Detection of Pleuropulmonary Blastoma at an Early Stage Using Vision Transformer Model

Sahar Almenwer, Dr. Hoda El-Sayed
Bowie State University
Bowie, MD, USA
salmenwer@bowiestate.edu, helsayed@bowiestate.edu

Dr. Md Kamruzzaman Sarker
Bowie State University
Bowie, MD, USA
ksarker@bowiestate.edu

Abstract—**Childhood cancer is the second most common cause of death in children under the age of fifteen, according to the American Cancer Society, and the incidence of diagnosis is rising. One common cancer is pleuropulmonary blastoma (PPB), which affects newborns to six-year-old children. Clinical diagnosis is through imaging, which is speedy and economical and does not require specialized equipment or laboratory tests. Still, it can be challenging to analyze PPB early using only imaging, and identifying clinical signs may also pose a challenge due to the numerous possible differential diagnoses. Clinical methods are unreliable for fast and accurate results, time-consuming, and prone to errors. Detecting PPB at an early stage is essential for its proper treatment, as it can be fatal if left untreated. In the last few years, convolutional neural networks (CNNs) have become the most prevalent technique for computer vision tasks. However, CNNs have a restricted local receptive field that may hinder their ability to learn about the global context. An alternative approach to CNNs that looks promising is the Vision Transformer (ViT). ViT utilizes self-attention between image patches to process visual information. This experiment uses the ViT base Model, an advanced deep-learning algorithm, to overcome these difficulties. ViT not only reduces the computation but also achieves better results than CNN. Our experiments with (LIDC-IDRI), include different models of medical imaging, such as CT, DX, and CR, and consist of 244,527 images. The proposed model evaluates the cancerous cells in the histopathological images to determine and detect PPB disease. From the result of the experiment, the efficiency of the proposed ViT model is verified and compared with other traditional clinical models and the DCNN model to evaluate the performance. The outcome shows that the accuracy and sensitivity of the method proposed in this research reach 99.47% and 99.9% for the medical imaging dataset.**

I. Introduction

Childhood cancer can be devastating and result in an overwhelming amount of concern and sadness for their families, friends, and even their communities. The American Cancer Society (ACS) reports that cancer is still the second most common cause of death in children aged fifteen and under [1]. In 2023, an estimated 9,910 children were diagnosed with cancer in the USA, and an estimated 1,040 children died from this disease [1]. The number of patients continues to rise yearly, as we unfortunately continue to lose many of our lovely and valuable children.

This ongoing issue underscores the need for faster and more accurate diagnostic equipment, as early detection significantly improves cancer treatment outcomes and survival rates [2]. Regrettably, almost 50% of cancer patients receive their diagnosis in advanced stages, when the cancer has already metastasized [3]. Early detection needs to be incorporated into health care systems to result in evidence-based early treatments, either to help slow the disease spread, cure it, or significantly affect survival. Thanks to the ever-deepening biological understanding and the rapid advancement of technology [2], researchers have reached a turning point in the study of early cancer diagnosis and its application to the goal of early curative therapies and improved cancer survival: computerized diagnosis.

Computer-aided diagnosis (CAD) systems, also called computer-aided detection (CAD) systems, assist in clinical diagnosis by doctors who may sometimes fail to detect disease early using conventional methods. This new technology helps address a challenging problem in oncology: pediatric lung cancers.–Pleuropulmonary blastoma (PPB)–primarily affects newborns and young children and forms in the lung's tissue and covering. It is tough to identify early based solely on imaging investigations, and clinical signs might be previously challenging due to the extensive range of possible differential diagnoses [4]. Since their accuracy is relatively poor, we need to improve the diagnostic accuracy of various CAD models, such as radiotherapy, image processing, and monotherapy.

In computer science, artificial intelligence (AI) and its branches, which are machine learning (ML) and deep learning (DL), have made a difference in the field of cancer diagnosis. AI is gaining popularity in enhancing patient outcomes and medical precision. It is currently being used to predict and automate the diagnosis of various types of cancer. Machine learning allows computers to learn from training data. ML has demonstrated significant predictive power for multiple cancers, including brain, liver, prostate, breast, and lung. ML techniques were integrated with medical imaging, and the result was a commonly used method for cancer diagnosis. With feature extraction as the initial stage, several strategies were researched and applied for various kinds of cancer [5]. Nevertheless, feature extraction methods have limitations that delay CAD system performance improvement. Recent emphasis has been on representation learning instead of feature extraction [6] [7].

Deep learning is a representation-learning technique that generates high-level feature representations from raw images [7]. It has achieved enormous success in many fields using Graphic Processing Units (GPUs) for massive parallel architecture. For example, convolutional neural networks (CNNs), one of the most common deep learning algorithms used in oncology, have shown promise in different types of pediatric cancer [8]. Hence, this research aims to provide an overview of popular new deep learning models that detect and diagnose PPB.

PPB is the leading cause of hereditary pediatric lung cancers arising from tissues lining the lungs and chest cavity (pleura) or pulmonary structures [9]. This is a highly aggressive and rare pulmonary malignancy that primarily occurs in children younger than six years old [10], and it is often incorrectly characterized as a respiratory tract infection until it progresses to the easily treatable stages. PPB can spread to adjacent organs through blood like other types of cancer [11].

According to the World Health Organization research, the classification of lung tumors is based on the histologic types and their subtypes. Histologic types include epithelial tumors, lymph histiocytic tumors, mesenchymal tumors, tumors of ectopic origin, or metastatic tumors. There are three types of PPB, including type I, type II, and type III, representing the purely cystic, mixed cystic solid, and purely solid [12] [13]. The detection and diagnosis of PPB types vary because Type I appears as the cystic without a nodular, Type II as the tumor, and Type III as the solid with no cystic [14]. Owing to the diversified appearance of PPB in childhood, researchers and clinicians need to discriminate the disease among various solid and cystic lung masses. Type I tumors progress over time, resulting in mutating to type II and type III tumors.

Identifying and diagnosing PPB poses challenges due to its similar patterns in related diseases. Children with PPB may exhibit atypical clinical symptoms such as shortness of breath, respiratory distress, flushing, and fever, often misdiagnosed as respiratory tract infections, pneumothorax, or pneumonia [15] [16]. In such cases, a chest radiograph is the standard imaging examination performed upon admission to diagnose this lung inflammation-like condition [15]. Chest radiographs often show reduced lung transparency, frequently misdiagnosed as pneumonia when combined with the children's symptoms [17]. Chest CT scans help diagnose PPB but provide limited diagnostic information [17]. Physicians [13] have examined the medical history, histopathology, and multimodal radiological features to identify PPB among various tumors. Observing the related malignancies in the same patient or their blood relatives, scientists have determined the association between the PPB and similar malignancies to help differentiate benign and malignant tumors. Early identification and precise differentiation of PPB are essential for accurate diagnosis and efficient treatment.

