






Article

Diagnosis of Lung Cancer Using Endobronchial Ultrasonography Image Based on Multi-Scale Image and Multi-Feature Fusion Framework

Huitao Wang ¹, Takahiro Nakajima ², Kohei Shikano ³, Yukihiro Nomura ⁴ and Toshiya Nakaguchi ^{4,*}

¹ Department of Medical Engineering, Graduate School of Science and Engineering, Chiba University, Chiba 263-8522, Japan; wanghuitao@chiba-u.jp

² Department of General Thoracic Surgery, Dokkyo Medical University, Mibu 321-0293, Japan; t-nakajima@dokkyomed.ac.jp

³ Department of Respiriology, Graduate School of Medicine, Chiba University, Chiba 260-8670, Japan; shika.v.a.r.1107@gmail.com

⁴ Center for Frontier Medical Engineering, Chiba University, Chiba 263-8522, Japan; ynomura@chiba-u.jp

* Correspondence: nakaguchi@faculty.chiba-u.jp; Tel.: +81-43-290-3296

Abstract: Lung cancer is the leading cause of cancer-related deaths globally and ranks among the most common cancer types. Given its low overall five-year survival rate, early diagnosis and timely treatment are essential to improving patient outcomes. In recent years, advances in computer technology have enabled artificial intelligence to make groundbreaking progress in imaging-based lung cancer diagnosis. The primary aim of this study is to develop a computer-aided diagnosis (CAD) system for lung cancer using endobronchial ultrasonography (EBUS) images and deep learning algorithms to facilitate early detection and improve patient survival rates. We propose M³-Net, which is a multi-branch framework that integrates multiple features through an attention-based mechanism, enhancing diagnostic performance by providing more comprehensive information for lung cancer assessment. The framework was validated on a dataset of 95 patient cases, including 13 benign and 82 malignant cases. The dataset comprises 1140 EBUS images, with 540 images used for training, and 300 images each for the validation and test sets. The evaluation yielded the following results: accuracy of 0.76, F1-score of 0.75, AUC of 0.83, PPV of 0.80, NPV of 0.75, sensitivity of 0.72, and specificity of 0.80. These findings indicate that the proposed attention-based multi-feature fusion framework holds significant potential in assisting with lung cancer diagnosis.

Keywords: lung cancer; endobronchial ultrasonography (EBUS); deep learning; convolutional neural network; multi-scale image; multi-feature fusion



Academic Editor: Emilio Quaia

Received: 16 December 2024

Revised: 23 February 2025

Accepted: 24 February 2025

Published: 27 February 2025

Citation: Wang, H.; Nakajima, T.; Shikano, K.; Nomura, Y.; Nakaguchi, T. Diagnosis of Lung Cancer Using Endobronchial Ultrasonography Image Based on Multi-Scale Image and Multi-Feature Fusion Framework. *Tomography* **2025**, *11*, 24.

<https://doi.org/10.3390/tomography11030024>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide [1–3]. Most lung cancer deaths occur because the disease is typically diagnosed at an advanced stage [3], by which point the optimal window for treatment has already passed. This delay in detection is a key factor contributing to lung cancer's exceptionally high mortality rate. Consequently, early diagnosis and screening are critical, playing a pivotal role in improving overall survival rates for patients [4]. Lung cancer is broadly categorized into two types: small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC) [5]. SCLC is characterized by rapid tumor growth, a short doubling time, and early metastasis, making early screening, diagnosis, and treatment crucial for improving prognosis. Non-small-cell lung cancer

accounts for approximately 85% of all lung cancer cases. Each type often requires distinct treatment approaches. Traditional diagnostic methods for lung cancer primarily rely on medical imaging, such as magnetic resonance imaging (MRI) [6,7], computed tomography (CT) [8], dual-modality positron emission tomographic (PET)–CT imaging [9], and histopathology [10,11]. However, the differences in imaging techniques mean that each method serves unique diagnostic purposes and requires specific interpretation methods [5]. Clinically, doctors typically diagnose lung cancer based on medical imaging, considering factors such as tumor size and location. However, visual inspection is subjective, and traditional imaging-based diagnoses often lead to missed cases. This can result in delayed treatment, causing patients to miss the optimal treatment window [12].

Commonly used medical imaging techniques for lung cancer screening include CT, MRI, and tissue histopathology images. Studies have shown that low-dose CT [8,13,14] is an effective method for the early diagnosis of lung cancer and can significantly improve patient survival rates. However, its drawbacks include radiation exposure and a relatively high rate of false positives in the diagnostic results [8]. In the diagnosis of brain metastases from NSCLC, MRI has shown a higher diagnostic rate compared to CT [15]. However, this has not translated into improved patient survival rates. This highlights the significant challenges that remain in using CT or MRI for early lung cancer diagnosis. The gold standard for lung cancer diagnosis remains pathological evaluation, where tumor tissue is obtained through invasive procedures like biopsies to determine the tumor's histology and stage. However, biopsy procedures have limitations and associated risks, making them less ideal for certain patients [16]. For example, tissue biopsy procedures may carry the risk of complications, such as infection, bleeding, or damage to surrounding organs [17].

In recent years, with the rapid advancement of computer technology, cancer diagnostic systems based on artificial intelligence (AI) algorithms have begun to gain prominence in medical image diagnosis. These systems have shown great potential in various areas, such as skin disease diagnosis [18,19], heart disease diagnosis based on ECG [20,21], eye disease diagnosis using fundus images [22], nodule screening from CT images [23,24], and brain tumor diagnosis from MRI [25]. AI has also played a crucial role in lung cancer diagnosis [26], demonstrating significant promise in improving diagnostic accuracy and efficiency.

Raza et al. [5] introduced a CT scan image-based lung cancer classification network called Lung-EffNet, which employs transfer learning for model training. Results indicate that this framework outperforms other classification models in terms of both time complexity and diagnostic performance for lung cancer, achieving strong results even on imbalanced datasets. Teramoto et al. [27] designed a lung cancer diagnostic support system based on a deep convolutional neural network, achieving accuracy comparable to that of pathologists. In their experiments, cytological image data were used to train the network. Due to the large size of the $\times 40$ microscope images, they first cropped the images into 768×768 pixel squares, and then downsample them to 256×256 pixels. These adjusted cell images were then used for training the deep learning network. Ardila et al. [28] developed an end-to-end CT-based lung cancer diagnostic system with the key advantage of allowing direct input of CT data to both locate lesions and diagnose lung cancer. This end-to-end system first trains a cancer ROI detection model to identify three-dimensional (3D) candidate cancer regions within the CT volume. A 3D CNN-based diagnostic network is then trained to assess these ROI regions for cancer. The system achieved diagnostic performance comparable to that of radiologists, significantly improving the accuracy of lung cancer screening.

In addition to deep learning models, machine learning models are also widely applied in lung cancer detection and classification. Singh et al. [29] proposed a supervised machine

learning algorithm for classifying CT scan images. Before training the model, they extracted texture and statistical features from the images, using these features to train the machine learning algorithms. Seven different classifiers were tested, and experiments showed that, compared to the other six algorithms, the multilayer perceptron (MLP) achieved the best performance in lung cancer detection. Radiomics is also a crucial tool in medical image processing, allowing high-throughput feature extraction to uncover detailed information from medical images. This approach provides valuable decision support for medical image analysis by revealing additional insights beyond what is visible to the human eye [30]. Hyun et al. [31] explored the use of PET-based radiomics features combined with machine learning algorithms to classify lung cancer subtypes. Experimental results indicated that PET-based radiomics features can aid clinicians in accurately diagnosing lung cancer subtypes, supporting improved clinical decision-making.

However, classifying lung cancer based on CT, MRI, PET, or PET-CT imaging usually requires additional confirmation through cytology or histology [32]. Endobronchial ultrasound (EBUS) is a relatively new bronchoscopic method that plays a significant role in lung cancer diagnosis and staging. EBUS is particularly effective for guiding biopsies or performing transbronchial needle aspiration, offering a higher diagnostic accuracy compared to traditional bronchoscopy [32]. Given the advantages of endobronchial ultrasound (EBUS), including its minimally invasive nature, real-time imaging, lack of radiation [32,33], and high detection rate, AI-assisted diagnostic systems using EBUS images have been proposed to support physicians in diagnosing lung cancer. Li et al. [34] proposed a multimodal fusion framework based on convex probe endobronchial ultrasound (CP-EBUS) for lung cancer diagnosis. This framework utilizes CP-EBUS images from three modes and constructs a three-branch multi-feature fusion module. Each branch is an improved version of SqueezeNet [35], enhanced with a squeeze-and-excitation (SE) attention [36] mechanism to extract more effective features. Finally, the three different feature sets are fused using a feature weighting approach [37]. Chen et al. [33] proposed a CNN-SVM hybrid framework, where a CNN model is used as a feature extractor to generate image features, which are then used to train an SVM classifier. This approach achieved an accuracy of 85%. Research has shown that AI-based diagnostic systems using EBUS images are highly effective for lung cancer detection. However, in cases of highly imbalanced data, even deep learning algorithms may not yield optimal results. This has been confirmed in our previous studies [38]. According to the experimental results, even with data augmentation, the CNN model attained an accuracy of merely 0.63 on the imbalanced dataset. This outcome suggests that data imbalance hinders the CNN model's ability to effectively learn and extract pertinent features from the data.

This study proposes a new multi-feature fusion network, called M³-Net. In the name "M³-Net", each "M" represents a key concept in the model: the first "M" stands for the multi-branch framework, the second "M" represents multi-feature fusion, and the third "M" refers to multi-scale images. The main idea behind the network is to construct a multi-branch framework, where each branch uses a different EBUS image. The features from the multiple branches are then fused through a feature fusion module. Finally, a classifier is constructed, and the fused features are fed into the classifier layer to output the predicted results. In summary, the main contributions of this paper are as follows:

- We developed a novel multi-feature fusion framework M³-Net for lung cancer diagnosis using EBUS images.
- We propose a method for generating multi-scale EBUS images from a single EBUS image.
- We constructed three attention-based multi-feature fusion modules to enhance the effectiveness of feature fusion.

