existing depth-domain adaptive transfer methods are poorly effective and interpretable for the laryngeal cancer problem [11]. Therefore, we explore for the first time the effectiveness of four different cancer pathology image transfers to laryngeal cancer according to CPESA and provide interpretable transfer laws.

Deep learning algorithms have already achieved some success in recognizing and grading other cancer pathology images [12]. However, In the traditional CNN model, the model only pays attention to the relationship between local pixel blocks due to the characteristics of the local receptive field of CNN [43]. Obviously, the spatial relationship of cells in the pathological whole glass slide image that we are studying is at the pixel level. CNN lacks the information representation of a single pixel, and the traditional depth model obtained by stacking CNN layers will amplify this information representation deviation between pixels. Secondly, we know

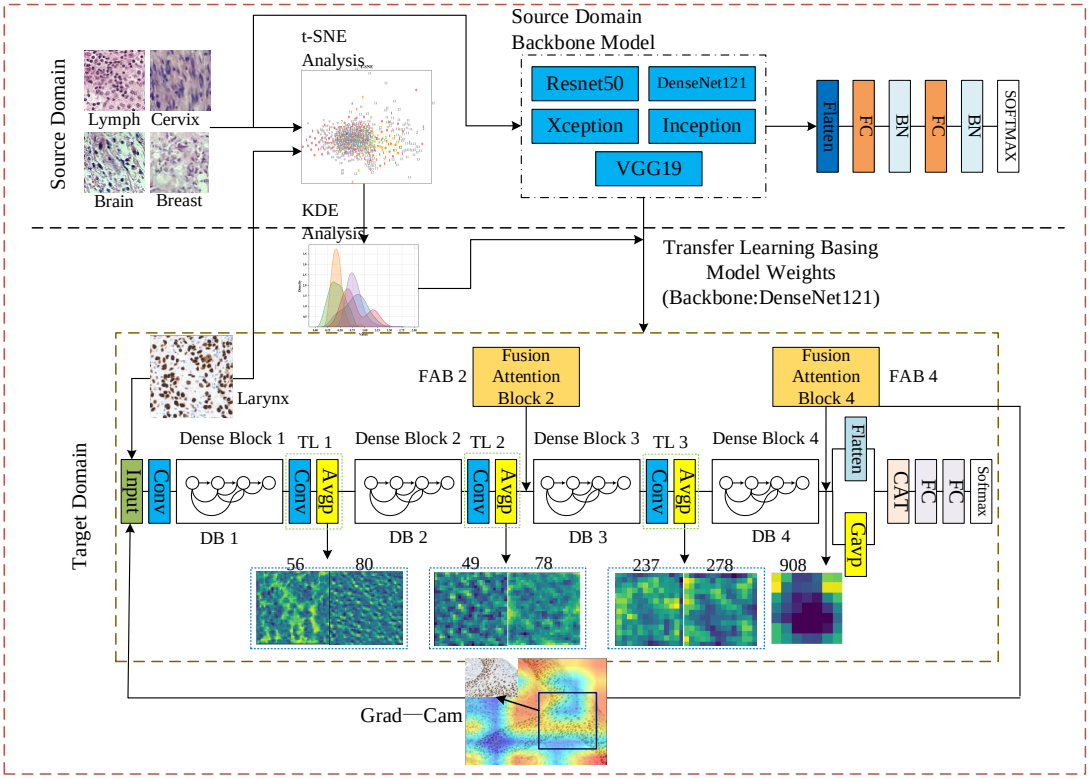

Fig 1. The overall structure of FABNet, which consists of four parts: CPESA, a DenseNet121-based backbone, a fusion attention block (FAB), and a classifier. We found through a large number of ablation experiments that inserting the FAB module proposed in this article at positions 2 & 4 is the best. The model shown in Fig 1 is the optimal model, so it is not shown where insert at positions 1 & 3. Moreover, the fusion rule used by FAB is the Average Fusion Attention Block (AFAB) shown in Fig 8.a[1].Grading pathological images of laryngeal cancer is difficult and creates a heavy workload because identifying pathological images depends very much on the subjective experience of pathologists, and young doctors have biased diagnosis due to their lack of experience. Second, most of the pathological images of suspected patients present with cancer-free manifestations[10]. Therefore, there is a great need to propose a computer-aided diagnostic algorithm to assist pathologists and clinicians in LCT grading and provide them with more accurate lesion regions of interest (LROIs) and interpretable grading results.

that the data in the original data is non-uniform sparse, and the CNN depth model we built is uniform sparse, which leads to some deviations between the feature representation we have learned and the pattern of the original data[42, 44]. Naturally, the depth model of stacking CNN layers cannot identify LROIs that are more conducive to grading. Moreover, the attention mechanism can effectively solve the above problem, the channel attention can learn more non-uniform sparse features, the position attention is able to learn the spatial information between pixels , and thereby effectively improve the recognition ability of the model for accurate LROIs. However, the following problems exist when using the attention mechanism to improve the depth model: 1) inserting the attention mechanism module in different positions of the depth model will produce different effects; 2) the spatial relationship between nuclei and the non-uniform sparse feature representation in the pathological tissue images are both very effective information for tumor classification, but using only the channel attention mechanism will miss the information of the spatial relationship between nuclei, such as Intercellular bridge information, while using the position attention mechanism alone will miss the information of the non-uniform sparse feature representation, for example,the pathological area that the pathologist pays attention to is often non-uniform sparse. To address this situation, we propose the fusion attention block (FAB), which can simultaneously learn the spatial relationships between nuclei and non-uniform sparse feature representations in images and significantly improve the recognition ability of deep models for accurate LROIs in laryngeal cancer pathology images.

Motivated by the above observations, we propose a novel fusion attention block network (FABNet) algorithm to assist physicians in grading LCTs accurately. First, we apply the model transfer method, clinical a priori experience, and sample analysis to solve negative transfer and poor interpretability caused by blind attempts. The FAB module is used to substantially improve the ability to identify the precise LROI in LCT images; the module incorporates both channel and position attention. The FAB-based gradient-weighted class activation mapping (Grad-CAM) [13] is used to mark the lesions that are ROIs to the FAB and can intuitively and effectively assist physicians in precise LCT grading. Our main contributions are summarized as follow:

(1) We use deep learning for the first time in laryngeal cancer P63 IHC histopathology images to design computer-aided diagnostic algorithms to automate the classification of LCTs and provide visualization results to pathologists.

(2) We propose a novel FABNet model. 1) We adopt a method based on CPESA to explore the effect of transferring different cancer pathology images to laryngeal cancer and combine clinical a priori experience and statistical knowledge to analyze the transfer ability between source and target domains, thereby effectively solving negative transfer and the heavy workload caused by nonempirical transfer learning. 2) Thus, we design a novel fusion attention block (FAB) module to effectively improve the recognition of accurate LROIs in LCT images without increasing the depth model parameters; this module can both provide advanced non-uniform sparse representation of images and analyze the spatial relationship between cell nuclei. 3) In

addition, we explore the effect of inserting the FAB at different positions on accurate LROI recognition by using the output of convolutional pooling to make corresponding assumptions and interpretations, which can provide references for subsequent related work.

(3) We adopt the Grad-CAM visualization map which based FAB, and use it to mark the LROIs of FAB , and compare them with LROIs of clinicians and pathologists; thus, this map can provide a more intuitive and effective aid to clinical decision-making and medical teaching.

The remainder of this article is organized as follows. In Section II, we briefly introduce the application of transfer learning and attention mechanisms in cancer pathology image grading. In Section III, we present the details and principles of FABNet implementation in detail. In Section IV, we introduce our extensive and comparative experiments and report the results of our proposed FABNet and related methods. In Section V, we discuss and visually analyze our method. We summarize our work in Section VI.

II. RELATED WORKS

To the best of our knowledge, considering that no deep learning classification method is publicly available for larynx histopathology images yet, this subsection introduces related work on only other cancer pathology images.

A. Transfer Learning for Cancer Histopathology Image Classification

M. Talo [14] proposed pretrained ResNet-50 and DenseNet-161 models based on ImageNet; these models achieved good results on both grayscale and color images of the Kimia Path24 dataset. In an experimental study, V. G. Buddhavarapu et al. [15] used ImageNet pretrained on five CNN models to classify thyroid histopathology images. U. A. H. Khan et al. [16] proposed an approach based on transferring breast cancer pathology images to prostate cancer pathology images with some transfer effect. S. Boumaraf et al. [17] proposed a transfer learning depth model based on ImageNet for classifying multiscale breast cancer pathology images. T. Xia et al. [18] proposed a domain adaptive approach to adjust the differences between various cancer pathology images to facilitate transferring cancer data with labeled bulk data to the target cancer. R. Singh et al. [19] proposed a pretrained VGG19 model based on ImageNet to identify unbalanced breast cancer pathology images.

B. Attention Mechanism for Cancer Histopathology Image Classification

H. Yang et al. [20] proposed a guided attention mechanism-improved neural network model to classify breast cancer pathology tissue images. A. BenTaieb et al. [21] proposed a circular attention mechanism-improved InceptionV3 model and performed experiments on three cancer pathology tissue image public datasets with good results. N. Tomita et al. [22] used an attention mechanism to improve ResNet18 to classify esophageal cancer pathology images. J. W. Yao et al. [23] proposed a Siamese network based on an attention mechanism and a fully connected layer for multi-instance learning; this network effectively solved poor overall recognition due to the absence of labels in cancer pathology patch images. P. Chen et al. [24] proposed a fused attention mechanism to improve the CNN to improve the searchability of ROIs in whole-slide images (WSIs) of cancer pathology automatically recognized by neural networks. P. X. Wu et al. [25] proposed a CNN module with spatial convolution and channel convolution separation and used an attention mechanism to improve the module, and experiments were conducted using the Breast Cancer Histopathological Image Classification (BreakHis) dataset with good results.

