(Received 13 March 2025, Accepted 27 April 2025, Scheduled 29 April 2025)

Automatic Identification of Aspiration Pneumonia Based on Bronchoscope Images and Deep Learning

Dawei Gong[1, †], Ke Cui[2, †], Weidong Wang[3], Xiaobo Chen[4, 5], Chao Zhang[6], Haifei Xiang[6], Shaohua Zhang[6], and Sailing He[4, 1, 7, *]

[1] *MedEngInfo Collaborative Research Center*
Zhejiang Engineering Research Center for Intelligent Medical Imaging, Sensing and Non-invasive Rapid Testing
Taizhou Hospital of Zhejiang Province, Zhejiang University, Taizhou, China
[2] *Intensive Care Medicine Department*
Taizhou Hospital of Zhejiang Province affiliated to Wenzhou Medical University, Taizhou 317000, China
[3] *Zhejiang UE Medical Instrument Co., LTD, Zhejiang, China*
[4] *Center for Optical and Electromagnetic Research, National Engineering Research Center for Optical Instruments*
College of Optical Science and Engineering, Zhejiang University, Hangzhou 310052, China
[5] *School of Opto-Electronic Engineering*
Changchun University of Science and Technology, Changchun 130022, China
[6] *Enze HospitalTaizhou Enze Medical Center(Group)*
Taizhou Hospital of Zhejiang Province affiliated to Wenzhou Medical University, China
[7] *Taizhou Agility Smart Technologies Co., Ltd., Taizhou, China*

ABSTRACT: Aspiration pneumonia is a type of lung infection caused by the accidental inhalation of foreign substances into the respiratory tract. It is commonly seen in the elderly, young children, and individuals who are unconscious or have difficulty swallowing. Early detection and diagnosis of aspiration pneumonia are beneficial for improving patient outcomes and reducing the medical burden. In this study, we collected bronchoscopic video data from 25 patients in two hospitals. After image preprocessing and expert annotation, we obtained 2830 images from some patients for training and 1215 images from the other patients for validation. We selected three deep learning methods for training. The experimental test results for the identification of aspiration pneumonia showed that ResNet-50, which is based on convolutional operations, gave the best performance in the automatic identification of aspiration pneumonia, with a precision of 97.82%, a recall of 91.82%, an F1 score of 94.73%, and an overall accuracy of 95.88%. The experiments demonstrated that deep learning methods can be used for the automatic identification and diagnosis of aspiration pneumonia from bronchoscope images and deep learning is reported here for the first time for diagnosing aspiration pneumonia from bronchoscope images.

1. INTRODUCTION

Aspiration pneumonia is a type of pneumonia that occurs when substances such as oropharyngeal secretions, gastric contents, or other foreign materials (e.g., food, liquids, vomit, or chemical agents) are inadvertently inhaled into the lower respiratory tract due to impaired swallowing function, altered consciousness, or other underlying conditions. It can lead to pulmonary inflammation or infection and is clinically classified into chemical pneumonitis or bacterial pneumonia [1]. Community-acquired pneumonia (CAP) is highly prevalent worldwide, and studies suggest that aspiration may be a contributing factor to CAP, particularly among the elderly population [2]. Several reports indicate that aspiration pneumonia is commonly observed in CAP patients, with an estimated incidence ranging from 5% to 15% [3–5]. In China, the incidence of aspiration pneumonia among hospitalized elderly patients aged ≥ 55 and ≥ 60 years has been reported to be 13.4% and 14.0%, respectively [6]. However, the diagnostic criteria for aspiration pneumonia (AP) have not been universally established. Therefore, an effective method for diagnosing aspiration pneumonia could contribute significantly to advancing research and treatment strategies for CAP.

Chemical pneumonitis is a form of chemical injury caused by the inhalation of sterile gastric contents, and it is also referred to as Mendelson's syndrome. It typically occurs in patients with impaired consciousness, such as those experiencing drug overdose, seizures, cerebrovascular accidents, or following the administration of anesthetics. Clinically, it presents with sudden-onset dyspnea, hypoxemia, tachycardia, and diffuse wheezing or crackles upon auscultation. Chest radiographs often reveal abnormalities, and patients may develop acute respiratory distress syndrome (ARDS). Bacterial pneumonia is an infectious process resulting from the aspiration of oropharyngeal secretions colonized by pathogenic microorganisms. It is commonly observed in individuals with dysphagia, particularly the elderly and residents of long-term care facilities. Clinical manifestations typically include fever, cough, sputum production, and physical signs of pneumonia [7].

Currently, there is no universally accepted definition for the diagnosis of aspiration pneumonia (AP). Diagnosis is primar-

* Corresponding author: Sailing He (sailing@zju.edu.cn). † Dawei Gong and Ke Cui contributed equally to this work.

ily based on clinical history, risk factors, and pulmonary imaging findings [7]. Clinically, patients often present with recurrent fever, cough, sputum production, and signs of pulmonary consolidation or moist rales on auscultation [8]. Imaging studies typically reveal new inflammatory infiltrates in the lungs, with the distribution of lesions closely related to the patient's position at the time of aspiration [9–11]. In the early stages, radiographic findings may be negative. Characteristic imaging features include infiltrates in gravity-dependent lung segments (such as the superior segment of the lower lobe or the posterior segment of the upper lobe in supine patients, and the basal segments of the lower lobes in upright patients) with a higher prevalence in the right lung [9–11]. Fiberoptic bronchoscopy is a valuable diagnostic tool for aspiration pneumonia [12]. The identification of aspirated materials (such as food particles, vomitus, foreign bodies, gastric fluid, or bile) during bronchoscopy provides strong evidence for the diagnosis of AP. Additionally, the presence of mucosal abnormalities such as hyperemia, edema, hemorrhage, or erosion in the trachea, mainstem bronchi, or segmental bronchi, when combined with known risk factors (e.g., impaired consciousness, dysphagia, periodontal disease, or poor oral hygiene) [13–17], and interpreted in reference to established diagnostic criteria [18], may also support a diagnosis of aspiration pneumonia. Due to its complex clinical presentation and overlap with other types of pneumonia, along with the presence of multiple predisposing factors (whose exact contributions to the onset of AP remain unclear) aspiration pneumonia is frequently underdiagnosed or misdiagnosed, posing challenges for early and accurate diagnosis [19]. Some studies have explored the use of biomarkers in diagnosing AP. For instance, El-Solh et al. [20] investigated serum procalcitonin levels as a potential marker but found that it failed to distinguish between patients with aspiration pneumonia confirmed by positive quantitative bronchoalveolar lavage (BAL) cultures and those with negative cultures. Suzuki et al. [21] found no significant correlation between elevated amylase levels in BAL and aspiration risk factors, and thus could not establish a definitive link between this phenomenon and aspiration pneumonia. Diagnostic criteria for differentiating between bacterial pneumonia and chemical pneumonitis have been proposed by the Infectious Diseases Society of America (IDSA) and the British Thoracic Society (BTS) [22, 23]. However, a study by Delforge et al. [24] concluded that these criteria were suboptimal for making such distinctions and may have adverse implications for patient management in subsequent stages of treatment.