- To address the impact of data imbalance, we customized the loss function weights, training the model with a weighted loss function to mitigate the influence of data imbalance on the experimental results.

The structure of this paper is organized as follows: Section 2 introduces the materials and methods used, providing a detailed description of data acquisition, data preprocessing, multi-scale image generation, the M³-Net framework, and the construction of feature fusion modules. Section 3 presents the experiments and results, detailing the experimental setup, evaluation metrics, and classification results. Section 4 discusses the experimental results. Section 5 presents the conclusions drawn from this study.

2. Materials and Methods

This section provides a detailed description of the processes involved in data collection, preprocessing, and dataset construction for training the lung cancer classification model using EBUS images. After data collection, initial preprocessing is applied to the raw data, as they cannot be directly used for model training. Additionally, due to the limited amount of data and significant class imbalance, data imbalance poses a challenge. To address this, we employ both data augmentation techniques and a weighted loss function in model training. The proposed M³-Net integrates multiple features within a multi-branch framework, offering more efficiency than single-feature approaches. Details on preprocessing and the M³-Net framework will be discussed in the following sections.

2.1. Data Acquisition

The EBUS images used in this study were collected at Chiba University Hospital. This study was conducted under the Declaration of Helsinki and approved by the Institutional Ethics Committee of Chiba University Graduate School of Medicine (M10393). The data include cases from two periods, September 2019 to April 2020, and January 2022 to May 2022, involving patients who underwent Endobronchial Ultrasound-guided Transbronchial Lung Biopsy (EBUS-TBLB) examination. The data collection was conducted using a bronchial ultrasound system (EU-M30; Olympus Co., Tokyo, Japan) equipped with a 20 MHz radial probe (UM-S20-20R; Olympus Co., Tokyo, Japan). Under the guidance of the ultrasound probe, radiologists progressively approached the lesion area, taking sufficient samples for biopsy preparation. The process of the EBUS probe approaching the lesion area was recorded as a MOV video format on the monitor. Each case corresponds to a video file, from which the radiologist removes frames without lesion information, keeping only the relevant frames for model training. Additionally, the label information for each case was determined based on the biopsy results.

2.2. Data Preprocessing

After thorough selection and screening by radiologists, a total of 95 lung tumor cases were included in this study. Of these, 13 were benign cases, while the remaining 82 were malignant cases. The EBUS case data used in this study were selected through a rigorous screening process by radiologists. The selection criteria applied during the screening are as follows: First, only cases with clearly visible lesion areas and minimal artifacts were selected. Second, all cases included in this study had prior informed consent from the patients. Lastly, cases without diagnostic results were excluded. For each case, a MOV-format video file was generated, which was reviewed frame by frame by the radiologists. Only the frames containing lesion information were retained. This study focuses on developing a lung cancer CAD system based on individual EBUS images. After obtaining MOV video files, which have been pre-screened by radiologists, we first convert the video into individual EBUS image frames. Since the relevant information about the lesion is typically located

around the ultrasound probe in the EBUS images, and the images also contain surrounding text, it is necessary to perform a central crop on the original images. This removes irrelevant noise outside the central lesion area. In our previous research (under review), we proposed a scale line-based image cropping (SLBIC) strategy to achieve this [39]. The SLBIC strategy is inspired by the scale markings present in the original image. Using these markings, we determine the coordinates along the image's diagonal and use the ultrasound probe as the center for cropping, which addresses variations in central crop sizes in the EBUS images. By identifying the top left and bottom right coordinates for the crop area, we standardize the size of the cropped images. Through the SLBIC cropping strategy, we ultimately obtain four EBUS image blocks of different sizes: 4×4 cm (Type 1), 3×3 cm (Type 2), 2×2 cm (Type 3), and 1×1 cm (Type 4).

In addition, this study randomly divided the 95 lung tumor cases into training, validation, and test sets, maintaining the same proportion for each category. Table 1 shows the number of benign and malignant cases included in each set (training, validation, and test sets). Additionally, it describes how many EBUS images were selected from the cases in each data subset.

Table 1. The table shows the dataset division, detailing how many benign and malignant cases were included in each subset. It also indicates the number of EBUS images selected from the included cases.

	Benign		Malignant	
	Image Nums	Case Nums	Image Nums	Case Nums
Train	270	6	270	40
Valid	150	3	150	21
Test	150	4	150	21
Overall	570	13	570	82

2.3. Multi-Scale Image Generation

In the image preprocessing stage, four different sizes of EBUS images were obtained based on the SLBIC image cropping strategy. Additionally, the previously proposed image transformation method called polar coordinate transformation was introduced [39]. Previous studies have shown that training the model with transformed EBUS images can significantly improve the experimental results, so we decided to adopt this approach. Using this method, the four differently sized EBUS images were converted based on the coordinate system and uniformly resized to 224×224 pixels. The transformed images were named polar type 1, polar type 2, polar type 3, and polar type 4, respectively. This resulted in a total of eight EBUS images: four obtained through SLBIC cropping and the other four through the coordinate system-based image conversion method. Using SLBIC cropping, four types of EBUS images were obtained. Based on the multi-scale image generation method, three multi-scale images were generated: MS (multi-scale) 1-2-3, MS 1-2-4, and MS 2-3-4. Additionally, according to the four polar EBUS image types, three multi-scale images were also generated: PMS (polar multi-scale) 1-2-3, PMS 1-2-4, and PMS 2-3-4.

By placing different EBUS images into the R channel, G channel, and B channel, a composite image containing information from three types of EBUS images is formed. This idea also serves as the inspiration for creating multi-scale images. The process of generating multi-scale images is illustrated in Figure 1. Similar research on generating multi-scale images can be found in the study by Bazargani et al. [40]. In their work, they obtain case images at different resolutions, each corresponding to images of different scales. The detailed process for generating multi-scale EBUS images is as follows: First, three EBUS images are randomly selected from the four available types. Each selected image is then

converted from an RGB format to gray-scale and resized to 224×224 pixels. Finally, these three single-channel EBUS images are combined into a single multi-scale image by merging them across the RGB channels.

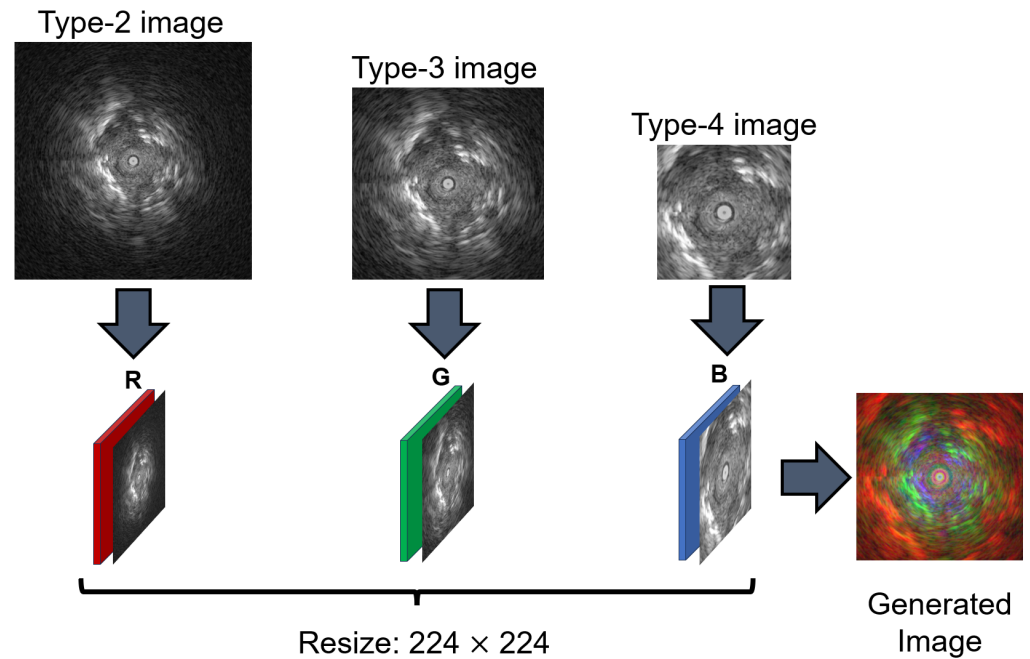


Figure 1. This figure illustrates the detailed process for generating multi-scale EBUS images: First, three EBUS images are selected from the four available types. Each selected EBUS image is then converted from RGB to gray-scale and resized to 224×224 pixels. Finally, the three resized gray-scale images are merged to form a single multi-scale RGB image. This approach utilizes varying image resolutions to effectively capture multi-scale features within one composite image.

Table 2 presents the combinations of multi-scale images generated using the four types of original EBUS images. Table 3 shows the combinations derived from the EBUS images transformed using the coordinate system-based image conversion method. In this study, six different types of multi-scale images were created: three based on the original EBUS images and another three derived from the coordinate system-based image conversion method.

Table 2. This table displays the various combinations of multi-scale (MS) images generated using the four types of EBUS images. Each combination represents different sets of three EBUS images selected from the four available types, merged into a single RGB image to capture multi-scale features from diverse resolutions. This approach aims to enhance the model's ability to detect patterns across scales, improving classification effectiveness.

Generated Multi-Scale Image	Type 1	Type 2	Type 3	Type 4
MS 1-2-3	Used	Used	Used	-
MS 1-2-4	Used	Used	-	Used
MS 2-3-4	-	Used	Used	Used

Table 3. This table presents the polar multi-scale (PMS) image combinations generated from four types of EBUS images using the coordinate system-based image conversion method. PMS stands for polar multi-scale. Each combination represents a selection of three different EBUS image types from the four available, which are then merged into a single RGB image. This configuration is designed to capture multi-scale features across various resolutions, enhancing the model’s ability to recognize patterns across scales and thus improving classification performance.

