Ingénierie des Systèmes d'Information
Vol. 30, No. 6, June, 2025, pp. 1425-1435
Journal homepage: http://iieta.org/journals/isi

Feature Dimensionality Reduction for Lung Tumor Classification Using Transformer Deep Learning and Yolo Model

Aliya Thaseen[1,2*], Sheshikala Martha[2], Durgesh Nandan[2]

[1] Department of CSE(AI&ML), Vidya Jyothi Institute of Technology, Hyderabad 500075, India
[2] School of Computer Science and Artificial Intelligence, SR University, Warangal 506371, India

Corresponding Author Email: aliya.tehseen86@gmail.com

https://doi.org/10.18280/isi.300602

Received: 19 January 2025
Revised: 9 May 2025
Accepted: 27 May 2025
Available online: 30 June 2025

Keywords:
lung cancer, feature dimensionality reduction, feature ranking, recursive feature elimination, transformer model, YOLOv5, classification

ABSTRACT

Healthcare ranks among the most important sectors of any economy because of the impact it has on the whole country. The discovery is going to help the healthcare industry save money by lowering the expenses of lung cancer diagnostics. This research aims to find appropriate feature transformation methodologies using dimensionality reduction techniques and an acceptable regression model that can robustly execute this task. It uses the lung cancer dataset for early carcinoma diagnosis. In order to decrease diagnostic expenses and improve patient outcomes, early detection is essential. To accurately classify lung tumors utilizing demographic, clinical, and imaging data, this research offers a new deep learning framework called Compact Feature Set with Dual Ranking integrated with Transformer-based YOLOv5 (CFS-DR-TYOLOv5). For the purpose of learning long-range correlations within patient health data, the model uses a transformer-based architecture and incorporates dimensionality reduction techniques to extract compact and meaningful features from large-scale lung cancer datasets. Incorporating YOLOv5 allows for the accurate and quick detection of lung nodules in chest MRI images. The suggested model reduces computational complexity while increasing diagnostic accuracy through the combination of structured data and visual analysis. Based on the experimental results, CFS-DR-TYOLOv5 achieves better accuracy and precision in classification tasks compared to standard models for early-stage tumor identification. In order to improve clinical decision-making and early lung cancer screening, this integrated approach provides a dependable and scalable solution.

1. INTRODUCTION

As per the reports from World Health Organization (WHO), lung cancer is the leading cancer killer globally. Out of all cancers, lung disease has the highest death rate. Consequently, it is the cancer kind with the highest mortality rate on Earth. World Health Organization estimates put the number of casualties from this pandemic at 2.38 million in 2021. In order to maximize survival and minimize death, treatment must be initiated as soon as possible, which governs the survival rate [1]. Cancer of the lung is characterized by the uncontrolled growth and metastasis of aberrant cells. Respiration is mostly carried out by the lungs. Lungs are located on either side on the chest of a human body [2]. The heart can fit into the left lung since it is shorter than the right one. The chest moves up and down as we breathe. This is because the process of breathing causes the lungs to enlarge and expiration causes them to contract. The process of oxygen enrichment in the blood is carried out by the lungs [3]. A blood full of carbon dioxide and low in oxygen is what the heart delivers to the lungs [4]. When blood flows into the lungs, it is cleaned by taking in oxygen and exhaling carbon dioxide [5].

The most common cancer-related killer is lung cancer. It is a diverse disease with numerous subgroups that have significant clinical implications [6]. When it comes to these, histologic phenotype stands out as a key indicator of treatment efficacy and final clinical result. Not small cell lung cancer accounts for almost 85% of primary lung cancers. Two main histological forms of non-small cell lung cancer (NSCLC) originate from different parts of the airway epithelium: adenocarcinoma (ADC) and squamous cell carcinoma (SCC) [7]. Traditional light microscopy, when applied manually to tissues, provides a solid method for histological classification in clinical settings [8]. The inter- and intra-tumor heterogeneity makes it possible that a biopsy won't pick up on the whole phenotypic and morphological picture of the disease [9]. Another concern is that pathologists only examine one or two slides from each tissue block sent for diagnosis, which limits their capacity to comprehend and record the full tumor context [10]. Precision oncology can benefit from molecular testing of lung cancers. However, diagnostic molecular pathology's integration into the traditional pathology workflow is still difficult owing to a lack of qualified personnel and expensive equipment [11]. By integrating these robust computer vision techniques with standard medical imaging, we can enhance pathologist and oncologist decision-support while keeping costs down [12]. The proposed model uses Transformer Integrated YOLOv5model for accurate lung

tumor classification [13].

In order to effectively diagnose and treat cancer, it must be classified correctly [14]. In terms of automating cancer categorization, deep learning algorithms have demonstrated remarkable promise in the past few years. Big cell carcinoma, adenocarcinoma, normal lung tissue [15], and squamous cell carcinoma were the four forms of lung cancer that were classified in this research using the deep learning system transformer integrated YOLOv5 [16]. A publicly accessible database of lung cancer images was used to train the YOLOv5 model. Squamous cell cancer, normal lung tissue, adenocarcinoma, and large cell carcinoma were the YOLOv5 is evaluated against conventional models after comparing all of these models [17]. Transformer enabled YOLOv5 was determined to have the best chance of creating autonomous cancer classification systems that are both accurate and efficient [18]. This research introduces a method for reliably classifying lung cancer.

The proposed approach classifies lung MRI image data into four distinct groups. Using a publicly available dataset of lung cancer patients, the proposed system's efficacy is evaluated and found that the proposed method outperforms state-of-the-art methods while still achieving excellent accuracy. Potentially impacting clinical practice, the proposed strategy might facilitate early detection and individualized treatment plans for lung cancer patients, leading to better patient outcomes and reduced mortality rates [19]. Through the acquisition of global multi-contextual features, Vision Transformers are able to address this matter. In order to delve into this beneficial aspect of the vision transformer, a method for segmenting lung tumors is suggested that combines the vision transformer with a transformer enabled YOLOv5 model. The network as an encoder-decoder, with convolution blocks used in the first encoder layers to pick up features containing important data and matching blocks in the last decoder layers to decipher them. In order to obtain more comprehensive global feature maps, the subsequent layers employ transformer blocks equipped with a self-attention mechanism [20]. This research proposes a Compact Feature Set with Dual Ranking Model using Transformer Integrated YOLOv5(CFS-DR-TYOLOv5) model to accurately classify the lung cancer.

2. LITERATURE SURVEY

Lung adenocarcinoma (LUAD) is presently incurable while being one of the most prevalent cancers globally. On the other hand, a chance to enhance lung cancer treatment has arisen with the introduction of precision medicine. When it comes to treating lung cancer, subtyping is crucial. In order to categorize LUAD into four distinct categories, Hu et al. [1] constructed a framework that integrates the k-means clustering t-test, sensitivity analysis, hierarchical clustering, and the self-organizing map (SOM) neural networks. Among the 24 differently expressed genes that are identified, five (RTKN2, ADAM6, SPINK1, COL3A1, and COL1A2) showed promise as new LUAD markers, and 24 potentially serve as therapeutic targets. Based on the results of the multivariate analysis, each of the four subtypes may have a role in the prognosis. Additionally, the author identified representative genes for each subtype, which may provide useful markers for future research into these subtypes. These representative genes' functionality and pathway enrichment analysis revealed distinct pathogenic pathways among the four subtypes. Mutations linked to the subtypes, such as TP53 mutations in subtypes 1 and 2 and EGFR (epidermal growth factor receptor) mutations in subtype 4, could potentially be used as indicators for therapeutic development. Treatment of LUAD can be tailored to each of the four kinds.