The existing methods do not consider non-uniform sparse of the representation feature and the spatial position relationship of the nucleus at the same time.

III. METHODOLOGY

The structure of our proposed FABNet is shown in Figure 1. Our FABNet consists of four main components: a model transfer algorithm based on a CPESA, a backbone based on DenseNet121 [26], a fusion attention block (FAB) and a classifier. In this section, we first introduce the implementation and general features of FABNet, which is a method of fusing channel attention [27] and position attention [44] and inserting them into the appropriate positions of DenseNet121, and the weights of the lymphoma-trained DenseNet121 model are used as initialization weights of FABNet. We will describe CPESA, the selection of the FAB insertion position and the method of FAB fusion in detail later.

A. Overview of the Dataset

Biopsy sections were formalin fixed, paraffin embedded, and specimens were HEstained for histological tumor grade and stage assessment and IHC stained for p63 expression. IHC staining was performed not automatically, but by a standard streptavidin peroxidase method, using a Dako Production, ready-to-use, monoclonal mouse antibody clone DAK-p63/IsotypeIgG2a, kappa[1].

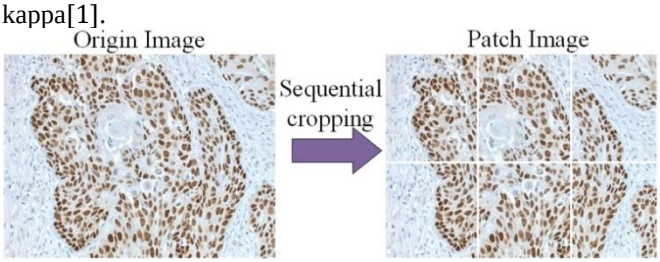

Fig 2. Sequential cropping of the original ROI images.

All areas of tumor were examined under the microscope and the histopathologist selected and marked the most heterogeneous region on each patient's p63-stained slide. From this region four images on average, with p63 nuclei expression were digitized, using a Leica DM2500 light microscope, equipped with a Leica DFC420C digital camera. Images were captured at magnification of 4009 and resolution of 1728 9 1296 9 24bit (see Fig. 1b) and stored on the hard disk of a dedicated host computer, which was connected to the digital camera of the microscope. Image capturing and storage were executed by means of the Leica Application Suite (LAS) program, which also was responsible for automatically regulating image-capturing parameters such as focus, exposure time, gamma value, and white balance. All areas of tumor on the slide were examined and the most heterogeneous area selected, If you want to use the data, please apply to the organization (https://medisp.bme.uniwa.gr/hicl/) [1].

LCTs are a type of squamous cell carcinoma, and the grade of an LCT is distinguished mainly by the degree of staining, number, morphology and interrelationship of atypical squamous cells [1,8,9]. LCTs are divided into three main grades. In Grade I LCTs (as shown in Figure 3b), atypical squamous cells are less than 25% of the tumor cells, the cancerous tissue invades into the dermis and does not exceed the level of sweat glands, and the basal cell

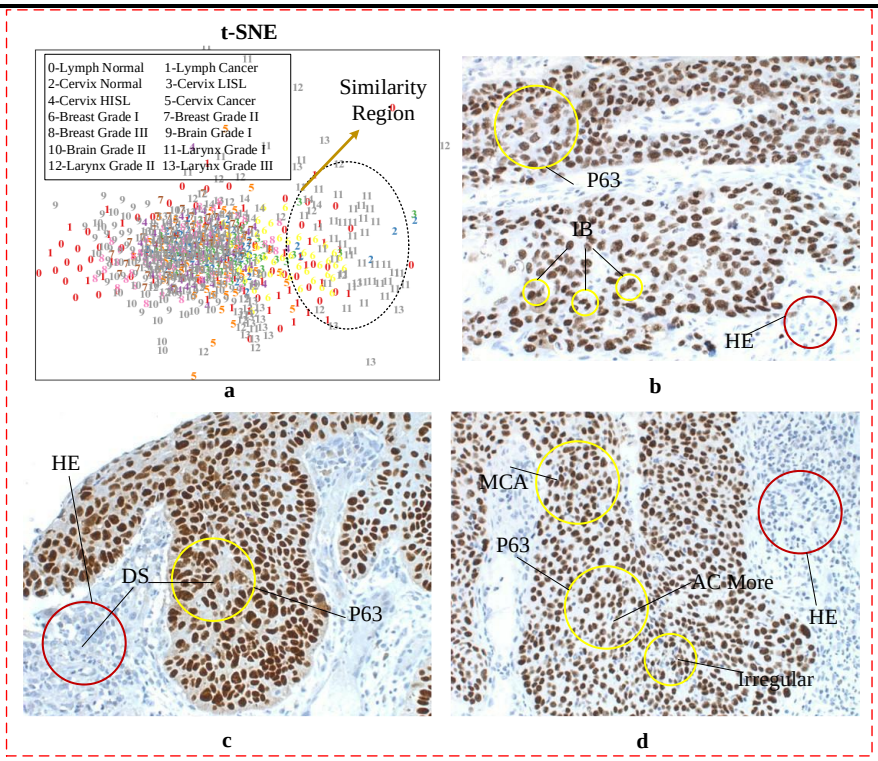

Fig 3. Sample distribution of different cancer pathology images and schematic diagram of laryngeal cancer pathology images. Intercellular bridge (IB), atypical cells (AC), more chaotic arrangement (MCA), and darker staining (DS) Existing methods based on the transfer of natural images to pathological images result in poor transfer effects because there are differences in the task requirements and data samples. However, transfer screening work is more arduous with the method of migration from similar cell images due to the lack of prior experience and analysis of sample distribution.

arrangement is still intact at the edge of the cancer mass in some parts, while in other parts, the arrangement is disordered or even without basal cells. At this point, there is no obvious demarcation between the cancer cells and the surrounding mesenchyme, and the cells of the cancer tissue are irregularly arranged. In Grade II LCTs (as shown in Figure 3c), the cancer tissue invades downward and reaches the deep dermis. The boundary between the cancer cell mass and the surrounding mesenchyme is unclear. Additionally, compared to Grade I LCTs, Grade II LCTs have more atypical squamous cells (which are approximately 25%-50% of the tumor cells) with light keratinization and only a few keratinized beads, and incomplete keratinization is seen mostly in the cell center. Grade III LCTs (as shown in Figure 3d) have many atypical squamous cells (approximately 50%-70%), and keratinization is either not obvious or not seen at all. No keratinized beads are seen, and individual poorly keratinized cells are visible.The nuclei are atypical with significant mitotic signs, and the surrounding inflammation is not obvious, thus indicating that the tissue has become unresponsive to the cancer cells.

We sequentially cropped the ROI images (Lymph, Cervix, Brain, and Breast) identified by doctor (as shown in Figure 2) to fit the training of the depth model. The ROI image resolution of Lymph is 96x96, so further cropping into patch images is not required. We divide each cancer image dataset into a training set, a validation set, and a test set. The training set is mainly used to train the weight of the model, and the validation set is mainly used to prevent overfitting and select hyper-parameters such as, learning rate, and training times, and the test set is used to evaluate the performance of the final model.(It shown in Table 1).

B. Overview of FABNet

1) Structure: Figure 1 shows that FABNet consists of two main parts: the source domain and target domain. In the source domain, the sample space distribution and probability distribution of five cancer pathology images are analyzed by t-distributed stochastic neighbor embedding (t-SNE) [28] and kernel density estimation (KDE) [29] algorithms, and the optimal possible source domain samples with transfer effect are analyzed by combining the a priori experience of clinicians and pathologists. We found that lymph cancer images [30] transferred to laryngeal cancer images [1] with the best effect and brain cancer images [31] transferred with the second best effect; then, we select lymph data to train the model weights as initialization weights for target domain training. In the target domain section, we partitioned DenseNet121 into four major blocks, and we chose to insert the fusion attention block (FAB) after transition layer 2 (TL 2) and dense block 4 (DB4) to obtain the best larynx tumor grading effect, where the FAB uses the average fusion attention block (AFAB) to fuse both channel and position attention. Then, we use the flatten layer and global average pooling layers (Gavps) behind FAB 4 to serially fuse the extracted multidimensional tensor into a two-dimensional tensor with two full-connect (FC) layers and softmax for hierarchical recognition. Finally, we visualize and analyze the feature maps after each TL to provide useful reference information for the insertion location selection, and we use the Grad-CAM map to identify lesion regions of interest (LROIs) on the feature maps output from the attention module in FAB4. The LROI is compared with the LROIs provided by clinicians and pathologists to verify the effectiveness of our proposed algorithm.

2) Analysis: Our proposed FABNet has the following features:

First, our proposed model transfer algorithm based on CPESA can better explore the degree of similarity and transfer ability between different pathological images while being more convenient to operate, more generalizable, and transplantable to other forms of medical image recognition.

Second, our proposed FAB method can retain the ability of channel attention to identify the importance of feature map channels and can use position attention to identify the spatial relationship between cell nuclei. The improved deep model can identify more accurate LROIs.