Generated Multi-Scale Image	Polar Type 1	Polar Type 2	Polar Type 3	Polar Type 4
PMS 1-2-3	Used	Used	Used	-
PMS 1-2-4	Used	Used	-	Used
PMS 2-3-4	-	Used	Used	Used

2.4. M^3 -Net: Multi-Feature Fusion Framework

The proposed framework for lung cancer diagnosis is illustrated in Figure 2. Our M^3 -Net model is designed with three primary components. **Feature extraction module:** M^3 -Net follows a multi-branch architecture, where each branch is dedicated to a specific type of EBUS image. Each EBUS image type is processed by a selected CNN encoder that extracts feature maps (F_1 , F_2 , and F_3) specific to that image type. The choice of CNN encoder is data-driven, based on experimental results detailed in Section 3.3. **Feature fusion module:** This module efficiently combines features from the different EBUS images, leveraging an attention mechanism to assign higher weights to features that contain critical information. This enables the model to prioritize the most relevant features. In this study, we experimented with three different fusion module designs, which are described in detail in Section 2.5. **Classifier module:** After feature fusion, the combined features are sent to the classifier module for the final prediction. The classifier consists of three main layers: an adaptive average pooling layer, which adjusts feature maps for connection to subsequent fully connected layers; a batch normalization layer, which enhances model stability through standardization; and a fully connected layer, which outputs the predicted probabilities.

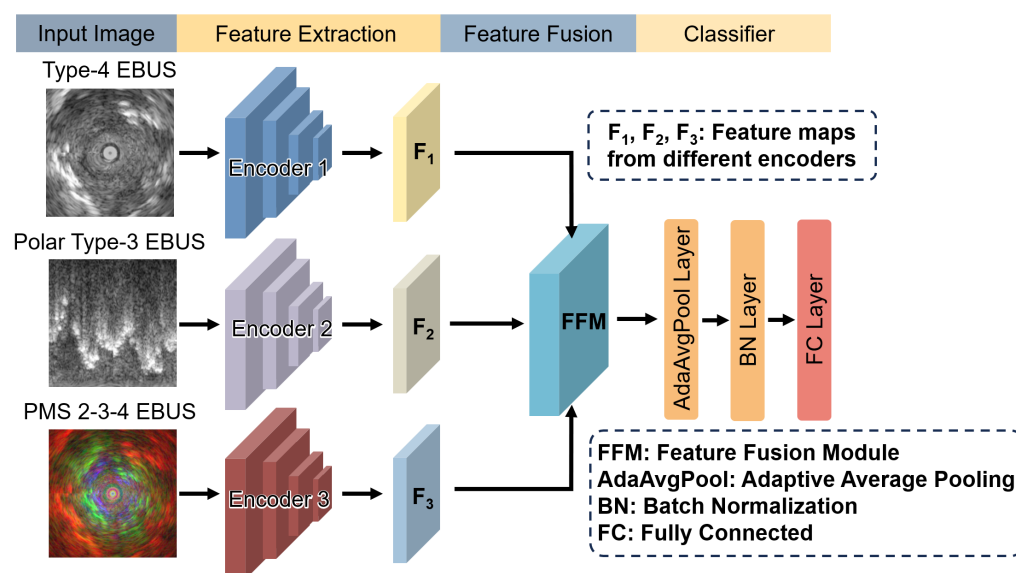


Figure 2. This figure illustrates the architecture of our proposed M^3 -Net, which is composed of three key components: a feature extraction module, a feature fusion module, and a classifier. Each branch in the M^3 -Net architecture corresponds to a specific type of EBUS image, with each image type processed

by an individual CNN encoder for feature extraction. The extracted features are then sent to the feature fusion module, where they are combined. Finally, the fused features are passed into the classifier to generate the final class prediction. The feature extraction module extracts distinct features, F_1 , F_2 , and F_3 , from different input images. These extracted features are then processed through an attention-based feature fusion module, where they are integrated. Finally, the fused features are passed into the classifier to generate probability predictions for the target categories.

2.5. Feature Fusion Modules

To enhance multi-feature fusion, we designed three distinct attention-based fusion modules, each aimed at effectively integrating multiple feature types. Generally, compared to traditional single-feature CNN models, multi-feature fusion models offer improved classification performance [41], which is the core principle behind the M^3 -Net framework. This approach allows M^3 -Net to leverage complementary information across different feature sets, contributing to a more robust and accurate diagnostic model. Additionally, numerous studies have explored the development of attention-based feature fusion modules [42–45], as attention mechanisms assign greater weight to key features, enabling the network to focus more effectively on valuable information. Next, the details of each feature fusion module will be described in detail.

The specifics of feature fusion module version 1 (FFM-v1) are shown in Figure 3. Here, F_1 , F_2 , and F_3 represent feature maps extracted from different encoders. Each of these feature maps is then processed through a CBAM (Convolutional Block Attention Module) attention [46] layer, which assigns weights to the feature maps based on their significance. A residual connection [47] is incorporated into the attention module for two main reasons: to mitigate the gradient vanishing problem as the network depth increases and to ensure that information from previous layers is effectively transferred to subsequent layers without loss due to the introduction of attention. After weighting, we obtain feature maps F'_1 , F'_2 , and F'_3 . These are then concatenated into a unified feature map. This concatenated feature map is subsequently fed into a multi-head self-attention module [48], selected for its ability to allow the model to focus on information from multiple subspaces at various positions simultaneously. The final output is the fused feature map F'' .

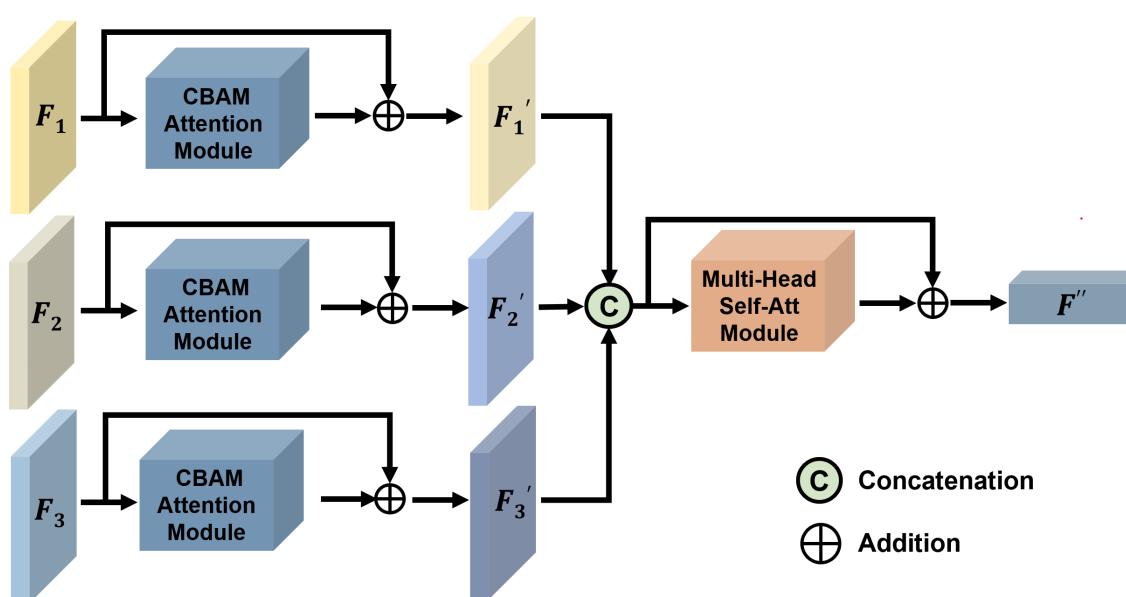


Figure 3. This figure illustrates the framework of feature fusion module version 1 (FFM-v1).

The details of feature fusion module version 2 (FFM-v2) are illustrated in Figure 4. FFM-v2 is the simplest among the three fusion modules. It builds upon the concatenation

of features from different branches by incorporating a multi-head self-attention mechanism. Feature maps F_1 , F_2 , and F_3 are extracted from different encoders and concatenated. This concatenated feature map is then used as the input to the multi-head self-attention mechanism, where attention weights are applied. Finally, a residual connection merges the pre-attention input feature map with the post-attention feature map through an addition operation, resulting in the final feature map F' .

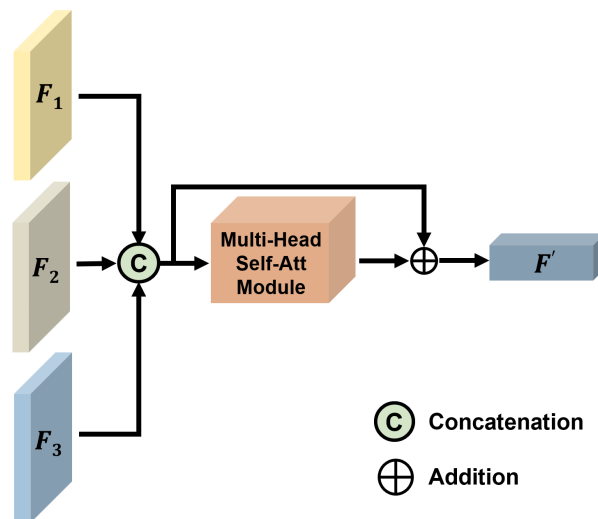


Figure 4. This figure illustrates the framework of feature fusion module version 2 (FFM-v2).

The details of feature fusion module version 3 (FFM-v3) are shown in Figure 5. The concept behind FFM-v3 is to effectively integrate feature maps with fully connected (FC) layer features. Feature maps generally retain local spatial information of the image, while FC layer features represent global information. This fusion enables the model to capture both local and global image features, enhancing overall performance. The detailed process is as follows: F_{1-FM} and F_{1-FL} are derived from Encoder 1, F_{2-FM} and F_{2-FL} from Encoder 2, and F_{3-FM} and F_{3-FL} from Encoder 3. Here, FM denotes feature map features, and FC denotes fully connected layer features. The handling of both types of features is generally similar. First, the features are concatenated and fed into a multi-head self-attention module. Then, a residual connection merges the concatenated features with the output from the attention module. Finally, another concatenation operation fuses the processed feature map features F'_1 with the processed FC layer features F'_2 after attention processing.