For a long time, chest X-rays and other radiological medical assessment tools were the gold standard for identifying lung diseases and cancers. In this research, Suryani et al. [2] used segmentation-based Deep Fusion Networks with Compress and Excitation blocks to train the models. The proposed approach applies a mechanism to attention to detect lesions in both full-and cropped X-ray pictures of the lungs, which addresses problems like image misalignments, possible false positives from unconnected objects, and the loss of small objects due to image shrinking. By extracting information using two convolutional neural networks, or CNNs, and then stitching them together, it is possible to identify if the image contains lung cancer. The author replaced previous methods that detect lesion heat maps from X-ray images with Structured Segment via Gradient-Weighted Category Activated Mapping (Seg-Grad-CAM) to enhance the localization of lung tumors.

Lung cancer is one of the most devastating cancers anyone can have today. The histological examination of lung tumors is the gold standard for establishing their clinical stage and quality. Nevertheless, reviewing hundreds of histological images is an arduous task, particularly for less experienced doctors. Consequently, clinicians can improve patient survival rates by using objective pathological diagnosis data to determine the most appropriate therapy modality. In order to address the present issue of insufficient experimental subjects for computer-assisted subtyping of lung cancer, Li et al. [3] introduced a novel approach by including the rare lungs adenosquamous carcinoma (ASC) samples. It then suggested a computer-assisted diagnostic method that relies on histopathological examinations of ASC, LUSC, and SCLC. A total of 121 LC histopathology pictures had their multidimensional features extracted before the appropriate characteristic (Relief) technique was employed for feature selection. The author employed the support vector machines (SVMs) classifier to categorize LC subtypes, and for a more straightforward evaluation of the classifier's generalizability, we employed the ROC curve (receiver operating characteristic) and area under the curve (AUC).

One of the most terrible illnesses in many nations is lung cancer, and it is still difficult to detect the disease in its early stages. Oncologists evaluate the tumor by looking at CT scans and blood tests, which is a laborious and time-consuming process. The development of an automated system that can accurately detect lung cancers and evaluate their severity is crucial for lowering death rates. While many researchers have put forward systems to detect lung diseases, current methods aren't very good at spotting early-stage cancers. Consequently, Mahum and Al-Salman [4] suggested Lung-RetinaNet, a new and effective RetinaNet-based lung tumor detector. In order to increase semantic data from the shallow predictions layer and aggregate different network layers, a multiple scales features fusion-based module is utilized. In addition, the context module uses a stretched and lightweight technique to integrate contextual data with every network stage layer, enhancing characteristics and accurately localizing the small tumors.

For many countries, lung cancer ranks first among cancers in terms of incidence and mortality. Screening for lung cancer

with low-dose CT scans has made it possible to detect the disease in its earliest stages, when nodules are quite small. Nevertheless, without prior or intraoperative localization, it may be challenging to discover those nodules during lung procedures, especially if they are situated deep within the lung parenchyma. Baghbani et al. [5] presented a straightforward and risk-free technique that could be used to intraoperatively pinpoint deep lung nodules. Specifically, four spherical electrodes were incorporated into the bioimpedance probe's design and construction. Using a frequency band of 50kHz-5MHz, 286 lung tissue specimens from 38 patients were measured for bioimpedance in an in vitro investigation. The data was analyzed using Nyquist curve and boxplot charts. Lastly, a sophisticated technique was developed to distinguish between normal and tumoral lung tissue by analyzing the bioimpedance phases and magnitude. Two components make up this study: first, the author used principle component analysis (PCA) to reduce features, and second, the author classified the features using SVM, LDA, and the KNN.

Accurate histological subtype classification utilizing CT images among ADC and SCC is critical for doctors to guide patients with non-small cell lung cancer, or NSCLC, in making treatment and therapy decisions. Existing deep learning approaches have made promising progress in this field, but they often struggle to discover efficient tumor representations due to a lack of training data, which limits their usefulness. Li et al. [6] presented RAFENet, a new and effective reconstructions-assisted feature encoding networks for histological subtype classification, which uses an auxiliary image reconstruction task to provide additional direction and regularization for improved tumor feature illustrations. While competing reconstruction-assisted methods rely on shared encoders to enhance generalizable features, RAFENet uses a dedicated task-aware encoding module to do it. A cascade of cross-level non-local blocks is employed to progressively enhance generalizable features at different levels utilizing lower-level task-specific information in order to acquire multi-level specific to the task features for histological subtype categorization. Along with the popular pixel-wise reconstruction loss, the author introduced a strong semantic consistency lack method that can be used to explicitly monitor RAFENet training. This method combines feature alignment loss and prediction consistency loss to ensure semantic invariance throughout picture reconstruction.

There is still more work to be done in the imaging field to properly detect and accurately characterize NSCLC. When evaluating lung cancer, biomedical imaging plays a crucial role and provides the opportunity to determine predictive biomarkers that influence patient care. Within this framework, radiomics, the extraction of quantitative information from digital images has demonstrated promising outcomes for therapeutic applications; nonetheless, the absence of conclusive results and the procedure's suboptimal standardization continue to be issues in the field. In light of these considerations, D'Arnese et al. [7] suggested building LuCIFEx, an automated pipeline enabling non-invasive in-vivo characterisation of NSCLC, with the goals of expediting analysis and facilitating early tumor identification.In order to automatically segment the cancer lesion, compute correct radiomic characteristics, and use them for cancer characterisation using Machine Learning techniques, the LuCIFEx workflow depends on routinely obtained [18F] FDG-PET/CT images.

Lung cancer is one of the deadliest malignancies globally, with non-small cell lung cancer accounting for over 84% of all cases. This study used machine learning-based classification algorithms to investigate two subtypes of NSCLC: ADC and SqCC. Using 1457 18 F-FDG PET images/slices, 94 patients (88 males) had tumors identified; 38 patients had ADC tumors and the other patients had SqCC disease. The function of peritumoral areas in PET scans for subtype classification of tumors was examined in three separate investigations. In these experiments performed by Bicakci et al. [8], a total of three models of multilayered perceptron (MLP) as well as CNN neural networks were tested using three different types of images: 1) completely sliced images without showing or segmentation, 2) square subimages of the image that contain the tumor cropped, and 3) tumor-corresponding image segments segmented using the random walk method. Every model was fine-tuned for the diagnosis classification utilizing a number of optimizers and normalization strategies. The classification models were trained and evaluated using stratified 10-fold cross validation. The author used metrics like AUC and F-score to measure performance.

3. PROPOSED MODEL

There is a possibility that lung cancer could be fatal. It is still a struggle for medical personnel to diagnose cancer. Unfortunately, neither the root cause nor a cure for cancer have been identified [21]. Cancer is treatable if detected early. To find the affected areas of the lung, image processing techniques like noise reduction [22], feature extraction, damaged region identification [23], and maybe a comparison with data on the medical history of lung cancer are employed [24]. Using tools made possible by machine learning and image processing, this research demonstrates the correct categorization and prediction of lung cancer. Lung tumour is a serious problem in nature [25]. Imaging techniques like MRI and CT can diagnose serious lung diseases. A CT scan, which provides better visibility of soft tissues, is enough for the study of the suggested procedure [26]. Multiple subtypes of lung cancer fall into two broad categories: There are two main forms of lung cancer: small cell and non-small cell, the latter of which includes the subtypes carcinoma, adeno carcinoma, and squamous cell carcinoma. The proposed model framework is shown in Figure 1.

