Imperial College London

Faculty of Medicine

Department of Metabolism, Digestion and Reproduction

Scoring radiological features for lung CT scans

Author: Chang Liu
Supervisor: Dr Elsa Angelini

Submitted in partial fulfilment of the requirements for the MRes degree in Biomedical Research (Data science) of Imperial College London
March 2021

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, conducted during the current year of the MRes in Biomedical Research at Imperial College London. Any ideas or quotations from the work of other people, published or otherwise, or from my own previous work are fully acknowledged in accordance with the standard referencing practices of the discipline.

For the COVID19-20 dataset, the annotation of the dataset was made possible through the joint work of Children's National Hospital, NVIDIA and National Institutes of Health for the COVID-19-20 Lung CT Lesion Segmentation Grand Challenge.

The Python code that I work with (for lung segmentation, Fisher score and AFS feature selection models) is based on open-source packages available on GitHub.

Abstract

Understanding the important radiological characteristics of normal and disease scans is essential for precision medicine to obtain imaging biomarkers and scoring systems. Imaging biomarkers are challenging to define, and it is still an exploratory field. The main aim of this study is to design the extraction of radiomics biomarkers and quantify their stabilities. We used an open-source python package to extract imaging signatures and compared different feature selection methods, including Fisher score, auto-encoder and universal representation transformer. We encoded 93 features into 16 features and achieved better performance in several evaluation metrics. The dimensionality of the encoded features is significantly smaller than that of the original feature space, which not only saves computations but also reduces the impact of noise and outliers. This research mainly focuses on signature building, further studies can explore other lung segmentation methods and classification methods.

Table of Contents

Abbreviations

AE	Auto-Encoder
AFS	Attention-based Feature Selection
CAP	Community-Acquired Pneumonia
CNN	convolutional neural network
COVID-19	Coronavirus Disease 2019
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
DSC	Dice Similarity Coefficient
FN	False Negative
FP	False Positive
GGO	Ground-Glass Opacity
GLCM	Gray Level Co-occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Gray Level Run Length Matrix
GLSZM	Gray Level Size Zone Matrix
IBSI	Image Biomarker Standardisation Initiative
KL	Kullback-Leibler
NGTDM	Neighbouring Gray Tone Difference Matrix
PET-CT	Positron emission tomography CT
RIDER	Reference Image Database to Evaluate Therapy Response
RMSE	Root-Mean-Square Error
ROI	Region of Interest
RT-PCR	Reverse-Transcription Polymerase Chain Reaction
SPGC	Signal Processing Grand Challenge
TN	True Negative
TP	True Positive
t-SNE	t-Distributed Stochastic Neighbour Embedding
URT	Universal Representation Transformer
VAE	Variational Auto-Encoder
WBC	White Blood Cell
WHO	World Health Organization

1 Introduction

Imaging is an important technology in medical science. In the past decade, the field of medical image analysis has grown exponentially. Medical imaging benefits medical science and patients by non-invasively assessing the characteristics of human tissue. It is expected to contain complementary and interchangeable information with other sources. Computed tomography (CT) imaging is a common technique that can obtain high-resolution anatomical images. Due to its availability and rapidity, lung CT imaging has been routinely used in oncology research for prediction, screening, and treatment response assessment [1].

1.1 Radiomics

With the increase in the use of medical imaging technologies and the development of precision medicine, it has become more essential to understand the radiological signature and define appropriate imaging biomarkers. The field of introducing and investigating imaging biomarkers is called radiomics, which refers to mining and extracting quantitative features from medical images with the objective of modelling diagnostic or prognostic results [2]. Extracted features describe a wide range of characteristics of the ROI. To ensure the constructed models contain adequate mineable information with high fidelity and high throughput, the main workflow in radiomics is divided into the following five sections [3].

1.1.1 Image Acquisition

In clinical CT image acquisition, imaging characteristics such as image resolution (pixel size and slice thickness), patient position and reconstruction algorithm vary widely. The lack of standardization of these protocols across medical imaging centres with different scanners is usually not a problem when radiologists try to identify

radiological features. However, variations that are not caused by latent biological effects may be introduced by differences in acquisition and image reconstruction parameters when quantitative analysis of images is performed [4], [5].