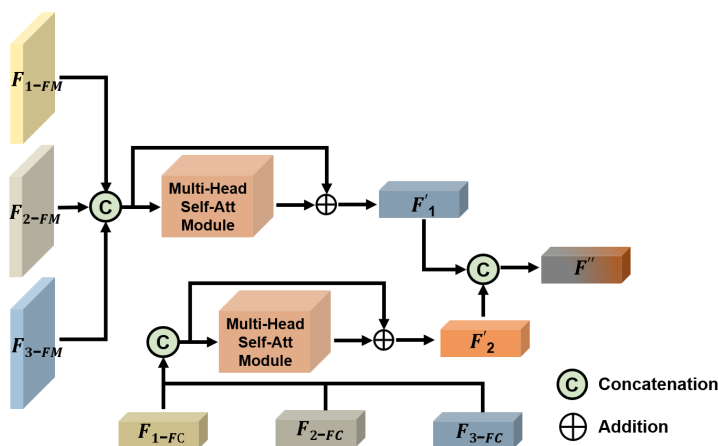


Figure 5. This figure illustrates the framework of feature fusion module version 3 (FFM-v3).

2.6. EBUS Image Encoder Architecture

The encoders used for EBUS image feature extraction are built upon the concept of transfer learning [49] and are fine-tuned based on several commonly used CNN models. This approach leverages pre-trained models to adapt their knowledge for the specific task of EBUS image analysis. By fine-tuning the classifier of these models, the framework effectively repurposes them for feature extraction in this specialized domain. This study utilized 12 CNN models pre-trained on the ImageNet dataset [50]: ResNet-18, ResNet-50, ResNet-101 [47], ResNext-50, ResNext-101 [51], DenseNet-121, DenseNet-169 [52], DenseNet-201, MobileNet-V2 [53], ShuffleNet-V2 [54], EfficientNet-B0 [55], and EfficientNet-V2 [56]. For each CNN model, only the final classifier was replaced. The details of the CNN encoder (DenseNet-121) are shown in Figure 6. The new classifier consisted of three fully connected layers, with batch normalization and Dropout layers added after every two fully connected layers. These additions served to accelerate model convergence and prevent overfitting, respectively. Before being used as encoders for feature extraction, the CNN models underwent fine-tuning on our dataset. The initial weights for each CNN model were sourced from their pre-trained weights on the ImageNet dataset. After fine-tuning, the CNN models were frozen, meaning their weights remained unchanged during feature extraction. This ensured stable and consistent feature extraction for subsequent stages of the framework.

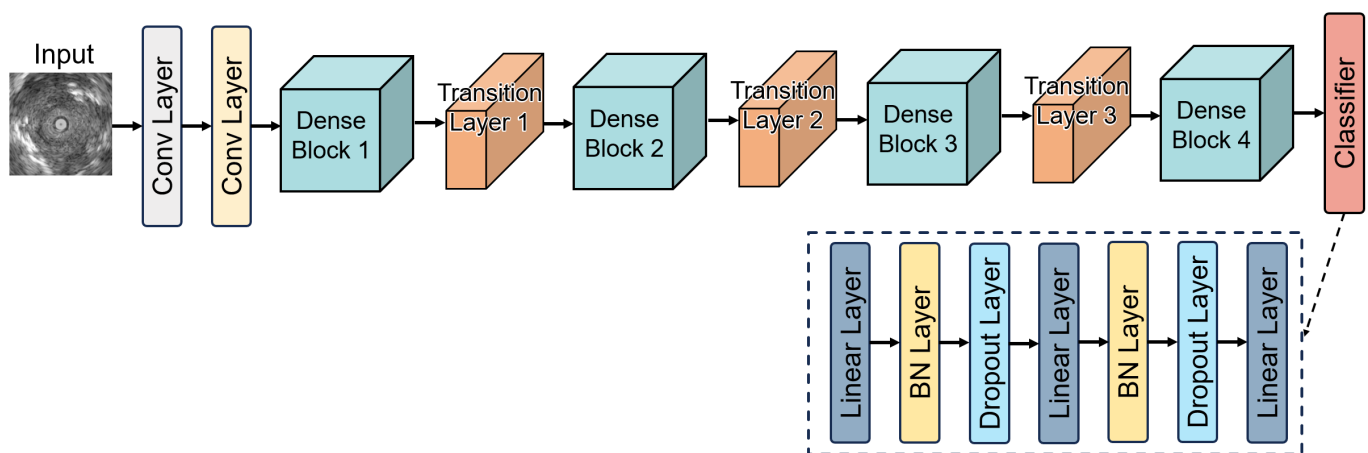


Figure 6. This figure represents the encoder framework based on the fine-tuned DenseNet-121 model.

2.7. Weight Definition for Loss Function

In the field of medical image diagnosis, data imbalance is particularly common. When a dataset is highly imbalanced, the classification performance of algorithms can be significantly affected [57]. Moreover, as the degree of imbalance increases, the performance loss escalates rapidly [58]. Our study also encountered severe data imbalance, as detailed in Table 1. The dataset contained approximately 6 to 7 times more malignant cases than benign ones. The cause of this situation lies in the fact that malignant lung tumors typically present with obvious clinical symptoms, requiring aggressive treatment. During the treatment process, patients undergo numerous tests, diagnoses, and treatments, resulting in the generation of extensive clinical data. In contrast, many patients with benign lung tumors experience mild or even no symptoms, and their condition is often discovered incidentally during routine check-ups or other examinations. As a result, these patients tend to seek medical attention and undergo diagnosis less frequently, as their benign tumors do not show significant symptoms. To mitigate the impact of this imbalance, we introduced a weighted loss function to compensate for the training loss bias of minority classes. Additionally, when constructing the dataset, we ensured that the training, validation, and test sets contained an equal number of images from each class as much as possible. This

approach and the weighted loss function during CNN model training further balanced the distributional bias. The weight calculation process is as follows: (1) Determine the number of benign and malignant cases, N_b and N_m , in the training set during model training. (2) Normalize these counts to obtain $N_{b_{norm}}$ and $N_{m_{norm}}$. (3) Compute their reciprocals, ensuring that the weights are inversely proportional to the number of cases: $\frac{1}{N_{b_{norm}}}$ and $\frac{1}{N_{m_{norm}}}$. (4) Normalize these reciprocals again to obtain the final weights used for training.

3. Experiment and Results

This section provides a detailed discussion of the experimental setup, including the specifications of the experimental platform, implementation details of the model, hyperparameter settings, definitions of evaluation metrics, and a presentation of the experimental results.

3.1. Experimental Setting

This study was conducted using the PyTorch 2.0.1 deep learning framework, with all experiments running on a workstation with an Intel Core i5-13600K 3.50 GHz six-core processor (Intel Corporation, Santa Clara, CA, USA) and NVIDIA TITAN graphics processing unit (GPU) (NVIDIA Corporation, Santa Clara, CA, USA) equipped with 24 GB of GPU memory. We utilized the Adam optimizer with weight decay, setting the decay coefficient to 5×10^{-4} . The network was trained for 200 epochs, with a batch size of 32, and input images were uniformly resized to 224×224 pixels. We used weighted cross-entropy as the loss function and set the initial learning rate to 1×10^{-3} . To prevent overfitting, early stopping was applied, with a patience parameter of six epochs. During training, data augmentation techniques were employed, including horizontal and vertical flips, 90-degree rotations, and random adjustments to brightness, contrast, and saturation. Additionally, all images were normalized before input, scaling pixel values from $[0, 255]$ to the range $[0, 1]$. The hyperparameters were determined through ablation studies, with the selected configuration shown to yield optimal performance.

The implementation details of the M³-Net framework are as follows. First, a two-stage training strategy is adopted. In the first stage, CNN models are trained separately on different EBUS image modalities to obtain CNN encoders for feature extraction. In the second stage, the trained CNN models are used as feature extractors. It is important to note that during feature extraction, the weights of each encoder remain frozen and are not updated. When training the feature fusion module and classifier, the CNN encoders do not participate in further training.

3.2. Evaluation Metrics

Seven metrics were used to evaluate the performance of the lung cancer diagnostic system: Accuracy, F1-score, Area Under the Curve (AUC), Positive Predictive Value (PPV), Negative Predictive Value (NPV), sensitivity (SEN), and specificity (SPEC). The definitions of these evaluation metrics are as follows: (1)–(6). In addition, this study employed four-fold cross-validation to ensure that the model evaluation results were more reliable and rigorous.

$$\text{Accuracy}(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$F1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

$$\text{Positive Predictive Value}(PPV) = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Negative Predictive Value}(NPV) = \frac{TN}{TN + FN} \quad (4)$$

$$\text{Sensitivity}(\text{Sen}) = \frac{TP}{TP + FN} \quad (5)$$

$$\text{Specificity}(\text{Spec}) = \frac{TN}{TN + FP} \quad (6)$$

Additionally, among the seven evaluation metrics mentioned above, AUC is a comprehensive measure derived from the ROC (receiver operating characteristic) curve. The other metrics are computed at a specific threshold. In this study, AUC is considered the most important evaluation metric. Particularly when dealing with class imbalance or when different evaluation metrics need to be balanced, AUC provides a more objective model evaluation. Therefore, AUC is prioritized in this study to comprehensively assess the model's overall performance across various classification thresholds, avoiding potential bias that could arise from using a single threshold. This approach offers a more global performance evaluation.

3.3. Classification Results

The experiments in this study are divided into two main parts. The first part involves training a CNN-based encoder for feature extraction from EBUS images. The results confirm the effectiveness of the weighted cross-entropy loss function. Subsequently, one optimal EBUS image type is selected from each of the four major categories: (1) four types of EBUS images obtained after cropping, (2) four types of EBUS images after coordinate system transformation, (3) three MS-EBUS (multi-scale) images generated from cropped EBUS images, and (4) three PMS-EBUS (polar multi-scale) images generated from transformed EBUS images. Although it is technically feasible to use all 14 types of EBUS images, this approach would create numerous combinations and significantly increase the time complexity. The primary aim of this study is to integrate different EBUS image types. Thus, based on experimental results, the most effective image from each category is chosen for feature fusion, ensuring both feature diversity and reduced experimental complexity. In the second part of the experiment, the four selected EBUS image types are used with the trained CNN model as the feature extraction encoder within the proposed M³-Net framework. The experimental results confirm the effectiveness of this framework.