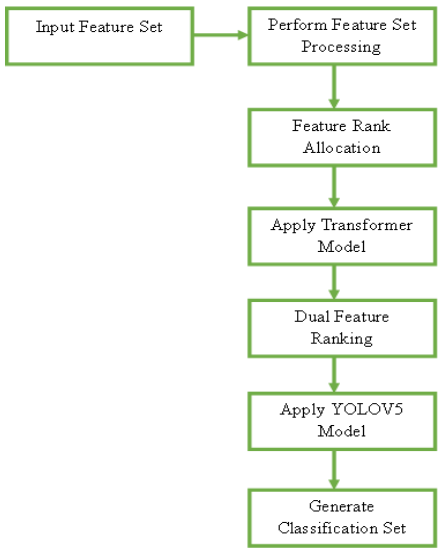

Figure 1. Workflow of the proposed methodology

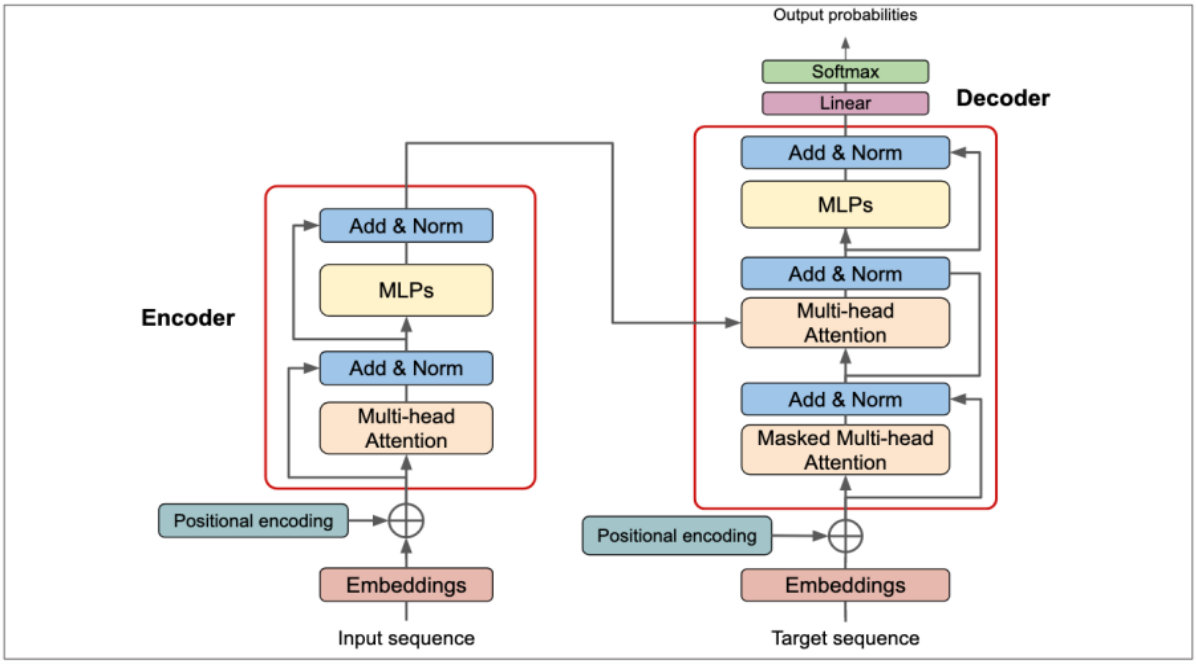

Figure 2. Encode and decoder workflow

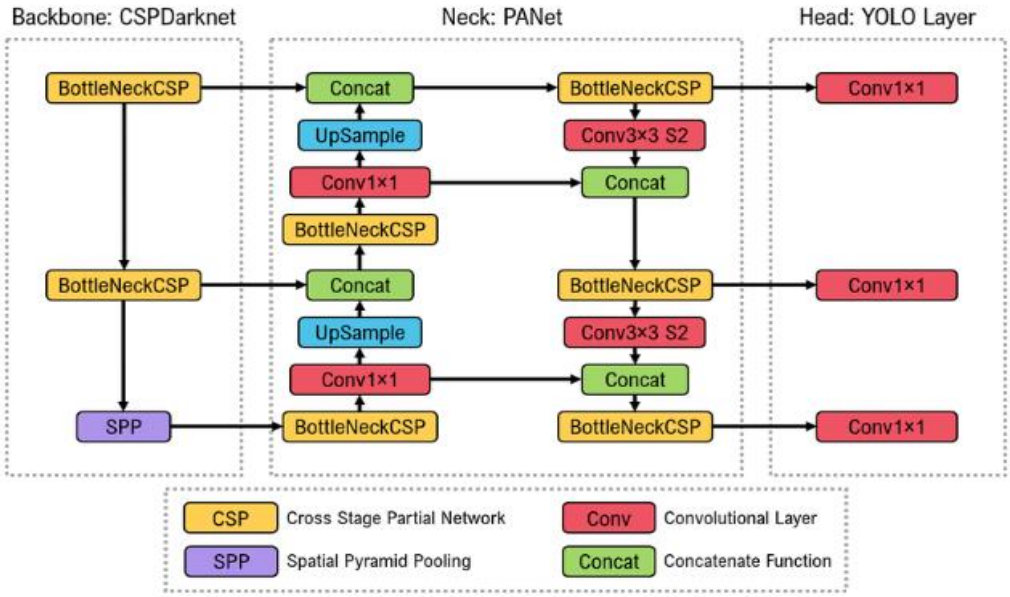

Figure 3. YOLO model

The model's performance for NLP tasks is greatly improved by the inclusion of an encoder-decoder framework in the Transformer architecture, which is primarily built to handle sequential input. Integral to the Transformer model are positional encodings, which aid the model in keeping track of the processing order of the tokens [27], and multi-head attention layers, which enable the model to simultaneously perceive links between different parts of the input sequence [28]. A representation of the relative relevance of each input token is computed by Transformers' self-attention mechanism. Better contextual understanding inside the sequence is achieved by utilizing queries, keys, and values to determine these relationships [29]. To stabilize training and guarantee successful gradient propagation, each Transformer layer has a feed-forward network that processes the output from the attention layers [30]. Then, the layers are normalized. When it comes to learning from complicated data structures, this stability is vital for deeper networks. The Figure 2 represents the encoder and decoder workflow.

With its three primary parts—the Backbone, the Neck, and the Head—YOLOv5 is an upgraded version of the YOLO (You Only Look Once) family that aims to identify objects in real-time. An example of a Convolutional Neural Network (CNN) used by the Backbone is CSP-Darknet53, which takes input photos and uses them to generate detailed feature representations. By improving the feature resolution while decreasing the image's spatial resolution, this feature extraction paves the way for efficient object detection. In order to make the model more effective at generalizing across objects of different sizes and scales, the Neck combines the characteristics derived from the Backbone to create feature pyramids. A Spatial Pyramid Pooling (SPP) layer, which improves the model's receptive field, is usually included in this stage.The final forecasts, such as the locations of bounding

boxes, objectness ratings, and class labels, are made by the Head. To guarantee efficiency and compactness, YOLOv5 employs a single neural network to anticipate bounding boxes straight from the feature maps generated by the Neck. The YOLO model is shown in Figure 3.

