IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 July 2024, accepted 21 August 2024, date of publication 23 August 2024, date of current version 15 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3449230

RESEARCH ARTICLE

Transformer-Based Hierarchical Model for Non-Small Cell Lung Cancer Detection and Classification

MUHAMMAD IMRAN[1], BUSHRA HAQ[2], ERSIN ELBASI[3], AHMET E. TOPCU[3], (Member, IEEE), AND WEI SHAO[1], (Member, IEEE)

[1]Department of Medicine, University of Florida, Gainesville, FL 32610, USA
[2]Department of Computer Science, FICT, BUITEMS, Quetta 87300, Pakistan
[3]College of Engineering and Technology, American University of the Middle East, Kuwait 15453, Kuwait

Corresponding author: Muhammad Imran (imran.muet@gmail.com/muhammad.imran@medicine.ufl.edu)

ABSTRACT Lung cancer is the leading cause of cancer-related deaths worldwide. Early diagnosis significantly improves the 5-year survival rate from 6% in patients with metastatic cancer to 60% in those with localized cancer. Histopathological examination is the gold standard for lung cancer diagnosis, but analyzing whole slide images (WSI) is time-consuming and prone to error for pathologists. This study aims to enhance the classification accuracy of non-small cell lung cancer (NSCLC) histopathological images by proposing a novel deep-learning architecture that integrates convolutional neural networks (CNNs) and vision transformers (ViTs). The model classifies NSCLC into three categories: normal, adenocarcinoma, and squamous cell carcinoma. CNNs are employed to capture local features, while ViTs are used to understand long-range relationships between image patches. We trained and validated our model on the LC25000 dataset, a benchmark dataset for NSCLC histopathology image classification. Our proposed model demonstrated superior performance, achieving an accuracy of 0.988, an F-1 score of 0.980, specificity of 0.991, recall of 0.982, and precision of 0.980, outperforming existing state-of-the-art methods. Additionally, our model achieved a low inference time of 1.816 ms, highlighting its potential for real-world applications where both accuracy and speed are critical. Our code is now publicly available at https://github.com/ImranNust/LungCancerDetection to facilitate further research and validation of our findings.

INDEX TERMS Non-small cell lung cancer, neural network, vision transformers, convolutional neural networks, classification, adenocarcinoma, squamous cell carcinoma.

I. INTRODUCTION

Cancer is regarded as the leading cause of death worldwide compared to other diseases [1], [2]. According to a World Health Organization (WHO) report [3], over 10 million cancer-related deaths were documented in 2020. Lung cancer ranks second among the various types of cancers in terms of prevalence and first in terms of fatalities. According to the American Cancer Society [4], lung cancer – with a total expected case count of 236,740 for 2022 – has become the second most common cancer in the United States. Of those lung cancers, over 55% are predicted to be fatal and would not respond to any treatment. Statistics are similar outside of the U.S. For example, according to the WHO [3], 2.21 million lung cancer cases were reported globally in 2020, making it the second most common cancer overall after breast cancer (2.26 million instances). However, lung cancer surpasses all other cancer types in terms of fatalities, with an estimated 1.8 million deaths worldwide in 2020. Lung cancer incidence and prevalence are influenced by various factors, including radon exposure and air pollution. Smoking, however, is one of the main causes since it accounts for over 80% of lung

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeswari Sundararajan.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

cancer incidences [4]. Lung cancer cases are slowly dropping as people become more aware of the harmful effects of smoking; nevertheless, it still tops the list of fatal cancers. Late diagnosis is one of the primary causes of this low survival rate [2], [5].

Based on histological heterogeneity and molecular subtypes, lung cancer is divided into two types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) [6]. Of these, NSCLC is the most commonly occurring lung cancer, accounting for more than 80% of cases [5]. NSCLC is further subdivided into adenocarcinoma, squamous cell carcinoma, and large cell carcinoma; all three subtypes fall under NSCLC because they follow the same course of treatment and prognosis. Each of these subtypes is discussed briefly below:

- **Adenocarcinoma:** The most prevalent histologic subtype of non-small cell lung cancer is adenocarcinoma, which accounts for about 40% of all occurrences [7]. Although it is the most common type of lung cancer in non-smokers, it primarily affects current or former smokers. Obesity and large body mass are additional factors influencing the prevalence of adenocarcinoma [8].
- **Squamous Cell Carcinoma:** The second most common type of NSCLC is squamous cell carcinoma. It frequently develops in the central lung or central airway, such as the left or right bronchus. Alcohol and tobacco use are the main causes of lung squamous cell carcinoma [8].
- **Large Cell Carcinoma:** Large Cell Carcinoma, which affects the outer portions of lungs, is the third kind of NSCLC. Compared to other types of NSCLC, it only accounts for approximately 10% of cases [9].

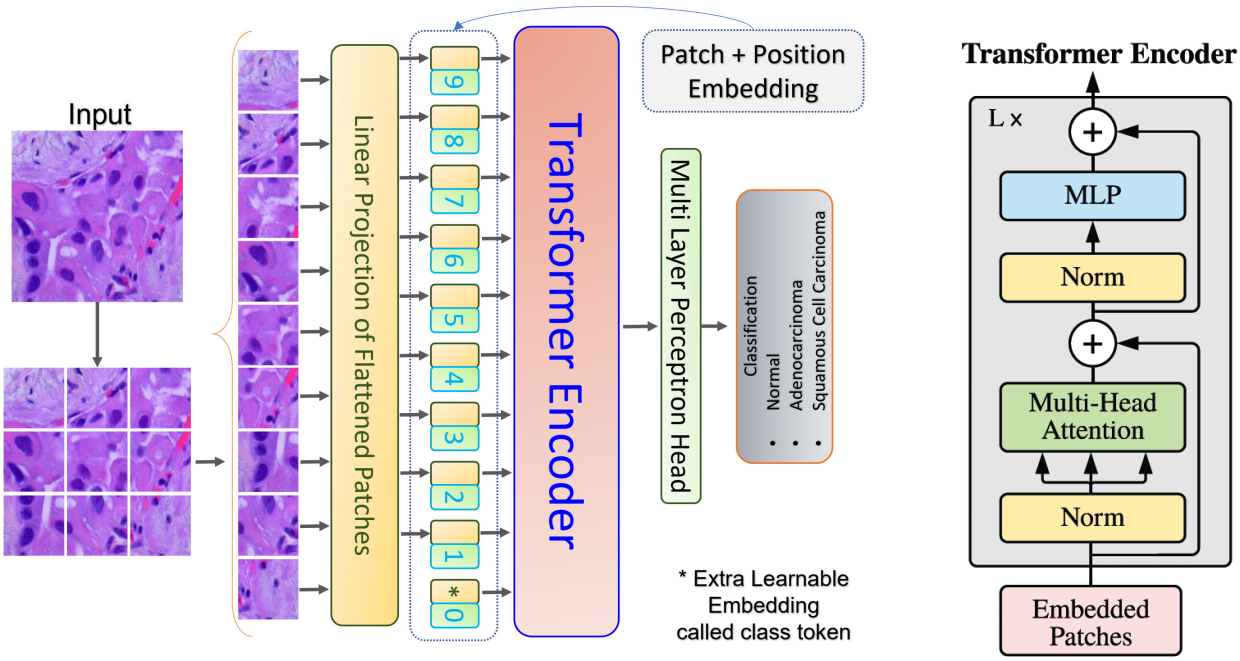
A. RELATED WORK

Medical images, such as X-ray [10], [11], computed tomography (CT) [12], [13], [14], [15], [16], and magnetic resonance imaging (MRI) [17], [18], are widely used for lung cancer diagnosis and classification. However, these methods cannot differentiate between cancerous and non-cancerous cells. The most reliable way to diagnose malignancies is to perform a biopsy and analyze the cells under a microscope [19]. This histopathological analysis, however, has some limitations [20]. For example, a biopsy may not capture the whole spectrum of the disease due to tumor heterogeneity. Also, only a few slides are examined from each tissue sample, which may not reflect the tumor context. Furthermore, it takes about ten days for the pathologists to report the results, which can delay the treatment [19]. Artificial intelligence (AI) techniques can help pathologists determine the type and stage of lung cancer. Among various AI techniques, deep learning has emerged as the most effective one for image processing tasks, including NSCLC classification. In this paper, we focus on deep learning methods for NSCLC classification on histopathological images. Javier et al. [19] proposed a deep neural network for detecting and classifying NSCLC into three sub-types using histopathological images. Their network consisted of two parts: one for grayscale images and one for color images. Both parts used the same architecture of convolutional and feedforward layers, but they applied different preprocessing techniques. Grayscale images were enhanced by histogram equalization, while color images were transformed to YUV format, which separates brightness from color components. The network achieved satisfactory results, but it required preprocessing of the input images. Using the existing Inception v3 model [24], researchers [20] fine-tuned it for three classification problems: tumor/normal, adenocarcinoma/squamous cell carcinoma, and normal/adenocarcinoma/squamous cell carcinoma. Despite the complexity and inefficiency of the model, they obtained promising results. Researchers [25] used the same approach as [20] and trained the existing EfficientNet model [26] by Google to classify carcinoma and non-neoplastic tissue images. They tiled 3,554 high-dimensional WSI images into patches of size 512×512 to create a larger dataset of small images. Wang et al [27] used deep learning and random forest to classify lung cancers, such as small cell, non-small cell, adenocarcinoma, and normal. Their architecture has three parts: discriminative patch prediction, context-aware block, and random forest. The first part predicts cancer risk in WSIs, the second part considers spatial context, and the third part aggregates features for WSI-level prediction. The method proposed in [52] uses CNNs and the enhanced Light Gradient Boosting Model (LightGBM) classifier to classify lung cancer histopathology images into normal lung tissue, lung adenocarcinoma, and lung squamous cell carcinoma. The method applies CNNs for feature extraction and LightGBM for classification on the LC25000 dataset. Another method [53] uses InceptionV3 architecture to classify non-small cell lung cancer (NSCLC) histopathology images into squamous and non-squamous subtypes. The method trains and optimizes the model on image patches of the biopsies. Likewise, the technique [54] uses a two-step approach to detect lung cancer from histopathology images. The first step is image preprocessing, which involves contrast enhancement, histogram equalization, median filtering, adaptive thresholding, and image segmentation. The second step is image classification, which uses a simple CNN model with three convolutional layers, two max-pooling layers, and one fully connected layer. The method trains the model on a large dataset of histopathology images from different sources. Authors in [46] also uses a deep learning model based on an encoder-decoder architecture for segmentation to detect and segment lung cancer on chest radiographs. The method utilizes both normal and inverted chest radiographs to improve robustness and performance. The method achieves high sensitivity and low false positive rate in detecting lung cancers on chest radiographs. However, the method has some limitations, such as not being compared with other state-of-the-art methods. In [14], authors propose an intelligent lung cancer detection system that employs multilevel