In the first part of the experiment, 12 different CNN models were tested on the polar type 3 EBUS dataset, with all results reported as cross-validation averages. The models were trained using both unweighted cross-entropy and weighted cross-entropy as loss functions. As shown in Figure 7, the results indicate that over 70% of the CNN models showed performance improvements when trained with the weighted cross-entropy. Notably, the best performance was achieved with the DenseNet-169 model, reaching an AUC of 0.79.

According to the experimental results shown in Figure 7, training CNN models with weighted cross-entropy effectively enhances performance. Currently, there are four major categories of EBUS image datasets, comprising a total of 14 types of EBUS images. From each category, the most effective EBUS image dataset will be selected for training within the M³-Net framework. Using weighted cross-entropy as the loss function, 12 CNN models were trained on each of the 14 EBUS image datasets. The best results for each dataset are presented in Figure 8, with all results representing cross-validation averages. For the detailed performance of the best-performing models in each image type and specific information on the selected CNN models, please refer to the appendix in Appendix A. Based on the experimental results shown in Figure 8, the highest AUC of 0.79 was achieved on the polar type 3 EBUS image datasets. Additionally, within each of the four major EBUS image categories, the best-performing datasets were type 4, polar type 3, MS 2-3-4, and PMS 2-3-4, with an AUC of 0.72, 0.79, 0.72, and 0.78, respectively. Based on the experimental results, the four selected EBUS image types from the major categories are type 4, polar type

3, MS 2-3-4, and PMS 2-3-4. Details of the CNN models used to achieve the highest AUC for each dataset, along with other evaluation metrics, are provided in Table 4.

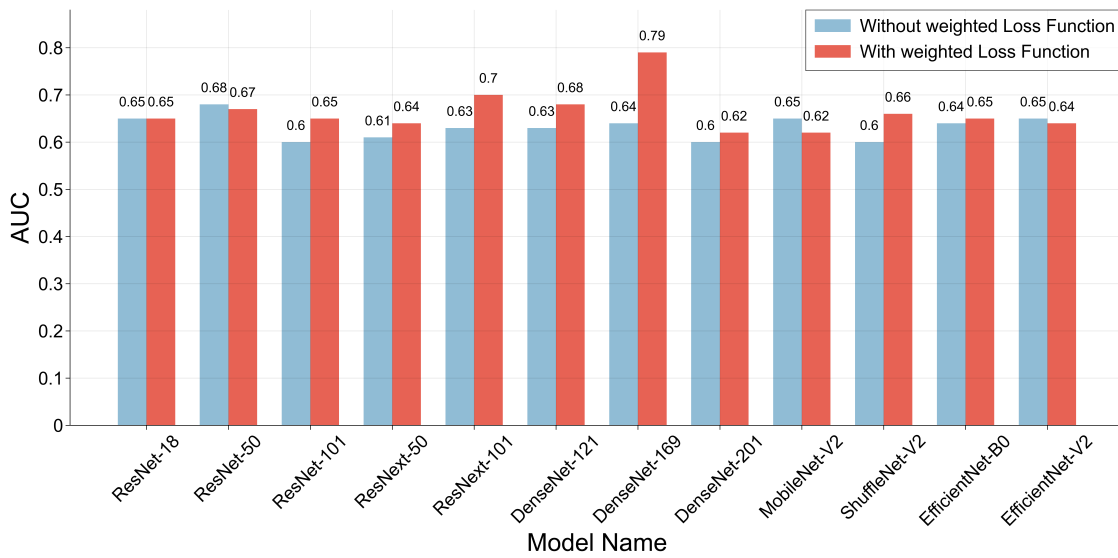


Figure 7. This figure illustrates 12 CNN models that were trained on the polar type 3 EBUS dataset using both unweighted cross-entropy and weighted cross-entropy as loss functions to evaluate their AUC.

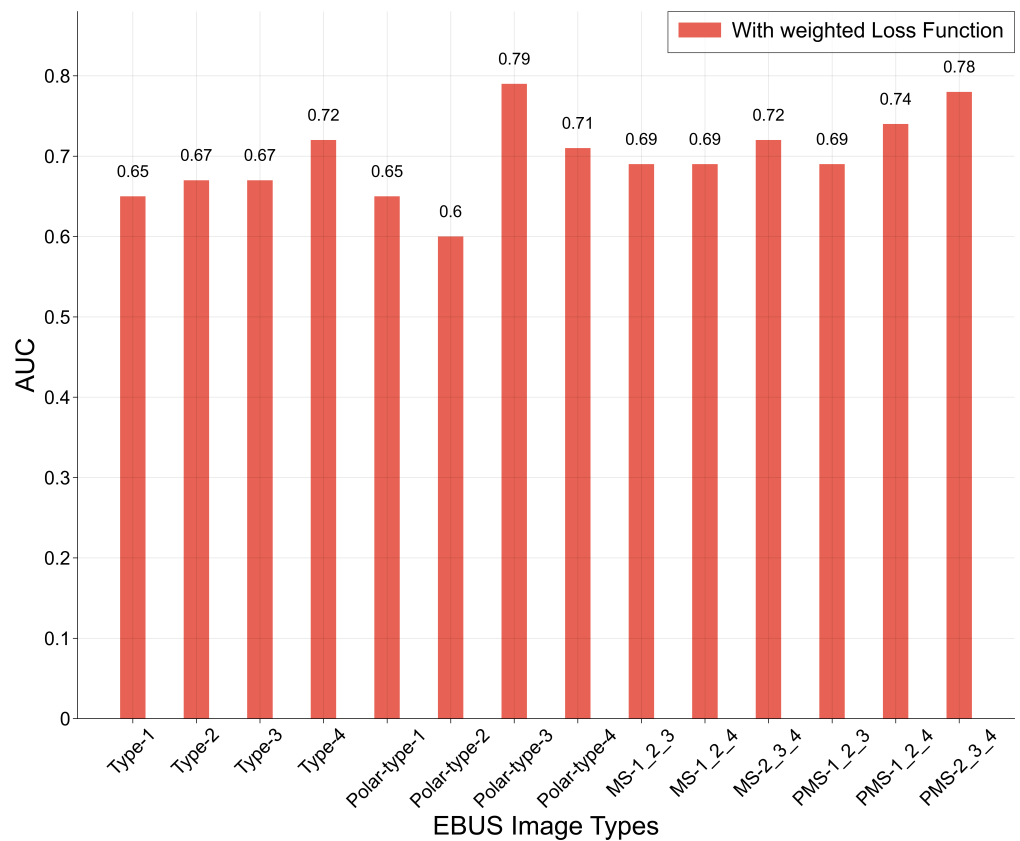


Figure 8. This figure shows the best AUC achieved when training 12 CNN models with weighted cross-entropy loss on each of the 14 EBUS image datasets.

Table 4. This table shows the best results achieved on the selected four types of EBUS image datasets—type 4, polar type 3, MS 2-3-4, and PMS 2-3-4. The details of these networks and evaluation metrics other than accuracy are summarized. The experimental result of each model is the average result after the cross-validation.

Image Type	CNN Model	AUC	Acc	F1-Score	PPV	NPV	Sen	Spec
Type 4	DenseNet-121	0.72	0.70	0.66	0.77	0.67	0.59	0.81
Polar Type 3	DenseNet-169	0.79	0.74	0.72	0.79	0.74	0.70	0.77
MS 2-3-4	DenseNet-201	0.72	0.71	0.70	0.71	0.73	0.71	0.72
PMS 2-3-4	MobileNet-V2	0.78	0.74	0.72	0.79	0.70	0.66	0.81

Based on the experimental results, the most effective EBUS images from each category and their corresponding CNN model were selected as encoders for feature extraction. These were then used in the proposed M³-Net framework for multi-feature fusion. As shown in Table 4, four types of EBUS images were selected for the multi-feature fusion stage within the proposed M³-Net framework. During this stage, there are multiple ways to combine these four image types. The fusion options include the following: (1) combining two types of EBUS images, (2) combining three types, and (3) combining all four types. Only a selection of representative experimental results is presented here. All experimental results are the final averages obtained through four-fold cross-validation.

The experimental results for feature fusion using the proposed M³-Net multi-feature fusion framework, combining type 4 and polar type 3 features, are shown in Table 5. The results showed varying degrees of improvement when using three different fusion modules. Specifically, with FFM-v2, the accuracy improved by up to 2%, and the AUC increased by up to 4% compared to using the single modality polar type 3.

Table 5. This table shows the experimental results of the fusion of type 4 and polar type 3 based on the M³-Net framework and the comparison result with the single modality.

	Image Type	CNN Encoder	Feature Fusion Module	AUC	Acc	F1-Score	PPV	NPV	Sen	Spec
Single Model	Type 4	DenseNet-121	-	0.72	0.70	0.66	0.77	0.67	0.59	0.81
	Polar Type 3	DenseNet-169	-	0.79	0.74	0.74	0.79	0.74	0.70	0.77
M ³ -Net Framework	Type 4+ Polar Type 3	DenseNet-121+ DenseNet-169	FFM-v1	0.81	0.75	0.72	0.80	0.74	0.70	0.79
			FFM-v2	0.83	0.76	0.75	0.80	0.75	0.72	0.80
			FFM-v3	0.82	0.77	0.76	0.79	0.78	0.76	0.77

Table 6 presents the experimental results of the proposed M³-Net multi-feature fusion framework, integrating features from polar type 3 and multi-scale PMS 2-3-4 images. The results show that using the three different fusion modules led to varying degrees of improvement. Specifically, with FFM-v3, the AUC improved by up to 3%, compared to the single-modality polar type 3.

Table 7 presents the experimental results of the proposed M³-Net multi-feature fusion framework, integrating features from type 4, polar type 3, and PMS 2-3-4 images. The results show that using FFM-v3 increased the AUC by 8% compared to the single-modality type 4 and by 1% compared to the single-modality polar type 3.

Table 8 presents the experimental results of the proposed M³-Net multi-feature fusion framework, combining features from type 4, polar type 3, MS 2-3-4, and PMS 2-3-4 images. The results show that using FFM-v3 improved the AUC by 5% compared to the single-modality type 4 and MS 2-3-4. However, the AUC decreased by 2% and 1% compared to the single-modalities polar type 3 and PMS 2-3-4, respectively.