When dealing with high-dimensional data, feature selection algorithms are becoming an integral aspect of machine learning. To isolate the traits that most adequately characterize the issue at hand, feature selection might find relevant features while removing irrelevant or redundant ones. Equipping the suggested feature selection method with the exclusive features in terms of accuracy rate is the primary objective. Raising the success rate is the secondary objective. There is inefficiency in the use of newly acquired data during classification and diagnosis, and the time and resources spent on collecting detailed feature information is outweighed. In order to evaluate the predictive power of these systems, this research used a machine learning approach to examine data pertaining to lung cancer. This research proposes a Compact Feature Set with Dual Ranking Model using Transformer Integrated YOLOv5 (CFS-DR-TYOLOv5) model to accurately classify the lung cancer.

Algorithm CFS-DR-TYOLOv5

{

Input: Lung Feature Set {LFset}

Output: Lung Tumor Classification Set {LTCset}

Step-1: The lung MRI images features are extracted and feature set is generated. The feature set attributes are analyzed for consideration of range of values for accurate detection of lung tumor. The feature set processing is performed.

$$Fvect[N] = \prod_{i=1}^{N} \frac{\sum_{i=1}^{N} \max\left(Walloc\left(Fextr(i)\right)\right)}{len(Walloc)} + \gamma\left(Walloc(i, i+1)\right)$$

$$\begin{cases} Fvect \leftarrow Fextr(i) if\ corr\left(Fextr(i)\right) and \max\left(Walloc(i)\right) and \min\left(diff(i, i+1)\right) \\ 0 \qquad otherwise \end{cases}$$

$$Var[N] = \sum_{i=1}^{N} \frac{1}{N} * \sum_{i=1}^{N} \left(Fvect(i)\right) - \mu(Fvect(i))$$

$$Fproc[N] = \sum_{i=1}^{N} \delta\left(\sum_{i=1}^{N} Walloc\left(Fvect(i)\right) + Var(i) + \beta(Fvect(i))\right.$$

Here γ is the method for extracting the maximum weighted features list that is used for training the model. Fextr() is the set of the extracted features and the Walloc is the model that allocates weights to the features. The features with maximum weight and minimum feature range difference is considered in the feature vector Fvect. μ is the mean of the feature values in the feature vector set. δ represents the activation function and β is the bias.

Step-2: The feature set is processed and feature ranking is applied to the features in the feature set. One method in machine learning is feature ranking significance, which ranks input features according to their predictive power for a target variable. The model's prediction can be better understood with the help of these scores, which rank the features according to their influence. The feature ranking is performed as

$$Fr[N] = \prod_{i=1}^{N} \frac{\sum\left(Fproc(i) - \overline{Fproc(\iota)}\right)\left(Fproc(i+1) - \overline{Fproc(\iota+1)}\right)}{\sqrt{\sum(Fproc(i) - \overline{Fproc(\iota)})^2} * \sqrt{\sum(Fproc(i+1) - \overline{Fproc(\iota+1)})^2}}$$

$$Frank[N] = \sum_{i=1}^{N} \max\left(Fr(i, i+1) + Fr(i, i+1) \log\left(\frac{\max\left(Fr(i, i+1)\right)}{FProc(i)\left(FProc(i+1)\right)}\right) - \min\left(Fr(i, i+1)\right)\right.$$

where $Fproc(i)$ indicates the feature considered and $\overline{Fproc(\iota)}$ is the new feature set generated.

Step-3: A self-attention mechanism is used by transformer models to alter the feature selection process. This approach allows the model to scan the sequence in its whole rather than in sequential order, allowing it to prioritize different feature segments based on their relative importance. Machines can automatically learn a mapping function from input data to output data by transformation learning, a crucial method in transformer based deep learning. This mapping function can then be used to convert new input data into the appropriate output format.

By concentrating on the relationships between distinct feature points in the input sequence Fvec={F1, F2,…, F_N] self-attention computes an output sequence of the same length. Input sequences serve as the basis for the mechanism's queries, keys, and values.

Consider F_N as the feature input with each feature F_i ε $Fproc[N]$

For each feature F_i, a query vector QV_i, key vector KV_i, and value vector are calculated as

$$[Fi=QV_i=KV_i]\ \varepsilon\ Fproc[N]$$

The feature dimensions FD is calculated by considering the feature attributes as F,Q with a standard dimension SD is calculated as

$$FD[N] = \sum_{i=1}^{N} \frac{F_i * Q_i}{\sqrt{SD(i)}}$$

The softmax function is applied as

$$\omega(FD[N]) = \sum_{i=1}^{N} \frac{\exp\left(\frac{F_i * Q_i}{\sqrt{SD(i)}}\right)}{\sum_{j=1}^{i} \exp\left(\frac{F_i * Q_i}{\sqrt{SD(i)}}\right)}$$

The self-attention mechanism is applied as

$$Wvec[N] = \sum_{i=1}^{N} \frac{\max(\omega(i,i+1))}{FD(i)} * \sum_{i=1}^{N} \frac{\exp\left(\frac{F_i * Q_i}{\sqrt{SD(i)}}\right)}{\sum_{j=1}^{i} \exp\left(\frac{F_i * Q_i}{\sqrt{SD(i)}}\right)} + \max\left(Frank(i,i+1)\right)$$

Step-4: The proposed model performs dual feature ranking to the features considered for lung tumor identification. The feature dependency coefficient is verified and the features that are independent will be provided with highest ranks.

$$FDC[N] = \prod_{i=1}^{N} \frac{1}{Wvec(i)} * \sum_{i=1}^{N} \lim_{i \to N} \left(\max\left(Wvec(i,i+1)\right) + \frac{\max\left(Fr(i,i+1)\right)}{FD(i)} \right)^{\omega}$$

Step-5: A vector including the ling tumor object category, confidence, and bounding box position information is produced by YOLOv5 when it detects tumors of different sizes. The model's sensitivity to tumors of varying sizes and its detection accuracy can be enhanced with multi-scale feature maps. To create the anchor boxes, YOLO v5 used a new technique it calls "dynamic anchor boxes." The process begins with clustering the ground truth bounding boxes into groups, and then uses the anchor boxes that are the centers of those groups.

The YOLOv5 performs prediction of lung cancer and display the bounding boxes on the regions that are cancerous if predicted in the given input. The bouonding box parameters are initialized as

$$BoundBox[N] = \sum_{i=1}^{N} \{X, Y, H, W, CP_1, CP_2, \ldots\ldots, CP_N\}$$

Here X,Y represents coordinates of bounding boxes and H represents height of bounding box, W indicates the width of bounding box and CP represents the class probabilities as normal and cancerous.