Third, we used the Grad-CAM algorithm to visualize and analyze the FAB module feature map and labeled the LROI that FAB focuses on. Compared with the LROI that pathologists and clinicians focus on, our method is more effective in assisting in clinical diagnosis.

C. Model transfer algorithm based on a clinical priori experience and sample analysis

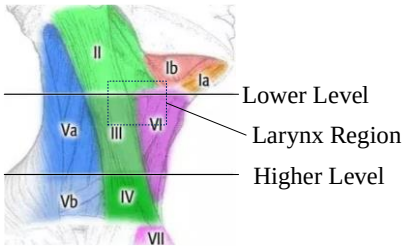

Fig 4. Distribution of lymph nodes and larynx locations. (for the purpose of illustrating the location of lymph nodes and larynx.)

The focus of transfer learning is to find the parts where the target domain is similar to the source domain. Based on this principle, transfer learning is divided into four main categories: sample-based transfer, relationship-based transfer, model-based transfer, and feature-based transfer [32]. The more mainstream transfer learning method at this stage is model-based transfer, which pretrains a model by the source domain and then transfers the model weights to the target domain model. We propose model transfer based on a priori clinical experience and sample analysis (shown in Figure 1) to explore the effect of four types of source domain data (i.e., four types of cancer pathology images: Cervix Cancer [33-34], Breast Cancer [35], Brain Cancer, and Lymph Cancer) on transferring to the target domain (laryngeal cancer cancer pathology images).

How to quickly and effectively determine whether there is transfer value in the source domain is a key issue. We propose a hypothesis based on doctors' a priori experience, visualize the sample distribution by using t-distribution stochastic neighbor embedding (t-SNE), use kernel density estimation (KDE) estimation to map the empirical distribution of the samples, and finally validate the proposed hypotheses with model cross-validation classification results.

According to the a priori empirical analysis of clinicians, the cervix, breast and brain are far from the larynx; the differences in cell differentiation are large; and the similarity between pathological images is relatively small. However, the larynx is surrounded by many lymph nodes, and the laryngeal and

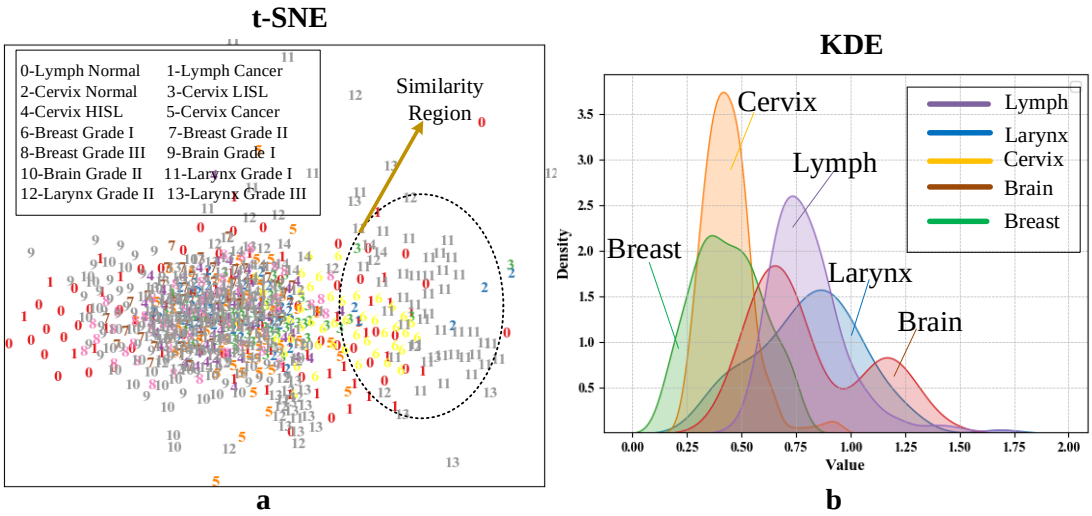

Fig 6. Visualization of t-SNE and KDE distribution of cancer image samples.

TABLE 1. Data tables related to pathological images of five types of cancer.

Parameter / Datasets	Native resolution	Magnification	Croped resolution	Tumor Grading	Patches sets	Training set	Validation set	Test set
Larynx[1]	534x400	40x	200x200	2	897	559	168	170
Lymph[30]	96x96	40x	—	3	4500	3600	450	450
Cervix[33-34]	3456x4608	40x	800x800	4	8845	7095	875	875
Brain[31]	534x400	40x	200x200	7	4733	3771	481	481
Breast[35]	534x400	40x	200x200	3	974	558	181	235

lymphatic tissues are intertwined, so from a biomedical point of view, the laryngeal and lymphatic tissues are highly similar in tissue and cellular morphology.

Based on the clinical a priori experience, as shown in Figure 4, we initially found that the similarity between laryngeal cancer and lymphoma was probably the greatest among the five cancer images. Therefore, we further characterized the pathological images by pathologists: as Figure 5 shows, the cells of early laryngeal cancer lesions appeared oval and more regular and have darker nuclei. The lymphoma cells also appear oval and regular, have dark nuclei, and can easily distinguish between the cell stroma and the nucleus. Brain cancer cells have a distinctly different morphology; they have a more irregular, elongated cell shape and vary more greatly in size. Cervical cancer cells and nuclei are chaotically distributed, and the nuclei are irregularly shaped and have bands and varying sizes. Breast cancer cells and nuclei differ more from the first four types of cancer cells in that breast cancer cells and nuclei have an angular shape and less chromatin in the nucleus, thus resulting in lighter staining.

Based on the a priori experience of clinicians first and then pathologists, we will focus on exploring the effect of lymphatic image transfer to laryngeal cancer later.

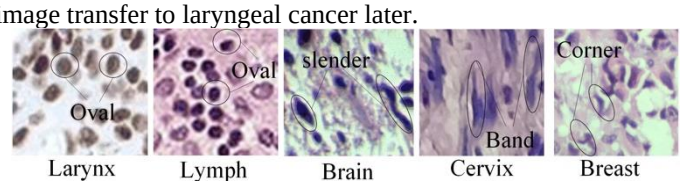

Fig 5. Early lesions pathological images of five cancer types.

Concerning sample distribution visualization, the image samples in the paper are multidimensional data, so it is difficult to visualize their data distribution directly, and now the sample distribution visualization analysis is performed mainly by dimensionality reduction. The optimal sample distribution visualization method is t-SNE, whose main principle is to model the distribution of the nearest neighbors of each data point, where the nearest neighbors are the set of data points close to each other, as in equation (1) [28]. In the original high-dimensional space, we model the high-dimensional space as a Gaussian distribution, while in the two-dimensional output space, we can model this space as a t-distribution in Eq. (2) [28]. The goal of the process is to find the transformation that maps the high-dimensional space to the two-dimensional space and minimizes the gap between these two distributions for all points, as in Eq. (3) [28]. In this paper, the t-SNE algorithm is used to visualize and analyze five cancer image samples, as shown in Figure 6a. In addition, the KDE algorithm is used to map the empirical distribution of the samples based on the data after the dimensionality reduction of t-SNE, as shown in Figure 6b.

Figure 6a clearly shows that the differences between the levels in different cancers are small, intertwined and very difficult to distinguish, which indicates to some extent that it is difficult to grade cancer pathological images; in addition, Figure 6b clearly shows that the empirical distributions of laryngeal cancer, lymphoma and brain cancer are more similar, but the shape of the empirical distribution of brain cancer is double peaked, while the empirical distribution of laryngeal cancer and lymphoma is single-peaked; therefore, through the empirical distribution, we can initially verify that the similarity between the images of laryngeal cancer and lymphoma is greater and that more information can be transferred, and the effect of subsequent transfer should be better.

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma^2\right)} \tag{1}$$

In formula (1), $p_{j|i}$ denotes the conditional probability of Euclidean distance between high-dimensional sample points, x_i denotes a high-dimensional sample point, x_j denotes a nearest neighbor of x_i in the high-dimensional data, and σ is the standard deviation of the Gaussian distribution.

$$q_{j|i} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)} \tag{2}$$

In formula (2), $q_{j|i}$ denotes the conditional probability of the Euclidean distance between the low-dimensional sample points, y_i denotes the low-dimensional sample points, and y_j is denoted as the nearest neighbor of x_i in the low-dimensional data.

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}} \tag{3}$$

In formula (3), C minimizes the KL scatter between two distributions (Kullback-Leibler divergence).

D. Fusion Attention Block

Two main problems must be solved to improve the deep network model using an attention mechanism: the first problem is where to insert the attention mechanism module in the deep network and how many attention mechanism modules should be inserted. The second problem is how to fuse both channel[27] and position attention[44].

To solve the above two problems, we use the weights of lymphoma as the initialization weights based on the experience obtained from transfer learning in Section B. DenseNet121 is used as the backbone model for this experiment. First, we choose four dense block outputs of DenseNet121 as the experimental points of four insertion locations, as shown in Figure 7, and 15 insertion combination methods are shown in Table 3. Then, we propose three fusion attention mechanisms (shown in Figure 8): the average fusion attention block (AFAB), splicing fusion attention block (SFAB), and concatenate fusion attention block (CFAB).

In Section D, we introduce and discuss the location and number of attention mechanism insertions in detail; in Section E, we introduce the fusion principle and fusion method of the attention mechanism in detail.