Table 6. This table shows the experimental results of the fusion of polar type 3 and PMS 2-3-4 based on the M³-Net framework and comparison result with the single modality.

	EBUS Image Type	CNN Encoder	Feature Fusion Module	AUC	Acc	F1-Score	PPV	NPV	Sen	Spec
Single Model	Polar Type 3	DenseNet-169	-	0.79	0.74	0.74	0.79	0.74	0.70	0.77
	PMS 2-3-4	MobileNet-V2	-	0.78	0.74	0.72	0.79	0.70	0.66	0.81
M ³ -Net Framework	Polar Type 3+ PMS 2-3-4	DenseNet-169+ MobileNet-V2	FFM-v1	0.80	0.75	0.74	0.77	0.74	0.72	0.78
			FFM-v2	0.80	0.74	0.75	0.74	0.75	0.76	0.72
			FFM-v3	0.82	0.75	0.75	0.78	0.74	0.72	0.79

Table 7. This table shows the experimental results of the fusion of type 4, polar type 3, and PMS 2-3-4 based on the M³-Net framework and the comparison result with the single modality.

	EBUS Image Type	CNN Encoder	Feature Fusion Module	AUC	Acc	F1-Score	PPV	NPV	Sen	Spec
Single Model	Type 4	DenseNet-121	-	0.72	0.70	0.66	0.77	0.67	0.59	0.81
	Polar Type 3	DenseNet-169	-	0.79	0.74	0.74	0.79	0.74	0.70	0.77
	PMS 2-3-4	MobileNet-V2	-	0.78	0.74	0.72	0.79	0.70	0.66	0.81
M ³ -Net Framework	Type 4+ Polar Type 3+ PMS 2-3-4	DenseNet-121+ DenseNet-169+ MobileNet-V2	FFM-v1	0.78	0.74	0.74	0.76	0.73	0.72	0.76
			FFM-v2	0.76	0.72	0.73	0.73	0.75	0.75	0.70
			FFM-v3	0.80	0.75	0.74	0.77	0.75	0.73	0.78

Table 8. This table shows the experimental results of fusing four different EBUS images based on the M³-Net framework, compared with the results using a single modality.

	EBUS Image Type	CNN Encoder	Feature Fusion Module	AUC	Acc	F1-Score	PPV	NPV	Sen	Spec
Single Model	Type 4	DenseNet-121	-	0.72	0.70	0.66	0.77	0.67	0.59	0.81
	Polar Type 3	DenseNet-169	-	0.79	0.74	0.74	0.79	0.74	0.70	0.77
	MS 2-3-4	DenseNet-201	-	0.72	0.71	0.70	0.71	0.73	0.71	0.72
	PMS 2-3-4	MobileNet-V2	-	0.78	0.74	0.72	0.79	0.70	0.66	0.81
M ³ -Net Framework	Type 4+Polar Type 3+MS 2-3-4 +PMS 2-3-4	DenseNet-121+Dense Net-169+DenseNet -201+MobileNet-V2	FFM-v1	0.72	0.68	0.70	0.66	0.73	0.77	0.59
			FFM-v2	0.75	0.71	0.73	0.69	0.74	0.77	0.66
			FFM-v3	0.77	0.73	0.73	0.73	0.73	0.74	0.71

Table 9 presents the comparison of experimental results between the proposed M³-Net multi-feature fusion framework and previously published methods. Due to the unavailability of data from related studies, we replicated the experiments using our own dataset, strictly following the procedures described in the original papers. Cross-validation was employed in all experiments. According to the experimental results, our proposed M³-Net shows an improvement of approximately 9% in accuracy and 30% in AUC compared to the previously published methods. Compared to our previous research, the accuracy increased by about 10%, and the AUC improved by more than 11%.

Table 9. This table presents a comparative analysis of the experimental results between the proposed M³-Net multi-feature fusion framework and existing published methods.

	Methods	AUC	Acc	F1-Score	PPV	NPV	Sen	Spec
Published methods	CNN-SVM [33]	0.54	0.68	0.80	0.88	0.14	0.74	0.33
	Single image-based model [59]	0.52	0.67	0.79	0.87	0.12	0.73	0.31
Previous work [38]	Majority Voting	0.66	0.66	0.59	0.76	0.62	0.52	0.79
	Output Average	0.67	0.67	0.62	0.75	0.64	0.56	0.77
	Performance Weighting	0.72	0.66	0.60	0.76	0.62	0.53	0.78
Our proposed	M ³ -Net	0.83	0.76	0.75	0.80	0.75	0.72	0.80

4. Discussion

M³-Net is designed to integrate features from multiple types of images. It first trains separate CNN models on different EBUS image types, which are then used as feature encoders. These encoders extract features from various EBUS images, which are subsequently fused through an attention-based feature fusion module. The final output is the prediction result. Experimental results after feature fusion demonstrate that the proposed M³-Net multi-feature fusion framework effectively improves lung cancer diagnostic performance.

In our study, a total of 95 cases were included, comprising 13 benign and 82 malignant cases, with each case corresponding to an EBUS video. During dataset preparation, unlike other studies [33,59] that typically select a single EBUS image per case, we opted not to follow this approach. Selecting only one image per case would have resulted in an extremely imbalanced and small dataset, which could lead to poor experimental results. In our study, we selected EBUS images evenly from each case video based on its length, ensuring that the numbers of benign and malignant EBUS images included in the training, validation, and test sets were balanced, as shown in Table 1. While the number of EBUS images for each class was equal, there was a significant disparity in the number of cases within each class. This discrepancy motivated our proposal to define custom weights for the loss function based on the number of cases in the dataset. By training CNN models using a weighted loss function, as illustrated in Figure 7, the experimental results show that this approach can further improve the accuracy of CNN models. Coupled with ensuring an equal number of EBUS images per class in the dataset, this strategy enhances data balance. According to the results in Table 4, CNN models trained with the weighted loss function on single-modal data, such as polar type 3 and MS 2-3-4 EBUS images, achieved similar sensitivity and specificity. If CNN models are trained on highly imbalanced datasets without any balancing strategy, the resulting models often exhibit significantly higher sensitivity than specificity. This further confirms the effectiveness of the weight definition for the loss function strategy in ensuring data balance and improving model performance.

Tables 5 and 6 present the experimental results of integrating two types of EBUS image features using the proposed M³-Net multi-feature fusion framework. The results demonstrate that employing any of the three feature fusion modules yields notable performance improvements. Furthermore, the results indicate that the sensitivity and specificity achieved through M³-Net are remarkably close, suggesting that the proposed framework effectively mitigates the impact of imbalanced data. As shown in Table 5, the sensitivity observed for type 4 EBUS images in single-modality experiments is unexpectedly lower than the specificity. These results were obtained through four-fold cross-validation, with the final outcome calculated as the average across all folds. Although sensitivity may exceed specificity in some folds, the averaged sensitivity remains 22% lower than specificity. This discrepancy could be attributed to over-correction during model training caused by the weighted loss function.

Similarly, Table 6 reveals a comparable issue in single-modality experiments conducted on PMS 2-3-4 EBUS images, where sensitivity is 15% lower than specificity. This discrepancy might be due to the limited number of cases in the dataset, leading to distribution imbalances among training, validation, and test sets in each fold. To address these issues, future efforts will focus on collecting more data to enhance the framework's robustness and generalization capabilities. Additionally, a dynamic weighting strategy for the loss function will be explored, replacing the current method, which relies solely on the proportion of benign and malignant cases in the training set to determine the weight.

Table 7 presents the experimental results of integrating three EBUS image features using the proposed M³-Net multi-feature fusion framework. The results show that the performance improvement is limited in some cases. For instance, when using the FFM-v3

fusion module, accuracy and AUC improved by only 1% compared to the single-modality polar type 3 results, which deviates from expectations. When integrating type 4 and polar type 3 EBUS images through the M³-Net framework, the accuracy and AUC achieved maximum improvements of 3% and 4%, respectively, compared to single-modality polar type 3. Similarly, when fusing polar type 3 and PMS 2-3-4 EBUS images, the accuracy and AUC improvements were 1% and 3%, respectively. However, when combining features from type 4, polar type 3, and PMS 2-3-4, the accuracy and AUC showed only a 1% improvement, which falls short of expectations. Theoretically, integrating multiple features should enhance the information available for determining class boundaries, leading to improved results. However, during feature fusion, hidden information loss may occur [60], as feature extraction for each component is performed independently before fusion. This loss could explain the limited performance gains observed in the experiments. This phenomenon may also explain why the performance decreases rather than improves when we use M³-Net to fuse four types of EBUS image features. For detailed results on fusing four types EBUS image feature, please refer to Table 8.

To assess the contribution of individual components to model performance, we conducted an ablation study evaluating the impact of the weighted loss function, feature fusion module, and data augmentation. The results are summarized in Table 10. When none of the components were included, the model achieved an AUC of 0.74 and an accuracy of 0.68, serving as the baseline performance. Introducing the weighted loss function alone resulted in a slight decrease in AUC to 0.72, while accuracy remained at 0.68. This suggests that the weighted loss function may not directly improve AUC when used in isolation, but could have a positive impact when combined with other techniques. Incorporating the feature fusion module alone increased the AUC to 0.76 and accuracy to 0.71, indicating that feature fusion effectively integrates complementary information, thereby enhancing model performance. When only data augmentation was applied, the AUC remained at 0.74, but accuracy dropped to 0.64, suggesting that data augmentation alone introduced greater variability without yielding a significant performance improvement. When all three components were incorporated, the model achieved the highest AUC (0.83) and accuracy (0.76), confirming the complementary nature of these techniques. This further demonstrates that synergy among multiple strategies effectively enhances overall performance.

The specific details of the feature fusion module (FFM-v1) are illustrated in Figure 3. Features F_1 , F_2 , and F_3 , represent the feature maps extracted from different encoders. After being processed by the CBAM attention layer, these features are concatenated and fed into a multi-head self-attention module. In contrast, the FFM-v2 is simpler: it concatenates the features from different encoders and directly feeds them into the multi-head self-attention module. Compared to FFM-v1, the early fusion of features in FFM-v2 is more effective. The feature maps after the CBAM attention module may alter the model's representation, and using two different attention mechanisms can increase the training complexity of the model. This could be one reason why the experimental results of FFM-v1 are not as good as those of FFM-v2.