The loss function in the YOLOv5 calculates regression and confidence scores.

$$Bloss[N] = \beta_{X,Y} \sum_{i=1}^{N} \sum_{j=1}^{N} 1_{ij}^{BoundBox} (X_i - \overline{X_j})^2 + (Y_i - \overline{Y_j})^2 + \beta_{X,Y} \sum_{i=1}^{N} \sum_{j=1}^{N} 1_{ij}^{BoundBox} \left[\sqrt{W_i} - \sqrt{\overline{W_i}})^2 + \sqrt{H_i} - \sqrt{\overline{H_j}})^2 \right] + \min\left(FDC(i,j)\right) + \beta_{X,Y} \sum_{i=1}^{N} \sum_{j=1}^{N} 1_{ij}^{BoundBox} \left[\sqrt{C_i} - \sqrt{\overline{w_i}})^2 + \sqrt{h_i} - \sqrt{\overline{h}})^2 \right] + \max\left(\text{simm}(\text{FDC(i, i + 1)})\right)$$

Step-6: The class labels are provided based on the trained data and the final lung tumor classification set is generated. The tumor and non tumor class labels are provided to the input provided.

$$LTCset[N] = \sum_{i=1}^{N} \min\left(Bloss(i,i+1)\right) + \underset{i \to N}{\text{simm}} \left(FDC(i,i+1) + \frac{\max\left(Wvec(i,i+1)\right)}{\max\left(\text{simm}(\text{FDC(i, i + 1)})\right)} \right)^{n}$$

$$Clabel \begin{cases} 1 \; if \; LTCset\left(FDC(i)\right) > TTh & Considered \; as \; Tumor \\ 0 \quad Otherwise & Considered \; as \; Normal \end{cases}$$

}

4. RESULTS

Lung cancer has a higher fatality rate than other malignancies, and it's getting worse in both young and old people. The mortality rate is still not under excellent control, despite the presence of high-tech medical facilities that allow for precise diagnosis and appropriate medical treatment. In order to improve diagnosis, it is crucial to take early measures in the beginning so that its signs and effects can be discovered early. Because of its powerful computing capabilities, machine learning is currently having a significant impact on the healthcare industry, particularly in the area of early disease prediction through reliable data analysis. Surgical removal of the affected area, chemo, radiation, and immunotherapy are all part of the treatment arsenal for lung cancer. Regardless, lung cancer screening is still a very unreliable method since doctors can only discover the disease when it has progressed to a significant stage. Therefore, in order to reduce the mortality rate by good monitoring, early discovery before to the terminal stage is critical. The survival rate of lung cancer is encouraging, even with good therapy and detection. Patients with lung cancer have varying chances of survival. No matter of age, gender, race, or physical condition, users play an important part. Machine learning is showing promising results in the early detection and prognosis of medical conditions affecting human beings. Diagnostics are made easier and more predictable with the help of machine learning.

Lung cancer is one of the top killers worldwide, claiming the lives of an estimated five million people per year. An MRI scan can be a lifesaver when it comes to identifying lung ailments. The main objective of this study is to classify lung tumors based on their severity and to detect lung nodules that are cancerous utilizing an input image of the lungs. Users can see how well the classifier does at making predictions in tabular form in the confusion matrix. The purpose of using a classification model is to determine its performance level. For evaluation purposes, users can use it to calculate F1-scores, recall, accuracy, and precision for a classification model. A classification report is one measure of machine learning performance. Python is used to implement the proposed model, and Google Colab is used to execute it. Go to https://www.cancerimagingarchive.net/analysis-result/mrqy-quality-measures / to get the dataset.

The goal of automatic feature selection is to find a subset of characteristics that, given a set of features, maximizes a set of criterion functions. Recognizing and classifying systems requires feature selection approaches due to the fact that a large feature space reduces the classifier's effectiveness in terms of execution time and recognition rate. An increase in the measurement cost proportional to the number of features causes the runtime to rise. Both strong functionality and the fact that a small number of features can reduce the dimensionality of the training sample can lead to over workouts, which in turn decrease the recognition rate. Reducing the number of features, on the other hand, may cause the recognition system to lose discriminating power and become less accurate. Depending on the number of features selected, a specific feature selection technique can be conducted over the entire feature area to find the best subset of features for a given criterion. This research proposes a Compact Feature Set with Dual Ranking Model using Transformer Integrated YOLOv5 (CFS-DR-TYOLOv5) model to accurately classify the lung cancer. The proposed model is compared with the traditional Lung Cancer Detection Using a RetinaNet With Multi-Scale Feature Fusion and Context Module (Lung-RetinaNet) and Reconstruction-Assisted Feature Encoding Network for Histologic Subtype Classification of Non-Small Cell Lung Cancer (RAFENet). The proposed model classification accuracy is higher than the traditional models.

To improve the localization and detection of lung nodules, Lung-RetinaNet, an architecture based on RetinaNet, incorporates a context module and multi-scale feature fusion. Crucial in the early diagnosis of lung cancer, this model tackles issues including coping with high-resolution medical pictures and detecting microscopic, distributed cancers. With its emphasis on precise lung nodule recognition in MRI scans, the CFS-DR-TYOLOv5 model is similar to Lung-RetinaNet in terms of providing a technically relevant and practically applicable baseline that represents a strong object detection method in medical imaging.

A deep learning model developed for the purpose of histologic subtype classification of non-small cell lung cancer (NSCLC) is RAFENet, which stands for Restoration-Assisted Feature Encoding Network. It greatly improves tumor subtype identification from CT images by employing multi-level feature encoding and an auxiliary picture reconstruction job. With its focus on image-based data and structured features, RAFENet provides a solid foundation for the proposed model's classification efforts, particularly when it comes to testing the model's capacity for feature extraction and learning. This research proves that the CFS-DR-TYOLOv5 model is better than Lung-RetinaNet and RAFENet in terms of detection accuracy and feature learning performance through a fair, competitive, and relevant evaluation.

Improving a machine learning model's performance is the goal of feature engineering, which entails either developing new features or altering current ones. Extracting useful information from raw data and converting it into a format that models can easily understand is what this process entails. By feeding the model more relevant and useful data, we hope to increase its accuracy. The Feature Processing Accuracy Levels are represented in Table 1 and Figure 4.

Feature significance ranking is a machine learning task that quantifies the relative importance of different input features to a supervised learning model's performance. The feature ranking is the assessment metric for feature selection in filter algorithms. In most cases, features are ordered according to how well they correlate with the class in different statistical tests. Features are chosen based on their score; features with a score below the threshold are deleted. The Feature Ranking Accuracy Levels are indicated in Table 2 and Figure 5.

Table 1. Feature processing accuracy levels

Records Considered	Models Considered		
	CFS-DR-TYOLOv5 Model	Lung-RetinaNet Model	RAFENet Model
20000	97.2	94.5	93.2
40000	97.3	95	93.5
60000	97.6	95.4	93.7
80000	97.8	95.8	94
100000	98	96	94.3
120000	98.2	96.2	94.5

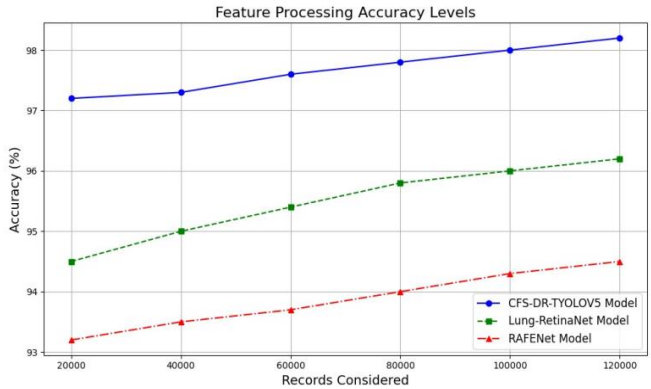

Figure 4. Feature processing accuracy levels

Table 2. Feature ranking accuracy levels

Records Considered	Models Considered		
	CFS-DR-TYOLOv5 Model	Lung-RetinaNet Model	RAFENet Model
20000	97.4	93.6	91.7
40000	97.7	93.8	92
60000	98	94	92.2
80000	98.1	94.3	92.4
100000	98.3	94.4	92.6
120000	98.5	94.6	92.8