E. Insertion Position

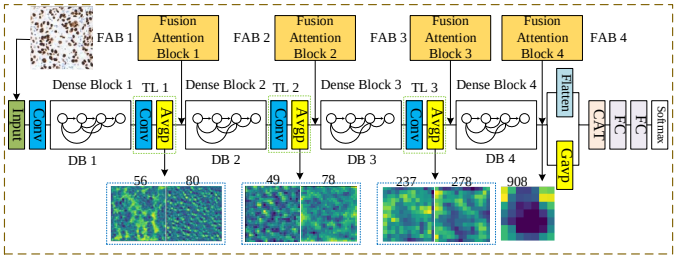

Fig 7. Block diagram of the attention mechanism insertion position.

1) Structure: We use DenseNet121 as the backbone model.

DenseNet has four dense blocks (DBs): Dense Block 1 (six 1x1 convolutional layers and six 3x3 convolutional layers), Dense Block 2 (twelve 1x1 convolutional layers and twelve 3x3 convolutional layers), Dense Block 3 (twenty-four 1x1 convolutional layers and twenty-four 3x3 convolutional layers) and Dense Block 4 (sixteen 1x1 convolutional layers and sixteen 3x3 convolutional layers). Dense Block 3 (twenty-four 1x1 convolutional layers and twenty-four 3x3 convolutional layers) and Dense Block 4 (sixteen 1x1 convolutional layers and sixteen 3x3 convolutional layers). The first three dense blocks are followed by a transition layer (TL), which consists of a 1x1 convolutional layer (Conv) and an average pooling layer (Avgp). In addition, we choose to insert the fusion attention block (FAB) after the first three TL modules and the last DB module. The fusion method and detailed structure of the FAB are shown in Figure 8. To guide us in studying the schematic study of inserting the FAB at each layer, we performed feature map visualization analysis on the outputs of the first three TL modules and the last DB module, and the output feature maps are shown as 56, 80, 49, 78, 237, 278, and 908 in Figure 7 (these numbers represent the number of feature map channels output at different locations). Finally, we add a global average pooling (Gavp) and flatten layer to the output of the backbone model, use serial fusion to fuse the tensor output from these two layers, and use two full-connect (FC) layers and softmax functions for classification.

2) Analysis: Based on the large differences in the feature maps output by different levels of convolutional pooling, we propose the hypothesis that inserting FAB blocks at different locations will

produce different effects. Inserting the blocks generally has the following pattern: the bottom convolutional block identifies hyperlocal features (lines, edges, etc.), the middle convolutional block identifies local features (texture, grayscale distribution, the regions of interest (ROIs) of lesions, etc.), and the top convolutional block identifies global features (e.g., identifying high-level non-uniform sparse representations or sparse representations of objects).

F. Fusion Methods

F.a Average Fusion Attention Block

1) Structure: The average fusion attention block (AFAB) consists of channel attention and position attention, and the structure of the AFAB is shown in Figure 8a. AFAB is a two-branch structure, where the input 4-dimensional tensor $T_{A-in} \in R^{(B,M,N,F)}$ has to go through both channel attention (CA) and position attention (PA) for the corresponding operations. In addition, we define B, M, N, F uniformly, where B denotes the number of images of the input model, M is the height of each feature map in the tensor, N is the width, and F is the dimension of the feature channels of the tensor.

First, we use Equation (4) to calculate the tensor $T_{A-CAB} \in R^{(B,M,N,F)}$ obtained after CA:

$$T_{A-CAB} = RS\left(\delta\left(\delta\left(Avgp(T_{A-in}) \otimes w_{fc1}\right) \otimes w_{fc2}\right)\right) \quad (4)$$

In Eq. (4), δ denotes the excitation function, we use the ReLU function, $Avgp$ represents the average pooling layer, $w_{fc1} \in R^{(F,F/2)}$ is the weight of the first FC layer, $w_{fc2} \in R^{(F/2,F)}$ is the weight of the second FC layer, RS is the layer

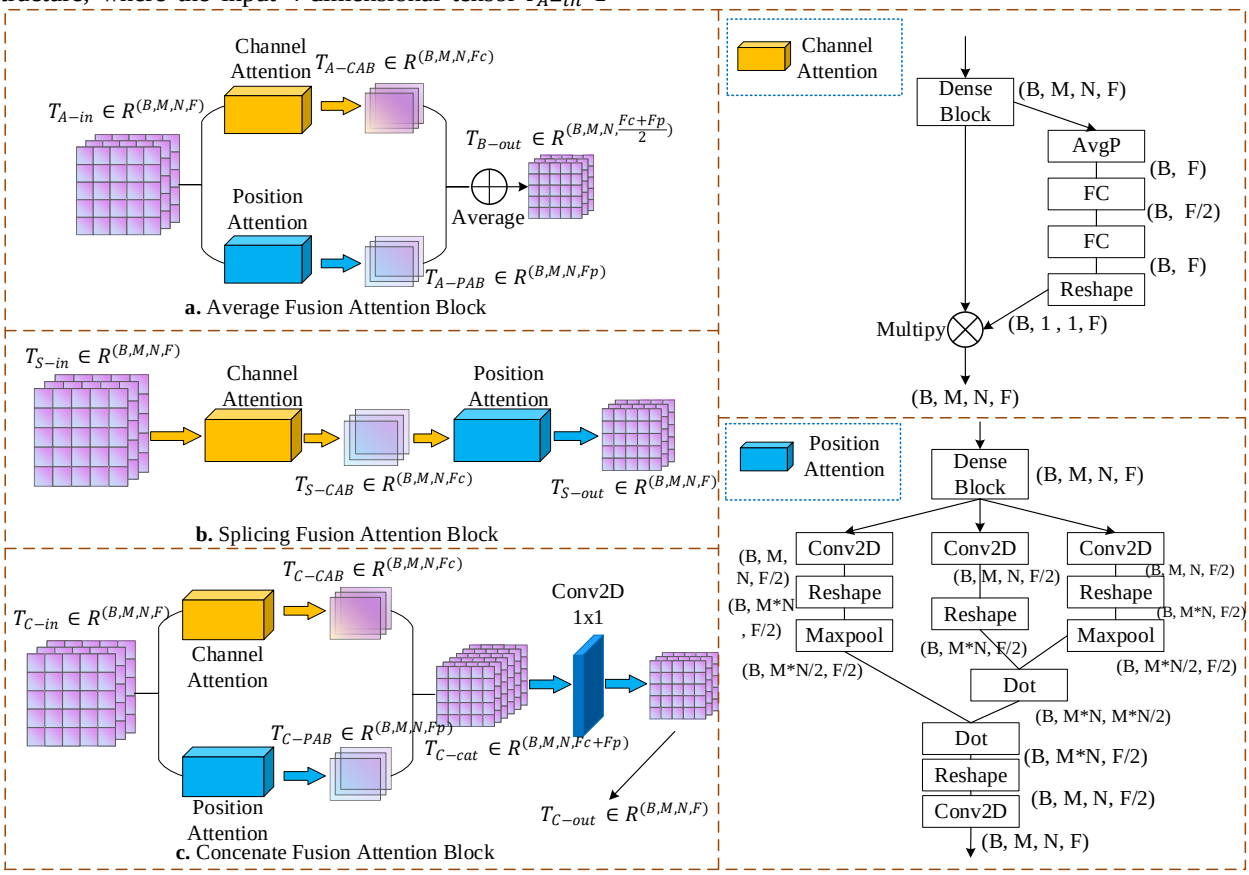

Fig 8. Block diagram of fusion attention mechanism.

that we define to change the tensor size, and $\otimes$ denotes elementwise multiplication.

Additionally, we use by equation (5) to compute the tensor $T_{A-PAB} \in R^{(B,M,N,F)}$ obtained after passing through PA.

$$T_{A-PAB} = \delta\left\{Conv\left[RS\left(\left(MP\left(RS\left(\delta\left(Conv(T_{A-in})\right)\right)\right) \odot RS\left(\delta\left(Conv(T_{A-in})\right)\right)\right) \odot MP\left(RS\left(\delta\left(Conv(T_{A-in})\right)\right)\right)\right)\right\} \quad (5)$$

In equation (5), MP denotes the maximum pooling layer; $\odot$ denotes the vector dot product operator; $Conv$ denotes the convolution operator; the convolution kernel is 1x1; δ denotes the excitation function; and we use a linear function, whose main role is to change the dimension of the feature channels of the tensor.

Finally, we obtain the output tensor $T_{A-out} \in R^{(B,M,N,F)}$ by averaging the fusion by using Equation (6).

$$T_{A-out} = \frac{T_{A-PAB} + T_{A-CAB}}{2} \quad (6)$$

2) Analysis: Channel attention uses mainly the learned channel attention weights to select the importance of the tensor feature channels, thereby retaining the useful feature channels while removing the redundant ones. Position attention indicates mainly the importance degree of spatial location in the feature map by the learned position attention weights. Our proposed AFAB can retain the features of these two attention mechanisms, both to identify the importance degree of feature channels and to analyze the spatial relationship between cell nuclei in laryngeal cancer pathological tissue images. In addition, to a larger extent, this fusion approach does not lose the useful information learned by either attention mechanism, so the AFAB fusion approach may be the most favorable for the final grading results.

F.b Splicing Fusion Attention Block

1) Structure: The splicing fusion attention block (SFAB)

consists of channel attention and position attention, and the structure of SFAB is shown in Figure 8b. SFAB is a single branch structure, where the input 4-dimensional tensor $T_{S-in} \in R^{(B,M,N,F)}$ is first processed by CA and then PA for the corresponding operations.