TABLE 1. Comprehensive picture of the literature review.

Model	Modality	Performance	Benefits	Limitations
Javier et al. [19]	Histopathological images	Satisfactory results	Detects and classifies NSCLC into three sub-types	Requires preprocessing of input images
Inception v3 Model [24]	Histopathological images	Promising results	Fine-tuned for three classification problems	Complexity and inefficiency
EfficientNet Model [26]	Histopathological images	Promising results	Classifies carcinoma and non-neoplastic tissue images	Requires tiling of high-dimensional WSI images
Wang et al. [27]	Histopathological images	Effective	Considers spatial context	Complexity of architecture
CNN + LightGBM Classifier [52]	Histopathological images	Effective	Combines CNN for feature extraction and LightGBM for classification	Requires large dataset
InceptionV3 Architecture [53]	Histopathological images	Effective	Classifies NSCLC into squamous and non-squamous subtypes	Requires image patches of biopsies
Two-Step Simple CNN Approach [54]	Histopathological images	Effective	Uses simple CNN model for classification	Requires large dataset
Encoder-Decoder Architecture [46]	Chest radiographs	High sensitivity and low false positive rate	Detects and segments lung cancer	Not compared with other state-of-the-art methods
Intelligent Lung Cancer Detection System [14]	CT scan images	Effective	Employs multilevel brightness-preserving techniques, enhanced deep neural network, hybrid spiral optimization, and ensemble classifier	Complexity of system
WS-LungNet [47]	Medical images	Improved performance	Utilizes semi-supervised learning and cross-attention mechanism	Addresses label scarcity and inconsistency
Improved Gabor Filter + E-DBN [48]	CT scan images	Efficient	Enhances feature extraction and classification accuracy	Requires dataset of CT scan images
CovidViT [49]	Chest X-rays	High accuracy	Rapid and economical detection tool	Specific to Covid-19 detection
Deep Dual Attention Network (D2 ANet) [50]	Chest CT scans	High accuracy	Analyzes chest CT scans for Covid-19 detection	Specific to Covid-19 detection
CNN model [57]	Chest CT scans	>90%	Analyzes different types of Lung Cancer	Complexity
SVD model on HOF features [58]	PET/CT scans	96%	Validated with large dataset	Computationally intensive
Deep learning approach [60]	CT scans	72.42%	Accurate and automated detection	Low accuracy
weighted graph convolutional network [61]	CT scans	77.27%	Various data types	Sample size
PCA-SMOTE-CNN model [63]	multi omics data	77.27%	Integrating multi-omics data with deep learning	Deep learning models may overfit
Multi-modal Heterogeneous Graph Forest [64]	CT scan	promising	heterogeneous data sources	practical challenges

brightness-preserving techniques for image enhancement, an enhanced deep neural network for region segmentation, a hybrid spiral optimization intelligent-generalized rough set approach for feature selection, and an ensemble classifier for cancer classification. The last method [47] proposes a weakly-supervised lung cancer detection and diagnosis network (WS-LungNet) that utilizes semi-supervised learning for nodule segmentation and a cross-attention mechanism for patient-level malignancy evaluation. WS-LungNet leverages unlabeled data and explores correlations among detected nodules, achieving improved performance in both nodule detection and patient-level malignancy evaluation. WS-LungNet addresses the limitations of label scarcity and inconsistency in nodule annotations. Another method [48] introduces novel methods, such as an improved Gabor filter and an Enhanced Deep Belief Network (E-DBN), to enhance feature extraction and improve lung cancer classification accuracy. These innovative approaches reduce processing time and offer more relevant feature selection, leading to more efficient lung cancer detection and classification. The technique applies the improved Gabor filter and the E-DBN to a dataset of CT scan images of lung nodules. These methods are examples of the recent advances in deep learning for the diagnosis of lung cancer using various modalities, such as histopathology images, chest radiographs, and CT scans. Researchers in [49] developed a new method, CovidViT, to detect Covid-19 from chest X-rays using a deep learning approach and achieved high accuracy, potentially offering a rapid and economical detection tool. Building on the prior research using chest X-rays, the study [50] proposes a Deep Dual Attention Network (D2 ANet) that analyzes chest CT scans for Covid-19 detection. D2 ANet incorporates attention mechanisms to identify relevant regions and subtle variations in the scans, potentially leading to even more accurate diagnosis of Covid-19. There are many good review papers, which have summarized the latest development of deep learning for the diagnosis of lung cancer using various modalities [51], [55], [56]. Vij and Kaswan [57] used a dataset of 1000 chest scan images for various types of lung cancer, including Adenocarcinoma, Large Cell Carcinoma, and Squamous Cell Carcinoma. Multiple machine learning algorithms were compared, and it was confirmed that CNN is among the best for predicting accuracy. Reference [58] processed CT images and extracted HOG features from them. These features are then fed into an SVM classifier to identify the cancer type.The accuracy for the test set is found to be 96%. Xu [59] reimplemented CNN model based on ResNet to improve low learning efficiency. Additionally, three different CNNs are compared, demonstrating that deeper neural networks have better learning efficiency and higher accuracy in classifying lung CT images. References [60] and [61] applied deep learning algorithms to detect lung cancer. These algorithms have up to 77% accuracy. Reference [62] develops a multimodal late fusion approach that combines hand-crafted features from radiomics, pathomics, and clinical data to predict radiotherapy outcomes for non-small-cell lung cancer patients. The proposed method, tested on a cohort of 33 patients, achieved an AUC of 90.9%, outperforming unimodal approaches and demonstrating the potential of data integration to enhance precision medicine. Reference [63] developed an innovative deep-learning model for lung cancer detection by integrating mRNA, miRNA, and DNA methylation markers. Reference [64] proposed a Multi-modal Heterogeneous Graph Forest (MHGF) approach to extract deep representations of LNM from multi-modal data. The results indicate that the graph method effectively explores relationships between different feature types for LNM prediction. Shi et al. [65] proposed a high-dimensional kernel non-negative matrix factorization (NMF) method that integrates multi-modal information. Experimental results demonstrate that the proposed NMF method outperforms traditional NMF in terms of stability, decomposition accuracy, and robustness. These review papers provide a comprehensive overview of the state-of-the-art methods, the challenges, and the future directions in this field. The Table 1 provides a comprehensive summary of the key features, performance metrics, benefits, and limitations of previously used models for lung cancer detection and classification.



(a) Vision Transformer (b) Transformer Encoder

FIGURE 1. The general architecture of the vision transformer and multihead unit (Figure Inspiration: [33]).

In medical images, the symptoms of any disease are not only subtle but distributed across an image; therefore, to detect and classify any disease using medical images, we need a network that can capture the subtle features and the long-range relationship simultaneously for better generalization and data efficiency. CNNs are translation invariant, which makes them a defacto choice for most image processing tasks, particularly involving detection or classification. The translational invariance of CNNs makes it easy to detect and subsequently classify objects irrespective of their position within an image. However, CNNs are not good at deriving the long-range relations among pixels or objects in an image due to their locally restricted receptive field. On the other hand, ViT is permutation invariant, which helps obtain the long-range contextual relationship between pixels or objects in an image. However, ViT cannot directly process grid-structured data; they need data sequences. Therefore, to utilize the advantages of both CNNs and ViT, we propose Transformer Based Hierarchical Model for Non-Small Cell Lung Cancer Detection & Classification. The experiments and results demonstrate our proposed architecture's effectiveness and its superiority over other models designed for the same objective. Our major contributions are outlined below.

- Pre-processing (such as histogram equalization, converting to other color formats, or grayscale) or post-processing (such as random forest or other machine learning algorithms for classification) are used in the most currently available approaches. In contrast, we created an end-to-end deep learning architecture that does not need any pre- or post-processing.
- Our model is straightforward and efficient since we conducted an ablation study to preserve only the layers, components, and hyperparameter settings essential to achieving the intended results.
- Our novel model leverages the capabilities of both CNNs (inductive bias and constrained receptive field) and vision transformers (long-range relationship).
- Our proposed method can accurately diagnoses the presence of lung cancer within a few milliseconds.