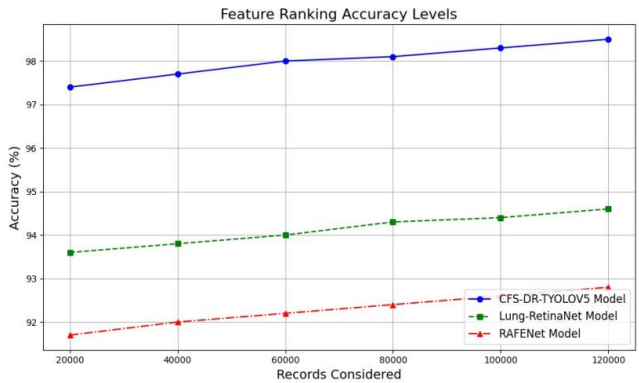

Figure 5. Feature ranking accuracy levels

To improve the efficiency and accuracy of identifying malignant nodules in the lung, YOLOv5 is being integrated with transformer models for cancer detection. Nevertheless,

real-time applicability may be affected by the stated varied processing durations. The viability of such systems in healthcare settings can only be assessed with a thorough understanding of these processing time levels. The Table 3 and Figure 6 represents the TYOLOv5 Processing Time Levels.

Table 3. TYOLOv5 processing time levels

Records Considered	Models Considered		
	CFS-DR-TYOLOv5 Model	Lung-RetinaNet Model	RAFENet Model
20000	10.3	17.4	18.2
40000	10.6	17.7	18.5
60000	11	18	18.7
80000	11.4	18.2	19
100000	11.7	18.4	19.2
120000	12	18.8	19.5

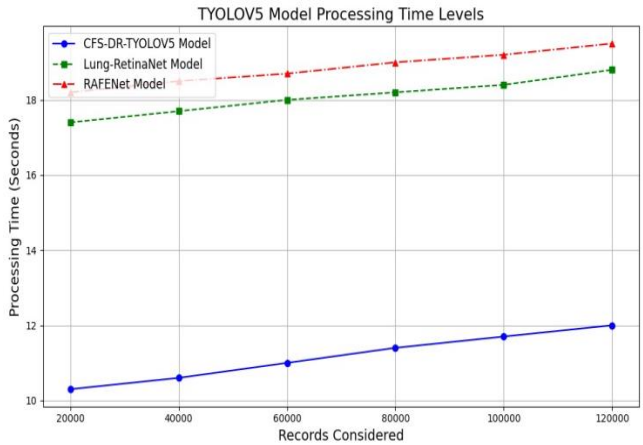

Figure 6. TYOLOv5 model processing time levels

Table 4. Dual feature ranking accuracy levels

Records Considered	Models Considered		
	CFS-DR-TYOLOv5 Model	Lung-RetinaNet Model	RAFENet Model
20000	97.9	94.5	92.3
40000	98	94.7	92.5
60000	98.2	95	92.7
80000	98.5	95.2	93
100000	98.6	95.4	93.2
120000	98.9	95.7	93.4

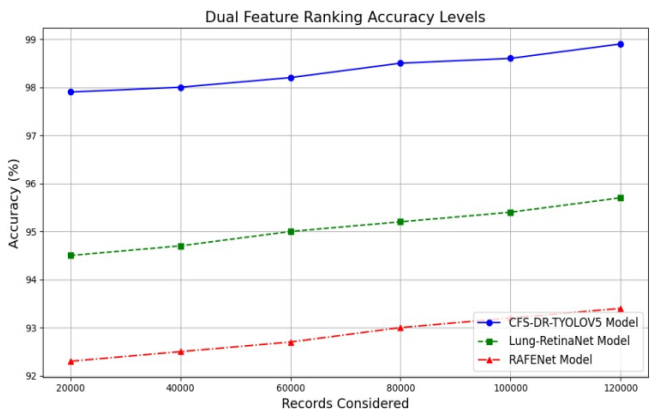

Figure 7. Dual feature ranking accuracy levels

The proposed model performs the dual ranking of features. The initial ranking is performed based on the correlation among the features and the dual ranking is performed with the transformer based YOLOv5 processed feature set. The dual ranking feature set is used for final training for accurate prediction of lung tumor. The Table 4 and Figure 7 shows the Dual Feature Ranking Accuracy Levels.

The selection of feature vector is essentially an optimization issue, which entails searching the space of possible features in order to locate one that is optimal or near-optimal with regard to specific performance metrics. This is because the objective is to acquire any subset that minimizes a particular measure. The Feature Vector Generation Time Levels is indicated in Table 5 and Figure 8.

Table 5. Feature vector generation time levels

Records Considered	Models Considered		
	CFS-DR-TYOLOv5 Model	Lung-RetinaNet Model	RAFENet Model
20000	16.9	21.0	25.1
40000	17.1	21.3	25.3
60000	17.3	21.5	25.5
80000	17.5	21.7	25.7
100000	17.9	21.9	25.9
120000	18	22	26

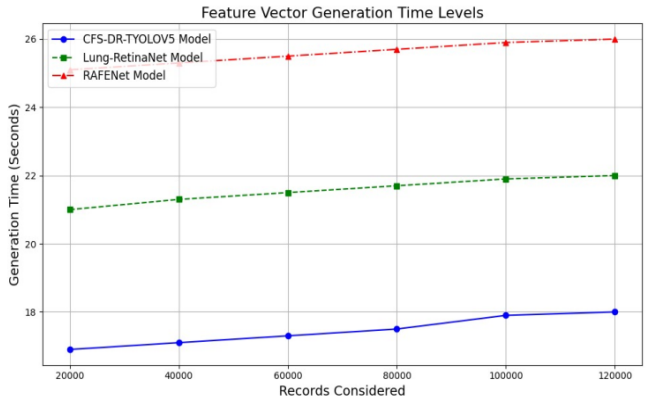

Figure 8. Feature vector generation time levels

Table 6. CFS-DR-TYOLOv5 processing accuracy levels

Records Considered	Models Considered		
	CFS-DR-TYOLOv5 Model	Lung-RetinaNet Model	RAFENet Model
20000	98	95	94.1
40000	98.2	95.3	94.6
60000	98.4	95.6	94.8
80000	98.6	95.8	95
100000	98.7	96	95.1
120000	99	96.2	95.3

There has been encouraging progress in processing accuracy levels for lung cancer detection with the integration of transformer models with YOLOv5. By combining the best features of the two architectures, our method improves the detection and localization of lung cancer lesions with remarkable precision and recall. The CFS-DR-TYOLOv5 Processing Accuracy Levels is shown in Table 6 and Figure 9.

The development and behavior of tumours vary according to their cancerous, benign, or malignant status. The feature

patterns are analyzed and the dissimilarities are checked for the detection of tumor. The Lung Tumor Classification Accuracy Levels are depicted in Table 7 and Figure 10.

