RESEARCH PAPER 1

**TITLE:** Lung Cancer Prediction Using Electronic Claims Records: A Transformer-Based Approach

**Abstract:**

Electronic claims records (ECRs) offer large-scale, longitudinal data on individuals' medical service actions, standardized across sites, unlike electronic medical records (EMRs). Recent studies show the potential of claims data for various medical applications, though many fail to consider strict cohort exclusion criteria or focus on early-stage cancer prediction. This study presents a lung cancer prediction framework utilizing a transformer-based model with rigorous exclusion criteria. We apply it to the entire population in Taiwan, achieving strong results with a predictive power of over 2.1, PPV of 5, and an AUC of 0.668 for all-stage lung cancer. For early-stage prediction, we report a predictive power of 2.0, PPV of 1, and an AUC of 0.645. Sub-cohort analysis aids in prioritizing clinical examinations, and onset analysis verifies exclusion criteria effectiveness. This framework advances lung cancer prediction using ECRs and can be applied broadly to disease risk prediction.

**Introduction:**

Digital collections of patient records, including EMRs and ECRs, have become valuable data sources for healthcare. EMRs provide detailed information on a patient’s health status, treatments, and test results, while ECRs focus on recording medical actions for insurance claims. ECRs offer a unique advantage over EMRs due to their standardization and broader scope, extending across sites and allowing longitudinal tracking of medical service behaviors. Machine learning models have shown promise in predicting diseases and analyzing clinical trends using large-scale healthcare data. However, most studies focus on EMRs and lack rigorous cohort exclusion criteria. This work introduces a transformer-based framework that leverages the spatial-temporal information in ECRs to predict lung cancer, particularly focusing on early detection. We also emphasize cohort design to avoid overestimating model performance and aim for a framework that can be generalized for disease prediction across health systems.

**Methodology:**

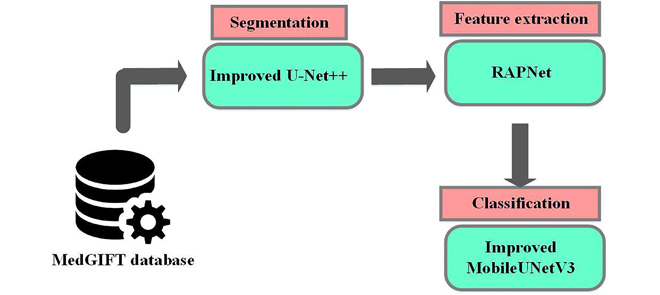
The dataset used is from Taiwan's National Health Insurance Research Database (NHIRD), which includes demographic, diagnostic, and medication data for the entire population. Four subsets are used: Population Set, Million Subset, Catastrophic Subset, and Local Development Set, which combine for model training and evaluation. The cohort focuses on individuals aged 45-65 with at least three years of claim history, excluding those with prior lung cancer diagnoses to focus on new incidences. The prediction model targets early-stage lung cancer (stage I and II), leveraging a transformer-based framework.

**Conclusion:**

This study introduces a novel transformer-based framework for lung cancer prediction using electronic claims records. By applying strict cohort exclusion criteria, we ensure that the model focuses on new cancer incidences without prior indications, avoiding inflated performance results. The proposed method demonstrates strong predictive power and AUC scores, even in early-stage lung cancer prediction, and proves effective at a population level. Sub-cohort analysis further refines clinical prioritization, while onset analysis validates the exclusion criteria. Our approach represents a significant step toward using non-intrusive claims data for large-scale disease prediction and has the potential to be generalized for various diseases across different healthcare systems. Future work will focus on improving encoding methods, expanding the framework to other diseases, and testing its application in different countries.

RESEARCH PAPER 2:

TITLE : Segmentation and Classification of Interstitial Lung Diseases Based on Hybrid Deep Learning Network Model



**Abstract:**  
Interstitial lung diseases (ILD) encompass over 200 disorders characterized by interstitial fibrosis and inflammation, significantly affecting lung disease morbidity and mortality. Early classification of ILD is crucial, but manual identification of regions of interest (ROI) from High-Resolution Computed Tomography (HRCT) images is time-consuming, and clinical similarities across disorders complicate diagnosis. This research proposes a hybrid deep learning network model for ILD classification. The lung portion of HRCT images is segmented using an improved U-Net++ model with multi-scale feature extraction, followed by feature extraction via a Refined Attention Pyramid Network (RAPNet). Finally, a MobileUNetV3 network classifies ILD into five categories. The proposed method's performance is tested on an ILD database, showing significant improvements over existing deep learning models.