The remainder of the article is structured as follows: In Section II, we discuss multihead units, layer normalization, vision transformers, attention processes, and the proposed architecture. The proposed algorithm's experiments, findings, and comparison with other cutting-edge methods are covered in Section II-A4. Section III is our conclusion.

II. TRANSFORMER BASED HIERARCHICAL MODEL FOR NON SMALL CELL LUNG CANCER DETECTION AND CLASSIFICATION

CNNs have long been the default choice for computer vision tasks, especially to capture the inherent correlation among pixels in a local neighborhood for detection and classification tasks. The properties of CNNs – in terms of spatial localization and translation invariance, which help map an image's features for later tasks, such as recognition and feature extraction – are the reasons behind their vast success across various applications. Therefore, for most existing deep learning architectures [28], [29], [30], [31], [32], CNNs serve as the backbone.

However, in [33], authors demonstrated that, when trained on large datasets (like ImageNet), vision transformers perform better than well-known CNN architectures, such as ResNet [28]. A vision transformer (ViT) is an extension of the original concept of the transformer [34], initially proposed for natural language processing applications. In the paper, we combined both CNNs and ViTs to utilize their properties effectively. Below, we briefly discuss the working principle of ViT; for a detailed explanation, readers can refer to the original paper [34].

A. VISION TRANSFORMER (ViT)

A vision transformer (ViT), as shown in Fig. 1a, contains a patch extractor layer, an encoder layer, a transformer encoder, a multilayer perceptron (MLP) head, and a classification head. An input image, $I \in \mathbb{R}^{H,W,C}$, is split into N patches, each with a size of $P \times P$, by the patch extractor layer before being flattened to create a sequence, $\mathbf{x}_p \in \mathbb{R}^{N \times P^2 C}$, where $N = \frac{HW}{P^2}$. H, W, and C stand for the height, width, and number of channels of the input image, respectively. The flattened sequence is then passed through a fully connected dense layer to produce a projected sequence (called patch embedding) of size $N \times \mathcal{D}$, where $\mathcal{D}$ is the latent size of the fully connected layer. The patch encoder layer, to retain the positional information of each patch in the original input image, adds positional embeddings to the patch-embedding sequence and an extra learnable parameter called the class token to produce encoded patches, $\hat{\mathbf{x}}_p \in \mathbb{R}^{(N+1) \times P^2 C}$. The encoded patches pass through the transformer encoder. As shown in Fig. 1b, the transformer encoder consists of multiple copies of a multihead self-attention (MHA) unit, a multilayer perceptron (MLP) unit, normalization layers, and residual connections. The transformer encoder processes the input it receives from the patch-embedding layer and produces an output for the MLP head. The MLP head comprises a few fully-connected-dense and dropout layers, where each dense layers uses a Gaussian Error Linear Unit (GELU) [35] as the activation function. We chose GELU for two reasons: First, researchers employed it in the original ViT study [33]. Second, because GELU combines the qualities of dropout, zoneout, and rectified linear unit (ReLU), offering better generalization than ReLU or other well-known activation functions [35]. In the subsequent subsections, we will briefly discuss the working principles of an MHA in ViT.

1) MULTI HEAD SELF ATTENTION (MHA)

A vision transformer has numerous layers, including add and normalization, skip connections, a multihead attention unit (MHA), and a multilayer perceptron (MLP) block, as illustrated in Fig. 1b. To understand how an MHA functions, let us first comprehend how deep learning's self-attention mechanism functions. Considering an input $\mathbf{X} \in \mathbb{R}^{batch \times tokens \times \mathcal{D}}$ and its corresponding trainable weights $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{\mathcal{D} \times \mathcal{D}_k}$. The terms '*batch*', '*tokens*', and '$\mathcal{D}$' represent the 'number of batches,' 'the number of patches (also called tokens) in each image,' and the 'embedding dimension,' respectively. The input $\mathbf{X}$ is mapped to three distinct representations, referred to as the query (Q), the key (K), and the value (V), by passing them through different fully connected layers, each with a dimension or number of neurons equal to $\mathcal{D}_k$, as follows:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^V,$$

Afterward, we define a self-attention layer using a scaled dot-product, as shown in Fig 2a, to compute the attention weights as follows:

$$\begin{aligned}\mathbf{Y} &= Attention\,(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\ &= softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{\mathcal{D}_k}}\right)\mathbf{V}\end{aligned}$$

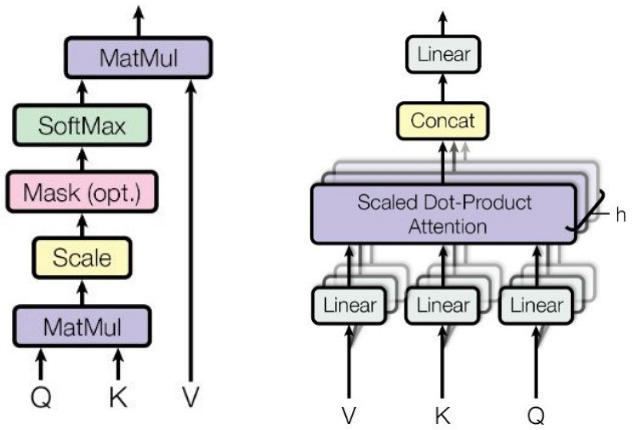
The higher attention weights indicate more similarity between patches and vice versa. For MHA, we independently and concurrently carry out the same actions we did for single-self-attention (described above). For instance, say the multihead unit has 'h' heads, as shown in Fig. 2b. We would compute each head's attention weights (scores) and concatenate them before moving through a subsequent dense layer, as illustrated in Fig. 3. The working of MHA can be explained mathematically as follows:

$$MultiHead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concatenate(\mathbf{Y}_1, \ldots, \mathbf{Y}_h) \quad (1)$$

where,

$$\mathbf{Y}_i = Attention\left(\mathbf{X}\mathbf{W}_i^Q, \mathbf{X}\mathbf{W}_i^K, \mathbf{X}\mathbf{W}_i^V,\right) \quad i = 1, \ldots, h \quad (2)$$

It is pertinent to mention that $\mathcal{D}$ represents the embedding dimension of the embedding layer. In contrast, $\mathcal{D}_k$ denotes the number of neurons used in the fully-connected layers used in multihead units.



(a) Scaled Dot-Product (b) Multi Head Attention

FIGURE 2. Multi head attention (Figure Credit: [34]).

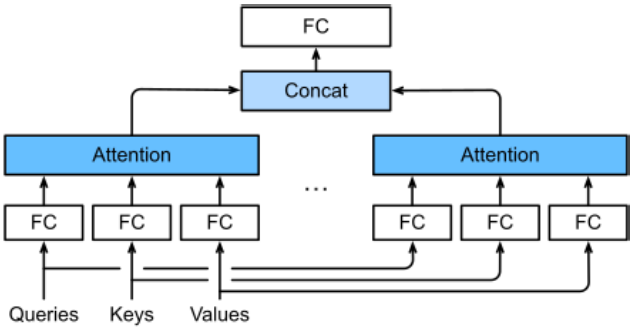


FIGURE 3. Multi-head attention unit (Source: [36]).

2) LAYER NORMALIZATION

The internal covariate shift that occurs in the deep neural networks during training causes a decreased training rate and worse model generalization. The concept of batch normalization was proposed [37] to overcome these problems, enhance computational efficiency, and obtain greater generalization accuracy. By using the mean and variance across all batches, batch normalization equalizes the distribution of the intermediate training data. Consequently, simple feedforward neural networks not only converge more quickly but also act as a regularizer, improving the model's ability to generalize. The batch normalization technique requires running means of the training data; therefore, it is unhelpful in networks where the sequence length frequently varies, like recurrent neural networks (RNN). Layer normalization [38], on the other hand, eliminates the need for batch statistics, and standardizes the training data across all features, which makes it suitable for RNN and attention-based models like transformers [39]. Therefore, the default choice for the vision transformer is also layer normalization. The algorithm [40] for layer normalization computation is illustrated in Algorithm 1.In the following subsection, we discuss our proposed deep learning architecture.

Algorithm 1 Layer Normalization

/* Normalizes layer activations $\mathbf{k}$. */
Input: $\mathbf{k} \in \mathbb{R}^{\mathcal{D}_k}$, neural network activations.
Output: $\hat{\mathbf{k}} \in \mathbb{R}^{\mathcal{D}_k}$, and normalizaed activations.
Parameters : $\gamma, \gamma \in \mathbb{R}^{\mathcal{D}_k}$, element-wise scale and offset.
$\mu \leftarrow \frac{1}{\mathcal{D}_k} \sum_{i=1}^{\mathcal{D}_k} \mathbf{k}_i$
$\sigma \leftarrow \frac{1}{\mathcal{D}_k} \sum_{i=1}^{\mathcal{D}_k} (\mathbf{k}_i - \mu)^2$
$\hat{\mathbf{k}} \leftarrow \frac{\mathbf{k}-\mu}{\sqrt{\sigma}} \odot \gamma + \beta$; /* $\odot$ represents element-wise multiplication. */
Return : $\hat{\mathbf{k}} \in \mathbb{R}^{\mathcal{D}_k}$, normalizaed activations.