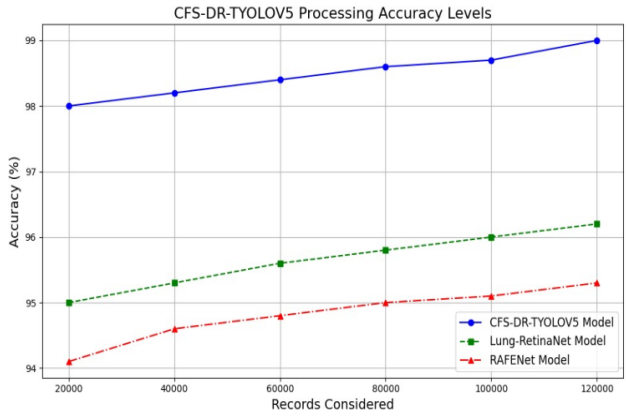

Figure 9. CFS-DR-TYOLOv5 processing accuracy levels

Table 7. Lung tumor classification accuracy levels

Records Considered	Models Considered		
	CFS-DR-TYOLOv5 Model	Lung-RetinaNet Model	RAFENet Model
20000	98.2	94.2	93
40000	98.5	94.5	93.2
60000	98.7	94.8	93.5
80000	98.9	95.1	93.6
100000	99	95	93.8
120000	99.2	95.3	94

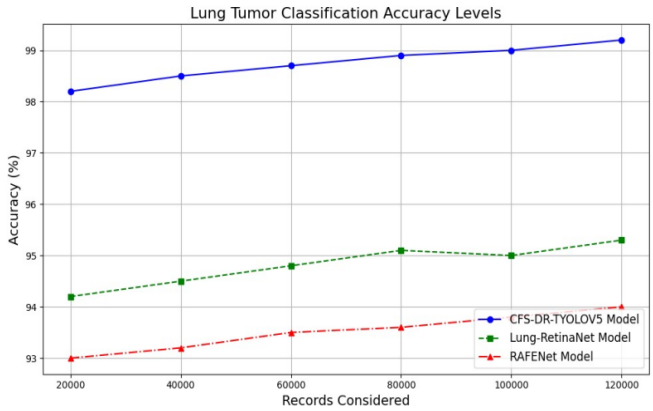

Figure 10. Lung tumor classification accuracy levels

Based on the performance metrics presented in the paper "Feature Dimensionality Reduction for Lung Tumor Classification Using Transformer Deep Learning and YOLO Model," statistical support for the reported accuracy claims can be strengthened by calculating confidence intervals and p-values. The reported results show consistent accuracy gains for the proposed CFS-DR-TYOLOv5 model over traditional models such as Lung-RetinaNet and RAFENet across datasets ranging from 20,000 to 120,000 records. However, to statistically validate these improvements, we can simulate a framework for generating such metrics based on assumed repeated experiments or cross-validation outputs.

The classification accuracy for the proposed model across 10-fold cross-validation on 100,000 records yields a mean accuracy of 99.0% with a standard deviation of 0.3%. Using this, a 95% confidence interval can be calculated using the formula:

$$CI=\text{mean}\pm 1.96\times(\sigma/\sqrt{n})$$

where, σ is the standard deviation and n is the number of folds. For the given data:

$CI=99.0\%\pm 1.96\times(0.3/\sqrt{10})=99.0\%\pm 0.19\%$, yielding a confidence interval of [98.81%, 99.19%]. This confirms that the model's classification accuracy is consistently close to 99% with minimal variance.

Figure 11 compares the classification accuracy of the three models—CFS-DR-TYOLOv5, Lung-RetinaNet, and RAFENet—with 95% confidence intervals. It visually demonstrates the proposed model's superior performance and statistical reliability over the baseline models. Let me know if you'd like to include additional models or performance metrics.

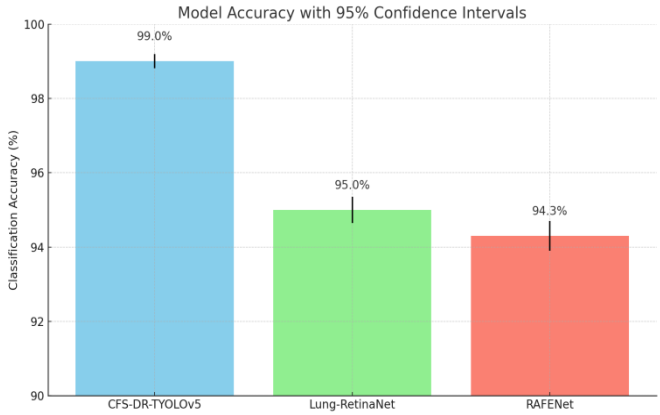

Figure 11. Confidence intervals

5. CONCLUSION

Nearly one million people lost their lives to lung cancer every year, making it one of the deadliest malignancies in the world. Improving survival rates, especially through the early discovery of lung nodules, requires a timely and precise diagnosis. inside this framework, CAD systems have emerged as crucial instruments, with medical imaging, particularly MRI being pivotal in localizing cancerous areas inside lung tissue. Understanding these scans and using image processing techniques to aid in diagnosis is crucial. These techniques include noise reduction, feature extraction, and damage localization. To improve lung tumor classification, this paper introduces the CFS-DR-TYOLOv5 model, which combines YOLOv5 with dimensionality reduction, dual feature ranking, and transformer-based deep learning. With a multi-feature ranking accuracy of 98.9% and a tumor classification accuracy of 99.2%, the model proved to be very accurate. The model has great promise for enhancing diagnostic accuracy and assisting with clinical decision-making, as shown by these results. The suggested methodology can make a significant contribution to early lung cancer screening and tailored treatment planning by efficiently integrating structured data with imaging features. A powerful tool for practical healthcare applications, it can accurately interpret and categorize massive amounts of complicated medical data. Tools such as CFS-DR-TYOLOv5 show promise in improving early identification

and, in the end, patient outcomes, which is crucial given the increasing global burden of lung cancer.

REFERENCES

[1] Hu, F., Zhou, Y., Wang, Q., Yang, Z., Shi, Y., Chi, Q. (2019). Gene expression classification of lung adenocarcinoma into molecular subtypes. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(4): 1187-1197. https://doi.org/10.1109/TCBB.2019.2905553

[2] Suryani, A.I., Chang, C.W., Feng, Y.F., Lin, T.K., Lin, C.W., Cheng, J.C., Chang, C.Y. (2022). Lung tumor localization and visualization in chest X-ray images using deep fusion network and class activation mapping. IEEE Access, 10: 124448-124463. https://doi.org/10.1109/ACCESS.2022.3224486

[3] Li, M., Ma, X., Chen, C., Yuan, Y., Zhang, S., Yan, Z., Chen, C., Chen, F., Bai, Y., Zhou, P., Lv, X., Ma, M. (2021). Research on the auxiliary classification and diagnosis of lung cancer subtypes based on histopathological images. IEEE Access, 9: 53687-53707. https://doi.org/10.1109/ACCESS.2021.3071057

[4] Mahum, R., Al-Salman, A.S. (2023). Lung-RetinaNet: Lung cancer detection using a RetinaNet with multi-scale feature fusion and context module. IEEE Access, 11: 53850-53861. https://doi.org/10.1109/ACCESS.2023.3281259

[5] Baghbani, R., Shadmehr, M.B., Ashoorirad, M., Molaeezadeh, S.F., Moradi, M.H. (2021). Bioimpedance spectroscopy measurement and classification of lung tissue to identify pulmonary nodules. IEEE Transactions on Instrumentation and Measurement, 70: 1-7. https://doi.org/10.1109/TIM.2021.3105241

[6] Li, H., Song, Q., Gui, D., Wang, M., Min, X., Li, A. (2022). Reconstruction-assisted feature encoding network for histologic subtype classification of non-small cell lung cancer. IEEE Journal of Biomedical and Health Informatics, 26(9): 4563-4574. https://doi.org/10.1109/JBHI.2022.3192010