**Introduction:**  
Interstitial lung disease (ILD) refers to over 200 types of chronic inflammation that affect lung tissue, posing a significant risk to breathing and increasing the likelihood of lung cancer. Early diagnosis is critical to guiding treatment decisions and improving patient outcomes. However, classifying ILD is challenging due to similar clinical symptoms and visual features across ILD classes, making precise detection difficult. HRCT images are often used to identify ILD, but manual extraction of features is time-consuming. Deep learning models, particularly convolutional neural networks (CNNs), have shown promise in medical image analysis by automating feature extraction and improving classification accuracy. However, existing models struggle with ILD due to the high intra-class variance and low inter-class differences. This research proposes a hybrid deep learning network model that uses improved U-Net++ for lung segmentation, RAPNet for feature extraction, and MobileUNetV3 for ILD classification, focusing on enhancing segmentation accuracy and classification performance.

START

// Step 1: Load HRCT Images

Input: HRCT images of lungs (entire dataset)

FOR each HRCT image IN dataset:

// Step 2: Lung Segmentation using Improved U-Net++

SegmentedLungImage = U-Net++(HRCT image)

// Step 3: Feature Extraction using RAPNet

Features = RAPNet(SegmentedLungImage)

// Step 4: ILD Classification using MobileUNetV3

ILDClass = MobileUNetV3(Features)

// Step 5: Store the predicted ILD class for each image

PredictedClass[image] = ILDClass

END FOR

// Step 6: Evaluate Performance

Evaluate PredictedClass using accuracy, precision, recall, and F1 score against ground truth labels

// Step 7: Report Results

Output:

- Accuracy, precision, recall, F1 score

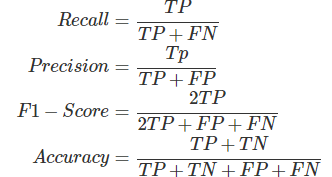
- Confusion matrix showing classification results for six ILD classes:

(consolidation, micronodules, ground glass, fibrosis, emphysema, normal)

- Segmentation performance of U-Net++ model

END

FORMULAE:



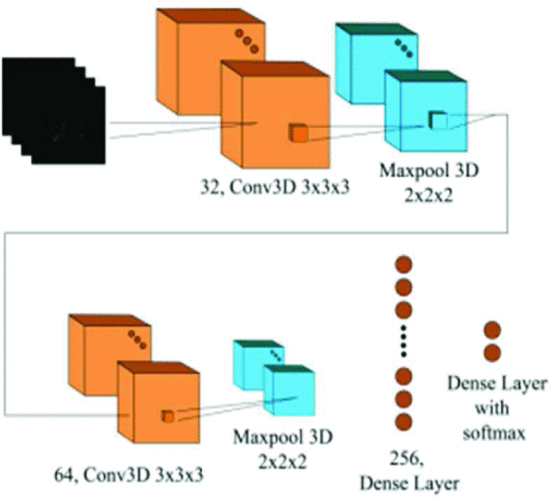
**Conclusion:**  
A novel hybrid deep learning network model is proposed for ILD classification using HRCT images. The model segments lung regions using an improved U-Net++ model, followed by deep feature extraction via RAPNet. These features are then classified into six ILD categories—consolidation, micronodules, ground glass, fibrosis, emphysema, and normal—using MobileUNetV3. The proposed method achieves 99.10% accuracy, outperforming existing deep learning-based models. By optimizing deep learning performance at each stage, the model enhances overall classification accuracy. Future work will focus on reducing feature set dimensionality using feature reduction techniques to further improve efficiency.

RESEARCH PAPER-3

TITLE: Lung Cancer Detection using 3D-Convolution Neural Network

**Abstract:**

Our paper uses an innovative 3D Convolutional Neural Network to determine lung cancer from patients' Computed Tomography (CT) scans because Convolutional Neural Networks (CNN) are useful for extracting important characteristics from images. This project aims to analyze CT scan slices (the data) and create a machine learning model based on the analysis. In this case, 3D Convolutional Neural Networks determine whether a person has cancer by evaluating the data and using a preprocessing technique. By utilizing this device, one can identify and eliminate cancerous cells at their earliest stages



**Introduction**

Lung cancer is the leading cause of cancer-related deaths, responsible for approximately 1.8 million deaths annually . Computed Tomography (CT) scans are essential in detecting lung cancer, as they generate detailed images by combining multiple X-ray slices taken from different angles of the lungs.

Deep learning, particularly using Convolutional Neural Networks (CNNs), has significantly improved the ability to classify medical images. CNNs excel in image analysis due to their weight-sharing feature, which allows them to efficiently capture important image details while minimizing memory usage . Unlike traditional deep networks, CNNs leverage convolutional layers and pooling techniques to extract relevant features from images, making them particularly effective in medical imaging .

This project aims to evaluate CT scan slices using various preprocessing techniques and implement 3D Convolutional Neural Networks (3D CNN) to develop a model capable of accurately predicting the presence of lung cancer.