1.1.2 Image Segmentation

When radiomics is used to detect whether a tumour is malignant or benign, the lesion contours need to be extracted as the region of interest (ROI). Since our purpose in this study is to detect certain diseases in the entire lungs, we define our ROI as individual lobes, individual lungs or the 2 lungs together. Segmentation is a critical step in the radiomics workflow because radiomics features are extracted from segmentation, and different segmentation methods will generate different values of extracted features. Manual annotation is time-consuming, and it varies between radiologists, which leads to seeking to develop automated segmentation pipelines that can also produce highly accurate and reproducible masks.

1.1.3 Feature Extraction and Quantification

Once the lung regions are segmented, pre-designed radiomics features can be extracted from them. These features describe the histogram of the grey intensity values, the shape, and the characteristics of texture patterns in the ROI. Extracted features will be discussed in detail in Subsection 2.2.

"First Order Features" describe the distribution of grey-level values by converting the 2D or 3D ROI into single histograms and basic metrics (e.g., mean, skewness, energy). They allow us to explore properties such as asymmetry. Shape-based features can be used to determine the surface-to-volume ratio of the ROI, which is an important characteristic. Texture-based features provide information about the correlation between neighbour voxels, which is crucial to detect tissue heterogeneity [6]. Different subtype of texture-based features can be calculated by different matrices where global

and local texture information has been encoded. These matrices include Gray Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM), Neighbouring Gray Tone Difference Matrix (NGTDM) and Gray Level Dependence Matrix (GLDM) [7], [8].

There are several factors that will affect radiomics features quantification. Technical influences such as reconstruction parameters and segmentation algorithm are the most important [4]. Some features are found to be unstable between scans acquired within even a short time [9]. Besides, the choice of image intensity quantization scheme such as the value of bin width will result in different feature values [5].

1.1.4 Feature Reduction

As described above, a large number of radiomics features could be extracted. However, not all these features are informative for some specific classification tasks and most of the features are highly correlated. The risk of overfitting will increase if too many irrelevant features are considered in a predictive model, which will jeopardize model generalization to patients not previously evaluated. Besides, if the number of features is larger than the number of samples, some machine learning models do not accept it as an input. Therefore, reducing the dimensionality of the feature space is a crucial step in radiomics. The feature selection strategies aim to select the most robust subset of features based on training data.

Features could be selected by several criteria, including reproducibility, relevancy and non-redundancy [3]. Reproducibility refers to the robustness of the feature to image pre-processing and segmentation. Selected features should be insensitive to properties such as slice thickness and image rescaling. Features that are highly related to the target label are more informative. Redundant features are the ones with high correlations with other features, which should be excluded after feature selection.

1.1.5 Statistical Analysis and Classifier Modelling

Classifier modelling refers to the classification for input images using extracted radiological features. After obtaining a large number of high-quality data, we can discover underlying patterns in the dataset. Supervised and unsupervised machine learning methods, such as random forests and k-means clustering, could be used as the predictive or classification model. These methods use prior knowledge of the training set to predict the results of unseen samples.

Radiomics has already been used to solve different medical problems, such as nodule malignancy prediction [7], cancer diagnosis [10] and cardiac disease [11]. Besides, it has also been applied to the screening of coronavirus disease 2019 (COVID-19) [12], [13]. Potentially, radiomics could be used to provide a tool to detect disease in the whole scan or a sub-region for a given CT scan into a subtype of diseases. Although several challenges were raised because of lacked manual annotation and heterogeneity of datasets, it is expected to facilitate outcome prediction and improve personalised treatment planning in the near future.