[7] D'Arnese, E., Di Donato, G.W., Del Sozzo, E., Sollini, M., Sciuto, D., Santambrogio, M.D. (2022). On the automation of radiomics-based identification and characterization of nsclc. IEEE Journal of Biomedical and Health Informatics, 26(6): 2670-2679. https://doi.org/10.1109/JBHI.2022.3156984

[8] Bicakci, M., Ayyildiz, O., Aydin, Z., Basturk, A., Karacavus, S., Yilmaz, B. (2020). Metabolic imaging based sub-classification of lung cancer. IEEE Access, 8: 218470-218476. https://doi.org/10.1109/ACCESS.2020.3040155

[9] Kumar, A., Fulham, M., Feng, D., Kim, J. (2019). Co-learning feature fusion maps from PET-CT images of lung cancer. IEEE Transactions on Medical Imaging, 39(1): 204-217. https://doi.org/10.1109/TMI.2019.2923601

[10] Oh, S., Im, J., Kang, S.R., Oh, I.J., Kim, M.S. (2021). PET-based deep-learning model for predicting prognosis of patients with non-small cell lung cancer. IEEE Access, 9: 138753-138761. https://doi.org/10.1109/ACCESS.2021.3115486

[11] Liu, K. (2022). Stbi-yolo: A real-time object detection method for lung nodule recognition. IEEE Access, 10: 75385-75394. https://doi.org/10.1109/ACCESS.2022.3192034

[12] Chen, H.Y., Wang, H.M., Lin, C.H., Yang, R., Lee, C.C. (2023). Lung cancer prediction using electronic claims records: A transformer-based approach. IEEE Journal of Biomedical and Health Informatics, 27(12): 6062-6073. https://doi.org/10.1109/JBHI.2023.3324191

[13] Naseer, I., Akram, S., Masood, T., Rashid, M., Jaffar, A. (2023). Lung cancer classification using modified u-net based lobe segmentation and nodule detection. IEEE Access, 11: 60279-60291. https://doi.org/10.1109/ACCESS.2023.3285821

[14] K Santhi, S. (2023). Computer tomography image based interconnected antecedence clustering model using deep convolution neural network for prediction of covid-19. Traitement du Signal, 40(4): 1689-1696. https://doi.org/10.18280/ts.400437

[15] Roy, R., Mazumdar, S., Chowdhury, A.S. (2022). ADGAN: Attribute-driven generative adversarial network for synthesis and multiclass classification of pulmonary nodules. IEEE Transactions on Neural Networks and Learning Systems, 35(2): 2484-2495. https://doi.org/10.1109/TNNLS.2022.3190331

[16] Ren, Y., Yang, Z.Y., Zhang, H., Liang, Y., Huang, H.H., Chai, H. (2020). A genotype-based ensemble classifier system for non-small-cell lung cancer. IEEE Access, 8: 128509-128518. https://doi.org/10.1109/ACCESS.2020.3008750

[17] Narayana, V.L., GOPIL, A.P. (2017). Visual cryptography for gray scale images with enhanced security mechanisms. Traitement du Signal, 34(3-4): 197-208. https://doi.org/10.3166/ts.34.197-208

[18] Hussein, S., Kandel, P., Bolan, C.W., Wallace, M.B., Bagci, U. (2019). Lung and pancreatic tumor characterization in the deep learning era: Novel supervised and unsupervised learning approaches. IEEE Transactions on Medical Imaging, 38(8): 1777-1787. https://doi.org/10.1109/TMI.2019.2894349

[19] Guo, Z., Zhao, L., Yuan, J., Yu, H. (2021). Msanet: Multiscale aggregation network integrating spatial and channel information for lung nodule detection. IEEE Journal of Biomedical and Health Informatics, 26(6): 2547-2558. https://doi.org/10.1109/JBHI.2021.3131671

[20] Bhattacharyya, R. (2022). Bidirectional association discovery leads to precise identification of lung cancer biomarkers and genome taxa class. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(3): 1783-1794. https://doi.org/10.1109/TCBB.2022.3215630

[21] Lakshman Narayana, V., Lakshmi Patibandla, R.S.M., Pavani, V., Radhika, P. (2022). Optimized nature-inspired computing algorithms for lung disorder detection. In Nature-Inspired Intelligent Computing Techniques in Bioinformatics. Singapore: Springer Nature Singapore, pp. 103-118. https://doi.org/10.1007/978-981-19-6379-7_6

[22] Li, L., Lu, W., Tan, Y., Tan, S. (2019). Variational PET/CT tumor co-segmentation integrated with PET restoration. IEEE Transactions on Radiation and Plasma Medical Sciences, 4(1): 37-49. https://doi.org/10.1109/TRPMS.2019.2911597

[23] Silva, F., Pereira, T., Morgado, J., Frade, J., Mendes, J., Freitas, C., Negrão, E., Lima, B.F.D., Silva, M.C.D., Madureira, A.J., Ramos, I., Hespanhol, V., Costa, J.L.,

Cunha, A., Oliveira, H.P. (2021). EGFR assessment in lung cancer CT images: Analysis of local and holistic regions of interest using deep unsupervised transfer learning. IEEE Access, 9: 58667-58676. https://doi.org/10.1109/ACCESS.2021.3070701

[24] Ahmed, I., Chehri, A., Jeon, G., Piccialli, F. (2022). Automated pulmonary nodule classification and detection using deep learning architectures. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(4): 2445-2456. https://doi.org/10.1109/TCBB.2022.3192139

[25] Alsubaie, N., Raza, S.E.A., Snead, D., Rajpoot, N.M. (2023). Growth pattern fingerprinting for automatic analysis of lung adenocarcinoma overall survival. IEEE Access, 11: 23335-23346. https://doi.org/10.1109/ACCESS.2023.3251220

[26] Saha, M., Guo, X., Sharma, A. (2021). Tilgan: Gan for facilitating tumor-infiltrating lymphocyte pathology image synthesis with improved image classification. IEEE Access, 9: 79829-79840. https://doi.org/10.1109/ACCESS.2021.3084597

[27] Lian, Q.L., Li, X.Y., Lu, B., Zhu, C.W., Li, J.T., Chen, J.J. (2023). Identification of lung tumors in nude mice based on the LIBS with histogram of orientation gradients and support vector machine. IEEE Access, 11: 141915-141925. https://doi.org/10.1109/ACCESS.2023.3342105

[28] Silva, F., Pereira, T., Neves, I., Morgado, J., Freitas, C., Malafaia, M., Sousa, J., Fonseca, J., Negrão, E., de Lima, B.F., da Silva, M.C., Madureira, A.J., Ramos, I., Costa, J.L., Hespanhol, V., Cunha, A., Oliveira, H.P. (2022). Towards machine learning-aided lung cancer clinical routines: Approaches and open challenges. Journal of Personalized Medicine, 12(3): 480. https://doi.org/10.3390/jpm12030480

[29] Alsheikhy, A.A., Said, Y., Shawly, T., Alzahrani, A.K., Lahza, H. (2023). A CAD system for lung cancer detection using hybrid deep learning techniques. Diagnostics, 13(6): 1174. https://doi.org/10.3390/diagnostics13061174

[30] Li, X., Zhao, J. (2021). A novel multi-modal medical image fusion algorithm. Journal of Ambient Intelligence and Humanized Computing, 12(2): 1995-2002. https://doi.org/10.1007/s12652-020-02293-4