Computer vision is a rapidly growing image processing field involving automatic object recognition. Deep CNN has been beneficial in various fields like video processing, object recognition, and many more. With large datasets and hardware availability, innovative concepts like activation functions, regularization, optimization, and architecture have improved CNN performance [18]. New architectures have significantly increased the capacity for deep CNNs, enhancing computer vision [18]. Many pre-trained DL models such as VGG16, ResNet, DenseNet, and EfficientNet were based on CNN architecture that was trained on large image datasets and has dominated the expansive field of image recognition and computer vision tasks. The learned features of pre-trained models are a good starting point for significantly accelerating many custom vision applications, including disease detection and prediction. It is necessary to improve their interpretability and explainability to achieve real-time diagnosis using deep neural network traditional lung cancer detection models.

Despite the advantages, CNN models fail to capture the global or sequential correlation of objects in the images due to the inability to examine the long-term dependencies in the photos. To address these constraints, many DL applications extensively demonstrate a new type of deep learning system called vision transformer (ViT) architecture. It has been proven to work better than CNNs in tasks involving image classification, such as predicting non-small cell lung cancer (NSCLC) [19]. CNNs use a sequence of layers to extract features from images, whereas ViT uses a self-attention mechanism to look at the whole picture simultaneously. This helps ViT to understand big-picture connections in images and make more accurate predictions. Moreover, the Vision Transformer design can handle many parameters and be trained on extensive datasets, making the model more accurate. Using ViT allows for more than just looking at the position of the pixels; it also looks at how the pixels are related.

Furthermore, a ViT [20] [21] has been formulated for sequential image classification to recognize long-term dependencies in the image and emerge as a potential alternative to CNN. ViT has demonstrated its capability to learn high-quality image features and encode long-range dependencies of images. After seeing significant improvement in transformer models in natural language processing (NLP), which use self-attention mechanisms to model dependencies between words in a text, transformer architectures have been increasingly applied to computer vision (CV) applications. During the analysis of medical images, transformers play a vital role in various tasks of clinical applications, such as image reconstruction, image segmentation, image captioning, disease detection, and disease diagnosis [22] [23].

The diagnosis of childhood lung cancer, or PPB, can be a difficult task as early and accurate detection is challenging. Symptoms of PPB can be like other lung conditions, complicating the diagnosis process. Manually interpreting medical images can be slow, error-prone, and vary between observers.

Unfortunately, there is a lack of research on applying advanced sensing technologies to detect PPB early. Current models for diagnosing PPB suffer from lower detection accuracy, poor generalization, inadequate labeled data, data scarcity, high tumor variability, insufficient capturing of potential patterns, lack of attention mechanisms in CNN models, and difficulties in classifying benign and malignant patients at an early stage due to the complex patterns of inputs.

Addressing these challenges is crucial to enhancing the analysis, performance, and reliability of PPB detection. More importantly, there is a significant research gap in directly addressing the early detection of PPB using advanced technologies like DL models. Studies have shown that various sensors can detect different compounds with remarkable sensitivity. However, these technologies still need to be applied to improve the early diagnosis of PPB, presenting an opportunity for future research.

Further investigation is necessary to identify and overcome these obstacles, which will pave the way for the future of PPB detection. This can lead to a more precise discrimination and evaluation of a significant portion of PPB analysis with improved sensitivity, specificity, and accuracy.

The study uses medical image data to assess the potential of vision transformers (ViTs) in classifying pulmonary perivisceral nodules (PPB). The integration of ViTs in medical imaging presents an exciting opportunity for improved PBB detection. The research involves experiments with a pre-trained ViT model from PyTorch to determine its effectiveness in feature extraction and fine-tuning on large datasets. Customizing deep learning models is a critical aspect of the study, which includes preprocessing patient scans by converting them to grayscale, resizing, and standardizing using familiar image transformations. The dataset is then divided into training and validation sets for input into the ViT architecture, which consists of convolutional, attention, feedforward, normalization, and classification layers arranged in an encoder-decoder structure. The study also focuses on critical hyperparameters such as the AdamW optimizer, cross-entropy loss, and a batch size 32 for gradient updates.

The main objective of this research is to apply a new advanced deep learning model called vision transformers (ViTs) to identify and overcome these obstacles, which will pave the way for the future of PPB detection. Deep learning can lead to more precise discrimination and evaluation of a significant portion of PPB analysis with improved sensitivity, specificity, and accuracy. Overall, we emphasize that our research objectives are improving the analysis and performance of detection PPB at an early stage and comparing our proposed model with existing models.

II. Literature Review

Approximately half of the studied individuals used new teaching approaches, such as medical image analysis and VIT, to aid them. Accurately capturing global relations and context is paramount in medical image analysis. To achieve this, we need a tool that excels in this area. Therefore, this is where the ViT comes in - its unique capabilities make it the perfect solution for accurately capturing global relations and context in medical image analysis [24].

PPB diagnosis and imaging techniques are a vital area of research. In this regard, we aimed to examine some of the diagnostic methods currently used for PPB and imaging techniques. Our analysis revealed the nature of PPB, and the need for more sophisticated diagnostic tools in the healthcare field was apparent [25]. We also looked at a study by Shao and colleagues [26], who used a SEER database to explore rare malignant pulmonary tumors in children and adolescents. The research highlighted the significant impact of histology, differentiation grade, surgery, TNM stage, and therapeutic modalities on survival rates. The authors recommended increasing treatment experience for each tumor type to improve evidence-based practices. [26].

In 2021, Kunisaki and colleagues [27] studied pediatric lung lesions and their potential risk factors, focusing on pleuropulmonary blastoma (PPB). They used data from 521 databases across 11 pediatric hospitals in the US and employed a multivariable logistic regression algorithm to assess PPB's characteristics and risk factors. The algorithm was also applied to CT scans to enhance the detection of malignant PPB. The study showed that CT scans are ineffective in detecting malignant PPB, with a sensitivity rate of only 33.3%. However, the specificity rate is high at 98.8%, which is usually accurate when the scan indicates the absence of malignant PPB. The positive predictive value is 71.4%, meaning that when the scan shows the presence of malignant PPB, there is a moderate chance that it is accurate. The negative predictive value of 94.1% suggests that when the scan indicates the absence of malignant PPB, there is a high chance that it is correct.