In recent years, with the rapid advancement of deep learning [25–28] and computer vision technologies, computer-aided diagnosis (CAD) systems have been widely applied in medical image analysis. Dysphagia is recognized as a major cause of AP [29], and current CAD-based research aimed at predicting AP often focuses on the assessment of dysphagia. Zhuang et al. [30] developed multiple machine learning models for the identification of AP, achieving classification accuracies ranging from 69.2% to 88.5%. Sejdic et al. [31] proposed a Bayesian classification approach based on swallowing accelerometry recordings for the detection of dysphagia, reporting an accuracy exceeding 90% in distinguishing between healthy swallows and aspiration events. Weng et al. [32, 33] developed a system that utilizes convolutional neural networks (CNNs) to automatically detect signs of aspiration from flexible endoscopic evaluation of swallowing (FEES) videos, achieving an accuracy of over 92% and demonstrating its effectiveness in enhancing diagnostic accuracy among resident physicians. Miura et al. [34] introduced a method based on B-mode video ultrasonography (BV-US), assisted by image processing, to detect aspiration events, with an accuracy of 92%. Sarraf Shirazi et al. [35] applied an unsupervised fuzzy k-means clustering algorithm to classify post-swallowing respiratory sounds as aspiration or non-aspiration. When compared with assessments by speech-language pathologists using FEES/VFSS, the system achieved a silent aspiration detection accuracy exceeding 86%. Subsequently, in a follow-up study [36], respiratory sounds immediately after swallowing were analyzed using phase space reconstruction and then classified via an SVM model to identify patients at high risk of aspiration, yielding an accuracy of 86%. Lee et al. [37] proposed a novel approach using an Inflated 3D Convolutional Network (I3D) to detect the pharyngeal phase of swallowing in videofluoroscopic swallowing studies (VFSS), achieving an accuracy of over 93%.

2. DEEP LEARNING FOR IMAGE CLASSIFICATION

Deep learning-based image classification techniques play a pivotal role in the field of endoscopic imaging. In the task of classifying endoscopic images related to pneumonia, deep learning models, such as convolutional neural networks (CNNs) and attention-based mechanisms, can automatically learn discriminative features from medical images. For instance, by training on large datasets of annotated normal and pathological endoscopic images, these models can effectively identify differences in texture, color, and structural patterns of tissues, thereby enabling accurate differentiation between normal and abnormal images. Mainstream deep learning models for image classification include ResNet, EfficientNet, and DeiT. The following section provides a brief overview of each of these models.

ResNet [38] (Residual Network) pioneered the development of deep convolutional neural networks by introducing the concept of residual connections. Traditional deep neural networks often suffer from issues such as vanishing gradients and performance degradation as the number of layers increases. ResNet addresses these challenges by enabling the network to learn residual functions — i.e., the difference between the input and the desired output — rather than directly fitting a mapping from input to output. This approach effectively mitigates gradient vanishing problems and makes it feasible to train very deep networks. Structurally, ResNet is composed of multiple residual blocks, each consisting of convolutional layers, batch normalization, and nonlinear activation functions. These residual blocks can extract multi-level image features, ranging from low-level features such as edges and textures to high-level semantic and abstract representations. There are two types of residual blocks in ResNet, as illustrated in Figure 1. The basic residual block is used in shallower architectures like ResNet-18 and ResNet-34, while the bottleneck residual block is employed in deeper networks such as ResNet-50, ResNet-101, and

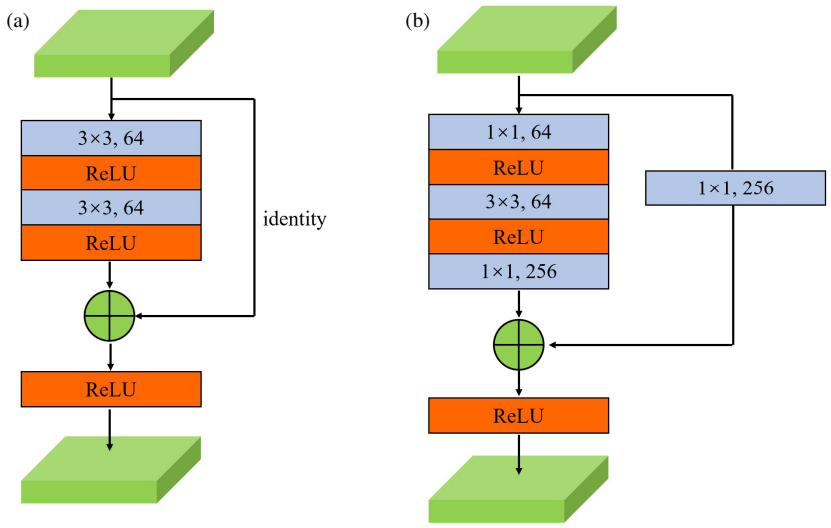

FIGURE 1. The structure of residual blocks in ResNet. (a) Basic residual block. (b) Bottleneck residual block.