3) PROPOSED HIERARCHICAL MODEL FOR NSCLC CLASSIFICATION

CNNs have inductive biases like translation invariance and a locally constrained receptive field that transformers do not have. The transformer, however, is by design permutation invariant. Therefore, to harness the key aspects of both great deep learning models, we attempted to integrate CNNs and transformers in this research. The result is the Transformer-Based Hierarchical Model, a novel hybrid NSCLC classification model, as shown in Fig. 4. The model comprises multiple layers, and each layer has two distinct blocks: feature selection and bottleneck blocks. We will discuss feature selection blocks in the forthcoming subsections; however, here, we will discuss the flow of the primary model, shown in Fig. 4. The suggested hierarchical model consists of five levels, which we have named $\mathcal{L}_0$ (topmost), $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$, and $\mathcal{L}_4$(bottommost), as shown in Fig. 4.

The input $I_0 \in \mathbb{R}^{B\times H\times W\times C}$ is fed to the patch extractor layer of $\mathcal{L}_0$, as shown in Fig. 4, and also passed to the feature selection block of $\mathcal{L}_1$, where B, H, W, and C represent, respectively, batch size, height, width, and number of channels of input I. The patch extractor layer of $\mathcal{L}_0$, with patch size P, produces a flattened sequence of size $\mathbf{x}_{p_0} \in \mathbb{R}^{B\times \frac{HW}{P^2}\times P^2C}$. For level $\mathcal{L}_1$, the input I flows through a bottleneck block to create an output, $I_1 \in \mathbb{R}^{B\times \frac{H}{2}\times \frac{W}{2}\times C}$, whose dimension is half that of the input. The output I_1 passes through the patch extraction layer of level $\mathcal{L}_1$ to produce a flattened output, $\mathbf{x}_{p_1} \in \mathbb{R}^{B\times \frac{HW}{P^2}\times \left(\frac{P}{2}\right)^2 C}$, with patch size $\frac{P}{2}$. The same procedure is followed on each level, and the output of the bottleneck block of each level is passed through the extraction layer of that level and also passed to the feature selection block of the next lower level. Every level uses the same process wherein the output of the bottleneck block is given to that level's extraction layer to produce a flattened sequence and passed to the next lower-level feature selection block. It is crucial to remember that the patch size for each subsequent lower-level extraction layer is half that of the layer above it, and each bottleneck block will reduce the receiving inputs' height and width by half. Each patch extractor layer's output is concatenated before passing through the patch-embedding layer. In our proposed model, we used twelve transformer encoders cascaded serially, with each transformer containing four MHA blocks. We tried various numbers for transformers and MHAs in the original vision transformer paper [33], and they observed better results with 12 transformers with 4 MHAs in each transformer. Therefore, we went with their figures and found highly satisfactory results.

4) FEATURE SELECTION AND BOTTLENECK BLOCKS

Detecting objects within an image may be an easy task for humans but not for machines due to the intricate relationships and similarities among pixels in an image. However, for medical images, where images look more alike, detecting diseases or other abnormalities becomes challenging for both humans and machines. Therefore, to extract the underlying features of medical images, we designed a block that we named feature selection block (FSB) for the detection and classification of NSCLC. As illustrated in Fig. 5, the FSB

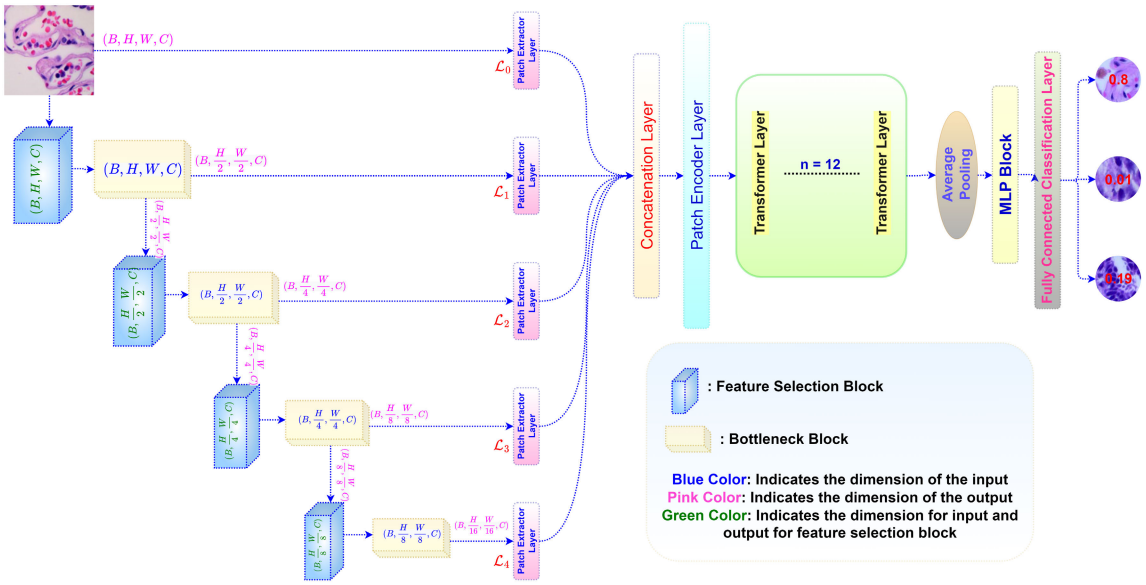


FIGURE 4. Transformer-based hierarchical model for small-scale cell lung cancer classification.

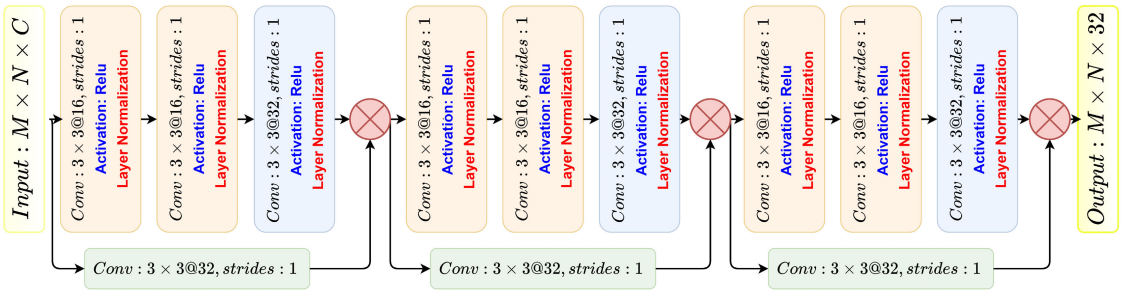


FIGURE 5. Feature selection block.

also consists of three sub-blocks, each of which includes a residual connection as well as convolutional, activation, and normalization layers. For activation, we chose rectified linear unit (RelU) due to its simplicity and computational efficiency, making the training stable. In addition, layer normalization was utilized to accomplish convergence and obtain generalizability while minimizing the covariate shift. Each sub-block in the FSB contains three convolution layers; for the first two layers, we used sixteen filters, while the last layer had 32 filters for each sub-block. Finally, the residual connection of all the sub-blocks had sixteen filters, and the size of each filter in all the convolutional layers was 3×3.

Using FSB, with 16, 16, and 32 filters in each sub-block, helps to select the prominent features of an image required to detect NSCLC correctly and subsequently classify it. As a result, complexity and training time both increase with the number of learnable parameters. Therefore, we created the bottleneck block to keep the learnable parameters to a manageable range and lessen the complexity. As shown in Fig. 6, the bottleneck block comprises three convolutional, activation, and normalization layers. Each of the first two convolutional layers includes 16 filters of sizes 3×3 with a stride of 1, while the last convolution layer has 32 filters of size 3×3 and a stride of 2. The output of the residual connection, which employs the same number of filters and sizes as the final convolution layer, is combined with it to produce an output half of the input size. Using these blocks with minimalistic numbers for filters keeps the model simple and computationally efficient and provides satisfactory outcomes, as our results corroborate this claim. We trained our proposed model using the well-known LC25000 dataset [41] on a shared GPU cluster controlled by a Slurm task scheduler. The cluster is equipped with multiple Tesla A100 and NVIDIA Tesla V100 GPUs. For writing the code, we used the open-source TensorFlow library. We used a batch size of 32 and trained our model for 50 epochs using a learning rate of 0.001. The Adam optimizer and categorical cross-entropy loss function were employed for training. In addition, we will make our code public for other researchers to verify our findings, extend our work, or apply it to other classification tasks. The code can be accessed via https://github.com/ImranNust/LungCancerDetection.

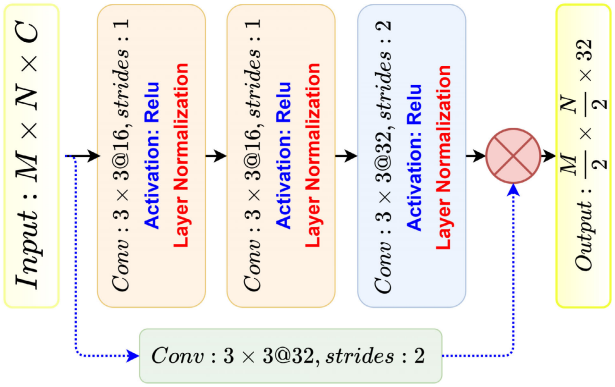


FIGURE 6. Bottleneck block.

In the following subsections, we discuss dataset, metrics, results, ablation analysis, and comparisons with other techniques.

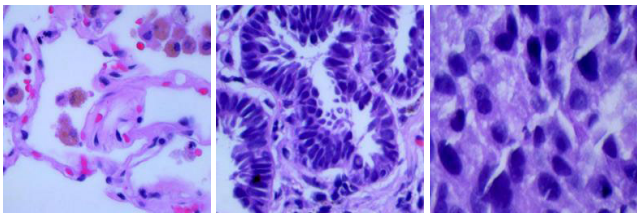


FIGURE 7. NSCLC cancer types: (a) normal (non-cancerous), (b) adenocarcinoma, and (c) squamous cell carcinoma.