The research group focused on the early diagnosis of PPB through various diagnostic tests and imaging methods. These imaging techniques include MRI, CT, and other commonly used methods for diagnosing PPB [28]. Engwall-Gill et al. [29] meticulously analyzed 477 CT scans and identified 40 cases that required extensive review. This research highlights the importance of thorough and accurate medical scan analysis, underscoring the need for continued research. The study found 9 cases (23%) had pathologically confirmed cystic PPB. The sensitivity of CT in detecting PPB was 58%, and the specificity was 83%. The overall accuracy rate for distinguishing benign and malignant lesions was 81%. Furthermore, these sources also introduced changes to allow for the appropriate diagnosis of PPB, the inconveniences of having to differentiate Pleuropulmonary Blastoma from other lung diseases, and the need to combine the image-based findings with the genetic information to have a complete diagnosis and appropriate treatment recommendations.

The groundbreaking discovery by Vu et al. [30] regarding the first-ever known case of type III late-stage PPB in a developing fetus has significantly contributed to the field of prenatal diagnosis. While deep learning models have been extensively studied, there is also a growing body of research on using VITs for diagnosing PPB. However, the literature still needs more exploration on how VITs can improve early detection rates and accurately diagnose more individuals with PPB. This study aims to bridge this gap by evaluating the potential of VITs to aid in developing more effective PPB detection techniques with utmost confidence.

To distinguish PPB from other medical conditions, doctors usually tend to apply advanced neural networks, specifically CNNs, containing state-of-the-art artificial intelligence capabilities that have accurately been proven to identify visual data patterns such as image classification and segmentation [31]. Different stages are involved in detecting PPBs; the first step would be gathering medical images marked with labels specifically related to PPBs [32]. These images can be found on websites like CT scans or

Histopathology. Bandi and Santhisri [33] used deep CNN with CT images and the DICER1 gene to classify PPB with 98.67% accuracy in CT image classification and 96% in DICER1 DNA analysis. Moreover, Helm et al. [34] employed CAD to discern pulmonary nodules on chest CT scans in the pediatric population. The study aimed to identify the efficacy of CAD in detecting pulmonary nodules in children and its feasibility as a tool for pediatric radiologists. Also, Tu et al. [35] and Chen et al. [36] applied machine learning algorithms to distinguish between harmless and malignant lung nodules on chest CT scans [35] and PET/CT scans [36].

For ViTs to be effective, they require large datasets. Amin et al. used the ISBI 2019 dataset containing thousands of images. Still, we need to refine it further. Research has shown that ViTs can be used to diagnose acute lymphoblastic leukemia (ALL) by sharing code and achieving an accuracy of 83.5% [37]. Priscilla and colleagues [38] reached 88.4% and 86.2% classification accuracies on an acute lymphoblastic leukemia dataset of 12,528 samples using the Vision Transformer and convolutional neural network models, respectively. Also, Tummala et al. successfully showed that The ViTs ensemble model has established an impressive performance in classifying brain tumors from MRI scans at a resolution of 384 × 384 [39]. It has achieved an overall test accuracy of 98.7% and a specificity of 99.4%, which is either at par or better than the previous CNN models [39]. Hence, using an ensemble of finetuned ViT models, computer-aided diagnosis of brain tumors from MRI can reduce the burden on clinical radiologists. Moreover, Liang and Zheng [40] utilized a transfer learning model to diagnose childhood pneumonia by using the ChestX-ray14 dataset. As a result, the experiment results show that the recall rate is 96% and the f1-score is 92.7%.

Research on diagnosing the rare pediatric lung complication PPB collectively highlights the challenge of obtaining an accurate and timely diagnosis. This study addresses the inherent difficulties of PPB diagnosis, especially its diverse and complex underlying images. The study's primary aim is to tackle the research difficulties related to PPB that have been thoroughly discussed, including enhancing imaging methods and developing new diagnostic procedures to ensure precise patient diagnosis [41]. The primary objective is to improve early detection methods and refine strategies for managing this rare malignancy. In screening trials, the primary aim is to determine whether different imaging techniques, such as CT scans, can differentiate PPB from other pulmonary complications [42]. Moreover, this study attempts to define radiologic and histopathologic subtypes based on three different classification methods expressing this highly complex disease.

III. DATASET

The LIDC-IDRI [43], [44] is a collection of different image modalities, such as thoracic CT (computed tomography), CR (computed radiography), and DX (digital radiography) images, that have been annotated to detect lesions, providing a valuable resource for the development, training, and evaluation of computer-assisted diagnostic methods for lung cancer detection and diagnosis. This dataset includes 244,527 images with a total image size of 125GB.

In total, it contains 1,018 cases with scans. Each case typically has images from a clinical thoracic scan of a single patient. Multiple radiologists annotate the scans to mark any observable lung nodules and rate their characteristics. This provides "ground truth" labels and diagnosis data for machine learning algorithms. Both malignant and benign lung nodules are included, along with non-nodule abnormalities. Approx 60% of cases have ≥ one nodule ≥ 3 mm. The dataset is one of the largest publicly available collections of lung scans with annotations. This makes it very valuable for research into computer-aided diagnosis of lung cancer.

IV. PROPSED MODEL ARCHITECTURE

Fig. 1 depicts the architecture of the proposed model. Before applying our proposed model, we must prepare our medical dataset. This system's proposed structure comprises an initial section that applies patches to the input images, a middle section that employs a multilayer Transformer encoder, and a concluding section that converts the resulting global representation into the output label. In our approach, each trial will run using the Vision transformers ViT, a base model with 16 × 16 image patch size (ViTb_16) architecture and will be provided in the torchvision package in PyTorch. However, we do some modifications on the original base ViT, such as replacing the multilayer perceptron (MLP) head with a linear layer for binary classification. Also, utilizing a pre-trained network, we will then analyze the effectiveness of the network when using it just as a feature extractor. This research employs a highly effective transfer learning method that fine-tunes pre-trained models on the model architecture to increase accuracy. Our approach yields superior results by leveraging the pre-trained ViT model from torchvision and refining it on the LIDC-IDRI dataset while freezing the base layers and modifying the new classifier layer. The pre-trained ViT Base-

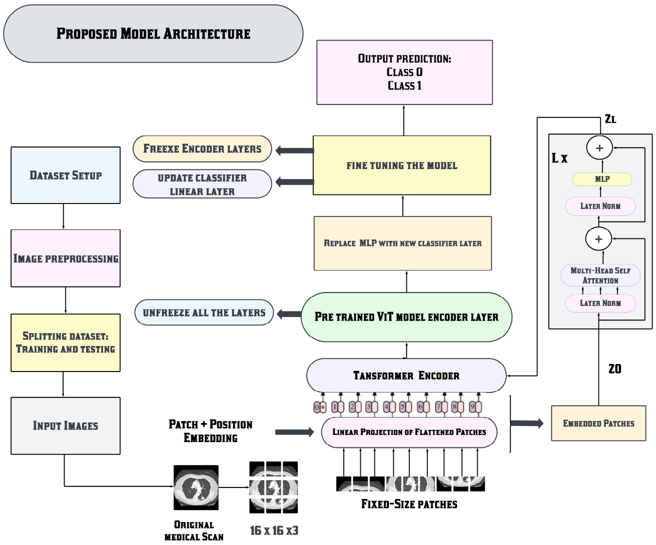


Fig. 1. Proposed model architecture

16 model is optimized to provide the mean for the unique representation, making it an exceptional choice for this work.