The feature fusion module FFM-v3 builds on FFM-v2 by adding additional fully connected layers. While the intention was to provide the model with more diagnostic information, the experimental results indicate that FFM-v3 performs worse than FFM-v2. This suggests that, although adding extra information through the fully connected layers might seem beneficial, it could also introduce noise, which in turn reduces model performance. Although our proposed multi-feature fusion framework, M³-Net, achieves satisfactory improvements compared to single-modality approaches, it has certain limitations. This study utilized over ten CNN models as feature extractors for EBUS images. CNNs primarily rely on local operations through convolutional filters. While convolution is advantageous for

developing efficient algorithms, its limited receptive field restricts the ability to capture long-range pixel relationships [61]. In recent years, Transformer-based techniques have gained increasing attention in the medical imaging domain, especially with the advent of ViT (Vision Transformer) [62]. This is largely because Transformers can effectively capture global contextual information. Considering these advantages, future research will explore using ViT-based models for feature extraction. Secondly, our study revealed that as the number of EBUS image modalities increased, the fusion results unexpectedly deteriorated. This could be attributed to misalignment in high-dimensional feature spaces. While CNN encoders extract more effective features from different EBUS image types, these features may lack alignment, and simple concatenation fails to address this issue. To resolve this, future research may explore the use of adversarially regularized encoders [63] to project features extracted from different EBUS images into a shared feature space. By aligning these features, we can train the classifier more effectively and improve the fusion results.

Table 10. Ablation study results evaluating the impact of the weighted loss function, feature fusion module, and data augmentation on model performance. The AUC and Acc are reported for each setting.

Weighted Loss Function	Ablation Setting			AUC	Acc
	Feature Fusion Module	Data Augmentation			
x	x	x		0.74	0.68
✓	x	x		0.72	0.68
x	✓	x		0.76	0.71
x	x	✓		0.74	0.64
✓	✓	✓		0.83	0.76

Additionally, this study lacks ablation experiments to verify the contribution of multi-scale fusion to the improvement of the framework’s classification performance. In future research, we will conduct ablation experiments to further demonstrate the effectiveness of multi-scale fusion. The amount of data included in the experiments is indeed insufficient, and external validation has not been conducted. In our future research, we will first collect more data and then perform external validation as well as multi-center validation to assess the model’s performance and generalization ability, ensuring its applicability in clinical settings. In future research, we plan to incorporate the gradient-weighted class activation mapping (Grad-CAM) method [64] to analyze the model’s decision-making process by visualizing key areas in the images that contribute significantly to the predictions of malignancy or benignancy. Additionally, we will use Shapley additive explanation (SHAP) values [65] to quantify the influence of individual features on the model’s output. This will enable us to gain a clearer understanding of how the model makes predictions and ensure that its decisions are aligned with clinical reasoning.

5. Conclusions

This study proposes a multi-scale image generation method and, based on the generated multi-scale images, constructs the M³-Net CAD framework for lung cancer diagnosis. To address the impact of data imbalance, the weight of the loss function was customized, and a weighted cross-entropy loss function was used to train the CNN model, alleviating the effects of the imbalance. The trained CNN model was then used as the feature extractor in the M³-Net multi-feature fusion framework, with the weights of the CNN model frozen during the training of the M³-Net multi-feature fusion framework. Additionally, three attention-based feature fusion modules were introduced to enhance the effectiveness of feature fusion. The experimental results show that FFM-v2 achieved the best performance,

with an accuracy of 0.76, F1-score of 0.75, AUC of 0.83, PPV of 0.80, NPV of 0.75, sensitivity of 0.72, and specificity of 0.80.

In conclusion, the deep learning-based computer-aided diagnostic system for lung cancer diagnosis has shown promising results. Although the dataset used in this study suffers from an imbalance issue, future work will include not only the use of data augmentation techniques but also the collection of more data to further validate the robustness of the M³-Net multi-feature fusion framework.

Author Contributions: Conceptualization, H.W. and T.N. (Toshiya Nakaguchi); methodology, H.W.; software, H.W.; validation, H.W.; formal analysis, H.W., Y.N. and T.N. (Toshiya Nakaguchi); investigation, H.W., K.S. and T.N. (Takahiro Nakajima); resources, H.W., K.S. and T.N. (Takahiro Nakajima); data curation, H.W., K.S. and T.N. (Takahiro Nakajima); writing—original draft preparation, H.W.; writing—review and editing, T.N. (Toshiya Nakaguchi), Y.N., T.N. (Takahiro Nakajima) and K.S.; visualization, H.W. and T.N. (Toshiya Nakaguchi); supervision, T.N. (Toshiya Nakaguchi); project administration, T.N. (Toshiya Nakaguchi); funding acquisition, T.N. (Toshiya Nakaguchi). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Ethics Committee of Chiba University Graduate School of Medicine (M10393, 9 September 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. This table presents the detailed results of the best-performing models for each type of EBUS image, along with the specific CNN backbone networks selected.

	Image Type	CNN Model	AUC	Acc	F1-Score	PPV	NPV	Sen	Spec
Original EBUS data	Type 1	ResNet-50	0.65	0.65	0.55	0.81	0.61	0.46	0.84
	Type 2	ResNet-50	0.67	0.67	0.61	0.74	0.63	0.53	0.81
	Type 3	ResNet-18	0.67	0.67	0.60	0.75	0.63	0.51	0.83
	Type 4	DenseNet-121	0.72	0.70	0.66	0.77	0.67	0.59	0.81
Polar EBUS data	Polar Type 1	ResNet-50	0.65	0.65	0.47	0.96	0.60	0.33	0.98
	Polar Type 2	ResNext-101	0.60	0.60	0.38	0.89	0.57	0.29	0.91
	Polar Type 3	DenseNet-169	0.79	0.74	0.72	0.79	0.74	0.70	0.77
	Polar Type 4	ResNext-101	0.71	0.71	0.67	0.82	0.66	0.58	0.84
MS EBUS data	MS 1-2-3	ShuffleNet-V2	0.69	0.66	0.62	0.71	0.63	0.55	0.76
	MS 1-2-4	ResNet-101	0.69	0.66	0.59	0.75	0.62	0.50	0.82
	MS 2-3-4	DenseNet-201	0.72	0.71	0.70	0.71	0.73	0.71	0.72
PMS EBUS data	PMS 1-2-3	ResNext-101	0.69	0.67	0.62	0.70	0.65	0.57	0.76
	PMS 1-2-4	ResNext-50	0.74	0.64	0.59	0.75	0.61	0.54	0.74
	PMS 2-3-4	MobileNet-V2	0.78	0.74	0.72	0.79	0.70	0.66	0.81

References

1. Leiter, A.; Veluswamy, R.R.; Wisnivesky, J.P. The global burden of lung cancer: Current status and future trends. *Nat. Rev. Clin. Oncol.* **2023**, *20*, 624–639. [[CrossRef](#)]
2. Thandra, K.C.; Barsouk, A.; Saginala, K.; Aluru, J.S.; Barsouk, A. Epidemiology of lung cancer. *Contemp. Oncol. Onkol.* **2021**, *25*, 45–52.
3. Nooreldeen, R.; Bach, H. Current and future development in lung cancer diagnosis. *Int. J. Mol. Sci.* **2021**, *22*, 8661. [[CrossRef](#)]
4. Huang, S.; Yang, J.; Shen, N.; Xu, Q.; Zhao, Q. Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. *Semin. Cancer Biol.* **2023**, *89*, 30–37. [[CrossRef](#)] [[PubMed](#)]
5. Raza, R.; Zulfiqar, F.; Khan, M.O.; Arif, M.; Alvi, A.; Iftikhar, M.A.; Alam, T. Lung-EffNet: Lung cancer classification using EfficientNet from CT-scan images. *Eng. Appl. Artif. Intell.* **2023**, *126*, 106902. [[CrossRef](#)]
6. Sim, A.J.; Kaza, E.; Singer, L.; Rosenberg, S.A. A review of the role of MRI in diagnosis and treatment of early stage lung cancer. *Clin. Transl. Radiat. Oncol.* **2020**, *24*, 16–22. [[CrossRef](#)] [[PubMed](#)]
7. Ireland, R.H.; Bragg, C.M.; McJury, M.; Woodhouse, N.; FICHELE, S.; Van Beek, E.J.; Wild, J.M.; Hatton, M.Q. Feasibility of image registration and intensity-modulated radiotherapy planning with hyperpolarized helium-3 magnetic resonance imaging for non-small-cell lung cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **2007**, *68*, 273–281. [[CrossRef](#)] [[PubMed](#)]
8. Hoffman, R.M.; Atallah, R.P.; Struble, R.D.; Badgett, R.G. Lung cancer screening with low-dose CT: A meta-analysis. *J. Gen. Intern. Med.* **2020**, *35*, 3015–3025. [[CrossRef](#)]
9. Antoch, G.; Stattaus, J.; Nemat, A.T.; Marnitz, S.; Beyer, T.; Kuehl, H.; Bockisch, A.; Debatin, J.F.; Freudenberg, L.S. Non-small cell lung cancer: Dual-modality PET/CT in preoperative staging. *Radiology* **2003**, *229*, 526–533. [[CrossRef](#)] [[PubMed](#)]
10. Yang, H.; Chen, L.; Cheng, Z.; Yang, M.; Wang, J.; Lin, C.; Wang, Y.; Huang, L.; Chen, Y.; Peng, S.; et al. Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: A retrospective study. *BMC Med.* **2021**, *19*, 1–14. [[CrossRef](#)] [[PubMed](#)]
11. Coudray, N.; Ocampo, P.S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A.L.; Razavian, N.; Tsirigos, A. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **2018**, *24*, 1559–1567. [[CrossRef](#)]
12. Sicular, S.; Alpaslan, M.; Ortega, F.A.; Keathley, N.; Venkatesh, S.; Jones, R.M.; Lindsey, R.V. Reevaluation of missed lung cancer with artificial intelligence. *Respir. Med. Case Rep.* **2022**, *39*, 101733. [[CrossRef](#)] [[PubMed](#)]
13. Lancaster, H.L.; Heuvelmans, M.A.; Oudkerk, M. Low-dose computed tomography lung cancer screening: Clinical evidence and implementation research. *J. Intern. Med.* **2022**, *292*, 68–80. [[CrossRef](#)]
14. Wu, H.; Sui, H. Effect of chest low-dose spiral CT scan for lung cancer screening among asymptomatic adults. *Chin. J. Radiol. Health* **2023**, *32*, 176–181.
15. Yokoi, K.; Kamiya, N.; Matsuguma, H.; Machida, S.; Hirose, T.; Mori, K.; Tominaga, K. Detection of brain metastasis in potentially operable non-small cell lung cancer: A comparison of CT and MRI. *Chest* **1999**, *115*, 714–719. [[CrossRef](#)] [[PubMed](#)]
16. Li, W.; Liu, J.B.; Hou, L.K.; Yu, F.; Zhang, J.; Wu, W.; Tang, X.M.; Sun, F.; Lu, H.M.; Deng, J.; et al. Liquid biopsy in lung cancer: Significance in diagnostics, prediction, and treatment monitoring. *Mol. Cancer* **2022**, *21*, 25. [[CrossRef](#)]
17. Liam, C.K.; Mallawathantri, S.; Fong, K.M. Is tissue still the issue in detecting molecular alterations in lung cancer? *Respirology* **2020**, *25*, 933–943. [[CrossRef](#)] [[PubMed](#)]
18. Li, Y.; Shen, L. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors* **2018**, *18*, 556. [[CrossRef](#)] [[PubMed](#)]
19. Han, S.S.; Kim, M.S.; Lim, W.; Park, G.H.; Park, I.; Chang, S.E. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J. Investig. Dermatol.* **2018**, *138*, 1529–1538. [[CrossRef](#)]
20. Rath, A.; Mishra, D.; Panda, G.; Satapathy, S.C. Heart disease detection using deep learning methods from imbalanced ECG samples. *Biomed. Signal Process. Control* **2021**, *68*, 102820. [[CrossRef](#)]
21. Liu, X.; Wang, H.; Li, Z.; Qin, L. Deep learning in ECG diagnosis: A review. *Knowl.-Based Syst.* **2021**, *227*, 107187. [[CrossRef](#)]
22. Abbas, Q. Glaucoma-deep: Detection of glaucoma eye disease on retinal fundus images using deep learning. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 41–45. [[CrossRef](#)]
23. Golan, R.; Jacob, C.; Denzinger, J. Lung nodule detection in CT images using deep convolutional neural networks. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; IEEE: New York, NY, USA, 2016; pp. 243–250.
24. Gu, Y.; Chi, J.; Liu, J.; Yang, L.; Zhang, B.; Yu, D.; Zhao, Y.; Lu, X. A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning. *Comput. Biol. Med.* **2021**, *137*, 104806. [[CrossRef](#)] [[PubMed](#)]
25. El-Dahshan, E.S.A.; Mohsen, H.M.; Revett, K.; Salem, A.B.M. Computer-aided diagnosis of human brain tumor through MRI: A survey and a new algorithm. *Expert Syst. Appl.* **2014**, *41*, 5526–5545. [[CrossRef](#)]