The lung cancer detection process involves three key steps:

1. **Data Pre-processing**: CT scans in DICOM format are loaded, resized, and standardized using libraries like OpenCV and pydicom. The inconsistent image sizes are corrected, and 3D images are created for model input.
2. **Data Computation via 3D-CNN**: A 3D Convolutional Neural Network (CNN) is used to train and test the model. The architecture is optimized by experimenting with various layer configurations. 1495 images are used for training, and 100 images for validation.
3. **Confusion Matrix**: A confusion matrix compares predicted and actual labels to evaluate model performance. With 0 representing no cancer and 1 for cancer, false positives and negatives are minimized to improve accuracy.

The model achieved 93% accuracy after training on 100 epochs with optimal layers.

PSUDOCODE:

START

# Step 1: Data Pre-processing

LOAD DICOM images from dataset

FOR each image in dataset:

Read image using pydicom

Resize image to standard dimensions (e.g., 50x50x20)

Normalize pixel values (0-255) to range (0-1)

Append processed image to the dataset array

END FOR

# Step 2: 3D-CNN Model Training

Initialize 3D-CNN model with:

Convolutional layers, Pooling layers, and Fully Connected layers

Define loss function (e.g., binary cross-entropy)

Define optimizer (e.g., Adam)

Divide dataset into training set (80%) and validation set (20%)

FOR epoch in range(number of epochs):

FOR batch in training data:

Feed batch into the 3D-CNN model

Compute forward pass, loss, and backpropagate to update weights

END FOR

Validate model on validation set

Store validation accuracy

END FOR

# Step 3: Confusion Matrix and Cancer Prediction

FOR each image in validation set:

Predict cancer (0: No cancer, 1: Cancer) using trained 3D-CNN model

Compare predicted labels with actual labels

Update confusion matrix (True Positive, True Negative, False Positive, False Negative)

END FOR

Calculate accuracy, precision, recall from confusion matrix

Display results

END

CONCLUSION

We used 3D Convolutional neural networks in our research to classify CT images of lung nodules as cancerous (malignant) or non-cancerous (benign). The Cancer Imaging Archive provided the dataset used. A pre-processing step ensures that CT images are of equal size and format before they are used to model a network. When we trained the model, we found it had an accuracy of 93%. In addition to measuring Precision, Recall, Kappa, and F1 score, we examine various metrics to evaluate the model's performance. Precision = 0.68669, Recall = 0.64384, Kappa score = 0.39733, and F1 score= 0.62699

RESEARCH PAPER 4

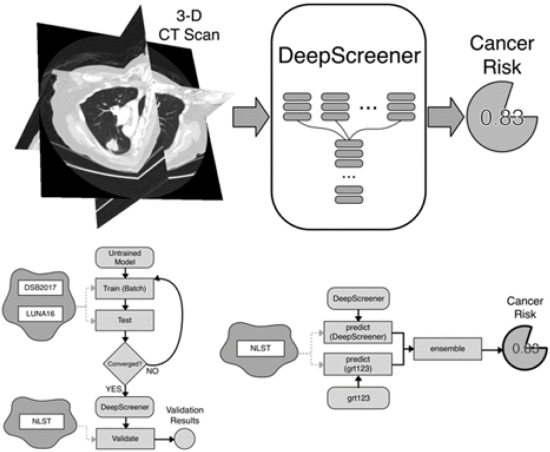
TITLE: Spatial Pyramid Pooling With 3D Convolution Improves Lung Cancer Detection