A. Patch embedding

The initial step in executing a ViTb-16 model entails dividing every image in the input dataset into a constant number of patches. These patches are subsequently projected linearly with the aid of learnable positional embeddings that facilitate the identification of the patch sequence. Following this, we utilize a transformer encoder and a novel classification linear layer for conclusive classification, as depicted in Fig. 2.

In the first step, the input image is divided into nonoverlapping patches because a transformer takes a 1D sequence of the token as input. Since images are usually in 2D format, to handle them, if we consider an input image X of size (H, W, C), Where H = image height, W = image width, C = number of channels (e.g., 3 channels for RGB), is embedded into a feature vector of shape (n+1, d), following a sequence of transformations. This corresponds to equation (1).

$$Z_0 = [\, X_{class};\, X_p^1 E, X_p^2 E, \dots, X_p^n E \,] + E_{pos} \tag{1}$$

Where:

$E = P^2 \times C, E_{pos} = (N+1) \times D, N = total\ number\ of\ patches,$
$D = model\ dimension, and$
$P =$ predefined parameter, in raster order (left to right, up to down).

To calculate N, specifying height (H) and width (W) both as P, distinct image patches of size P x P as in (2).

$$N = \frac{H \times W}{P^2} \tag{2}$$

B. Transformer encoder

Sequence Z0 passes through a transformer encoder architecture consisting of multiple blocks. Each block contains three major processing elements: layer norm, multi-head self-attention (MSA), two layers of multi-layer perceptron (MLP), and residual connections in between, as shown in Fig. 1.

Layer normalization is essential for stabilizing the dynamics of hidden states and reducing training time. It utilizes the scaling process's mean and standard deviation for each training example. The resulting features undergo multiplication by a scaling factor and addition to a shifting factor, both of which are trainable during the training process. This feature enables the independent normalization of input data within a given batch. It is not reliant on batch size, making it suitable for various batch sizes.

Residual connections provide alternative paths for gradients, effectively resolving the problem of vanishing gradients in very deep architectures.

The transformer model's multi-head self-attention (MSA) mechanism is highly effective for identifying significant image multiple regions, mainly for detecting lesions. It utilizes multiple attention mechanisms in parallel, each with its parameters. The MSA layer is based on the self-attention mechanism, enabling the model to assess the importance of different image regions when making predictions. By generating attention maps from embedded visual tokens, MSA can calculate the most relevant patches and discard the irrelevant ones. MSA consists of four layers: linear, self-attention, concatenation, and multiple heads that combine the output. Attention weights represent the attention mechanism calculated from a weighted sum of the sequence's values. The input sequence generates Q (query vector), K (key vector is number of heads), and V (value vector) three values by multiplying the elements (Q, K) with three learning matrices UQKV, making MSA an exceptional tool for identifying key features in data. To calculate the importance of one element concerning others in each input sequence, the value of the Q vector is multiplied by the dot product with the K vectors. The result is then scaled and passed to the SoftMax activation function to determine the high attention score patch's importance, as given mathematically in equation (3).

$$A\,(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{D_h}}\right) \tag{3}$$

$Where{:}\ D_h = {}^{D}/_{K}$

The Multi-Head Self-Attention (MSA) mechanism utilizes a collection of attention heads h instead of relying solely on individual values of Q, K, and V. By concatenating the outputs of each SA and projecting them through a feed-forward layer with learnable weights W, the MSA achieves robust and optimal feature selection.

Multi-layer perceptron (MLP) is an artificial neural network with multiple layers of artificial perceptron or neurons. It operates as a feedforward network, with information flowing linearly from the input to the output layer. The activation function for neurons in neural networks is the Gaussian Error Linear Unit (GELU), which is an activation function defined as f(x) = x * Φ(x), where Φ is the cumulative density function of a standard normal distribution. GELU is used to weigh the input layer.

MLP is a two-layer classification network with GELU at the end. The final MLP block, the MLP head, is the transformer's output. Applying SoftMax to this output can provide classification labels. After the patch embedding, the location and class embeddings are merged and fed into the transformer encoder. The resulting output of the transformer encoder is then processed by a structural system that includes an MLP unit. This unit incorporates an activation function and a fully connected layer, with the GELU (Gaussian Error Linear Unit) being the chosen activation function, as shown in (4).

$$GLEU\,(x) = 0.5x\,(1 + \tanh(\sqrt{\frac{2}{\pi}}\,(x + 0.044715\,x^3))) \tag{4}$$

C. Linear classifier layer

The base ViT architecture ends with an essential MLP head as a classification head. However, our proposed model replaces the MLP head classifier with a linear layer for binary classification. It consists of one hidden layer during pre-training and a single linear layer in the fine-tuning stage. We derive the model's input from the state of the classification token, located at the output of the Transformer encoder. The resulting output of the model comprises a sequence of logits, which confidently correspond to two different classes (class 0 and class 1). It has an out_features dimension equal to the number of classes in the small dataset.

The initial element Z_L^0 in the sequence is confidently directed to an external head classifier in the final layer of the encoder to predict the class label accurately. The output feature map Z_L from the top transformer encoder layers L_X is projected into a class probability distribution using a SoftMax over learnable linear classifier weights, as in (5).

$$Predition = SoftMax\,(W_h Z_L + b_h) \qquad (5)$$

D. Pretrain proposed model

Our model is explicitly trained for object detection and classification, enabling it to learn the features of objects in its training dataset successfully. Pre-training makes it highly adaptable to related tasks. As a result, our model trained to identify the presence of a mass or tumor in an image can be easily fine-tuned to identify infected or uninfected samples. Our model is pre-trained with image labels, a supervised model on a vast collection of 1.2 million images and 1k classes called ImageNet-1k, using the torchvision model. We use transfer learning, in which the pre-trained encoder layers remain fixed. As shown in Figure 4, we unfreeze (train) all the encoder model and classification layers (For pre-training, a 2-layer MLP is used). These layers learn robust and general features from a large dataset such as ImageNet. To prepare the images for the ViT models, we create a data pipeline that performs various transformations, including resizing them to 256 x 256 using interpolation=InterpolationMode.BILINEAR and apply a central crop of 224. We then convert them to tensors, rescale them to a range of [0, 1], and normalize them with a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225].