First, we use Equation (7) to calculate the tensor $T_{S-CAB} \in R^{(B,M,N,F)}$ obtained after CA:

$$T_{S-CAB} = RS\left(\delta\left(\delta\left(Avgp(T_{S-in}) \otimes w_{fc1}\right) \otimes w_{fc2}\right)\right) \quad (7)$$

In Eq. (7), δ denotes the excitation function, we use the ReLU function, $Avgp$ represents the average pooling layer, $w_{fc1} \in R^{(F,F/2)}$ is the weight of the first FC layer, $w_{fc2} \in R^{(F/2,F)}$ is the weight of the second FC layer, RS is the layer we define to change the tensor size, and $\otimes$ denotes elementwise multiplication.

Then, we use equation (8) to compute the output tensor $T_{S-out} \in R^{(B,M,N,F)}$ obtained after passing through PA.

$$T_{S-out} = \delta\left\{\mathrm{Conv}\left[\mathrm{RS}\left(\left(\mathrm{MP}\left(\mathrm{RS}\left(\delta\left(\mathrm{Conv}\left(T_{S-CAB}\right)\right)\right)\right) \odot \mathrm{RS}\left(\delta\left(\mathrm{Conv}\left(T_{S-CAB}\right)\right)\right)\right) \odot \mathrm{MP}\left(\mathrm{RS}\left(\delta\left(\mathrm{Conv}\left(T_{S-CAB}\right)\right)\right)\right)\right)\right]\right\} \quad (8)$$

In equation (8), MP represents the maximum pooling layer; $\odot$ represents the vector dot product operator; $Conv$ represents the convolution operator; the convolution kernel is 1x1; δ represents the excitation function; and we use a linear function, which here serves mainly to change the dimensionality of the feature channels of the tensor.

2) Analysis: Channel attention focuses on the importance analysis of tensor feature channels by learning weight probability distributions. Position attention focuses on exploring the importance of spatial location in images. Our proposed SFAB is processed by CA and then PA; therefore, the useful information learned by CA will be largely lost, and such fusion may lead to poor tumor grading in laryngeal cancer pathology images.

F.c Concatenate Fusion Attention Block

1) Structure: The concatenate fusion attention block (CFAB) consists of both channel and position attention, and the structure of the CFAB is shown in Figure 8c. CFAB is a two-branch structure, where the input 4-dimensional tensor $T_{C-in} \in R^{(B,M,N,F)}$ has to go through both CA and position attention (PA) for the corresponding operations.

First, we use Equation (9) to calculate the tensor $T_{C-CAB} \in R^{(B,M,N,F)}$ obtained after going through CA:

$$T_{C-CAB} = RS\left(\delta\left(\delta\left(Avgp(T_{C-in}) \otimes w_{fc1}\right) \otimes w_{fc2}\right)\right) \quad (9)$$

In Eq. (9), δ denotes the excitation function, we use the ReLU function, $Avgp$ represents the average pooling layer, $w_{fc1} \in R^{(F,F/2)}$ is the weight of the first FC layer, $w_{fc2} \in R^{(F/2,F)}$ is the weight of the second FC layer, RS is the layer we define to change the tensor size, and $\otimes$ denotes elementwise multiplication.

$$T_{C-PAB} = \delta\left\{\mathrm{Conv}\left[\mathrm{RS}\left(\left(\mathrm{MP}\left(\mathrm{RS}\left(\delta\left(\mathrm{Conv}\left(T_{C-in}\right)\right)\right)\right) \odot \mathrm{RS}\left(\delta\left(\mathrm{Conv}\left(T_{C-in}\right)\right)\right)\right) \odot \mathrm{MP}\left(\mathrm{RS}\left(\delta\left(\mathrm{Conv}\left(T_{C-in}\right)\right)\right)\right)\right)\right]\right\} \quad (10)$$

Additionally, we use equation (10) to calculate the tensor $T_{C-PAB} \in R^{(B,M,N,F)}$ obtained after passing the PA.

In Eq. (10), MP denotes the maximum pooling layer; $\odot$ denotes the vector dot product operator; $Conv$ denotes the convolution operator; and the convolution kernel is 1x1, which serves mainly to change the feature channel dimension of the tensor.Finally, we obtain the output tensor $T_{C-out} \in R^{(B,M,N,F)}$ by serial fusion in Eq. (11):

$$T_{C-out} = \delta\left(Conv\left(Cat(T_{C-CAB}, T_{C-PAB})\right)\right) \quad (11)$$

In equation (11), Cat denotes serial splicing in the feature channel dimension of the tensor; $Conv$ denotes the convolution operator; the convolution kernel is 1x1; δ denotes the excitation function; and we use a linear function, which here serves mainly to change the feature channel dimension of the tensor.

2) Analysis: Channel attention uses the learned channel attention weights mainly to select the importance of the tensor feature channels, thereby retaining the useful feature channels while removing the redundant ones. Position attention indicates mainly the importance degree of spatial location in the feature map by the learned position attention weights. Our proposed CFAB retains the features of these two attention mechanisms and the extracted useful information to the maximum extent, but the information extracted by this fusion approach is too redundant, thus leading to an increase in computational complexity and easily causing cumulative error propagation in deep networks. Therefore, the CFAB fusion approach may achieve good improvement in some insertion positions but overall should be less effective than AFAB.

G. Evaluation Metrics

We evaluate different models through quantitative indicators such as Grading accuracy, F1 score, Receiver Operating Characteristic Curve (ROC), Precision and Recall.

G.a Grading accuracy

We understand that the pathological images of laryngeal cancer are divided into three grades (Grade I, Grade II and Grade III), and the number of samples judged by the model is divided by the real number in this grade.

G.b Recall rate (Recall, R)

Table 2. Classification confusion matrix

Reality	Forecast result	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

When we automatically classify the pathological images of laryngeal cancer, various situations shown in the classification confusion matrix will appear. In order to analyze what proportion of the samples that are truly positive examples are predicted correctly, we use the recall rate to analyze, the formula is shown in (12)[45].

$$R = \frac{TP}{TP + FN} \quad (12)$$

G.c Precision (Percision, P)

As shown in Table 2, in order to analyze and identify the proportion of samples whose predictions are positive, we use the accuracy rate to classify, and the formula is shown in (13) [45].

$$P = \frac{TP}{TP + FP} \quad (13)$$

G.d F1 score

F1 score, also known as balanced F score (balanced F Score), is defined as the harmonic average of precision and recall. The formula is shown in formula (14) [45]:

$$F1 = 2 \times \frac{P * R}{P + R} \tag{14}$$

G.e ROC

We use ROC curve and AUC (Area Under ROC Curve, AUC) to evaluate the modeling effect of the algorithm. The samples are sorted according to the prediction results of the classifier, and the samples are used as positive examples to predict in this order, and the true positive rate (TPR) is calculated each time as shown in formula (15) [45] and false positive rate (False). Positive Rate, FPR) as in formula (16) [45], TPR is used as the vertical axis of the ROC curve, and FPR is used as the horizontal axis of the ROC curve[45].

$$TPR = \frac{TP}{TP + FN} \tag{15}$$

$$FPR = \frac{FP}{TN + FP} \tag{16}$$

In formulas (15-16), false negative examples, false positive examples , true positive examples and true negative examples are the the classification results in confusion matrix.

IV. EXPERIMENTS AND RESULTS

In this section, we introduce details of the experimental implementation in Section A, evaluation metrics of the algorithm in Section B, and various experimental results of FABNet in Section C.

A. Implementation Details

The input size of the depth model is 224x224x3, and the batch size is 16. The exception is Lymph, which is used to train the depth model; the input size is 96x96x3, and the batch size is 196. The epoch of the pretrained model is 100 (i.e., the first five rows in Table 3), while the other training epochs are 200. The learning rate is set as 1x10-4 for 0~50 epochs, 2x10-5 for 50~75 epochs, and 1x10-5 for 75~100 epochs for the pretraining model (i.e., the first five rows in Table 3). All the algorithms are implemented in Python, and the deep learning module is based on Keras 2.3.1 and TensorFlow 2.1. The GPUs are two RTX2080TI 11 GB.

B. Experimental Results

C.a Result of FAB insert position ablation Experiment

We visualized the feature maps of the output of the TL module and DB4, and some representative feature maps are shown in Figure 9. In Figure 9, the left RGB image is the original input laryngeal cancer pathology image, and the 4 columns on the right are the feature map visualization images. The 1st and 3rd feature maps output by TL1_Out clearly identify the boundary of the cell stroma by lines and edges, while the 2nd map identifies the boundary of the cell nucleus and the 4th map identifies the boundary of cell secretion; therefore, DB1 and TL1 identify mainly the hyperlocal features of laryngeal cancer pathological tissue images. Additionally, the 3,4 pictures of the TL2_Out output feature map present a texture phenomenon of cell secretion direction, while the 1,2 pictures present a texture phenomenon of cell nucleus distribution, so DB2 and TL2 identify mainly the texture and grayscale distribution of tissue images of laryngeal cancer pathology. Second, the 1-4 output feature maps of TL3_Out show that the convolution blocks now focus more on identifying the lesion salience region and the region of nonnormal cell aggregation, which is also the lesion that is the ROI to clinicians. Finally, the feature map output of DB4, which is clearly a kind of advanced non-uniform sparse representation or sparse representation, with more zero values in the feature map, by which the advanced sparse representation is beneficial to tumor grading.