1.2 COVID-19 Detection

COVID-19 is a respiratory disease caused by the SARS-Cov-2 virus [14], which is highly contagious [15] and has been declared to outbreak a pandemic by the World Health Organization (WHO). The typical clinical characteristics of COVID-19 include fever, pneumonia and decreased white blood cell (WBC) count [16]. The current gold standard for diagnosing a positive case is the reverse-transcription polymerase chain reaction (RT-PCR) test [17]. However, this method suffers from low sensitivity [18], [19], which leads some suspected patients with negative RT-PCR results not to be treated early. Medical imaging could play an important role as a complementary examination in diagnosing, because COVID-19 will lead to some typical imagery characteristics in lungs, such as ground-glass opacities (GGO) and pulmonary fibrosis

[20]. Despite its advantage of short diagnosis time, these imaging features are commonly shared between COVID-19 and other types of pneumonia [21]. This will lead to another problem distinguishing COVID-19 from other types of pneumonia.

In machine learning methods, lung CT scans are used more regularly than other imaging modalities, such as X-ray and Ultrasound scans, because the false-negative rate of CT scans is lower [18]. Recent studies have shown the potential of machine learning to diagnose COVID-19 using lung CT scans. Chen *et al.* [22] segment suspicious lesion region using a UNet++ model and predict the label based on the appearance of segmented regions. Zhang *et al.* [23] proposed a model which uses a U-Net model for lung segmentation and takes the segmentation results as the input of a 3D CNN COVID-19 probability predicting model. Li *et al.* [24] present a deep convolutional neural network (CNN) model (COVNet) using ResNet50 as the backbone, to extract visual features from volumetric chest CT exams for the detection of COVID-19. Their study suggests that this model is robust to distinguish COVID-19 from community-acquired pneumonia (CAP). Most recent studies are based on computer vision models using CNN architecture on input images directly, which are computationally expensive, time-consuming and not interpretable. On the other hand, the existing radiomics methods require radiologists to manually segment the lung lesion area.

1.3 Aim of Study

The essential hypothesis in radiomics is that the constructed descriptive matrices are capable of providing sufficient information of the underlying pathophysiology and therefore facilitate disease prediction, prognosis, and/or diagnosis. As mentioned above, this project aims to define a lung-based feature extraction strategy and develop a radiomics-based model that will encode global lung CT scans into subtypes of disease and provide a score to describe the deviation from normality. We mainly focus on

tumour, COVID-19 and CAP detection. We review the feature extraction technique so that the extracted features are robust to heterogeneous CT images. We quantify feature stabilities across various sources of variability. Subsequently, relatively informative features are selected for image clustering and classification, and selected features are projected onto 2-dimensional space to provide intuitive exploration. The potential of radiomics-based methods will be discussed, which may provide references for the future study of lung disease subtyping.

2 Materials and Methods

2.1 Datasets

	CT-RIDER	COVID19-20	SPGC-COVID19	SPGC-CAP	SPGC-normal
Image format	NIfTI	NIfTI	DICOM	DICOM	DICOM
Slice thickness	1.25mm	2mm or 5mm	2mm	2mm	2mm
Number of scans	59	199	171	60	76
Time of acquisition	N/A	N/A	Feb-Apr 2020	Apr 2018-Dec 2019	Jan 2019-May2020
Lesion type	Non-small cell lung cancer	COVID-19	COVID-19	CAP	healthy
Lesion masks	N	Y	N	N	N
Slice-level labels	N/A	N/A	55	25	N/A

2.1.1 CT-RIDER

The Reference Image Database to Evaluate Therapy Response (RIDER) Lung CT collection is a dataset designed to study the variability of non–small cell lung cancer [25]. It includes 59 Chest CT scans of 31 patients with non–small cell lung cancer. The size of reconstructed images is 512 × 512 pixels, and the slice thickness is 1.25mm. Each patient underwent two chest CT scan within 15 minutes using the same imaging protocol. There are 28 pairs of scan and rescan, and three unpaired scans. We used an in-house method to enhance the visual quality of the original data, and the enhanced data were called "enhanced" data. This dataset is provided in NIfTI format and used to assess the stability of radiomics features with respect to different CT scanner.

2.1.2 COVID19-20

This dataset is provided by The Multi-national NIH Consortium for CT AI in COVID-19 via the NCI TCIA public website, which is designed to evaluate emerging methods for the segmentation and quantification of lung lesions caused by SARS-CoV-2 infection from CT images [26]. It includes unenhanced chest CTs from 199 patients with positive RT-PCR for SARS-CoV-2 and ground truth annotations of COVID-19 lesions

in the lung. The scans are from a variety of sources with different imaging protocol and the slice thickness is 2mm or 5mm. The matrix size of the reconstructed images is 512 × 512.