B. DATASET

The LC25000 dataset consists of 15,000 histopathological images of lung tissue samples that are classified into three different types of NSCLC: normal, adenocarcinoma, and squamous cell carcinoma. Each type has 5,000 images, as shown in Fig. 7. Histopathological images are microscopic views of biopsied tissue that have been stained with special dyes to enhance the contrast between healthy and cancerous cells.

We split the dataset into three subsets for training, testing, and validation purposes. We used 80% of the images (4,000 per type) for training our proposed algorithm, 10% of the images (500 per type) for testing its performance, and 10% of the images (500 per type) for validating its generalization. The distribution of the dataset is summarized in Table 2.

C. EVALUATION OF PROPOSED ALGORITHM

We will first establish metrics typically employed to measure the effectiveness of any deep learning model created for classifications. We will then assess the effectiveness of the suggested architecture against all of these criteria. Later, we will compare our suggested model with some cutting-edge methods created recently for the same objective.

TABLE 2. Data Distribution for Training, Validation, and Testing.

Classes	Data Distribution			Total
	Train	Validation	Test	
Normal	4000	500	500	5000
Adenocarcinoma	4000	500	500	5000
Squamous Cell Carcinoma	4000	500	500	5000
Total	12000	1500	1500	15000

The proposed architecture will next be interpreted using results from ablation analysis.

1) ACCURACY

The first metric, used to measure the performance of any classification task, is accuracy. The accuracy $\mathcal{A}$, for a multiclass classification problem with K number of classes, is computed as follows:

$$\mathcal{A} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{TP_k + TN_k}{TP_k + TN_k + FP_k + FN_k}\right) \quad k = 1, \ldots K \tag{3}$$

In Eq. 3, TP_k, TN_k, FP_k, and FN_k denotes true positive, true negative, false positive, and false negative, respectively, for class k. There are two types of accuracies computed for multiclass classifications: micro accuracy and macro accuracy. While micro accuracy considers the contributions of all classes to calculate the micro accuracy, macro accuracy computes the metric independently for each class before taking the average and treating all classes equally. We have utilized macro accuracy in this study since it treats each class separately and independently.

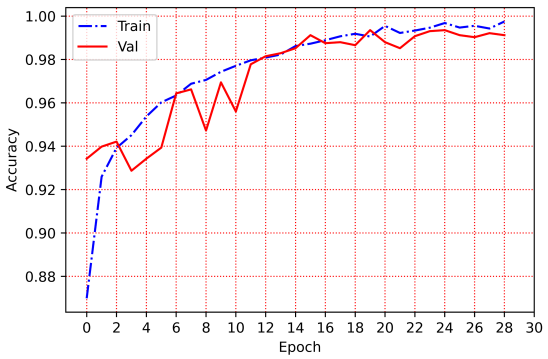


FIGURE 8. Train and validation accuracies for the proposed model $\mathcal{M}_1$.

The proposed approach achieved an accuracy of roughly 0.99 for both the training and validation datasets, as seen in the figure above (Fig. 8). It also generalizes well and prevents overfitting.

2) SPECIFICITY

A deep learning model's capacity to identify true negative classes as true negatives is quantified using a different parameter called specificity. Specificity, $0 \leq \mathcal{S} \leq 1$, can be computed using the following equation.

$$\mathcal{S} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{TN_k}{TN_k + FN_k}\right) \quad k = 1, \ldots K \tag{4}$$

In order to accurately identify false negatives, a model's specificity value should be greater; that is, close to 1. One

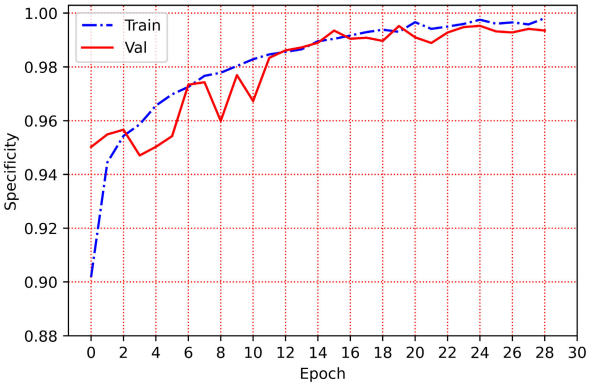


FIGURE 9. Train and validation specificities for the proposed model $\mathcal{M}_1$.

of the most important criteria for any classification model for medical data is specificity since labeling negative as positive can have serious repercussions. As seen in Fig. 9, our proposed model can classify training and validation datasets with a specificity exceeding 0.99.

3) PRECISION

In order to measure how many out of the total positive predicted classes are genuinely positive, we use a metric called precision. Precision, being the ratio of true positive and the sum of true-positive and false-positive values, takes a value between 0 and 1. We can compute the precision for a multiclass classification problem using the following equation.

$$\mathcal{P} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{TP_k}{TP_k + FP_k}\right) \quad k = 1, \ldots K \tag{5}$$

As with other measures, precision requires a higher value (closer to 1), and it is evident from Fig. 10 that our model was successfully precise.

4) RECALL

Sometimes, we are curious to learn how well a model can classify true-positive classes specifically, and its capacity to identify genuine positives as true positive. The following equation is used to compute recall, also known as sensitivity.

$$\mathcal{R} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{TP_k}{TP_k + FN_k}\right) \quad k = 1, \ldots K \tag{6}$$

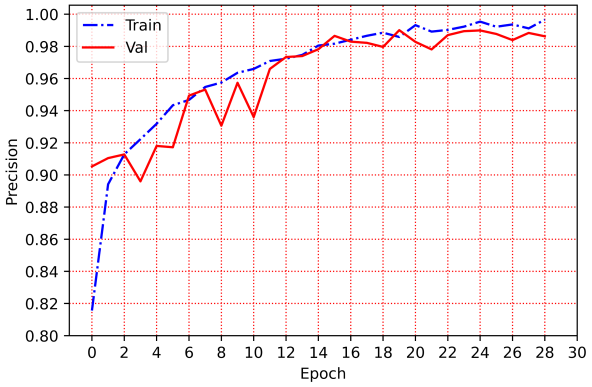


FIGURE 10. Train and validation precisions for the proposed model $\mathcal{M}_1$.

Recall, like specificity, is another crucial metric as it ensures that a model correctly classifies positive classes. It is clear from Fig. 11 that our model attained recall values above 0.98 for both the training and validation datasets.

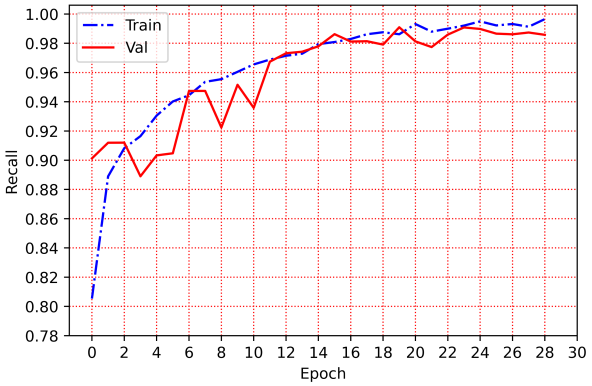


FIGURE 11. Train and validation recalls for the proposed model $\mathcal{M}_1$.

5) F1 SCORE

The F1-score is another metric that computes the harmonic mean of precision and recall and is used to assess how well a deep learning model performs when doing classification tasks. It can be computed as follows:

$$\mathcal{F} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{2 \times \mathcal{P}_k \times \mathcal{R}_k}{\mathcal{P}_k + \mathcal{R}_k}\right) \quad k = 1, \ldots K \tag{7}$$

$\mathcal{P}_k$ and $\mathcal{R}_k$ in the equation above represents precision and recall for class k, respectively. Finally, the graph shown in Fig. 12 corroborates that our proposed model can also achieve a reasonably good F1 score.

6) AUC-ROC CURVE

The AUC-ROC curve, where AUC stands for Area Under the Curve and ROC is an acronym for Receiver Operating Characteristics, is another metric we used to assess the effectiveness of our model. It serves as a summary of the ROC curve and measures the model's capacity to distinguish

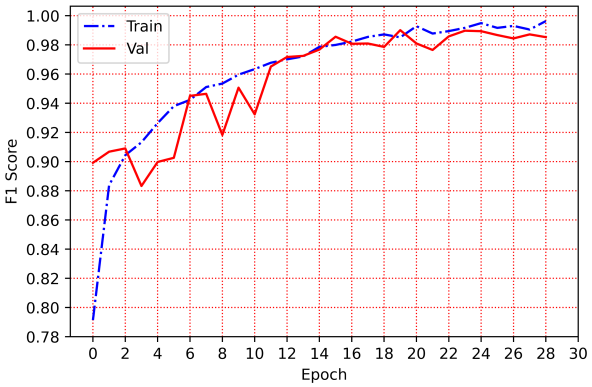


FIGURE 12. Train and validation F1 scores for the proposed model $\mathcal{M}_1$.

between various classes. The AUC-ROC curve for our model is shown in Fig. 13, and it can be clearly seen that our model's ability to distinguish different classes is extremely good. Although our model can categorize NSCLC with remarkable

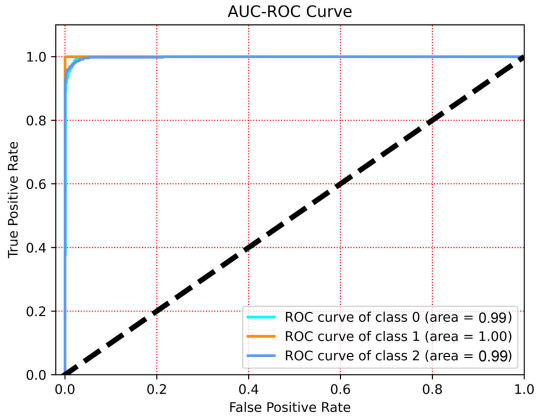


FIGURE 13. AUC-ROC Curves for three classes of lung cancer The graph shows that the proposed model can successfully distinguish each class from others.

metrics, its greatest accomplishment is being able to tell one class apart from another, as shown in Fig. 13.