E. Fine-tuning proposed model

Fine-tuning a vision transformer model is a crucial step that involves adapting a pre-trained model to a large-scale dataset and refining it for smaller downstream tasks. This process effectively utilizes the knowledge and representations learned by the pre-trained model, resulting in significant savings in both time and computational resources. During this step, the pre-trained step allows the model to utilize the knowledge learned during pre-training while allowing the new task-specific linear layer to be trained from scratch or fine-tuned on the target dataset, the LIDC-IDRI dataset by freezing the base layers and modifying the classifier layer. During fine-tuning, it is generally advisable to use a higher resolution than the one used for pre-training. This approach is unique in transfer learning, which leads to much more significant performance metrics and analysis of classification PPB.

V. Experiments and Discussion

A. Implementation details

All experiments in this paper were conducted on an Intel Core i9 24-Core, 13th generation processor with 2.2 GHz speed, 32 GB RAM, and an NVIDIA GeForce RTX 4080 GPU with 12 GB Graphics Double Data Rate 6. We used PyTorch, the most popular framework for implementing ViT, as the framework in this study.

B. Preprocessing medical images(data setup)

The first step in the data preparation process is to combine the file paths pointing to the lung scan images with unique subject identifiers for each scan. This metadata is compiled into a CSV file called metadata.csv. Next, the actual DICOM medical image files are loaded. These DICOM images contain the raw pixel data from the scans. This data then goes through a conversion and preprocessing step - the pixel values are transformed into a grayscale 8-bit image compatible with machine learning algorithms. Data augmentation techniques from the torchvision library are also applied to expand the diversity of the training data. Finally, the preprocessed scan images are saved as JPG files. The output is another CSV file called image_data, which contains the file paths linked to each preprocessed scan image, ready for input into the vision transformer model. Some of the image examples from the output image_data are shown below in Fig. 2. In summary, the data preparation converts the raw medical images into a standardized, augmented format that facilitates training.

C. Splitting dataset

We confidently load the original input images into training and validation Torch data loaders with an 80/20 split, 80% for the training data loader and 20% for the Test data loader, as in Fig. 2. The input image goes through the data loaders, which apply the necessary transformations of resizing and normalization specified by the auto transforms obtained from the pre-trained ViT model's weights.

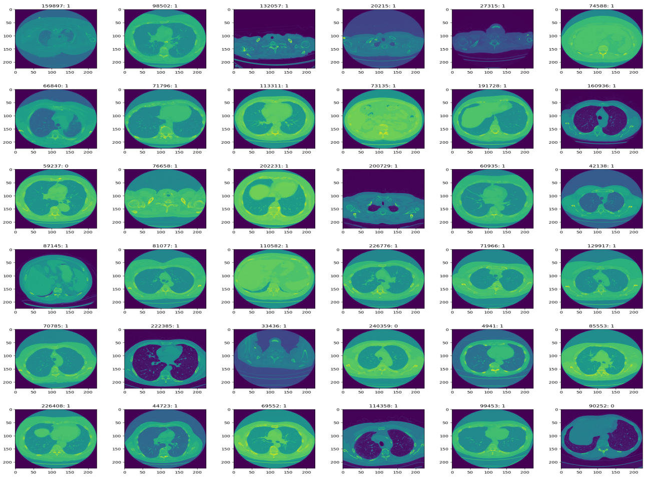


Fig. 2. Data preprocessing output images

D. Training process

The training process is initiated using the engine. Train function. It initially trains the model for a single epoch, which means it processes the entire training dataset simultaneously. During training, the model learns from the data and updates its parameters to minimize the loss. The cross-entropy loss function is used during the training process, with the initial learning rate set to 0.00001, and the AdamW optimizer is utilized. We set the batch size of the model to 32. The training time complexity is very high, so we have set the epochs to 50 to train the model as shown in Fig. 3. Also, during training, the model learns from the data and updates its parameters to minimize the loss. The training process updates the weights of the classifier layer we added to the pre-trained ViT model so it can accurately predict the target classes. We first define the loss function and optimizer. Cross-entropy loss is used as a standard for classification tasks. It measures the divergence between predicted class probabilities and proper labels. The AdamW optimizer adjusts the learning rate of each weight based on the gradient magnitude. Regularization is used to prevent overfitting. We set manual seeds to ensure reproducible results across runs. Then, we start a timer to track the total training time—the engine. The train function handles the training loop. On each epoch, it iterates through the entire training set in batches. For every batch, predictions are made and compared to the actual labels to compute the loss with cross-entropy. Gradients are calculated by backpropagation to determine which weights contributed most to the loss. Using these gradients, the AdamW optimizer adjusts the classifier weights slightly to reduce the loss. After every epoch, it calculates training, validation, loss, and accuracy statistics to monitor progress. During training for one epoch, we use early stopping if overfitting occurs.

The training process uses the loss as a feedback signal to tune the classifier weights for optimal performance on the validation data. This continued adjustment slowly builds a specialized mapping from images to predicted classes.

E. Model evaluation

After training the vision transformer model for PBB detection, as in Fig. 3, systematic quantitative analysis was done to validate real-world effectiveness. The proposed research work evaluates performance across several key metrics. First, a confusion matrix categorizes predictions as true/false positives/negatives. This underpins accuracy, i.e., the percentage of total correct predictions. While valuable, accuracy alone can be misleading if positive/negative classes are imbalanced. Additional class-specific metrics are calculated.

Precision reveals the percentage of correct PBB detections. High precision signifies that most identified regions represent actual PBB cases, not false alarms. Maximizing precision minimizes the time doctors waste investigating incorrect malignant warnings. However, high precision could come at the cost of missing PBB cases.

Therefore, recall, or sensitivity, is assessed - the proportion of total PBB cases correctly detected. Higher recall means fewer missed PBB cases, critical for lifesaving early diagnosis. However, maximizing recall risks overdiagnosis and overtreatment if some detections are wrong.

The F1 Score balances precision and recall into a singular metric critical for imbalanced classes. In the results, the F1 Score is 0.9967, indicating high precision without much PBB case detection miss rate, achieving a superior outcome. Since incorrectly flagged healthy patients as having cancer causes psychic trauma, specificity evaluation is prudent. Specificity measures the proportion of correctly cleared negative cases. Specificity causes false alarms, whereas high specificity assures truly healthy patients. Here, specificity is lower at 0.730, suggesting room for improvement in reducing false positives.

Finally, the ROC curve evaluates discrimination ability across all sensitivity/specificity levels attainable by adjusting the decision threshold. The area under this curve (AUC) provides an aggregate measure of model viability. An AUC of 1 represents perfect classification. The curve and AUC supplement the other metrics to deliver a comprehensive big-picture analysis.