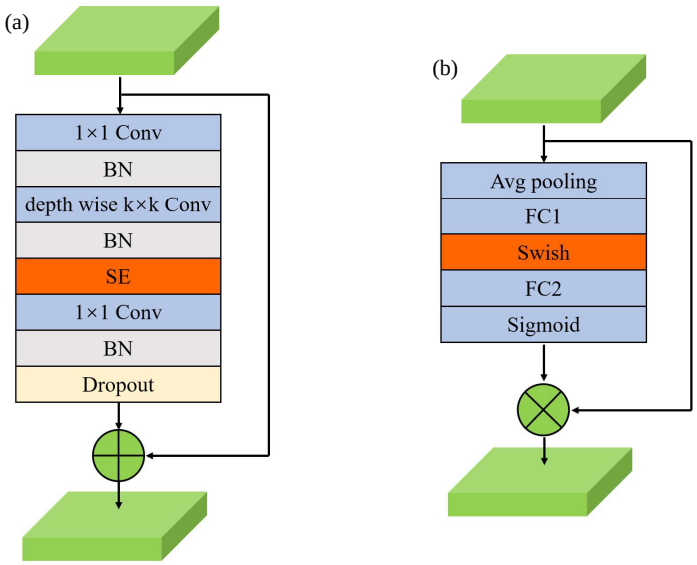

FIGURE 2. The MBConv structure and SE module. (a) MBConv. (b) Squeeze-and-Excitation (SE) module.

ResNet-152. In a basic residual block, the input feature map passes through two 3×3 convolutional layers with nonlinear activations, and the output is added element-wise to the original input feature map to form the final output. The bottleneck residual block is named for its characteristic "bottleneck" shape, where the number of channels is reduced at the beginning and end but expanded in the middle. In this structure, the input feature map first passes through a 1×1 convolution + ReLU activation, followed by a 3×3 convolution + ReLU, and finally through another 1×1 convolution to expand the number of channels. A parallel 1×1 convolution is also applied to the shortcut path to match the dimensionality before the element-wise addition of the two paths.

EfficientNet [39] is a high-performance image classification network known for its balanced scaling of network width, depth, and resolution to achieve optimal efficiency. Unlike traditional approaches that enhance performance by simply increasing the number of layers or the width of the network, EfficientNet employs a compound scaling method, which proportionally increases various dimensions of the network according to a fixed scaling factor. This approach enhances the accuracy of the model while maintaining its efficiency and scalability. EfficientNet-B0 serves as the baseline model. Variants such as B1 through B7 are derived by adjusting the scaling coefficients to accommodate different computational resources and task requirements. The network architecture of EfficientNet is realized by stacking multiple MBConv modules. MBConv is an inverted residual bottleneck structure with depth wise separable convolution. As shown in Figure 2, the input feature map first passes through a 1×1 convolutional layer followed by a batch normalization (BN) layer. It then goes through a depth wise convolution with a kernel size of $k \times k$ and a BN layer.

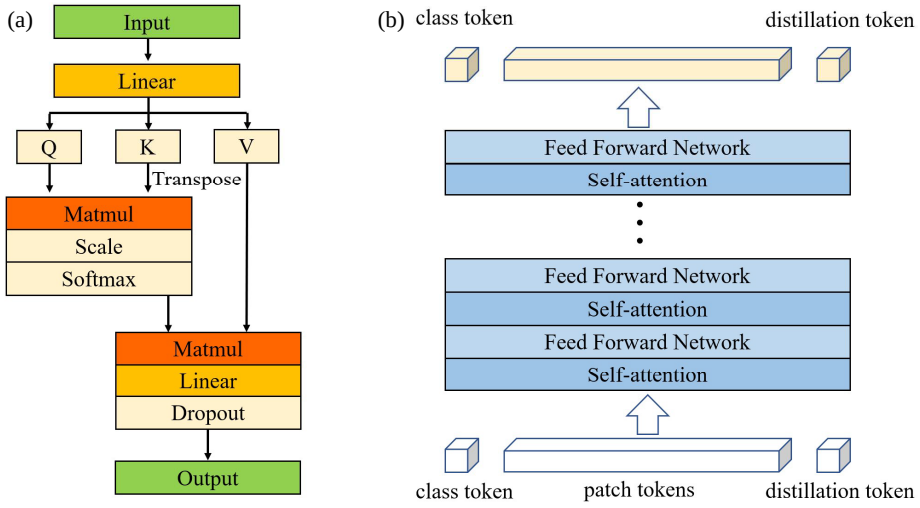

FIGURE 3. Self-attention mechanism and DeiT network architecture. (a) Self-attention. (b) Data-efficient image Transformers (DeiT).

The Squeeze-and-Excitation (SE) module is used to enhance the weights of the channels of interest. After that, the feature map passes through another 1×1 convolutional layer followed by BN and Dropout layers. Finally, the output feature map is added to the input feature map to obtain the final output. The Squeeze-and-Excitation (SE) module is widely used in convolutional neural networks to enhance the weights of the channels of interest. The feature maps input to the SE module are first subjected to average pooling to obtain a one-dimensional vector representing different channels. Subsequently, they pass through a fully connected layer 1, a Swish activation function, a fully connected layer 2, and a Sigmoid function to obtain the weights for each channel. Finally, these weights are multiplied with the feature maps input to the SE module to enhance the weights of the channels of interest.

With the successful application of self-attention mechanisms in natural language processing tasks, the Vision Transformer (ViT) emerged as the first image classification network to rely solely on self-attention, without any convolutional operations. The self-attention computation process is illustrated in Figure 3(a). The input data is projected through linear layers to generate the Query, Key, and Value vectors. The Query is then multiplied by the transposed Key, followed by a scaling operation to stabilize the computation. The result is passed through a softmax function and multiplied by the Value, then further processed through a linear layer and a Dropout layer to produce the final output. While traditional convolutional neural networks (CNNs) can achieve excellent performance with pre-training on the ImageNet dataset (containing 1.3 million images), ViT requires substantially larger datasets — such as JFT-300M, which is approximately 230 times the size of ImageNet — and more training iterations to reach comparable classification performance. Data-efficient image Transformers (DeiT) [40] represent an efficient image classification architecture that introduces a novel knowledge distillation mechanism, marking a significant breakthrough in the application of Transformer-based models to computer vision tasks. As shown in Figure 3(b), during training, a pre-trained teacher network (typically a CNN) is used to guide a smaller student network (DeiT). The teacher's prediction is derived from a distillation token, while the student's classification is determined by a class token. This approach does not introduce additional computational cost and enables DeiT to learn effectively even from small-scale datasets. DeiT partitions the input image into multiple fixed-size patches, which are then converted into embedding vectors via a linear projection (referred to as patch tokens in Figure 3(b)). These patch tokens are passed through a series of self-attention layers and feed-forward networks to extract high-level representations. The final classification result is obtained from the class token. DeiT generally employs multi-head self-attention mechanisms, which effectively capture global dependencies among image patches and allow the network to learn robust and expressive image representations.