7) PRECISION-RECALL CURVE

The Precision-Recall curve is the final metric we used to assess the effectiveness of our model. It serves as a summary of the precision and recall for different threshold values and measures the model's capacity to distinguish between various classes. The Precision-Recall curve for our model is shown in Fig. 14, and it can be clearly seen that our model's ability to distinguish different classes is extremely good.

D. ABLATION ANALYSIS

An ablation study, also known as an ablation analysis, is used to decipher deep learning or machine learning models [42] to reduce the computational complexity by avoiding unnecessary learnable parameters. Therefore, in this

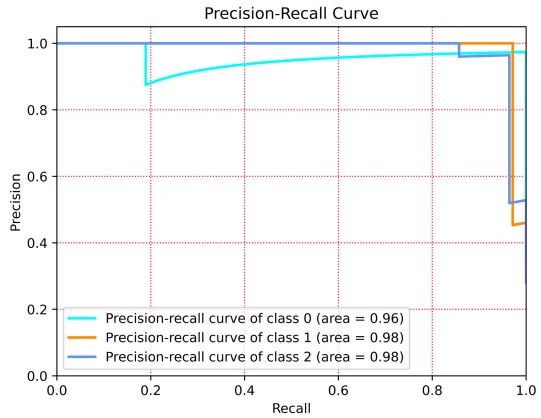


FIGURE 14. Precision-Recall Curves for three classes of lung cancer The graph shows that the proposed model can successfully distinguish each class from others.

research, we conducted an ablation study to interpret the proposed model. To achieve that, we created various models by omitting complex components from the suggested models and substituting them with simpler ones. We also evaluated our model's performance for various hyperparameters. The learning rate is one of the significant hyperparameters that play a crucial role in how fast a model converges and how stable the training is. In the proposed research, we utilized an improved version of Adam [44], proposed by Loshchilov and Hutter [43], which involves two terms: learning rate and weight decay. Since our proposed model contains convolutional networks and transformers, they train and converge at their own pace. Therefore, we had the intuition that our model would converge quickly and stably if we kept the learning rate and weight decay low. However, we began with somewhat higher values and kept lowering them until we achieved the desired outcomes. Table 3 displays the results of our model for various learning rates and weight decays. It is worth noting that we used the very same model shown in Fig. 4 for all of these values. The performance of

TABLE 3. Performance for Different Learning Rates.

Metrics	Different Learning Rate and Weight Decay Values		
	$\alpha = 1 \times 10^{-3}$ $\omega = 1 \times 10^{-5}$	$\alpha = 1 \times 10^{-4}$ $\omega = 1 \times 10^{-3}$	$\alpha = 1 \times 10^{-5}$ $\omega = 1 \times 10^{-6}$
Accuracy	0.5556	0.8751	**0.9875**
Specificity	0.6667	0.9069	**0.9906**
Precision	0.1111	0.8326	**0.9802**
Loss	1.0986	0.4028	**0.040**
Recall	0.3333	0.8143	**0.9815**
F1 Score	0.1655	0.7986	**0.9796**

our final proposed model, represented in Fig. 1, was further examined when various components were eliminated. For instance, we tested how well our model performed with some levels removed and how well it works if it contained only

transformers. The different architectures are summarized as follows:

1) **Original Model** ($\mathcal{M}$)**:** This is the final and the same hybrid model discussed in Section II-A3 and is shown in Fig. 4.
2) **Model Ver1** ($\mathcal{M}_1$) **:** Version 1, denoted as $\mathcal{M}_1$, contains only the first four levels ($\mathcal{L}_0$ to $\mathcal{L}_3$) and the transformer block, but not level ($\mathcal{L}_4$).
3) **Model Ver2** ($\mathcal{M}_2$) **:** Version 2, abbreviated as $\mathcal{M}_2$, only included the first three levels ($\mathcal{L}_0$ to $\mathcal{L}_2$) and the transformer block; $\mathcal{L}_3$ and $\mathcal{L}_4$ were not included.
4) **Model Ver3** ($\mathcal{M}_3$) **:** Only the first two levels ($\mathcal{L}_0$ and $\mathcal{L}_1$) and the transformer block were included in Version 3, represented as $\mathcal{M}_3$. The last three levels ($\mathcal{L}_2$ to $\mathcal{L}_4$) were not part of this model.
5) **Model Ver4** ($\mathcal{M}_1$) **:** Version 4, abbreviated as $\mathcal{M}_4$, is the final model, and it does not contain any convolutional layer; it is a purely transformer-based model.

Figure 15 displays how well each of the versions, as mentioned above, performed in relation to various objective measures. Although all models could perform reasonably well, the final model (M) surpassed all other models. The final model converges quickly, generalizes well, and the training is more stable - compared to other versions. Here, we want to make some observations we noticed during different training models. For smaller datasets, the final model (M) is preferred as it is good at extracting the crucial subtle underlying features of images, thereby determining the classification with better accuracy. However, for larger datasets, such as ImageNet [45], the performance will not suffer significantly if the last level (L4) or the last two levels (L3 and L4) are dropped. In terms of computational time,

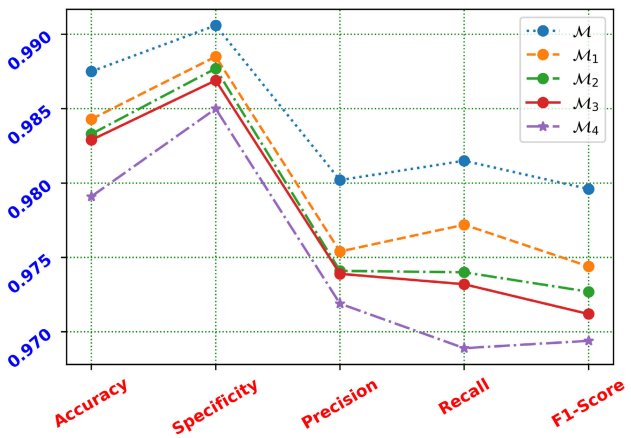


FIGURE 15. Ablation analysis of the proposed model.

the simpler model (with fewer layers or learning parameters) will train more quickly than a longer or more complex model. However, all of the above versions took almost six hours for training, and their prediction time varied in nanoseconds from one another. Therefore, computationally, the difference is not worth considering.

E. COMPARISON WITH STATE-OF-THE-ART TECHNIQUES

In the proposed study, we used deep learning to classify NSCLC into three classes using histopathological images: normal, adenocarcinoma, and squamous cell carcinoma. Therefore, for comparison, we have chosen only those techniques proposed for the same purpose using histological images based on deep learning or machine learning.

TABLE 4. Comparison with State-of-the-Art Techniques.

Metrics	Techniques							
	Ours	Javier et al. [19]	Coudray et al. [20]	Kanavati et al. [25]	Wang et al. [27]	Hamed et al. [52]	Page et al. [53]	Nannapaneni et al. [54]
Accuracy	**0.988**	0.974	0.932	0.962	0.961	0.945	0.956	0.949
Specificity	**0.991**	0.978	0.935	0.962	0.951	0.942	0.954	0.941
Precision	**0.980**	0.958	0.967	0.948	0.964	0.955	0.959	0.952
Recall	**0.982**	0.973	0.971	0.961	0.914	0.932	0.943	0.937
F1 Score	**0.980**	0.965	0.968	0.954	0.938	0.951	0.957	0.944

The comparison of the proposed algorithm with other contemporary techniques is shown in Table 4. It is visible that the proposed scheme outperforms other techniques in terms of all metrics. We could achieve these incredible results by effectively combining the all-time favorite CNNs and a recently introduced vision transformer.

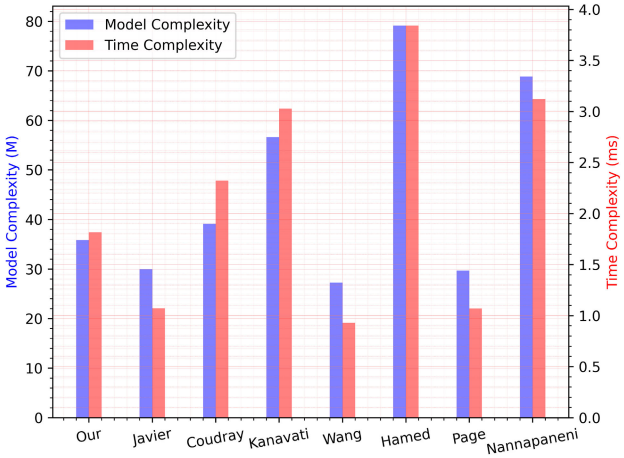


FIGURE 16. Comparison of the proposed technique with state-of-the-art techniques in terms of time and space complexities.

