Lung Classification

Introduction

For lung classification challenges, I present in this document a comprehensive overview of the methodology employed. The goal of the lung classification challenge is to predict overall survival (OS). To tackle this challenge, I have opted to implement the 3D wavelet transform method, combining it with a 3D CNN model for feature extraction in CT images. Additionally, a Proportional Hazards model is employed for feature selection, considering both CT image features and clinical features. Finally, a combined dataset, integrating clinical features and CT image volumetric features, will be used to train a Proportional Hazards Model.

Dataset:

The organizers have provided two main datasets for this project:

1. Image datasets:
   CT scan → Harmonization to guarantee unification in the image resolution across scans.
2. **Clinical datasets :**
   Gender (gender)\Age (age)\Smoking status (smoking_status)\ clinical category\ regional nodes category and metastasis category
3. Labels :
   Survival time (survival_time) → time from diagnosis to death or last follow-up
   Event (event) → if the patient suffered the event (death)

Preprocessing (CT images)

The initial step is consistently the implementation of a preprocessing method. To ensure uniformity in input image size before training the model, following the normalization of the data using mean and standard deviation values, all image volumes are resized to consistent dimensions: (512, 512, 512), with a length, height, and depth each set to 512. This process is employed to streamline the assurance of uniform input image sizes without the need for individual checks.

Feature extraction (CT images)

After the preprocessing process, I utilize the 3D wavelet transform method to extract the approximation coefficients.

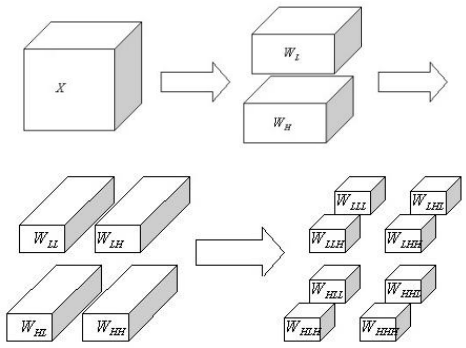

Fig. 1 The resolution of a 3-D signal is reduced in each dimension

The approximation coefficients capture the essential structural information of the signal or image, while discarding fine details. This is valuable in scenarios where the main trends and features of the

data are more critical than specific details. For example, in medical imaging, the general shape and structure of an organ might be more relevant than individual pixel-level details.

And then we apply a 3d CNN model to extract the volumetric feature, the layers of the CNN are shown below :

```
def get_model(depth=256, height=256, width=256):
    """Build a 3D convolutional neural network model."""

    inputs = keras.Input((depth,height,width, 1))

    x = layers.Conv3D(filters=16, kernel_size=3,
activation="relu")(inputs)
    x = layers.MaxPool3D(pool_size=2)(x)
    x = layers.BatchNormalization()(x)

    x = layers.Conv3D(filters=32, kernel_size=3, activation="relu")(x)
    x = layers.MaxPool3D(pool_size=2)(x)
    x = layers.BatchNormalization()(x)
    #x = layers.Dropout(0.3)(x)

    x = layers.Conv3D(filters=128, kernel_size=3, activation="relu")(x)
    x = layers.MaxPool3D(pool_size=2)(x)
    x = layers.BatchNormalization()(x)
    #x = layers.Dropout(0.3)(x)

    x = layers.Conv3D(filters=256, kernel_size=3, activation="relu")(x)
    x = layers.MaxPool3D(pool_size=2)(x)
    x = layers.BatchNormalization()(x)

    x = layers.GlobalAveragePooling3D()(x)
    x = layers.Dense(units=256, activation="relu")(x)
    x = layers.Dropout(0.4)(x)
    x = layers.Dense(units=128, activation="relu")(x)
    x = layers.Dropout(0.4)(x)



    outputs = layers.Dense(units=1, activation="linear")(x)

    # Define the model.
    model = keras.Model(inputs, outputs, name="3dcnn")
    return model
```

We extract the feature output from the `Dense` layer with units=128 and activation="relu" in the 3D CNN model. This results in 128 features being extracted from the model.

Feature selection

The feature in clinical data and image feature extracted by 3d wavelet and 3d CNN, the selection will use the Proportional Hazards model, with a threshold set at $p < 0.05$. so the feature which we keep :

	cmp to	z	p	-log2(p)
covariate				
age	0.00	2.12	0.03	4.89
feautre_cnn1	0.00	2.32	0.02	5.61
clinical_category	0.00	2.97	<0.005	8.40
metastasis_category_1	0.00	4.80	<0.005	19.27
metastasis_category_2	0.00	5.05	<0.005	21.13
reginal_nodes_category	0.00	2.41	0.02	5.95

Train model

Train the Proportional Hazards model by using the selected feature, the concordance index for training data sets:

Concordance = 0.73

Partial AIC = 1222.82

log-likelihood ratio test = 86.34 on 7 df

-log2(p) of ll-ratio test = 50.35

0.7349034308532312