3. LOSS FUNCTION AND EVALUATION METRICS

In image classification tasks, cross-entropy loss is the most used loss function. It measures the difference between the predicted probability distribution and the true probability distribution, serving as an indicator of the model's prediction accuracy. A smaller cross-entropy value indicates that the predicted distribution is closer to the true distribution, whereas a larger value suggests greater divergence between the two. Given a true distribution P (the ground truth) and a predicted distribution Q (the model's output), the cross-entropy is defined as:

$$H\left(P,Q\right) = -\sum\nolimits_{i=0} P\left(i\right)\log Q\left(i\right) \tag{1}$$

For binary classification tasks, the cross-entropy loss function can be simplified as:

$$\mathcal{L} = -\left[y\log\hat{y} + \left(1-y\right)\log\left(1-\hat{y}\right)\right] \tag{2}$$

Here, $\mathcal{L}$ denotes the loss; y represents the ground truth label, taking a value of 0 or 1; $\hat{y}$ represents the predicted probability, ranging between (0, 1). The closer $\hat{y}$ is to 1, the more accurate the network's prediction. When using the cross-entropy

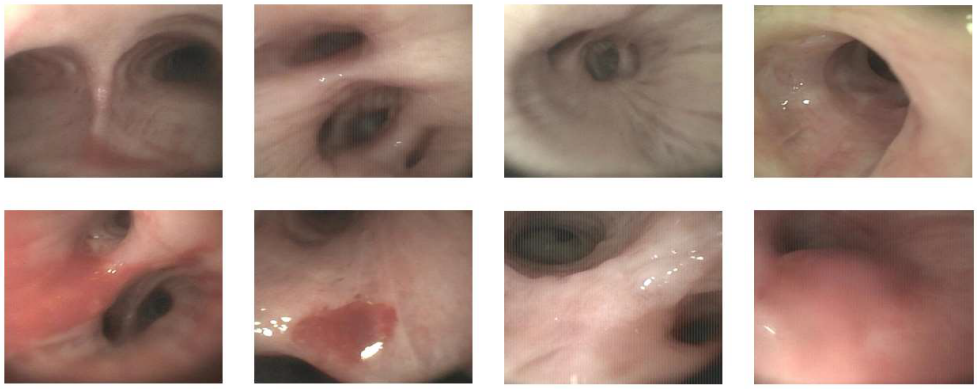

FIGURE 4. Bronchoscopic images of pulmonary mucosa: The top row shows normal images, while the bottom row displays images of pneumonia.

loss function, the neural network is encouraged to learn in a direction that maximizes the distinction between correct and incorrect labels. However, when the training data is limited and cannot sufficiently represent all sample characteristics, this can lead to overfitting. To address this issue, image classification networks based on the self-attention mechanism often adopt label smoothing loss, a regularization strategy. Label smoothing works by applying a "soft" one-hot encoding that introduces noise into the labels, reducing the weight assigned to the true class during loss computation. This technique helps to suppress overfitting by preventing the model from becoming overconfident in its predictions.

In image classification tasks, commonly used evaluation metrics include precision (P), recall (R), F1-score, and overall accuracy [41]. For predicted labels and ground truth labels, TP represents the number of correctly predicted positive samples, FP represents the number of incorrectly predicted positive samples, FN represents the number of incorrectly predicted negative samples, and TN represents the number of correctly predicted negative samples. For binary classification tasks, P, R, F1-score, and overall accuracy can be calculated using Equations (3) to (6).

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1-score} = \frac{2 \times P \times R}{P + R} \tag{5}$$

$$\text{Overall accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

4. EXPERIMENTAL RESULTS

In this experiment, we collected bronchoscopic video data from 25 patients at Taizhou Hospital and Enze Hospital, both affiliated with Taizhou Enze Medical Center in Zhejiang Province. The bronchoscopes used were the PL-F520 and EB260R models, manufactured and commercialized by ourselves (Zhejiang UE Medical Instrument Co., LTD, Zhejiang, China). Initially, the bronchial mucosal video data were preprocessed by extracting 3 frames per second and cropping the edges of the images to obtain clean bronchial mucosal images. Subsequently, a chief physician with over 15 years of clinical experience in respiratory medicine and a chief physician in critical care medicine classified and reviewed the images based on the patients' medical history, clinical symptoms, risk factors, laboratory tests, CT imaging, and bacteriological culture of bronchoalveolar lavage fluid, in accordance with clinical diagnostic criteria. A total of 4045 images were obtained. Figure 4 shows the bronchial mucosal images under the bronchoscope, with the first row displaying normal images and the second row showing images of patients with aspiration pneumonia. In the second row, the first two images of aspiration pneumonia show bleeding, while the last two images show edema.

In Table 1, there are a total of 1627 images of pulmonary mucosa with pneumonia and 2418 images of normal pulmonary mucosa. The dataset was randomly shuffled and divided according to a training-to-validation ratio of 7 : 3. After the split, the training set contains 2830 images, and the validation set contains 1215 images. In the training set, there are 1138 images of pulmonary mucosa with pneumonia and 1692 images of normal pulmonary mucosa. In the test set, there are 489 images of pulmonary mucosa with pneumonia and 726 images of normal pulmonary mucosa. The test set of images were from patients whose images have never been used for training but was solely employed to evaluate and test the model's performance.