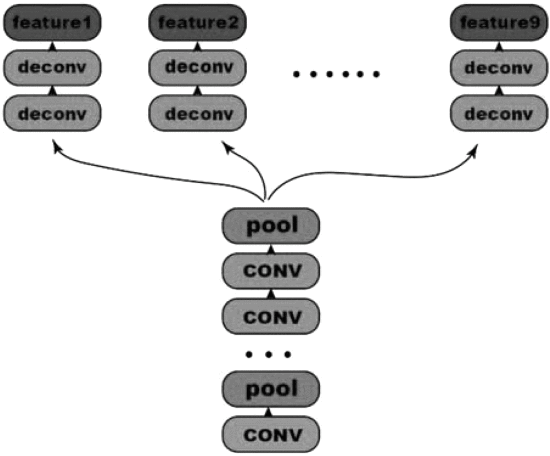
**Abstract**

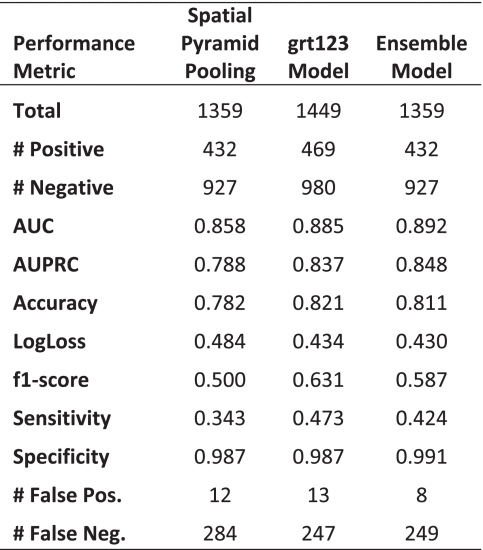
Lung cancer remains the leading cause of cancer-related deaths worldwide. While low-dose computed tomography (CT) screening has demonstrated a significant reduction in lung cancer mortality, it is often hindered by a high false-positive rate, leading to unnecessary diagnostic procedures and increased patient anxiety. To address this challenge, we introduce DeepScreener, an advanced deep learning algorithm designed to predict lung cancer status from volumetric CT scans. DeepScreener leverages Spatial Pyramid Pooling, a method that secured a top 1% ranking in the Data Science Bowl 2017, and integrates it with 3D Convolution for enhanced accuracy. Testing on an independent dataset of 1449 low-dose CT scans from the National Lung Screening Trial (NLST) cohort reveals that DeepScreener achieves an AUC of 0.892, surpassing the performance of previous state-of-the-art algorithms that relied solely on 3D Convolution. This advancement underscores the potential of deep learning techniques to improve lung cancer detection, reduce false positives, and ultimately enhance the efficacy of lung cancer screening programs.

**Introduction**

Lung cancer is the leading cause of cancer-related deaths globally, and early detection is critical to improving survival rates. Low-dose computed tomography (CT) scans have proven effective in reducing mortality by providing detailed imaging for early diagnosis. However, the high false-positive rate associated with low-dose CT screenings often leads to unnecessary follow-up tests and increased patient stress. Traditional methods of screening have struggled to address these issues effectively. Recent advancements in deep learning offer new avenues to enhance diagnostic accuracy and reduce false positives. In this study, we introduce DeepScreener, an advanced algorithm designed to predict lung cancer status from volumetric CT scans. DeepScreener utilizes Spatial Pyramid Pooling, a technique that achieved top rankings in the Data Science Bowl 2017, combined with 3D Convolution to improve diagnostic performance. Our evaluation on an independent dataset of 1449 CT scans from the National Lung Screening Trial (NLST) demonstrates DeepScreener’s high accuracy and effectiveness in enhancing lung cancer detection.







Psudocode:

1. Load and Preprocess Data

- Load volumetric lung CT scan dataset (NLST cohort)

- For each CT scan:

- Read and preprocess DICOM images

- Handle anomalies (e.g., inconsistent slice spacing, duplicate slices)

- Resize and normalize images as required

- Split dataset into training, validation, and test sets

2. Define Model Architecture

- Input Layer: 3D CT scan volume

- Apply 3D Convolutional Layers:

- Convolution3D(filters, kernel\_size, activation='relu', padding='same')

- Pooling3D(pool\_size, padding='same')

- Apply Spatial Pyramid Pooling (SPP) Layer:

- Spatial Pyramid Pooling with multiple levels

- Fully Connected Layers:

- Dense(units, activation='relu')

- Dropout(rate)

- Dense(units, activation='softmax') # or 'sigmoid' for binary classification

3. Train Model

- Compile Model:

- Loss function: Binary Crossentropy (for binary classification)

- Optimizer: Adam or similar

- Metrics: Accuracy, AUC

- Train Model with training set:

- Set epochs and batch size

- Monitor validation accuracy and loss

- Evaluate Model on validation set

4. Evaluate Model Performance

- Test Model with test set

- Calculate performance metrics:

- Accuracy

- AUC (Area Under Curve)

- Precision-Recall curve

- Generate confusion matrix:

- True positives (TP)

- True negatives (TN)

- False positives (FP)

- False negatives (FN)

5. Report Results

- Compare model performance with existing algorithms

- Highlight improvements in AUC and accuracy

- Discuss potential improvements and future work

6. Conclusion

- Summarize findings

- Discuss the impact of combining SPP and 3D Convolution on lung cancer detection

- Suggest further research or potential applications

**Conclusion**

DeepScreener represents a significant advancement in the field of lung cancer detection through CT imaging. By integrating Spatial Pyramid Pooling with 3D Convolution, the algorithm achieves an impressive AUC of 0.892, outperforming previous state-of-the-art methods that relied solely on 3D Convolution. The consistent high performance of DeepScreener on a large independent dataset highlights its robustness and potential for widespread application. This improvement in diagnostic accuracy not only enhances early detection but also has the potential to reduce the rate of false positives, leading to fewer unnecessary diagnostic procedures and lower patient anxiety. The successful application of deep learning techniques in DeepScreener underscores the transformative potential of these technologies in improving lung cancer screening and patient outcomes.