In addition to the accuracy, specificity, precision, recall, and F1 score metrics, we also compared our proposed technique with other state-of-the-art techniques in terms of time and space complexity. To compute the time complexity, we measured the inference time, which is the time it takes for a model to make a prediction on a new input image. For space complexity, we considered the number of learnable parameters in millions. Time and space complexity are important metrics to consider when deploying deep learning models in real-world applications. Figure 16 shows the

comparisons, and it is evident that the proposed technique achieves both high accuracy and fast inference time. This makes it a good choice for real-world applications where both accuracy and performance are important.

III. CONCLUSION

In this paper, we proposed a novel deep-learning architecture to classify NSCLC into three classes: normal, adenocarcinoma, and squamous cell carcinoma. Our model is a hierarchical model that combines CNNs and ViTs. We used CNNs to extract local features from histopathological images, and ViTs to capture and understand the long-range relationships between image patches. We trained and validated our model using the LC25000 dataset, a benchmark dataset of histopathology images for NSCLC classification. Our model achieved remarkable results, outperforming the current state-of-the-art methods in terms of various metrics: accuracy (0.988), F-1 score (0.980), specificity (0.991), recall (0.982), and precision (0.980). Our model also achieved comparatively the lowest inference time (1.816 ms), making it suitable for real-world applications where both accuracy and performance are important. We conducted an ablation study to understand the impact of different components and hyperparameters of our model. We showed that our model is relatively simple while being very effective, and that it can leverage the advantages of both CNNs and ViTs for NSCLC classification. We believe that our proposed model is a significant contribution to the field of medical image analysis, and that it can help pathologists in diagnosing lung cancer more efficiently and accurately.

ACKNOWLEDGMENT

The authors would like to thank Jessica Kirwan for kindly proofreading and editing this article.

REFERENCES

[1] C. Liu, X. Xiang, S. Han, H. Y. Lim, L. Li, X. Zhang, Z. Ma, L. Yang, S. Guo, R. Soo, B. Ren, L. Wang, and B. C. Goh, "Blood-based liquid biopsy: Insights into early detection and clinical management of lung cancer," *Cancer Lett.*, vol. 524, pp. 91–102, Jan. 2022, doi: 10.1016/j.canlet.2021.10.013.

[2] D. Mathios et al., "Detection and characterization of lung cancer using cell-free DNA fragmentomes," *Nature Commun.*, vol. 12, no. 1, pp. 1–20, Aug. 2021, doi: 10.1038/s41467-021-24994-w.

[3] (2022). *Cancer, World Health Organization*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[4] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *CA, A Cancer J. Clinicians*, vol. 72, no. 1, pp. 7–33, Jan. 2022, doi: 10.3322/caac.21708.

[5] C. Gridelli, A. Rossi, D. P. Carbone, J. Guarize, N. Karachaliou, T. Mok, F. Petrella, L. Spaggiari, and R. Rosell, "Non-small-cell lung cancer," *Nature Rev. Disease Primers*, vol. 1, no. 1, pp. 1–16, May 2015, doi: 10.1038/nrdp.2015.9.

[6] N. Howlader, G. Forjaz, M. J. Mooradian, R. Meza, C. Y. Kong, K. A. Cronin, A. B. Mariotto, D. R. Lowy, and E. J. Feuer, "The effect of advances in lung-cancer treatment on population mortality," *New England J. Med.*, vol. 383, no. 7, pp. 640–649, Aug. 2020, doi: 10.1056/nejmoa1916623.

[7] W. Lin, Y. Chen, B. Wu, Y. Chen, and Z. Li, "Identification of the pyroptosis-related prognostic gene signature and the associated regulation axis in lung adenocarcinoma," *Cell Death Discovery*, vol. 7, no. 1, pp. 1–10, Jun. 2021, doi: 10.1038/s41420-021-00557-2.

[8] B.-Y. Wang, J.-Y. Huang, H.-C. Chen, C.-H. Lin, S.-H. Lin, W.-H. Hung, and Y.-F. Cheng, "The comparison between adenocarcinoma and squamous cell carcinoma in lung cancer patients," *J. Cancer Res. Clin. Oncol.*, vol. 146, no. 1, pp. 43–52, Jan. 2020, doi: 10.1007/s00432-019-03079-8.

[9] G. Pelosi, M. Barbareschi, A. Cavazza, P. Graziano, G. Rossi, and M. Papotti, "Large cell carcinoma of the lung: A tumor in search of an author. A clinically oriented critical reappraisal," *Lung Cancer*, vol. 87, no. 3, pp. 226–231, Mar. 2015, doi: 10.1016/j.lungcan.2015.01.008.

[10] S. H. Bradley, N. L. F. Hatton, R. Aslam, B. Bhartia, M. E. J. Callister, M. P. T. Kennedy, and L. T. A. Mounce, "Estimating lung cancer risk from chest X-ray and symptoms: A prospective cohort study," *Roy. College Gen. Practitioners*, vol. 71, no. 705, pp. e280–e286, 2021.

[11] D. M. Ibrahim, N. M. Elshennawy, and A. M. Sarhan, "Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases," *Comput. Biol. Med.*, vol. 132, May 2021, Art. no. 104348, doi: 10.1016/j.compbiomed.2021.104348.

[12] P. M. Shakeel, M. A. Burhanuddin, and M. I. Desa, "Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks," *Measurement*, vol. 145, pp. 702–712, Oct. 2019, doi: 10.1016/j.measurement.2019.05.027.

[13] S. P. Primakov et al., "Automated detection and segmentation of non-small cell lung cancer computed tomography images," *Nature Commun.*, vol. 13, no. 1, p. 3423, Jun. 2022, doi: 10.1038/s41467-022-30841-3.

[14] P. M. Shakeel, M. A. Burhanuddin, and M. I. Desa, "Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier," *Neural Comput. Appl.*, vol. 34, no. 12, pp. 9579–9592, Jun. 2022.

[15] Y. Guo, Q. Song, M. Jiang, Y. Guo, P. Xu, Y. Zhang, C.-C. Fu, Q. Fang, M. Zeng, and X. Yao, "Histological subtypes classification of lung cancers on CT images using 3D deep learning and radiomics," *Academic Radiol.*, vol. 28, no. 9, pp. e258–e266, Sep. 2021, doi: 10.1016/j.acra.2020.06.010.

[16] M. Schwyzer, D. A. Ferraro, U. J. Muehlematter, A. Curioni-Fontecedro, M. W. Huellner, G. K. von Schulthess, P. A. Kaufmann, I. A. Burger, and M. Messerli, "Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks—Initial results," *Lung Cancer*, vol. 126, pp. 170–173, Dec. 2018, doi: 10.1016/j.lungcan.2018.11.001.

[17] A. J. Sim, E. Kaza, L. Singer, and S. A. Rosenberg, "A review of the role of MRI in diagnosis and treatment of early stage lung cancer," *Clin. Translational Radiat. Oncol.*, vol. 24, pp. 16–22, Sep. 2020, doi: 10.1016/j.ctro.2020.06.002.

[18] M. Kim, C. H. Suh, S. M. Lee, H. C. Kim, A. A. Aizer, T. K. Yanagihara, H. X. Bai, J. P. Guenette, R. Y. Huang, and H. S. Kim, "Diagnostic yield of staging brain MRI in patients with newly diagnosed non–small cell lung cancer," *Radiology*, vol. 297, no. 2, pp. 419–427, Nov. 2020, doi: 10.1148/radiol.2020201194.

[19] J. Civit-Masot, A. Bañuls-Beaterio, M. Domínguez-Morales, M. Rivas-Pérez, L. Muñoz-Saavedra, and J. M. Rodríguez Corral, "Non-small cell lung cancer diagnosis aid with histopathological images using explainable deep learning techniques," *Comput. Methods Programs Biomed.*, vol. 226, Nov. 2022, Art. no. 107108, doi: 10.1016/j.cmpb.2022.107108.

[20] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature Med.*, vol. 24, no. 10, pp. 1559–1567, Oct. 2018, doi: 10.1038/s41591-018-0177-5.

[21] J. A. ALzubi, B. Bharathikannan, S. Tanwar, R. Manikandan, A. Khanna, and C. Thaventhiran, "Boosted neural network ensemble classification for lung cancer disease diagnosis," *Appl. Soft Comput.*, vol. 80, pp. 579–591, Jul. 2019, doi: 10.1016/j.asoc.2019.04.031.

[22] N. Maleki, Y. Zeinali, and S. T. A. Niaki, "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," *Expert Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 113981, doi: 10.1016/j.eswa.2020.113981.

[23] X. Zhu, D. Dong, Z. Chen, M. Fang, L. Zhang, J. Song, D. Yu, Y. Zang, Z. Liu, J. Shi, and J. Tian, "Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer," *Eur. Radiol.*, vol. 28, no. 7, pp. 2772–2778, Jul. 2018, doi: 10.1007/s00330-017-5221-1.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[25] F. Kanavati, G. Toyokawa, S. Momosaki, M. Rambeau, Y. Kozuma, F. Shoji, K. Yamazaki, S. Takeo, O. Iizuka, and M. Tsuneki, "Weakly-supervised learning for lung carcinoma classification using deep learning," *Sci. Rep.*, vol. 10, no. 1, p. 9297, Jun. 2020, doi: 10.1038/s41598-020-66333-x.

[26] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.

[27] X. Wang, H. Chen, C. Gan, H. Lin, Q. Dou, E. Tsougenis, Q. Huang, M. Cai, and P.-A. Heng, "Weakly supervised deep learning for whole slide lung cancer image analysis," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3950–3962, Sep. 2020, doi: 10.1109/TCYB.2019.2935141.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[31] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.