TABLE 1. Image distribution of training and validation sets for pneumonia classification.

	training set	test set	total
pneumonia	1138	489	1627
normal	1692	726	2418
total	2830	1215	4045

The hardware experimental setup for training all models is as follows: the GPU used is NVIDIA GeForce RTX 4090 with 24 GB of memory, the CPU is Intel(R) Xeon(R) Silver 4210R with a clock speed of 2.4 GHz and 38 cores, and the memory is

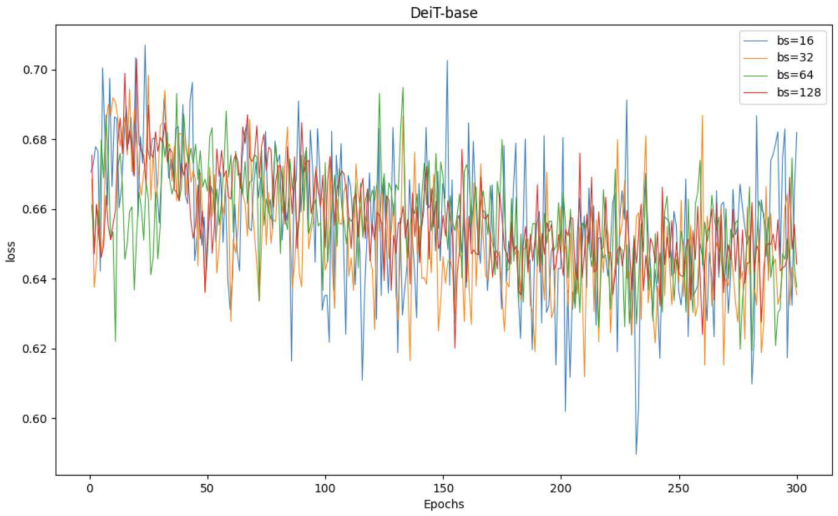

FIGURE 5. Loss curves during the training of DeiT-base with varying batch sizes (denoted as "bs").

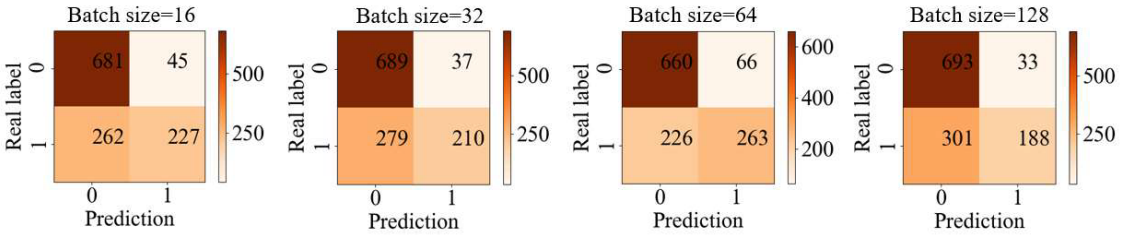

FIGURE 6. Confusion matrices of the optimal DeiT-base models trained with different batch sizes.

TABLE 2. Experimental results of the best-performing DeiT-base models under different batch sizes.

batch size	GPU memory/M	precision/%	recall/%	F1 score/%	overall accuracy/%
16	3958	83.46	46.42	59.66	74.73
32	5996	85.02	42.94	57.07	73.99
64	9558	79.94	**53.78**	**64.30**	**75.97**
128	17032	**85.07**	38.45	52.96	72.51

128 GB. The operating system is Ubuntu version 22.04.5, with Pytorch version 2.3.1 and CUDA version 11.8. The training code utilizes the mmpretrain library [42], version 1.2.0, and the model's hyperparameters are set using the default settings from the mmpretrain library. No pre-trained models were used during training; all models were trained from scratch. The mean and standard deviation of the three RGB channels of the entire dataset were calculated for data normalization, resulting in RGB means of 139.4, 102.87, 95.98, and RGB standard deviations of 47.83, 40.05, 36.73.

Due to the successful application of the self-attention mechanism in natural language processing and large models, this study first utilizes DeiT-base for training and validation. DeiT employs only the self-attention mechanism. During the data preprocessing stage, the original images are randomly scaled and cropped to a size of 224×224, with horizontal flipping applied as a data augmentation technique. The hyperparameters for the training process are as follows: the number of epochs is set to 300, the optimizer for model parameters is the cosine annealing method, the initial learning rate is set to 0.001, and the loss function used is LabelSmoothLoss. To maximize GPU memory utilization and select an optimal batch size, the batch sizes used for training EfficientNet-B2 are 16, 32, 64, 128. The loss curve is shown in Figure 5.

As shown in Figure 5, regardless of the batch size, the DeiT-base model fails to converge during training. This is primarily attributed to the limited size of the training dataset, which contains only 2830 images. As discussed in Section 2, Vision Transformer (ViT)-based models that rely solely on self-attention mechanisms typically require large-scale pretraining on massive datasets to achieve satisfactory performance. Although DeiT introduces a distillation strategy to alleviate this issue, the teacher network was not pretrained in this study. Consequently, training the DeiT-base model from scratch led to non-convergent behavior. Figure 6 illustrates the confusion matrices of the best-performing DeiT-base models trained with

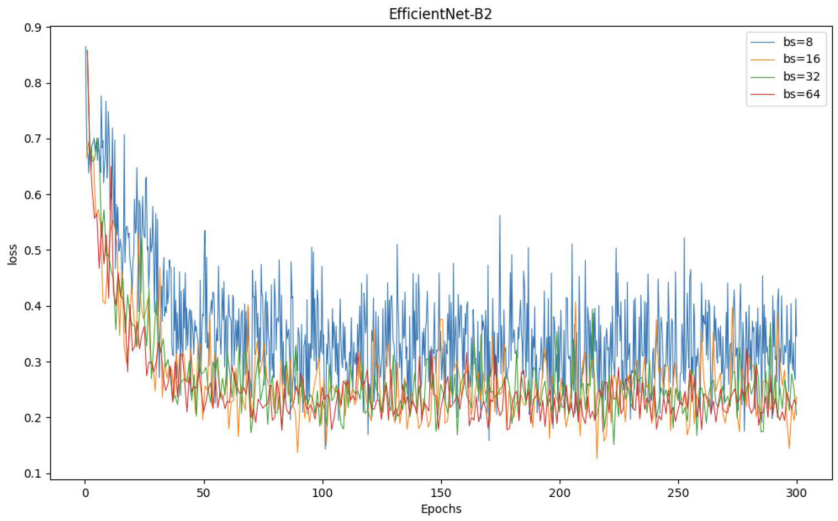

FIGURE 7. Loss curves of EfficientNet-B2 during training under different batch sizes (bs denotes batch size).