2.1.3 SPGC-COVID

The originally objective of Signal Processing Grand Challenge COVID (SPGC-COVID) dataset is to develop fully automated frameworks to identify/classify COVID-19 infections using lung CT scans [27]. This dataset includes three subsets of volumetric chest CT scans of 171 COVID-19, 60 CAP and 76 normal patients acquired with various imaging protocols from different medical institutes. This dataset is provided in Digital Imaging and Communications in Medicine (DICOM) format. The original data is provided in the Hounsfield Unit and the reconstructed image size is 512 × 512. The slice thickness of all scans in this dataset is 2mm. Slice-level infectious labels are provided for a subset of 55 COVID-19, and 25 CAP cases.

2.1.4 Data Pre-processing

To reduce the impact from different exposure dose, slice thickness, and the range of values in the Hounsfield Unit, we clip the intensity values to $[-1024, +1024]$ and rescale slice thickness to from 2mm to 5mm using average downsampling interpolation.

2.1.5 Training and Test Dataset

For the task of pairing scan and re-scan together in CT-RIDER dataset, we use 56 paired scans to evaluate which features are more stable for intra-scanner and inter-scanner variability. For the task of COVID-19 classification, we randomly select 45, 45, 22, 22 scans from CT-RIDER, COVID19-20, SPGC-COVID (COVID-19) and SPGC-COVID (normal) respectively as the training set, and 14, 14, 7, 7 as the test set. For the task of tumour and non-tumour classification, we randomly select 48, 8, 8, 16, 16 scans

from CT-RIDER, COVID19-20, SPGC-COVID (COVID-19), SPGC-COVID (normal) and SPGC-COVID (CAP) respectively as the training set, and 11, 2, 2, 3, 4 as the test set. For the task of CAP classification, we randomly select 16, 8, 8, 16, 48 scans from CT-RIDER, COVID19-20, SPGC-COVID (COVID-19), SPGC-COVID (normal) and SPGC-COVID (CAP) respectively as the training set, and 4, 2, 2, 3, 11 as the test set. By these settings of training and test composition, we ensure the datasets are balances between lesion and non-lesion classes and balanced for heterogeneous sources as well.

2.2 Lung Segmentation

Radiomics studies require a ROI, usually lesion segmentation, as one of its input, which needs numerous manual annotation work from radiologists. In order to avoid manual labelling, we use two lungs, individual lung or individual lobe as the ROI to extract radiomics features. In this experiment, "lungmask" [28], [29] is used for the segmentation process, which is an automatic open-source python tool. It takes a lung scan as its input and outputs a NIfTI file with labels from 1 to 5 for each lobe. We can further get the mask for two lungs, individual lung and individual lobe from the original output. The backbone of the segmentation method is a vanilla U-Net model, which is trained on two datasets respectively. The lung segmentation output from the model trained with "R231" dataset does not include the trachea and only distinguish the left or right lung. The model trained on a subset of the "LTRC" dataset segment individual lobe but yields limited performance with dense pathologies scans. Two models are combined together by fusing their output results, which means each incorrectly identified negative pixel from LTRC model will be assigned a label the same as its neighbour in model "R231" output, and the incorrectly identified positives from "LTRC" model are removed.

2.3 Feature Extraction

For the extraction of radiomics features, the open-source Python package Pyradiomics was used [7]. It takes 2D or 3D images along with their binary lung segmentation as its inputs. We evaluated two main categories of radiomics features, first-order features and texture-based features, extracted by Pyradiomics. Unlike tumour malignancy problems, the shape of the ROI in our study is relatively independent of the research objective. We rely on the image biomarker standardisation initiative (IBSI) to set the parameters of Pyradiomics.