In totality, multi-angle evaluation assesses real-world utilization accuracy, reliability in precisely detecting tumors, avoiding false alarms, and robustness against variability. Thorough vetting on varied factors inspires confidence in applying the model clinically to aid doctors in efficient and accurate PBB screening to save lives.

Our proposed model's effectiveness was evaluated using recall, F1 score, precision, and accuracy metrics. True positive (TP), false positive (FP), true negative (TN), and false negative (FN) examples are used to calculate the metrics:

- True Positive (TP): when both the actual and anticipated labels are positive.
- False Positive (FP): When the expected label is positive, but the actual label is negative.
- True Negative (TN): When the projected and correct labels are negative.
- False Negatives (FN): They are labels projected to be pessimistic but optimistic.
- Accuracy: This evaluation statistic identifies the model's overall performance. It is the ratio of the total number of predictions to the total number of correct predictions, as in (6).

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{6}$$

- Precision: This evaluation metric computes the proportion of all positive samples over the total of positive samples that were correctly or incorrectly categorized, as in (7).

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

- Recall: This assessment metric determines the proportion of all positive samples over the number

of correctly identified positive input samples, as in (8).
Recall, also known as Sensitivity, is the valid positive rate.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

- Specificity: This is another critical metric for evaluating classification models, particularly binary classification. It measures a model's ability to correctly identify negative instances, as in (9). Specificity is also known as the True Negative Rate.

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

- F1-score: This evaluation metric combines precision and recall while giving false positives and negatives more weight to determine the model's accuracy, as in (10).

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \tag{10}$$

F. Results

As in Fig. 4, we are pleased to present the results of our proposed model's performance evaluation, which demonstrated that our model is a reliable and accurate tool for identifying positive and negative instances. Our results indicate that our model correctly identified 43,116 cases of PPB, as in Figure 5, with a high precision rate of 99.78%. In addition, our model successfully identified almost all (99.89%) of the actual PPB cases, with an overall accuracy of approximately 99.37%. The F1 Score, the harmonic mean of precision and recall, indicated a good balance between the two metrics, providing additional confidence in the model's performance. Furthermore, the area under the curve (AUC) value of 0.99, as in Figure 6, representing the model's discrimination ability across all sensitivity/specificity levels, provides an aggregate measure of the model's viability. This value indicates that our model is reliable for accurately identifying positive and negative instances. We are confident that our proposed model's performance evaluation provides a comprehensive big-picture analysis of its reliability, accuracy, and viability and will help identify positive and negative instances in the future.

G. Comparison of the proposed model with state arts

Though the DCNN [33] model shows promise for improving PPB diagnosis, potentially leading to earlier detection and better outcomes for children, it fails to deliver with larger and unbalanced datasets. The ViT model has several advantages over it. Here is a comparison of some of the critical benefits of using a ViT over a CNN model for PBB detection:

1) Better modeling of global context: The self-attention mechanism in ViT models enables modeling longer-range dependencies in images. This allows for better incorporation of global context, which is crucial for medical image analysis. CNNs have a more localized receptive field.

2) Reduced need for data augmentation: ViT models are more robust to variations in input data, so they require fewer intensive data augmentation. The tokenization process also makes them invariant to low-level input transformations.

3) Enhanced transfer learning: Pre-trained ViT models can be fine-tuned with much less medical data and provide better feature representations. They offer more flexible generalizations compared to CNNs.

4) Reduced overfitting: The attention layers have fewer parameters, reducing the chance of overfitting, especially with smaller medical imaging datasets. ViTs generalize better with less tuning.

5) Interpretability: The attention weights provide some level of interpretability, allowing visualization of imaged regions that contribute most to predictions. This supports model transparency for medical use.

ViTs capture more meaningful global context, are more robust to image variations, provide effective transfer learning, avoid overfitting, and offer model transparency—all valuable attributes for analyzing complex medical scans like PBB detection. With the appropriate fine-tuning, ViTs can outperform conventional CNN models.

Table I compares proposed and existing DCNN models for detecting PPB using medical imaging. The vision transformer model achieved a precision of over 99.47% and a recall of 99.9% in identifying actual positive cases of PPB from the scans. The high F1 score of 0.99 further highlights the balance between precision and sensitivity achieved by the model, even for an imbalanced dataset.

TABLE I. PERFORMANCE COMPARISON OF CURRENT AND EXISTING MODELS

Models	Total Images	Recall, Sensitivity	Accuracy
DCNN	500	98.34	98.67
Proposed Model (ViT)	**2,45,000**	**99.89**	**99.37**

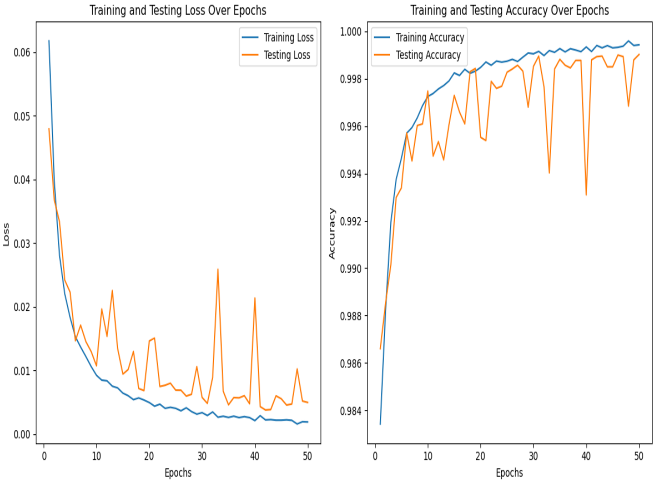
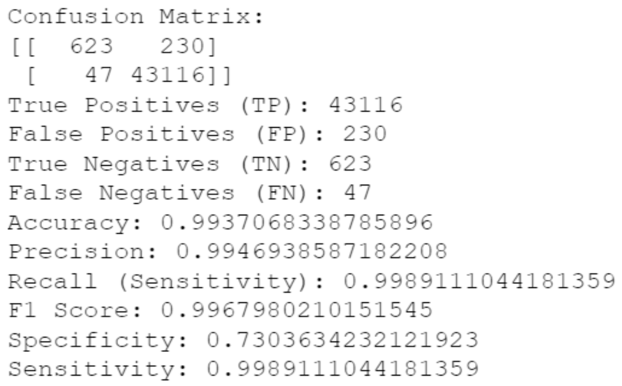