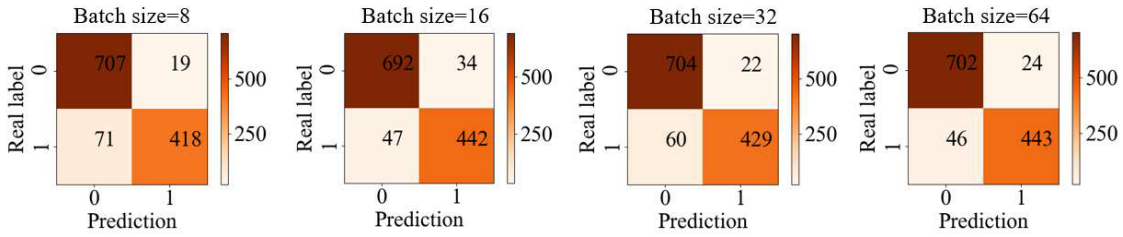

FIGURE 8. Confusion matrices of the best-performing EfficientNet-B2 models under different batch sizes during training.

TABLE 3. Experimental results of the best-performing EfficientNet-B2 models with different batch sizes.

batch size	GPU memory/M	precision/%	recall/%	F1 score/%	overall accuracy/%
8	2892	**95.65**	85.48	90.28	92.59
16	4630	92.86	90.39	91.61	93.33
32	8552	95.12	87.73	91.28	93.25
64	16432	94.86	**90.59**	**92.68**	**94.24**

various batch sizes. Based on the calculated performance metrics presented in Table 2, the optimal diagnostic performance for pneumonia was achieved when the batch size was set to 64, with a precision of 79.94%, recall of 53.78%, F1-score of 64.30%, and an overall accuracy of 75.97%.

EfficientNet-B2 was employed for model training and validation in this study. During data preprocessing, the original bronchoscope images were randomly scaled and cropped to 260×260 pixels, followed by horizontal flipping for data augmentation. The training process utilized SGD as the optimizer, with the number of epochs set to 300, an initial learning rate of 0.1, and the cross-entropy loss function. To maximize GPU memory utilization and explore optimal batch size configurations, batch sizes of 8, 16, 32, and 64 were tested. The corresponding loss curves are illustrated in Figure 7.

Figure 8 illustrates the confusion matrices of the best-performing EfficientNet-B2 models trained with different batch sizes. Based on the results, the evaluation metrics summarized in Table 3 were calculated. When the batch size was set to 64, the model achieved the best diagnostic performance for pneumonia, with a precision of 94.86%, recall of 90.59%, F1 score of 92.68%, and an overall accuracy of 94.24%.

Finally, ResNet-50 was employed for training and validation. During the data preprocessing stage, the original images were randomly scaled and cropped to a size of 224×224, and data augmentation was achieved by performing horizontal flipping. The hyperparameters for the training process were set as follows: the number of epochs was set to 300, the model parameter optimizer was SGD, the initial learning rate was 0.1, and the loss function was the cross-entropy loss function. To maximize the use of GPU memory and to better select the batch size, the batch size used during ResNet-50 training was set to 16, 32, 64, and 128, respectively. The loss curve is shown in Figure 9.

Figure 10 presents the confusion matrices of the optimal training models under different batch sizes during the training process of ResNet-50. After calculation, the evaluation metrics

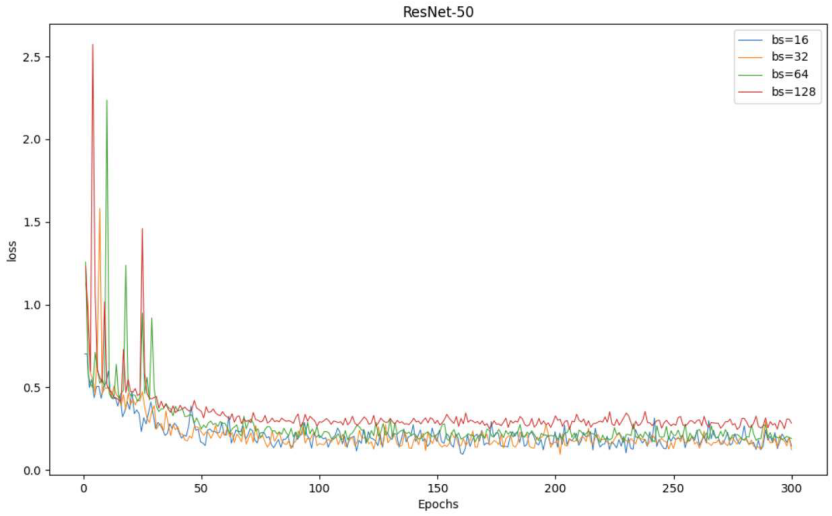

FIGURE 9. The loss curve of ResNet-50 training process, where bs denotes batch size.

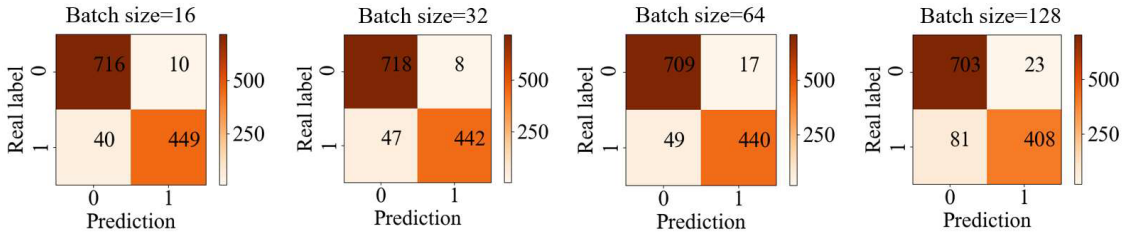

FIGURE 10. The confusion matrix of the optimal model under different batch sizes during the training process of ResNet-50.

TABLE 4. Experimental results of the optimal training models of ResNet-50 under different batch sizes.

batch size	GPU memory/M	precision/%	recall/%	F1 score/%	overall accuracy/%
16	2386M	97.82	**91.82**	**94.73**	**95.88**
32	3674M	**98.22**	90.39	94.14	95.47
64	6852M	96.28	89.98	93.02	94.57
128	13250M	94.66	83.44	88.70	91.44

shown in Table 4 were obtained. When the batch size was set to 16, the results for pneumonia diagnosis were optimal, with a precision of 97.82%, a recall of 91.82%, an F1 score of 94.73%, and an overall accuracy of 95.88%.