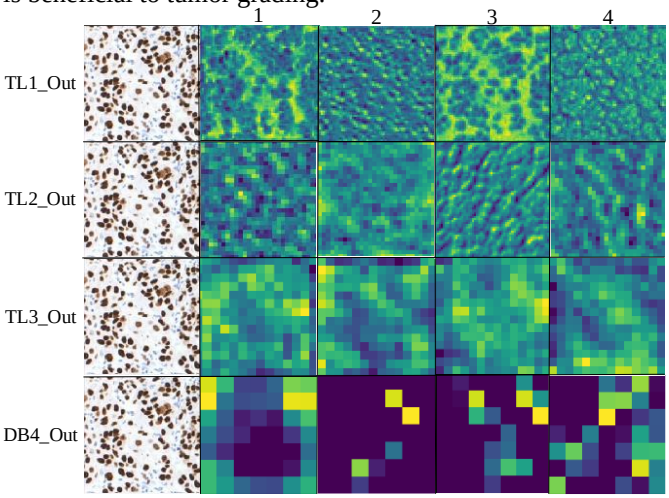

Fig 9. Visualization of feature channels with different insertion positions.

The above analysis shows that TL1_Out identifies mainly hyperlocal features, which are very useful for recognizing laryngeal cancer images; If the FAB module is inserted in TL1_Out, the line or edge information may be lost; this outcome would be disastrous to the subsequent mid-level and high-level non-uniform sparse representation. Second, we find that TL2_Out and TL3_Out identify specific LROIs, and inserting the FAB module in this block to remove non-ROI regions can clearly effectively improve the non-uniform sparsity of the subsequent feature maps. DB4_out is the most advanced non-uniform sparse feature extracted by the model. Adding the FAB module here should have good benefits.

C.b Result of Transfer Learning Experiment

We trained five models in the source domain model and transferred four cancer images and ImageNet training weights to LCTs for the experiments. Table 3 shows that DenseNet121 performs the best overall in the source and target domains; this model achieves 51.76%, 92.22%, 82.97%, 75.74%, and 98.3% in Larynx, Lymph, Cervix, Breast, and Brain, respectively, in the source domain training, and all of these performance rates are the best among those of the five deep models. For transfer learning, Lymph→Larynx training in the DenseNet121 model yielded the best accuracy (77.64%), which is an improvement of approximately 25%.

C.c Result of the FABNet Experiment

The results of the following experiments with three FAB fusion methods and 15 insertion positions plus one mainstream SOTA model with various insertion positions are shown in Table 4. The best recognition accuracy (87.65%) was achieved with the FAB2-FAB4 insertion method under AFABNet fusion;this accuracy rate is a 10% improvement over the rate

obtained after transfer and an approximately 35% improvement over the rate obtained before transfer.

C.d Results of the FABNet performance on other Histopathology Image Datasets

We validate the performance of our proposed FABNet model on each of the five cancer datasets, and the experimental results are shown in Table 5. We conclude that FABNet has the most significant improvement on the Larynx dataset with 35.9% improvement, followed by 12.8% improvement on the Breast dataset and 0.63% improvement on the Brain and Cervix datasets; the same result is obtained with the backbone model on the Lymph dataset.

C.e Results of FABNet and other SOTA Models

We compared four SOTA models of deep learning+attention on the Larynx dataset, and the experimental results are shown in Table 6. Only the Grade I accuracy of our proposed FABNet is lower than that of the DenseNet121+CBAM model, and the rest of the metrics are optimal.

TABLE 3. Experimental results of model transfer learning based on a priori clinical experience and sample analysis. Note: Lymph→Larynx indicates that the weights trained from the Lymph dataset are the initial weights of Larynx training.

Method \ Model	Resnet50[36]	DenseNet121[26]	Inception[37]	Xception[38]	VGG19[39]
Larynx[1]	42.35%	51.76%	44.11%	42.94%	39.41%
Lymph[30]	89.55%	92.22%	86.00%	90.22%	61.11%
Cervix[33-34]	77.94%	82.97%	81.25%	82.28%	67.19%
Brain[31]	96.67%	98.33%	97.08%	96.25%	74.84%
Breast[35]	74.46%	75.74%	54.46%	77.87%	36.17%
Lymph→Larynx	74.70% ↑32%	**77.64% ↑25%**	74.70% ↑33%	74.70% ↑32%	55.88% ↑16%
Cervix→Larynx	58.23% ↑16%	72.94% ↑21%	65.88% ↑22%	71.17% ↑28%	38.82% ↓0.6%
Brain→Larynx	67.64% ↑25%	72.35% ↑20%	70.58% ↑26%	73.52% ↑31%	48.82% ↑9.0%
Breast→Larynx	72.94% ↑30%	69.99% ↑18%	60.00% ↑16%	66.47% ↑24%	48.23% ↑9.0%
ImageNet→Larynx	68.82% ↑26%	67.05% ↑16%	72.94% ↑28%	72.94% ↑30%	41.76% ↑2.3%

TABLE 4. Experimental results of FAB module. Note: FAB 2-, FAB 3-, and FAB 4- mean to insert the FAB module after TL2, TL3, and DB4, respectively, of DenseNet121; please refer to Section III.

Insertion method \ Fusion Method	AFABNet	CFABNet	SFABNet	SENet[27]
FAB 1	73.52% ↓4%	73.52% ↓4%	68.82% ↓9%	74.47% ↓3%
FAB 2	85.29% ↑7%	85.29% ↑7%	68.23% ↓9%	72.35% ↓5%
FAB 3	77.64% ↑0%	83.52% ↑6%	65.29% ↓12%	73.52% ↓4%
FAB 4	77.64% ↑0%	75.88% ↓2%	80.00% ↑2%	72.94% ↓5%
FAB 1-FAB 2	77.05% ↓1%	85.29% ↑8%	71.76% ↓6%	75.29% ↓2%
FAB 1- FAB 3	77.05% ↓1%	79.41% ↑2%	68.82% ↓9%	75.88% ↓1%
FAB 1- FAB 4	80.58% ↑3%	78.82% ↑1%	71.17% ↓6%	72.35% ↓5%
FAB 2- FAB 3	86.47% ↑9%	77.64% ↑0%	65.29% ↓12%	74.11% ↓3%
FAB 2- FAB 4	**87.65% ↑10%**	81.76% ↑4%	75.29% ↓2%	69.99% ↓7%
FAB 3- FAB 4	79.41% ↑2%	82.35% ↑5%	74.11% ↓3%	75.88% ↓1%
FAB 1- FAB 2- FAB 3	78.23% ↑1%	82.35% ↑5%	65.88% ↓12%	70.58% ↓7%
FAB 1- FAB 2- FAB 4	72.35% ↓5%	69.41% ↓8%	70.58% ↓7%	75.29% ↓2%
FAB 1- FAB 3- FAB 4	85.29% ↑8%	70.58% ↓7%	77.11% ↑0%	71.64% ↓6%
FAB 2- FAB 3- FAB 4	85.88% ↑8%	82.94% ↑5%	65.88% ↓12%	73.52% ↓4%
FAB 1-FAB 2-FAB 3-FAB 4	85.88% ↑8%	84.70% ↑7%	57.64% ↓20%	68.23% ↓9%

TABLE 5. Experimental results of FABNet on different cancer pathology images.

Dataset \ Metric	Accuracy	Percision	Recall	F1	AUC
Lymph[30]	92.23% ↑0%	0.92	0.92	0.92	0.97
Cervix[33-34]	83.54% ↑0.63%	0.84	0.84	0.84	0.95
Brain[31]	98.96% ↑0.63%	0.99	0.99	0.99	1.00
Breast[35]	88.51% ↑12.8%	0.88	0.89	0.88	0.97
Larynx[1]	87.65% ↑35.9%	0.88	0.88	0.88	0.97

TABLE 6. Table of test results of different SOTA models (training and testing on the Laryngeal cancer pathology dataset).

Algorithm \ Metric	Grade I Accuracy	Grade II Accuracy	GradeIII Accuracy	Average Accuracy	P	R	F1	AUC
DenseNet121+Non-local [44]	75.00%	85.18%	64.58%	75.29%	0.76	0.75	0.75	0.90
DenseNet121+SENet [27]	76.47%	85.18%	64.58%	75.88%	0.76	0.75	0.75	0.88
DenseNet121+HIENet [40]	82.35%	77.77%	64.58%	75.88%	0.76	0.75	0.76	0.90
DenseNet121+CBAM [41]	88.23%	74.07%	75.00%	80.00%	0.80	0.80	0.79	0.93
FABNet (ours)	86.76%	88.88%	87.50%	87.65%	0.88	0.88	0.88	0.97

V. DISCUSSION AND VISUAL ANALYSIS

We will discuss the four aspects of FABNet's overall performance, transfer ability, precise LROI and the limitations of the method.

A. Overall performance of FABNet

Tables 4 and 5 show the following results: 1) Our proposed FABNet performs well in recognizing and grading five cancer pathology images. In particular, FABNet substantially outperforms the improved DenseNet121 model in recognizing laryngeal and breast cancer images. 2) All four SOTA models are not as effective as FABNet. There is one exception: the accuracy of the CBAM method on Grade I exceeds that of FABNet. Therefore, it is quantitatively demonstrated that our proposed model outperforms other SOTA models in identifying accurate laryngeal cancer LROIs and can identify accurate LROIs on different datasets.