Fig. 3. Training and testing loss and training and testing accuracy over 50 epochs



```
Confusion Matrix:
[[  623   230]
 [   47 43116]]
True Positives (TP): 43116
False Positives (FP): 230
True Negatives (TN): 623
False Negatives (FN): 47
Accuracy: 0.9937068338785896
Precision: 0.9946938587182208
Recall (Sensitivity): 0.9989111044181359
F1 Score: 0.9967980210151545
Specificity: 0.7303634232121923
Sensitivity: 0.9989111044181359
```

Fig. 4. Performance summary of Vit model

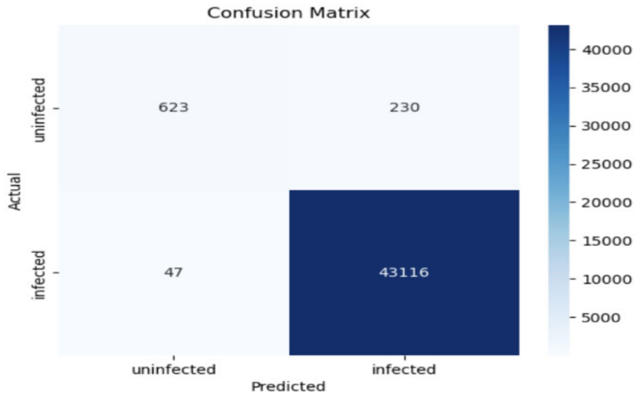


Fig. 5. Confusion matrix of evaluation model

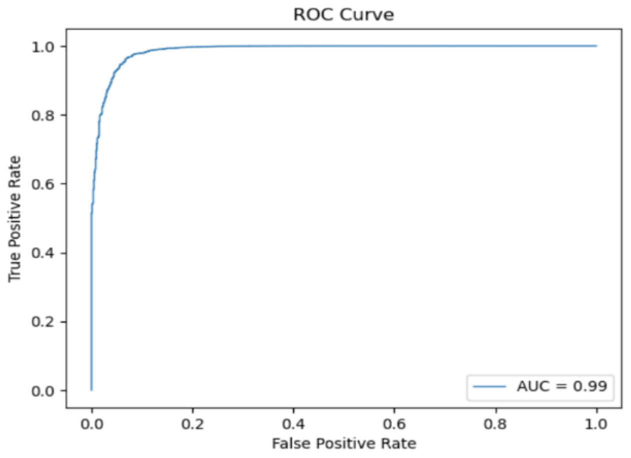


Fig. 6. ROC curve of the proposed model's performance evaluation

VI. CONCLUSION

The research focuses on applying the Vision Transformer (ViT) architecture for medical image classification, specifically for detecting PPB detection. The process involved several vital steps that demonstrated both the versatility of ViT models and their ability to achieve high performance on specialized tasks. The result was achieved with intelligent data preprocessing, which prepared the raw DICOM medical images for the model by converting them to standardized JPG files. This enabled the ViT model to ingest and interpret the data. Additional data splitting created dedicated training and test sets to evaluate model performance properly. The model implementation leveraged a pre-trained ViT network, demonstrating the power of transfer learning for medical applications where training data may be limited. By utilizing strong baseline feature representations known on natural images, the model could generalize effectively to PBB detection. The high-test accuracy increased above 99% after 50 epochs. The F1 score of 0.9939 demonstrates the excellent balance between precision and recall. The model excels at the most crucial task of identifying positive PPB diagnoses. As a specialized diagnostic tool, this ability to reliably flag potential cases can provide immense value. The proposed work displays how modern deep learning approaches like ViT can unlock strong performance on niche medical challenges, given careful implementation. The model shows immense promise as the foundation of a decision support system to aid clinicians in PPB detection. The ViT architecture could produce clinical impact and improve patient outcomes with further refinement of additional data. Further research into model optimization and combining radiomic/genomic modalities via neural architecture search could improve sensitivity, specificity, and efficiency for practical clinical decision support systems.

ACKNOWLEDGMENT

I sincerely thank all our professors who have served as our research supervisors. Their unwavering assistance, support, and constructive feedback have been invaluable in completing this research work. We are deeply grateful for their patience and enthusiastic encouragement, which have helped me to persevere through the challenges encountered during this research process. This research project would not have been possible without their expert guidance and meticulous critiques.

REFERENCES

[1] A. C. Society, "Cancer Facts & Figures 2020," 2020. [Online]. Available: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2020.html.

[2] D. Crosby et al., "Early detection of cancer.," *Science,* vol. 375, no. 6586, p. eaay9040, 2022.

[3] "U.S. Cancer Statistics: Lung Cancer Stat Bite | CDC," www.cdc.gov, 2023. [Online]. Available: https://www.cdc.gov/cancer/uscs/about/stat-bites/stat-bite-lung.htm.

[4] M. K. Dishop et al., "Fetal Lung Interstitial Tumor (FLIT): A Proposed Newly Recognized Lung Tumor of Infancy to Be Differentiated From Cystic Pleuropulmonary Blastoma and Other Developmental Pulmonary Lesions.," *The American Journal of Surgical Pathology,* vol. 34, no. 12, pp. 1762-1772, 2010.

[5] E. Gibson et al., "NiftyNet: a deep-learning platform for medical imaging," *Computer Methods and Programs in Biomedicine,* vol. 158, pp. 113-122, 2018.

[6] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 8, pp. 1798-1828, 2013.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-444, 2015.

[8] Y. Yang, Y. Zhang, and Y. Li, "Artificial intelligence applications in pediatric oncology diagnosis.," *Exploration of targeted anti-tumor therapy,* vol. 4, no. 1, p. 157–169, 2023.

[9] W. D. Foulkes, J. R. Priest, and T. F. Duchaine, "DICER1: mutations, microRNAs, and mechanisms. " *Nature Reviews Cancer,* vol. 14, no. 10, p. 662–672, 2014.

[10] J. C. Manivel et al., "Pleuropulmonary blastoma. The so-called pulmonary blastoma of childhood," *Cancer,* vol. 62, no. 8, p. 1516–1526, 1988.

[11] ". B. -. N. (. O. f. R. Disorders)", "Rare diseases," NORD (National Organization for Rare Disorders), 2015. [Online]. Available: https://rarediseases.org/rare-diseases/pleuropulmonary-blastoma/.

[12] J. R. Priest, M. B. McDermott, S. Bhatia, J. Watterson, J. C. Manivel, and L. P. Dehner,, "Pleuropulmonary blastoma," *Cancer,* vol. 80, no. 1, pp. 147-161, 1997.

[13] M. N. Koss, L. Hochholzer, and T. O'Leary, "Pulmonary blastomas.," *Cancer,* vol. 67, no. 9, pp. 2368-2381, 1991.

[14] L. P. Dehner, "Pleuropulmonary blastoma is THE pulmonary blastoma of childhood. " *PubMed,* vol. 11, no. 2, p. 144–151, 19954.