5. CONCLUSION

Aspiration pneumonia can be diagnosed through medical history (aspiration events), clinical manifestations (cough, fever), and imaging studies. In this study, we have collected bronchoscope images from 25 patients clinically using a fiberoptic bronchoscope. After data preprocessing and expert annotation, we obtained 2830 images for training and 1215 images for test. The test set was from patients whose images have never been used for training. In the experiments, three deep learning methods were employed: ResNet-50, EfficientNet-B2, and DeiT-base. Among them, ResNet-50 and EfficientNet-B2 are convolution-based networks, while DeiT-base is a network based on the self-attention mechanism. The experimental results showed that ResNet-50 achieved the highest overall accuracy of 95.88% for pneumonia diagnosis. EfficientNet-B2 had an overall accuracy of 94.24% for pneumonia diagnosis-more case images. For image classification algorithms based on the self-attention mechanism, such as DeiT, we need to collect more images and introduce a teacher network to improve the accuracy of pneumonia image classification. The experiments demonstrated that deep learning methods can achieve good results in the automatic identification of aspiration pneumonia from bronchoscope images, which can assist clinicians in making diagnoses. Early detection of aspiration pneumonia can help improve patient outcomes, reduce the medical burden, and enhance the quality of life.

ACKNOWLEDGEMENT

This work was supported by “Pioneer” and “Leading Goose” R&D Program of Zhejiang Province (No. 2023C03083), Science and Technology Plan Key Project of Taizhou City

(24gyz01) and Science and Technology Plan Project of Luqiao (Taizhou) District(2024G2009).

The authors are grateful to Dr. Julian Evans of Zhejiang University for valuable discussions.

REFERENCES

[1] Son, Y. G., J. Shin, and H. G. Ryu, "Pneumonitis and pneumonia after aspiration," *Journal of Dental Anesthesia and Pain Medicine*, Vol. 17, No. 1, 1, 2017.

[2] Yoshimatsu, Y., D. Melgaard, A. Westergren, C. Skrubbeltrang, and D. G. Smithard, "The diagnosis of aspiration pneumonia in older persons: A systematic review," *European Geriatric Medicine*, Vol. 13, No. 5, 1071–1080, 2022.

[3] Torres, A., J. Serra-Batlles, A. Ferrer, P. Jiménez, R. Celis, E. Cobo, and R. Rodriguez-Roisin, "Severe community-acquired pneumonia," *American Review of Respiratory Disease*, Vol. 144, No. 2, 312, 1991.

[4] Moine, P., J.-B. Vercken, S. Chevret, C. Chastang, and P. Gajdos, "Severe community-acquired pneumonia: Etiology, epidemiology, and prognosis factors," *Chest*, Vol. 105, No. 5, 1487–1495, 1994.

[5] Marrie, T. J., H. Durant, and L. Yates, "Community-acquired pneumonia requiring hospitalization: 5-year prospective study," *Reviews of Infectious Diseases*, Vol. 11, No. 4, 586–599, 1989.

[6] Wang, X., Y. Gao, Q. Wang, *et al.*, "Analysis and countermeasures of related factors for aspiration pneumonia in elderly patients," *Chinese Journal of Nosocomiology*, Vol. 24, No. 5, 1161–1162, 2014.

[7] Mandell, L. A. and M. S. Niederman, "Aspiration pneumonia," *New England Journal of Medicine*, Vol. 380, No. 7, 651–663, 2019.

[8] Marik, P. E., "Aspiration pneumonitis and aspiration pneumonia," *New England Journal of Medicine*, Vol. 344, No. 9, 665–671, 2001.

[9] Miyashita, N., Y. Kawai, T. Tanaka, H. Akaike, H. Teranishi, T. Wakabayashi, T. Nakano, K. Ouchi, and N. Okimoto, "Detection failure rate of chest radiography for the identification of nursing and healthcare-associated pneumonia," *Journal of Infection and Chemotherapy*, Vol. 21, No. 7, 492–496, 2015.

[10] Komiya, K., H. Ishii, K. Umeki, T. Kawamura, F. Okada, E. Okabe, J. Murakami, Y. Kato, B. Matsumoto, S. Teramoto, *et al.*, "Computed tomography findings of aspiration pneumonia in 53 patients," *Geriatrics & Gerontology International*, Vol. 13, No. 3, 580–585, 2013.

[11] Shan, K., D. Jia, and W. Guo, "Diagnosis of stroke-associated pneumonia: Expert consensus of the stroke-associated pneumonia study group," *Chinese Journal of Emergency Medicine*, Vol. 24, No. 12, 1346–1348, 2015.

[12] Darie, A. M. and D. Stolz, "Is there a role for bronchoscopy in aspiration pneumonia?" *Seminars in Respiratory and Critical Care Medicine*, Vol. 45, No. 6, 650–658, 2024.

[13] Maarel-Wierink, Van der C. D., J. N. O. Vanobbergen, E. M. Bronkhorst, J. M. G. A. Schols, and C. de Baat, "Meta-analysis of dysphagia and aspiration pneumonia in frail elders," *Journal of Dental Research*, Vol. 90, No. 12, 1398–1404, 2011.

[14] DiBardino, D. M. and R. G. Wunderink, "Aspiration pneumonia: A review of modern trends," *Journal of Critical Care*, Vol. 30, No. 1, 40–48, 2015.

[15] Woodhead, M., F. Blasi, S. Ewig, J. Garau, G. Huchon, M. Ieven, A. Ortqvist, T. Schaberg, A. Torres, G. Van Der Heijden, *et al.*, "Guidelines for the management of adult lower respiratory tract infections — Full version," *Clinical Microbiology and Infection*, Vol. 17, No. Suppl. 6, E1–E59, 2011.

[16] Faverio, P., S. Aliberti, G. Bellelli, G. Suigo, S. Lonni, A. Pesci, and M. I. Restrepo, "The management of community-acquired pneumonia in the elderly," *European Journal of Internal Medicine*, Vol. 25, No. 4, 312–319, 2014.

[17] Pace, C. C. and G. H. McCullough, "The association between oral microorgansims and aspiration pneumonia in the institutionalized elderly: Review and recommendations," *Dysphagia*, Vol. 25, 307–322, 2010.

[18] She, J., J. Ding, and J. Shen, "Expert recommendations on the diagnosis and treatment of aspiration pneumonia in adults," *International Journal of Respiratory*, Vol. 42, No. 2, 86–96, 2022.