26. Shin, H.; Oh, S.; Hong, S.; Kang, M.; Kang, D.; Ji, Y.g.; Choi, B.H.; Kang, K.W.; Jeong, H.; Park, Y.; et al. Early-stage lung cancer diagnosis by deep learning-based spectroscopic analysis of circulating exosomes. *ACS Nano* **2020**, *14*, 5435–5444. [[CrossRef](#)] [[PubMed](#)]
27. Teramoto, A.; Tsukamoto, T.; Kiriya, Y.; Fujita, H. Automated classification of lung cancer types from cytological images using deep convolutional neural networks. *BioMed Res. Int.* **2017**, *2017*, 4067832. [[CrossRef](#)] [[PubMed](#)]
28. Ardila, D.; Kiraly, A.P.; Bharadwaj, S.; Choi, B.; Reicher, J.J.; Peng, L.; Tse, D.; Etemadi, M.; Ye, W.; Corrado, G.; et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **2019**, *25*, 954–961. [[CrossRef](#)]
29. Singh, G.A.P.; Gupta, P. Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. *Neural Comput. Appl.* **2019**, *31*, 6863–6877. [[CrossRef](#)]
30. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images are more than pictures, they are data. *Radiology* **2016**, *278*, 563–577. [[CrossRef](#)] [[PubMed](#)]
31. Hyun, S.H.; Ahn, M.S.; Koh, Y.W.; Lee, S.J. A machine-learning approach using PET-based radiomics to predict the histological subtypes of lung cancer. *Clin. Nucl. Med.* **2019**, *44*, 956–960. [[CrossRef](#)] [[PubMed](#)]
32. Kokkonouzis, I.; Strimpakos, A.S.; Lampaditis, I.; Tsimpanakis, S.; Syrigos, K.N. The role of endobronchial ultrasound in lung cancer diagnosis and staging: A comprehensive review. *Clin. Lung Cancer* **2012**, *13*, 408–415. [[CrossRef](#)] [[PubMed](#)]
33. Chen, C.H.; Lee, Y.W.; Huang, Y.S.; Lan, W.R.; Chang, R.F.; Tu, C.Y.; Chen, C.Y.; Liao, W.C. Computer-aided diagnosis of endobronchial ultrasound images using convolutional neural network. *Comput. Methods Programs Biomed.* **2019**, *177*, 175–182. [[CrossRef](#)]
34. Li, J.; Zhi, X.; Chen, J.; Wang, L.; Xu, M.; Dai, W.; Sun, J.; Xiong, H. Deep learning with convex probe endobronchial ultrasound multimodal imaging: A validated tool for automated intrathoracic lymph nodes diagnosis. *Endosc. Ultrasound* **2021**, *10*, 361–371. [[PubMed](#)]
35. Iandola, F.N. SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
36. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
37. Vielzeuf, V.; Lechervy, A.; Pateux, S.; Jurie, F. Centralnet: A multilayer approach for multimodal fusion. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
38. Wang, H.; Shikano, K.; Nakajima, T.; Nomura, Y.; Nakaguchi, T. Peripheral pulmonary lesions classification using endobronchial ultrasonography images based on bagging ensemble learning and down-sampling technique. *Appl. Sci.* **2023**, *13*, 8403. [[CrossRef](#)]
39. Wang, H.; Takahiro Nakajima, K.S.Y.N.; Nakaguchi, T. Classification of Peripheral Pulmonary Lesions in Endobronchial Ultrasonography Image Using a Multi-branch Framework and Voting Ensemble. *Under review*.
40. Bazargani, R.; Fazli, L.; Gleave, M.; Goldenberg, L.; Bashashati, A.; Salcudean, S. Multi-scale relational graph convolutional network for multiple instance learning in histopathology images. *Med. Image Anal.* **2024**, *96*, 103197. [[CrossRef](#)]
41. Li, H.; Vong, C.M.; Wong, P.K.; Ip, W.F.; Yan, T.; Choi, I.C.; Yu, H.H. A multi-feature fusion method for image recognition of gastrointestinal metaplasia (GIM). *Biomed. Signal Process. Control* **2021**, *69*, 102909. [[CrossRef](#)]
42. Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; Jia, H. FFA-Net: Feature fusion attention network for single image dehazing. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11908–11915.
43. He, X.; Qi, G.; Zhu, Z.; Li, Y.; Cong, B.; Bai, L. Medical image segmentation method based on multi-feature interaction and fusion over cloud computing. *Simul. Model. Pract. Theory* **2023**, *126*, 102769. [[CrossRef](#)]
44. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional feature fusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 3560–3569.
45. Yan, Q.; Wang, B.; Zhang, W.; Luo, C.; Xu, W.; Xu, Z.; Zhang, Y.; Shi, Q.; Zhang, L.; You, Z. Attention-guided deep neural network with multi-scale feature fusion for liver vessel segmentation. *IEEE J. Biomed. Health Inform.* **2020**, *25*, 2629–2642. [[CrossRef](#)] [[PubMed](#)]
46. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
48. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
49. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A survey on deep transfer learning. In *Proceedings of the Artificial Neural Networks and Machine Learning—ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018, Proceedings, Part III 27*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 270–279.

50. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: New York, NY, USA, 2009; pp. 248–255.
51. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
52. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
53. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
54. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
55. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
56. Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10096–10106.
57. Thabtah, F.; Hammoud, S.; Kamalov, F.; Gonsalves, A. Data imbalance in classification: Experimental evaluation. *Inf. Sci.* **2020**, *513*, 429–441. [[CrossRef](#)]
58. Prati, R.C.; Batista, G.E.; Silva, D.F. Class imbalance revisited: A new experimental setup to assess the performance of treatment methods. *Knowl. Inf. Syst.* **2015**, *45*, 247–270. [[CrossRef](#)]
59. Khomkham, B.; Lipikorn, R. Pulmonary lesion classification framework using the weighted ensemble classification with random forest and cnn models for ebus images. *Diagnostics* **2022**, *12*, 1552. [[CrossRef](#)]
60. Imani, M.; Ghassemian, H. An overview on spectral and spatial information fusion for hyperspectral image classification: Current trends and challenges. *Inf. Fusion* **2020**, *59*, 59–83. [[CrossRef](#)]
61. Shamshad, F.; Khan, S.; Zamir, S.W.; Khan, M.H.; Hayat, M.; Khan, F.S.; Fu, H. Transformers in medical imaging: A survey. *Med. Image Anal.* **2023**, *88*, 102802. [[CrossRef](#)]
62. Dosovitskiy, A. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
63. Chen, W.; Yang, J.; Sun, Z.; Zhang, X.; Tao, G.; Ding, Y.; Gu, J.; Bu, J.; Wang, H. DeepASD: A deep adversarial-regularized graph learning method for ASD diagnosis with multimodal data. *Transl. Psychiatry* **2024**, *14*, 375. [[CrossRef](#)]
64. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
65. Lundberg, S. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.