2.3.1 First Order Features

First order features are sensitive to image rescaling and slice thickness and do not include the information of spatial relation between voxels across a given image. Besides, these features are highly affected by the number of histogram bins or bin width. Inappropriate choices of bin width will cause the histograms to not represent the true underlying distribution. One recent study on Positron emission tomography CT (PET-CT) by Leijenaar *et al.* [30] shows that fixed bin width has more stable reproducibility. By setting a fixed bin width, we assume the grey level values in every image have the same meanings. We chose the bin width (e.g., 50 for CT scans) so that the number of bins was between 30 and 130, which shows good reproducibility and performance in the study with fixed bin count [31].

2.3.2 Texture-based Features

Texture-based features are extracted based on five descriptive matrices mentioned above in section 1.1.3, each of which describes different relationships between image voxels:

- GLCM: each element represents the frequency of occurrences of each pair of intensity values appearing together in the image within distance δ (one voxel

distance by default) along angle θ ($0°, 45°, 90°, 135°$) [32]. This matrix is first calculated for each angle separately and then aggregated via a mean operation.

- GLDM: this matrix describes gray level dependencies in an image. Two neighbour voxels are dependent if the absolute difference between them is less than a threshold α (0 by default). For the (i, j) element, it is the number of gray-level i with j dependent voxels in its neighbourhood [33].
- GLRLM: the (i, j) element of this matrix is the number of appearances of adjacent pixels with intensity value i and length j on θ direction [34], [35]. The number of rows of this matrix depend on the number of discrete gray levels of the input.
- GLSZM: the (i, j) element of this matrix is the number of connected zones with gray level i and size j that appear in an image [36]. A connected zone is a set of voxels connected within distance 1 with infinity norm. The shape of the matrix depends on the intensity values instead of the shape of the image.
- NGTDM: this group of features is based on the visual characteristics of the image. For each gray-level value, the sum of the difference between this value and the value of its neighbour voxels are computed [37].

2.4 Feature Reduction Methods

High-dimensional data leads to computational and analytical complexities in many fields such as computer vision and machine learning. In order to extract useful information from huge amounts of data, feature reduction has become a key step for pre-processing high-dimensional data for machine learning tasks. There are two ways to reduce the number of features. On the one hand, the feature selection method aims to reduce the dimensionality of the data by identifying a subset of the relevant features in the dataset and then directly apply the subset of the relevant features to the learning task. They can retain the meanings of the original features and produce results with lower dimensionality that are more interpretable. On the other hand, standard

dimensionality reduction techniques, such as auto-encoder and independent components analysis, try to find a linear or non-linear projection that maps the high-dimensional data into a lower-dimensional feature space. These methods employ all features to obtain an optimal representation of data, and the resulting features do not necessarily have any clinical meanings and are usually hard to interpret [38]. Since the intrinsic dimensionality of data is often much lower than the original feature space, a simplified feature set usually improves the accuracy and interpretability of the model. It can also reduce the requirement of computational resources and avoid model overfitting to achieve better generalisation ability.

We aim to extract features that are able to not only predict the correct COVID-19 label for each scan but also pair scans/rescans together for the CT-RIDER dataset. The extracted features should be insensitive to slice thickness and image rescaling. Since various clinical settings lead to different intensity values for CT scans, the extracted features should be robust to heterogeneous datasets.

2.4.1 Fisher Score

Fisher score [39] is a filter-based features selection method, which ranks the features based on their significance according to some specific criterion as a pre-processing step prior to the learning task, and selects a subset of features with the highest ranking scores. The filter-based approaches assess correlations between features and the class label and select features that are efficient for discrimination. A high score is assigned to the feature with a large distance between data points for different classes and a small distance between data points for the same class. Since finding the global optimal solution is non-deterministic and polynomial-time (NP) hard, the feature scores are usually computed independently for each feature according to the pre-defined criterion and then ranked from high to low [40]. However, this approach can lead to suboptimal solutions, because it neglects the combination of features and

is not able to delete highly correlated features. Since filter-based approaches do not depend on learning models, they have much less computational complexity compared to the wrapper and embedded methods [38].