[15] J. Geiger et al., "Imaging findings in a 3-year-old girl with type III pleuropulmonary blastoma.," *PubMed,* vol. 21, no. 6, p. 1119–1122., 2008.

[16] Dmitry Bobylev, G. Warnecke, N. Dennhardt, and A. Horke, "Giant pleuropulmonary blastoma. " *European Journal of Cardio-Thoracic Surgery,* vol. 50, no. 6, p. 1215–1216, 2016.

[17] *Case Report: Pleuropulmonary Blastoma in a 2.5-Year-Old Boy: 18F-FDG PET/CT Findings,* 2021.

[18] D. Bhatt et al., "CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope.," *Electronics,* vol. 10, no. 20, p. 2470, 2021.

[19] A. Fanizzi et al., "Comparison between vision transformers and convolutional neural networks to predict non-small lung cancer recurrence," *Scientific Reports,* vol. 13, no. 1, p. 20605, 2023.

[20] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys,* vol. 54, no. 1557-7341, p. 1–41, 2022.

[21] K. Islam, "Recent Advances in Vision Transformer: A Survey and Outlook of Recent Work," *arXiv.org,* p. 2203.01536, 2023.

[22] F. Shamshad et al., "Transformers in Medical Imaging: A Survey," *Medical Image Analysis,* vol. 88, p. 102802, 2023.

[23] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with Transformers? A comparative review of critical properties, current progress, and future perspectives. " *Medical Image Analysis,* vol. 85, p. 102762, 2023.

[24] Chukwuemeka Clinton Atabansi, J. Nie, H. Liu, Q. Song, Y. Li, and X. Zhou, "A survey of Transformer applications for histopathological image analysis: New developments and future directions," *Biomedical Engineering Online,* vol. 22, no. 1, 2023.

[25] C. Wang, T. He, H. Zhou, Z. Zhang, and C. Lee, "Artificial intelligence enhanced sensors - enabling technologies to next-generation healthcare and biomedical platform," Bioelectronic Medicine, vol. 9, no. 1, p. 17, 2023.

[26] W. Shao et al., "Primary lung cancer in children and adolescents: Analysis of a surveillance, epidemiology, and results database," *Frontiers Media S.A.,* vol. 13, 2023.

[27] S. M. Kunisaki et al., "Pleuropulmonary blastoma in pediatric lung lesions," *Pediatrics,* vol. 147, no. 4, 2021.

[28] N. Adams, T. Victoria, E. R. Oliver, J. S. Moldenhauer, N. S. Adzick, and G. Colleran, "Fetal ultrasound and magnetic resonance imaging: a primer on how to interpret prenatal lung lesions," Pediatric Radiology, vol. 50, no. 13, p. 1839–1854, 2020.

[29] A. J. Engwall-Gill et al., "Accuracy of Chest Computed Tomography in Distinguishing Cystic Pleuropulmonary Blastoma From Benign Congenital Lung Malformations in Children.," *JAMA Network Open,* vol. 5, no. 6, p. e2219814, 2022.

[30] N. D. Vu et al., "A rare case of pleuropulmonary blastoma detected in the fetus. Radiology case reports," *Radiology Case Reports,* vol. 17, no. 9, p. 3251–3255, 2022.

[31] S. Solanki, C. M. Pandey, R. K. Gupta, and B. D. Malhotra, "Emerging trends in microfluidics-based devices," *Biotechnology Journal,* vol. 15, no. 5, p. 1900279, 2020.

[32] S. S. Leung, A. Donuru, V. Kandula, M. R. Parekh, and D. Saul, "Multimodality Imaging of Pleuropulmonary blastoma: pearls, pitfalls, and differential diagnosis," *Seminars in Ultrasound, CT and MRI,* vol. 43, no. 1, pp. 61-72, 2022.

[33] R. Bandi and T. Santhisri, "Implementation of a deep convolution neural network model for identifying and classifying Pleuropulmonary Blastoma on DNA sequences," *e-Prime - Advances in Electrical Engineering, Electronics and Energy,* vol. 5, p. 100233, 2023.

[34] E. J. Helm et al., "Computer-aided detection for the identification of pulmonary nodules in pediatric oncology patients: initial experience," *Pediatric Radiology,* vol. 39, no. 7, p. 685–693, 2009.

[35] S.-J. Tu, C.-W. Wang, K.-T. Pan, Y.-C. Wu, and C.-T. Wu, "Localized thin-section CT with radiomics feature extraction and machine learning to classify early-detected pulmonary nodules from lung cancer screening.," *Physics in Medicine & Biology,* vol. 63, no. 6, p. 065005, 2018.

[36] S. Chen et al., "Diagnostic classification of solitary pulmonary nodules using dual time 18F-FDG PET/CT image texture features in granuloma-endemic regions.," *Scientific Reports,* vol. 7, no. 1, p. 9370, 2017.

[37] Sifatul Amin, Samin Jawed, Rejuan Rashed Raj, Sabbir Ahmed Saimoon, Rakibuzzaman Rayhan, "Vision Transformer (ViT) Approach in Computer Aided Diagnosis of Acute Lymphoblastic Leukemia," *Brac University,* 2023.

[38] P. Cho, S. Dash, Aristeides Tsaris, and H.-J. Yoon, "Image transformers for classifying acute lymphoblastic leukemia," *Medical Imaging 2022: Computer-Aided Diagnosis,* vol. 12033, no. 2F, 2022.

[39] S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling," *Current Oncology,* vol. 29, no. 10, pp. 7498-7511. 2022.

[40] G. Liang and L. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," *Computer Methods and Programs in Biomedicine,* vol. 187, p. 104964, 2020.

[41] S. Solanki, C. M. Pandey, R. K. Gupta, and B. D. Malhotra, "Emerging trends in microfluidics-based devices," *Biotechnology Journal,* vol. 15, no. 5, 2020.

[42] S. H. Choi et al., "Quantitative computed tomographic imaging–based clustering differentiates asthmatic subgroups with distinctive clinical phenotypes," *Journal of Allergy and Clinical Immunology,* vol. 140, pp. 690-700.e8, 2017.

[43] Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., Kazerooni, E. A., MacMahon, H., Van Beek, E. J. R., Yankelevitz, D., Biancardi, A. M., Bland, P. H., Br, "Data From LIDC-IDRI [Data set].," The Cancer Imaging Archive., 2015. [Online]. Available: https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX.

[44] Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, Van Beeke EJ, Yankelevitz D, Biancardi AM, Bland PH, Brown MS, Engelmann RM, Laderach GE, Max D, Pais RC, Qing, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans.," *Medical Physics,* vol. 38, p. 915–931, 2011.