[19] Almirall, J., R. Boixeda, M. C. de la Torre, and A. Torres, "Aspiration pneumonia: A renewed perspective and practical approach," *Respiratory Medicine*, Vol. 185, 106485, 2021.

[20] El-Solh, A. A., H. Vora, P. R. K. III, and J. Porhomayon, "Diagnostic use of serum procalcitonin levels in pulmonary aspiration syndromes," *Critical Care Medicine*, Vol. 39, No. 6, 1251–1256, 2011.

[21] Suzuki, T., M. Saitou, Y. Utano, K. Utano, and K. Niitsuma, "Bronchoalveolar lavage (BAL) amylase and pepsin levels as potential biomarkers of aspiration pneumonia," *Pulmonology*, Vol. 29, No. 5, 392–398, 2023.

[22] Metlay, J. P., G. W. Waterer, A. C. Long, A. Anzueto, J. Brozek, K. Crothers, L. A. Cooley, N. C. Dean, M. J. Fine, S. A. Flanders, *et al.*, "Diagnosis and treatment of adults with community-acquired pneumonia. An official clinical practice guideline of the American Thoracic Society and Infectious Diseases Society of America," *American Journal of Respiratory and Critical Care Medicine*, Vol. 200, No. 7, e45–e67, 2019.

[23] Simpson, A. J., J.-L. Allen, M. Chatwin, H. Crawford, J. Elverson, V. Ewan, J. Forton, R. McMullan, J. Plevris, K. Renton, *et al.*, "BTS clinical statement on aspiration pneumonia," *Thorax*, Vol. 78, No. Suppl. 1, S3–S21, 2023.

[24] Delforge, Q., A. Gaudet, P. Boddaert, F. Wallet, B. Voisin, and S. Nseir, "Accuracy of the Infectious Diseases Society of America and British Thoracic Society criteria for acute pneumonia in differentiating chemical and bacterial complications of aspiration in comatose ventilated patients following drug poisoning," *Antibiotics*, Vol. 13, No. 6, 495, 2024.

[25] Zhu, H., J. Luo, J. Liao, and S. He, "High-accuracy rapid identification and classification of mixed bacteria using hyperspectral transmission microscopic imaging and machine learning," *Progress In Electromagnetics Research*, Vol. 178, 49–62, 2023.

[26] Weng, D., S. Dou, H. Wang, D. Gong, Q. Wang, and S. He, "Infrared image segmentation method based on DeepLabV3+ for identifying key components of power transmission line," *Progress In Electromagnetics Research C*, Vol. 138, 191–203, 2023.

[27] Xu, Z., Y. Jiang, J. Ji, E. Forsberg, Y. Li, and S. He, "Classification, identification, and growth stage estimation of microalgae based on transmission hyperspectral microscopic imaging and machine learning," *Optics Express*, Vol. 28, No. 21, 30 686–30 700, 2020.

[28] Zeng, Z., Q. Huang, and S. He, "Ai-based fast design for general fiber-to-waveguide grating couplers," *Progress In Electromagnetics Research M*, Vol. 119, 143–160, 2023.

[29] Almirall, J., L. Rofes, M. Serra-Prat, R. Icart, E. Palomera, V. Arreola, and P. Clavé, "Oropharyngeal dysphagia is a risk factor for community-acquired pneumonia in the elderly," *European Respiratory Journal*, Vol. 41, No. 4, 923–928, 2013.

[30] Zhuang, B., W. Zheng, and M. Zhang, "Construction of a prediction model for aspiration pneumonia in head and neck cancer patients receiving radiotherapy based on machine learning," *China Modern Medicine*, Vol. 26, No. 9, 56–61, 2024.

[31] Sejdić, E., C. M. Steele, and T. Chau, "Classification of penetration-aspiration versus healthy swallows using dual-axis swallowing accelerometry signals in dysphagic subjects," *IEEE Transactions on Biomedical Engineering*, Vol. 60, No. 7, 1859–1866, 2013.

[32] Weng, W., M. Imaizumi, S. Murono, and X. Zhu, "Expert-level aspiration and penetration detection during flexible endoscopic evaluation of swallowing with artificial intelligence-assisted diagnosis," *Scientific Reports*, Vol. 12, No. 1, 21689, 2022.

[33] Imaizumi, M., W. Weng, X. Zhu, and S. Murono, "Effectiveness of FEES with artificial intelligence-assisted computer-aided diagnosis," *Auris Nasus Larynx*, Vol. 51, No. 2, 251–258, 2024.

[34] Miura, Y., G. Nakagami, K. Yabunaka, H. Tohara, R. Murayama, H. Noguchi, T. Mori, and H. Sanada, "Method for detecting aspiration based on image processing-assisted B-mode video ultrasonography," *J. Nurs. Sci. Eng.*, Vol. 1, 2–20, 2014.

[35] Sarraf Shirazi, S., C. Buchel, R. Daun, L. Lenton, and Z. Moussavi, "Detection of swallows with silent aspiration using swallowing and breath sound analysis," *Medical & Biological Engineering & Computing*, Vol. 50, 1261–1268, 2012.

[36] Sarraf Shirazi, S., A. H. Birjandi, and Z. Moussavi, "Noninvasive and automatic diagnosis of patients at high risk of swallowing aspiration," *Medical & Biological Engineering & Computing*, Vol. 52, 459–465, 2014.

[37] Lee, J. T. and E. Park, "Detection of the pharyngeal phase in the videofluoroscopic swallowing study using inflated 3D convolutional networks," in *9th International Workshop on Machine Learning in Medical Imaging, MLMI 2018*, 328–336, Granada, Spain, Sep. 2018.

[38] He, K., X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778, Las Vegas, NV, USA, 2016.

[39] Tan, M. and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, 6105–6114, Long Beach, USA, Jun. 2019.

[40] Touvron, H., M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, 10 347–10 357, online, Jul. 2021.

[41] Farooq, S. and D. M. Zezell, "Diabetes monitoring through urine analysis using ATR-FTIR spectroscopy and machine learning," *Chemosensors*, Vol. 11, No. 11, 565, 2023.

[42] MMPreTrain Contributors, "Openmmlab's pre-training toolbox and benchmark," https://github.com/open-mmlab/mmpretrain, 2023.