2.4.2 Attention-based Feature Selection

Attention-based Feature Selection (AFS) [41] is a deep-learning-based feature selection method. AFS consists of two modules, an attention module which is responsible for computing the weights for all features, and a learning module that models the correlation between the weighted features and labels. The attention mechanism focuses on the most relevant information, instead of using all available information. The learning module in AFS can be replaced by other existing models. The two modules are detachable and can be trained separately. In order to find the correlation between features and labels, we need to study whether a feature should be selected for a specific label. Then, feature weights are calculated based on the distribution of representative feature selection patterns.

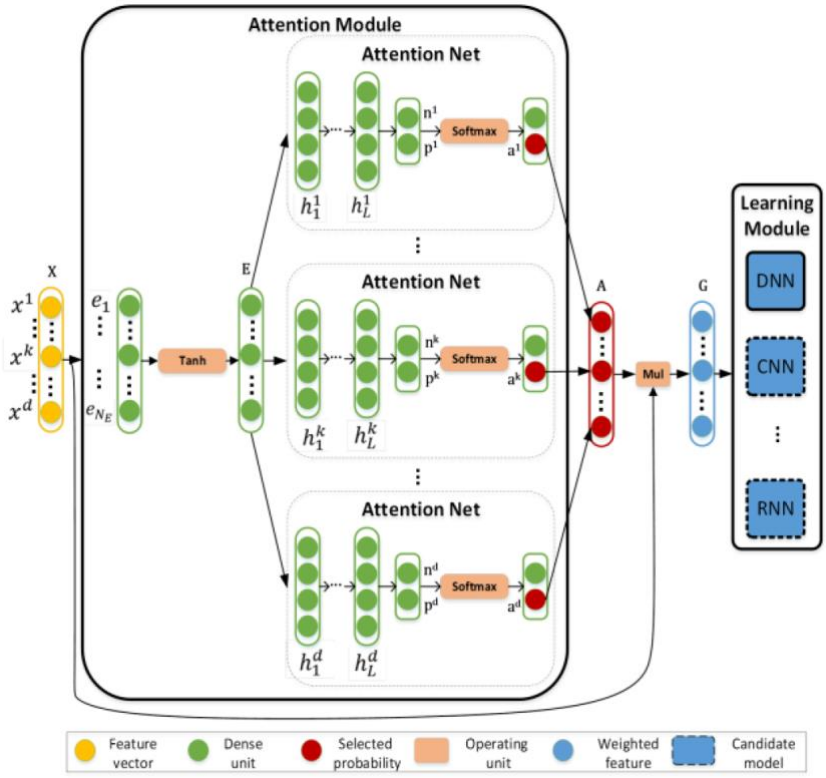

Figure 2.4.2. Architecture of AFS model. [41]

As shown in figure 2.4.2, there are two main processes in the attention module. Firstly, a dense network is employed to compress the original features space into a smaller subspace. It keeps a major part of the relevant information, while the impact from individual noises and outliers are reduced and certain redundant features are discarded. A non-linear activation function $\tanh(\cdot)$ is followed to rescale compressed features into $[-1,+1]$. Secondly, the attention net, which is a shallow neural network with L hidden layers, takes the compressed features as its input and outputs the probability of selection for each feature. A $\mathrm{softmax}$ function is used to ensure the sum of probabilities is equal to 1. The weighted feature can be obtained by multiplying the original features and corresponding weights.

2.4.3 Auto-encoder and Variational Auto-encoder

Auto-encoder (AE) is a type of neural network that can be used to learn a compressed feature representation. An AE is composed of two connected sub-models, an encoder and a decoder [42]. The encoder is a neural network that is trained to encode the raw input data into a lower-dimensional feature space. The decoder is another neural network that reconstructs the output from the encoded representation as close to the original data as possible. The encoded representation must retain enough information for the decoder to be converted back to the original data. Since the dimension of the encoded features is smaller than the original dimension, the encoder needs to learn to preserve as much of the relevant information as possible in the limited units while discarding the irrelevant information. The entire AE is trained together though backpropagation [43]. There are four hyper-parameters that need to be set before training: (1) number of nodes in the middle layer, which decides the dimension of the encoded representation, (2) number of layers, which decides the depth of the model, (3) number of nodes in each layer, (4) loss function, which measures the reconstruction loss between the input and the output data. After training, the decoder

is discarded, and the encoder can be used as a feature reduction model. AEs are considered as an unsupervised learning technique since explicit labels are not needed to train the model.