B. Discussion of transfer ability

The results in Table 3 show that Lymph→Larynx is the best, followed by Cervix→Larynx and Brain→Larynx, which also have good results. As our analysis in Chapter III shows, Lymph and Larynx are the most similar in terms of cell morphology, staining and location of onset. Additionally, Lymph and Larynx have the second most overlapping probability distributions and similar distribution shapes. Larynx, Brain and Cervix were less similar in cell morphology, degree of staining, and location of onset. Among the probability distributions, those of Brain and Larynx had the most overlapping area, but the distribution shapes were very different, with Brain having a bimodal distribution and Larynx having a single peak. Cervix had the least overlapping area with Larynx, although the distribution shapes were somewhat similar, and the distribution of Breast was worse in terms of both overlapping area and shape.

Therefore, we can draw some regular conclusions: the source and target cancer pathology images satisfy the conditions of similar locations of onset, similar morphologies of nuclei, similar texture distributions, comparable degrees of staining, similar shapes of sample probability distribution with large overlap area, etc. The transfer effect is better.

C. Discussion of Precision LROIs

An analysis of Table 4 reveals the following experimental phenomena: 1) The effect of using attention to improve DenseNet121 is ranked as AFABNet > CFABNet > SFABNet > SENet. This indicates that AFABNet can provide both an advanced non-uniform sparse representation of images and an analysis of spatial relationships between nuclei, while SENet alone can provide only a strong advanced non-uniform sparse representation and cannot provide a strong analysis of spatial relationships between nuclei, so the overall effect is poor. CFABNet is inferior to AFABNet because the information fused by CFABNet is overly redundant. The effect of SFABNet is also not particularly satisfactory because the model's position attention substantially destroys the information extracted by the channel attention module. 2) Inserting the FAB module after TL2 of DenseNet121 achieves good results, while inserting the FAB module in TL1 of DenseNet121 is less effective. The effect of inserting the FAB module in TL3 and TL4 of DenseNet121 is average. This finding fully verifies the results obtained from our analysis based on the feature maps output from convolutional pooling: TL1 convolutional blocks recognize hyperlocal features (lines, edges, etc.), which are very important for the subsequent high-level non-uniform sparse representation, and inserting FAB modules here will destroy the ability to recognize these features, thereby resulting in unsatisfactory results, as described in detail in subsection III.D. Therefore, it can be quantitatively concluded from the data level that our proposed FABNet can recognize more accurate LROIs.

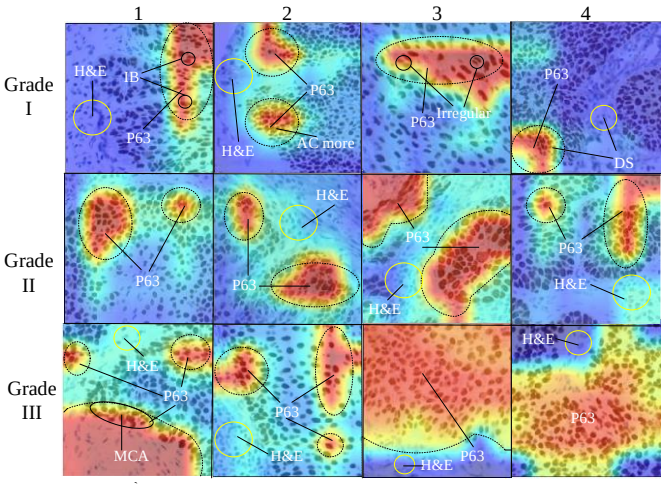

Fig 10. LCT Grad-CAM image based on FAB. Intercellular bridge (IB), atypical cells (AC), more chaotic arrangement (MCA), darker staining (DS).

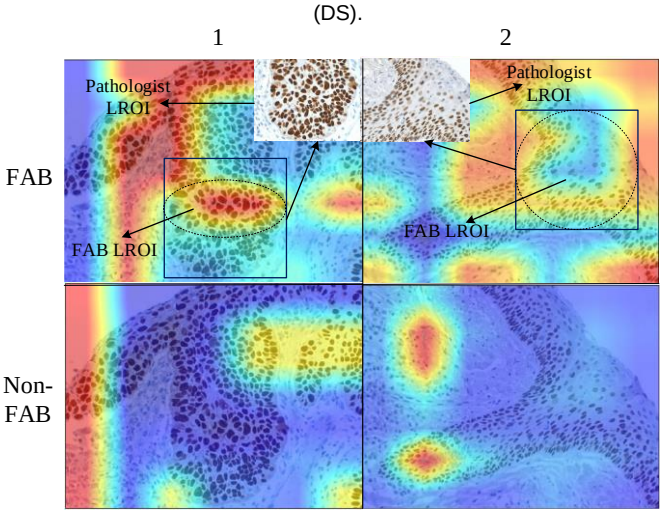

Fig 11. Comparison of the lesions that are regions of interest to the pathologist and the FAB and a schematic of the lesions that are regions of interest to DenseNet121 without the FAB module inserted.

The above discussed analyses are all quantitative and based on tables, and later, we will use FAB-based Grad-CAM plots for qualitative and visual analysis.

Figure 10 shows the following patterns: 1) Our proposed FABNet model focuses more on the P63 protein expression region, especially on the P63 staining presenting irregular and elongated cellular regions. In contrast, the FABNet model paid little attention to the nuclear regions of H&E cells that were not stained by P63. 2) The LROI focused on by the FABNet model had the following characteristics: deeper staining of nuclei, more atypical squamous cells per unit area, more irregular nuclei morphology, more chaotic nuclei arrangement, and more

intercellular bridges. Therefore, we can conclude that areas with deeper P63 staining, more atypical squamous cells, more irregular cell morphology and arrangement and more intercellular bridges are more favorable for accurate grading.

Further analysis of Figure 11 and a comparison of the LROIs identified by pathologists on the original LCT pathology images with those identified by the FAB module in our FABNet show much similarity between the LROIs. This finding indicates that our proposed FABNet can identify precise LROIs in LCT images well; thus, FABNet can substantially improve the accuracy of automated grading and substantially improve interpretable second opinions for clinicians and pathologists.

D. Limitations of the method

Through Table 5, FABNet's performance on different cancer pathology datasets is somewhat different. we add some explanation as follows: 1). The most fundamental reason is the difficulty of grading different cancers. The principle of grading order of magnitude is different. From the difficulty or principle of grading, it is obvious that the grading of adenocarcinoma is relatively easy, while the grading of squamous cell carcinoma is more difficult; in terms of the number of gradings, in our article, laryngeal cancer is divided into three levels, and the classification of cervical cancer is four, while Lymph is only divided into two levels, and Breast is three levels, so overall the more the classification, the more difficult the automatic identification will increase. 2). The number of datasets also has a certain impact on the final effect. For example, there are only more than 800 on Lymph, and the most cervix has more than 8,000. 3). The imaging equipment and technical methods of pathological images also have a certain effect on the final result. Because the data used in this manuscript comes from different countries and teams, there are certain differences in the imaging equipment and technology used, and different technologies or equipment processing processes will cause image degradation. Due to differences and image degradation problems are both have a certain impact on the final result.

Regarding the performance fluctuations of the model caused by the above problems, the performance difference caused by the medical grading magnitude is unavoidable. In addition, for the difference caused by the dataset or imaging problems, we give some suggestions: 1). First, collecting as much as possible complete dataset, so that the trained model fits the natural distribution of pathological images more. 2). Using image restoration or enhancement technology to pre-process the image to minimize the impact of degradation caused by the image imaging process on the model performance. 3). Normalizing or centralizing different cancer data sets to minimize image differences caused by different imaging devices.

Finally, the best grading accuracy of our FABNet was only 87.64%, which still leaves much room for improvement. The above situation occurs mainly because of the relatively low incidence of laryngeal cancer and the difficulty of sample collection; consequently, there are fewer samples of laryngeal cancer pathology images. In addition, there are few studies related to laryngeal cancer pathology image-assisted diagnosis algorithms, and studies using deep learning-based methods are especially scarce, thus leading to fewer comparative experiments and insufficient overall results.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a FABNet algorithm to transfer knowledge of lymph cancer pathology images to the grading task of LTC pathology images to reduce the negative transfer and the workload caused by inexperienced transfer training. The DenseNet121 model is improved by using the FAB module incorporating both channel and position attention to substantially improve the ability of the deep model to recognize accurate LROIs of LCT images. Our proposed FABNet is superior to other SOTA models, performs well and has good generality in grading different cancer pathology images. Our Grad-CAM map based on FAB can identify LROIs more accurately and will consequently provide visualization and resolvability references for clinical diagnosis and medical teaching.

However, due to limited data sample size, few relevant investigation methods, and small differences between levels of laryngeal cancer pathology images, the highest accuracy of our proposed FABNet algorithm in LTC image grading is only 87.64%, which obviously leaves much room for improvement. In our future research, we could improve the transfer learning algorithm to increase the transfer effect and adopt more advanced modules to enhance the depth model's ability to identify accurate LROIs by collecting more laryngeal cancer patient samples.

REFERENCE

[1] K. Ninos, S. Kostopoulos, I. Kalatzis, K. Sidiropoulos, P. Ravazoula, G. Sakellaropoulos, G. Panayiotakis, G. Economou, and D. Cavouras, "Microscopy image analysis of p63 immunohistochemically stained laryngeal cancer lesions for predicting patient 5-year survival," European Archives of Oto-Rhino-Laryngology, vol. 273, no. 1, pp. 159-168, 2016.

[2] G. Pruneri, L. Pignataro, M. Manzotti, N. Carboni, D. Ronchetti, A. Neri, B. M. Cesana, and G. Viale, "p63 in laryngeal squamous cell carcinoma: evidence for a role of TA-p63 down-regulation in tumorigenesis and lack of prognostic implications of p63 immunoreactivity," Laboratory investigation, vol. 82, no. 10, pp. 1327-1334, 2002.