[32] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 1–9.

[33] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, Matthias M., M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *ICLR*, 2021, pp. 1–22.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 1–15.

[35] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[36] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Attention mechanism and transformers," *Dive Deep Learn.*, vol. 1, pp. 409–468, Jul. 2022.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 1–24.

[38] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[39] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin, "Understanding and improving layer normalization," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 4383–4393.

[40] M. Phuong and M. Hutter, "Formal algorithms for transformers," 2022, *arXiv:2207.09238*.

[41] A. A. Borkowski, M. M. Bui, L. Brannon Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, "Lung and colon cancer histopathological image dataset (LC25000)," 2019, *arXiv:1912.12142*.

[42] L. Du, "How much deep learning does neural style transfer really need? An ablation study," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3139–3148, doi: 10.1109/WACV45572.2020.9093537.

[43] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[46] A. Shimazaki, D. Ueda, A. Choppin, A. Yamamoto, T. Honjo, Y. Shimahara, and Y. Miki, "Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method," *Sci. Rep.*, vol. 12, no. 1, p. 1727, Jan. 2022, doi: 10.1038/s41598-021-04667-w.

[47] Z. Shen, P. Cao, J. Yang, and O. R. Zaiane, "WS-LungNet: A two-stage weakly-supervised lung cancer detection and diagnosis network," *Comput. Biol. Med.*, vol. 154, Mar. 2023, Art. no. 106587, doi: 10.1016/j.compbiomed.2023.106587.

[48] E. A. Siddiqui, V. Chaurasia, and M. Shandilya, "Detection and classification of lung cancer computed tomography images using a novel improved deep belief network with Gabor filters," *Chemometric Intell. Lab. Syst.*, vol. 235, Apr. 2023, Art. no. 104763, doi: 10.1016/j.chemolab.2023.104763.

[49] H. Yang, L. Wang, Y. Xu, and X. Liu, "CovidViT: A novel neural network with self-attention mechanism to detect COVID-19 through X-ray images," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 3, pp. 973–987, Mar. 2023, doi: 10.1007/s13042-022-01676-7.

[50] Z. Lin, Z. He, R. Yao, X. Wang, T. Liu, Y. Deng, and S. Xie, "Deep dual attention network for precise diagnosis of COVID-19 from chest CT images," *IEEE Trans. Artif. Intell.*, vol. 5, no. 1, pp. 1–11, Jan. 2022, doi: 10.1109/TAI.2022.3225372.

[51] H. Ali, F. Mohsen, and Z. Shah, "Improving diagnosis and prognosis of lung cancer using vision transformers: A scoping review," *BMC Med. Imag.*, vol. 23, no. 1, p. 129, Sep. 2023, doi: 10.1186/s12880-023-01098-z.

[52] E. A.-R. Hamed, M. A.-M. Salem, N. L. Badr, and M. F. Tolba, "An efficient combination of convolutional neural network and LightGBM algorithm for lung cancer histopathology classification," *Diagnostics*, vol. 13, no. 15, p. 2469, Jul. 2023, doi: 10.3390/diagnostics13152469.

[53] A. Le Page, E. Ballot, C. Truntzer, V. Derangere, A. Ilie, D. Rageot, F. Bibeau, and F. Ghiringhelli, "Using a convolutional neural network for classification of squamous and non-squamous non-small cell lung cancer based on diagnostic histopathology HES images," *Sci. Rep.*, vol. 11, Jun. 2021, Art. no. 23912.

[54] D. Nannapaneni, V. R. S. V. Saikam, R. Siddu, V. M. Challapalli, and V. Rachapudi, "Enhanced image-based histopathology lung cancer detection," in *Proc. 7th Int. Conf. Comput. Methodologies Commun. (ICCMC)*, Erode, India, Feb. 2023, pp. 620–625, doi: 10.1109/ICCMC56507.2023.10084247.

[55] S. Dodia, A. B., and P. A. Mahesh, "Recent advancements in deep learning based lung cancer detection: A systematic review," *Eng. Appl. Artif. Intell.*, vol. 116, Nov. 2022, Art. no. 105490, doi: 10.1016/j.engappai.2022.105490.

[56] S. Huang, J. Yang, N. Shen, Q. Xu, and Q. Zhao, "Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective," *Seminars Cancer Biol.*, vol. 89, pp. 30–37, Feb. 2023, doi: 10.1016/j.semcancer.2023.01.006.

[57] A. Vij and K. S. Kaswan, "Prediction of lung cancer using convolution neural networks," in *Proc. Int. Conf. Artif. Intell. Smart Commun. (AISC)*, Greater Noida, India, Jan. 2023, pp. 737–741, doi: 10.1109/AISC56616.2023.10085058.

[58] A. K. Swain, A. Swetapadma, J. K. Rout, and B. K. Balabantaray, "A non-small cell lung cancer detection technique using PET/ CT images," in *Proc. 5th Int. Conf. Electr., Comput. Commun. Technol. (ICECCT)*, Feb. 2023, pp. 1–4, doi: 10.1109/icecct56650.2023.10179811.

[59] H. Xu, "Comparison of CNN models in non-small lung cancer diagnosis," in *Proc. IEEE 3rd Int. Conf. Power, Electron. Comput. Appl. (ICPECA)*, Shenyang, China, Jan. 2023, pp. 1169–1174, doi: 10.1109/ICPECA56706.2023.10075772.

[60] A. Bherje, A. Jidge, C. Roy, A. Hulke, M. A. Aswathy, V. Yadav, and K. V. Veenamol, "Design of deep learning-based approach to predict lung cancer on CT scan images," in *Proc. 5th Int. Conf. Innov. Trends Inf. Technol.*, Mar. 2024, pp. 1–5, doi: 10.1109/icitiit61487.2024.10580370.

[61] Q. Wu, J. Wang, Z. Sun, L. Xiao, W. Ying, and J. Shi, "Immunotherapy efficacy prediction for non-small cell lung cancer using multi-view adaptive weighted graph convolutional networks," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 11, pp. 5564–5575, Nov. 2023, doi: 10.1109/jbhi.2023.3309840.

[62] M. Tortora, E. Cordelli, R. Sicilia, L. Nibid, E. Ippolito, G. Perrone, S. Ramella, and P. Soda, "RadioPathomics: Multimodal learning in non-small cell lung cancer for adaptive radiotherapy," *IEEE Access*, vol. 11, pp. 47563–47578, 2023, doi: 10.1109/ACCESS.2023.3275126.

[63] T. I. A. Mohamed and A. E.-S. Ezugwu, "Enhancing lung cancer classification and prediction with deep learning and multi-omics data," *IEEE Access*, vol. 12, pp. 59880–59892, 2024, doi: 10.1109/access.2024.3394030.

[64] D. Hu, S. Li, N. Wu, and X. Lu, "A multi-modal heterogeneous graph forest to predict lymph node metastasis of non-small cell lung cancer," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 3, pp. 1216–1224, Mar. 2023, doi: 10.1109/JBHI.2022.3233387.

[65] Y. Shi, Z. Jin, J. Deng, W. Zeng, and L. Zhou, "A novel high-dimensional kernel joint non-negative matrix factorization with multimodal information for lung cancer study," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 2, pp. 976–987, Feb. 2024, doi: 10.1109/jbhi.2023.3335950.

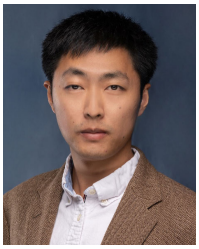
MUHAMMAD IMRAN received the bachelor's degree from Mehran University, Jamshoro, the M.S. degree from NUST, Islamabad, and the Ph.D. degree from Florida State University, USA. He is currently a Postdoctoral Associate of artificial intelligence with the College of Medicine, University of Florida. He is working on Tissue Image Analysis, particularly for the diagnosis of cancer. His research interests include computer vision and deep learning for medical image analysis.

BUSHRA HAQ received the M.S. degree from the Balochistan University of Information Technology, Engineering, and Management Sciences (BUITEMS), where she is currently pursuing the Ph.D. degree. Since 2014, she has been affiliated with BUITEMS as a Faculty Member with FICT. Her research interests include data science, machine learning, deep learning, remote sensing, and the IoT.

ERSIN ELBASI received the M.Sc. degree in computer science from Syracuse University, and the M.Phil. and Ph.D. degrees in computer science from the Graduate Center, The City University of New York.

He is currently with American University of the Middle East. His research interests include multimedia security, event mining in video sequences, and medical image processing.

AHMET E. TOPCU (Member, IEEE) received the B.Sc. degree from the Department of Electrical and Electronics Engineering, Middle East Technical University, Ankara, Turkey, in 1997; the M.Sc. degree in computer engineering from Syracuse University, Syracuse, NY, USA, in 2001; and the Ph.D. degree in computer science from Indiana University, Bloomington, IN, USA, in 2010. He is currently an Associate Professor with the College of Engineering and Technology, American University of the Middle East, Kuwait. His current research interests include distributed systems, machine learning, cloud computing, big data, the IoT, and blockchain.



WEI SHAO (Member, IEEE) is currently an Assistant Professor and the Assistant Director of AI imaging with the Intelligent Critical Care Center, University of Florida. He is the Principal Investigator of the UF Medical Imaging Research for Translational Healthcare with Artificial Intelligence (MIRTH AI) Laboratory, where he leads a team of researchers in developing and applying deep learning and medical image analysis methods for early disease diagnosis and clinical decision support.