Although the compressed features from AE can represent most of the information, the latent feature space may not be continuous. It is trained to encode and decode the features with as small loss as possible, regardless of the organisation of latent space. The data points that are close to each other on the original feature space are not necessarily close to each other on the latent space. A variational auto-encoder (VAE) introduce a regularisation of the latent space [44]. A single data point on the original feature space is encoded as a distribution $\mathcal{N}(\mu, \sigma^2)$ of latent variable over the latent space. Then we decode a data point that is sampled from the distribution to obtain a reconstruction. The model is still optimised by minimising the loss function via backpropagation. The loss function is composed of reconstruction error and regularisation of the latent variable distribution, namely, the Kullback-Leibler (KL) divergence [45].

The detailed architectures of AE and VAE models used in our experiment are shown in figure 2.4.3. There is a rectified linear activation function added after each hidden layer and a sigmoid activation function after the last layer before outputting the results. The dimension of the latent feature space is chosen to be 16, which is the number of nodes in the middle layer. The number of hidden layers is set to 3, and the number of nodes in the interim layer is set to 32. The loss function for the AE model is:

$$L = -\frac{1}{N}\sum_{i=1}^{N}(x_i \log \hat{x_i} + (1 - x_i)\log(1 - \hat{x_i}))$$

where x_i is the original value and $\hat{x_i}$ is the reconstructed value. The loss function for the VAE model is:

$$L + \frac{1}{2}\sum_{i=1}^{N}(\exp \sigma_i + \mu_i^2 - 1 - \sigma_i)$$

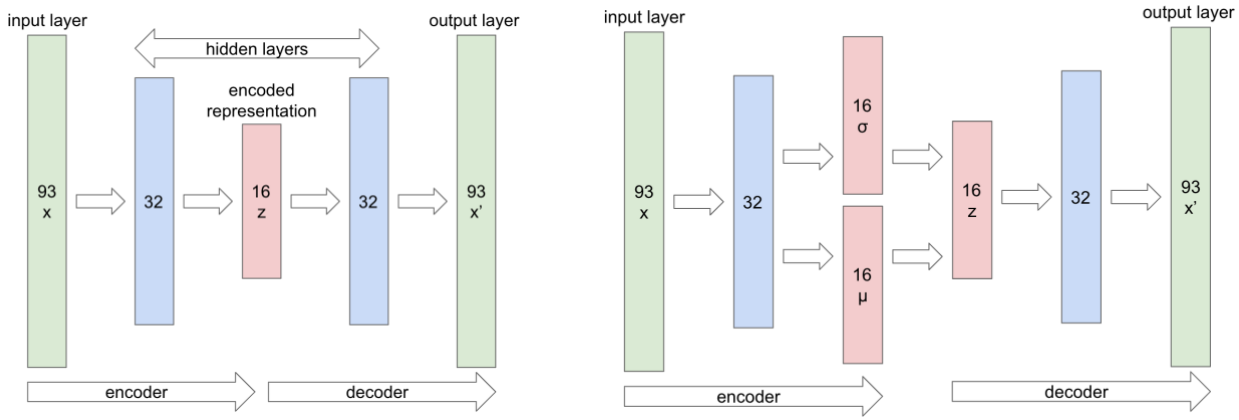

Figure 2.4.3. Architecture of AE (left) and VAE (right).

2.5 Classification Models - Random Forest

Classification refers to a predictive modelling problem where a class label is predicted for a given example of input data. A classification model uses the training dataset and tries to find the best mapping from input data to the corresponding class labels. The decision tree is a popular classification model, which attempt to partition the input space into a set of rectangles and fit a simple model in each one [46]. A recursive binary tree is interpretable, but when the input data is too complicated the decision tree will be deep and have the risk of overfitting. Random forest is an improvement for bootstrap aggregation of decision trees and consists of a number of decision trees where each tree is created from a different bootstrap sample of the training dataset [47]. A bootstrap sample is a sample of the same size as the original dataset sampled with replacement. Each individual tree outputs a class prediction and the class with the majority votes is the final prediction of the random forest model. Random forest reduces the variance of the decision tree to prevent overfitting.