[3] H.-R. Choi, J. G. Batsakis, F. Zhan, E. Sturgis, M. A. Luna, and A. K. El-Naggar, "Differential expression of p53 gene family members p63 and p73 in head and neck squamous tumorigenesis," Human pathology, vol. 33, no. 2, pp. 158-164, 2002.

[4] P. Nicolosi, E. Ledet, S. Yang, S. Michalski, B. Freschi, E. O'Leary, E. D. Esplin, R. L. Nussbaum, and O. Sartor, "Prevalence of Germline Variants in Prostate Cancer and Implications for Current Genetic Testing Guidelines," Jama Oncology, vol. 5, no. 4, pp. 523-528, Apr 2019.

[5] S. Srivastava, E. J. Koay, A. D. Borowsky, A. M. De Marzo, S. Ghosh, P. D. Wagner, and B. S. Kramer, "Cancer overdiagnosis: a biological challenge and clinical dilemma," Nature Reviews Cancer, vol. 19, no. 6, pp. 349-358, Jun 2019.

[6] S. Q. Bao, H. Q. Zhao, J. Yuan, D. D. Fan, Z. C. Zhang, J. Z. Su, and M. Zhou, "Computational identification of mutator-derived lncRNA signatures of genome instability for improving the clinical outcome of cancers: a case study in breast cancer," Briefings in Bioinformatics, vol. 21, no. 5, pp. 1742-1755, Sep 2020.

[7] J. Xu, L. Xiang, Q. S. Liu, H. Gilmore, J. Z. Wu, J. H. Tang, and A. Madabhushi, "Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images," Ieee Transactions on Medical Imaging, vol. 35, no. 1, pp. 119-130, Jan 2016.

[8] B. Miles, M. Ittmann, T. Wheeler, M. Sayeeduddin, A. Cubilla, D. Rowley, P. Bu, Y. Ding, Y. Gao, M. Lee, and G. E. Ayala, "Moving Beyond Gleason Scoring," Archives of Pathology & Laboratory Medicine, vol. 143, no. 5, pp. 565-570, May 2019.

[9] J. R. Wright, "Albert C. Broders, tumor grading, and the origin of the long road to personalized cancer care," Cancer Medicine, vol. 9, no. 13, pp. 4490-4494, Jul 2020.

[10] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot and B. Yener, "Histopathological Image Analysis: A Review," in IEEE

Reviews in Biomedical Engineering, vol. 2, pp. 147-171, 2009, doi: 10.1109/RBME.2009.2034865.

[11] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding Transfer Learning for Medical Imaging," p. arXiv:1902.07208Accessed on: February 01, 2019[Online].

[12] M. Veta, J. P. W. Pluim, P. J. van Diest and M. A. Viergever, "Breast Cancer Histopathology Image Analysis: A Review," in IEEE Transactions on Biomedical Engineering, vol. 61, no. 5, pp. 1400-1411, May 2014, doi: 10.1109/TBME.2014.2303852.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,"p. arXiv:1610.02391Accessed on: October 01, 2016[Online].

[14] M. Talo, "Automated classification of histopathology images using transfer learning," *Artif. Intell. Med.*, vol. 101, Nov 2019, Art no. 101743.

[15] V. G. Buddhavarapu and J. A. A. Jothi, "An experimental study on classification of thyroid histopathology images using transfer learning," *Pattern Recognition Letters*, vol. 140, pp. 1-9, Dec 2020.

[16] U. A. H. Khan, C. Stürenberg, O. Gencoglu, K. Sandeman, T. Heikkinen, A. Rannikko, and T. Mirtti, "Improving Prostate Cancer Detection with Breast Histopathology Images," Cham, 2019: Springer International Publishing, pp. 91-99.

[17] S. Boumaraf, X. B. Liu, Z. S. Zheng, X. H. Ma, and C. Ferkous, "A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images," *Biomedical Signal Processing and Control*, vol. 63, Jan 2021, Art no. 102192.

[18] T. Xia, A. Kumar, D. Feng and J. Kim, "Patch-level Tumor Classification in Digital Histopathology Images with Domain Adapted Deep Learning," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 2018, pp. 644-647, doi: 10.1109/EMBC.2018.8512353.

[19] R. Singh, T. Ahmed, A. Kumar, A. K. Singh, A. K. Pandey, and S. K. Singh, "Imbalanced Breast Cancer Classification Using Transfer Learning," *Ieee-Acm Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 83-93, Jan 2021.

[20] H. Yang, J. Y. Kim, H. Kim, and S. P. Adhikari, "Guided Soft Attention Network for Classification of Breast Cancer Histopathology Images," *Ieee Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1306-1315, May 2020.

[21] A. BenTaieb and G. Hamarneh, "Predicting Cancer with a Recurrent Visual Attention Model for Histopathology Images," in *Medical Image Computing and Computer Assisted Intervention - Miccai 2018, Pt Ii*, vol. 11071, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. AlberolaLopez, and G. Fichtinger, Eds. (Lecture Notes in Computer Science, 2018, pp. 129-137.

[22] N. Tomita, B. Abdollahi, J. S. Wei, B. Ren, A. Suriawinata, and S. Hassanpour, "Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides," *Jama Network Open*, vol. 2, no. 11, Nov 2019, Art no. e1914645.

[23] J. W. Yao, X. L. Zhu, J. Jonnagaddala, N. Hawkins, and J. Z. Huang, "Whole slide images based cancer survival prediction using attention uided deep multiple instance learning networks," *Medical Image Analysis*, vol. 65, Oct 2020, Art no. 101789.

[24] P. Chen, Y. Liang, X. Shi, L. Yang, and P. Gader, "Automatic whole slide pathology image diagnosis framework via unit stochastic selection and attention fusion," Neurocomputing, 2021/01/23/ 2021.

[25] P. X. Wu, H. Qu, R. R. Fi, Q. Y. Huang, C. Chen, D. Metaxas, and Ieee, "DEEP ATTENTIVE FEATURE LEARNING FOR HISTOPATHOLOGY IMAGE CLASSIFICATION," IEEE International Symposium on Biomedical Imaging, 2019, pp. 1865-1868.

[26] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," p. arXiv:1608.06993Accessed on: August 01, 2016[Online].

[27] J. Hu, L. Shen, S. Albanie, G. Sun, and E. H. Wu, "Squeeze-and-Excitation Networks," Ieee Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011-2023, Aug. 2020.

[28] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579-2605, Nov 2008.

[29] D. M. Asta, "Kernel density estimation on symmetric spaces of non-compact type," Journal of Multivariate Analysis, vol. 181, Jan 2021, Art no. 104676.

[30] Ehteshami Bejnordi et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA: The Journal of the American Medical Association, 318(22), 2199–2210. doi:jama.2017.14585.

[31] Glotsos, D., Kalatzis, I., Spyridonos, P., Kostopoulos, S., Daskalakis, A., Athanasiadis, E., Ravazoula, P., Nikiforidis, G., Cavouras, D.Improving accuracy in astrocytomas grading by integrating a robust least squares mapping driven support vector machine classifier into a two level grade classification scheme (2008) Computer Methods and Programs in Biomedicine, 90 (3), pp. 251-261.

[32] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," Journal of Big Data, vol. 3, no. 1, p. 9, 2016/05/28 2016.

[33] Huang, P.; Tan, X.; Chen, C.; Lv, X.; Li, Y. AF-SENet: Classification of Cancer in Cervical Tissue Pathological Images Based on Fusing Deep Convolution Features. Sensors 2021, 21, 122.

[34] P. Huang et al., "Classification of cervical biopsy images based on lasso and EL-SVM," IEEE Access, vol. 8, pp. 24219-24228, 2020.

[35] Kostopoulos, S., Glotsos, D., Cavouras, D., Daskalakis, A., Kalatzis, I., Georgiadis, P., Bougioukos, P., Ravazoula, P., Nikiforidis, G.Computer-based association of the texture of expressed estrogen receptor nuclei with histologic grade using immunohistochemically-stained breast carcinomas (2009) Analytical and Quantitative Cytology and Histology, 31 (4), pp. 187-196.

[36] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun, and Ieee, "Deep Residual Learning for Image Recognition," in 2016 Ieee Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," p. arXiv:1512.00567Accessed on: December 01, 2015[Online].

[38] F. Chollet and Ieee, "Xception: Deep Learning with Depthwise Separable Convolutions," in 30th Ieee Conference on Computer Vision and Pattern Recognition, 2017, pp. 1800-1807.

[39] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," p. arXiv:1409.1556 Accessed on: September 01, 2014[Online].

[40] H. Sun, X. X. Zeng, T. Xu, G. Peng, and Y. T. Ma, "Computer-Aided Diagnosis in Histopathological Images of the Endometrium Using a Convolutional Neural Network and Attention Mechanisms," Ieee Journal of Biomedical and Health Informatics, vol. 24, no. 6, pp. 1664-1676, Jun 2020.

[41] S. H. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in Computer Vision - Eccv 2018, Pt Vii, vol. 11211, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Lecture Notes in Computer Science, 2018, pp.

[42] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-Mixer: An all-MLP Architecture for Vision," p. arXiv:2105.01601Accessed on: May 01, 2021[Online].

[43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, "Going Deeper with Convolutions", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9

[44] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794-7803.

[45] Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. 2 (1): 37–63.