2.6 Blending of Backbones

The use of blending of backbones is inspired by the idea of inter-heterogeneity and intra-heterogeneity of lungs from Veeraraghavan *et al.* [48]. We extract features from each lobe to obtain the intra-site lobe heterogeneity, and then combine the

features by assigning different weights to each lobe to obtain inter-site lobe heterogeneity.

2.6.1 Universal Representation Transformer layer

The Universal Representation Transformer (URT) layer is inspired by Transformer networks [49] and uses a dot-product self-attention mechanism to weight the pre-trained backbones from different domains [50]. The backbones can be the four feature reduction methods mentioned in section 2.4. The URT layer uses meta-learning episodic training to learn how to combine the domain-specific backbones of a universal representation for classification tasks. This model supports a variety of strategies on how to retrieve backbones. If the learning task matches perfectly with the learning task from one of the domains, then the corresponding backbone will be retrieved from the set of backbones. On the other hand, if the learning task can benefit from more than one domain, then a better strategy might be to attempt some cross-domain generalization by blending many backbones together, even if none of the individual domain matches the learning task perfectly.

As shown in figure 2.6, universal representation $r(S)$ can be obtained by concatenating all features from each backbone. For each class, the class representation $r(S_c)$ can be calculated by averaging the universal representation of all data points with the corresponding class from the support set. The domain-class representation $r_i(S_c)$ can be calculated similarly in each domain. A scaled dot-product attention module is adopted to get the attention score of each backbone for each class respectively. In the scaled dot-product attention module, we first obtain a learnable query $q_c = W^q r(S_c) + b_q$ from a linear transformation of each class representation. Then we define a learnable key $k_{i,c} = W^k r_i(S_c) + b_k$ for each domain/class. The domain-class attention score $\alpha_{i,c}$ can then be calculated from the dot-product of q_c and $k_{i,c}$ followed by a softmax function. Then the weight for each backbone can be obtained by aggregating $\alpha_{i,c}$ via a mean

operation. The URT layer is trained by minimising the probability of a label y for a data point x given the support set of a task, and a regularisation term on the model parameters to avoid duplication of the attention scores.

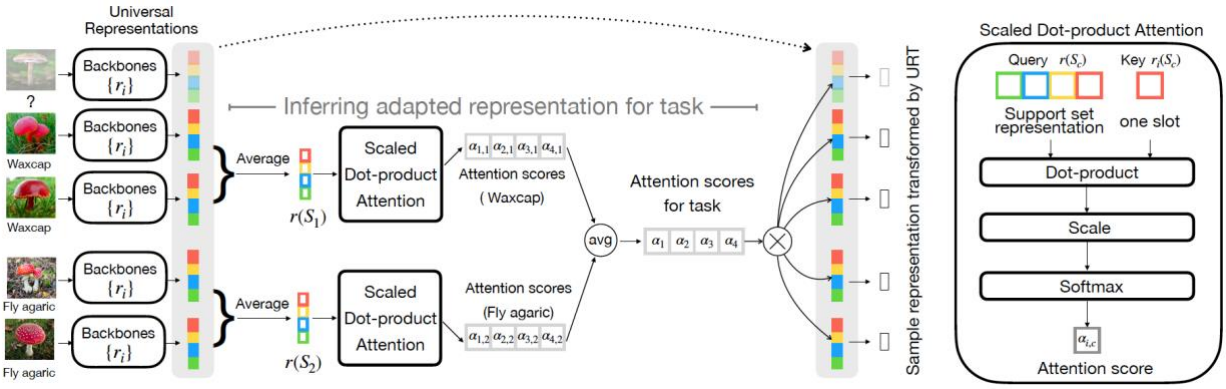

Figure 2.6. Architecture of single-head URT layer. [48]

2.7 t-SNE Projection

t-Distributed Stochastic Neighbour Embedding (t-SNE) [51] is a non-linear dimensionality reduction technique for visualising high-dimensional data. The algorithm aims to find a low-dimensional representation (2 or 3 dimensions) for the high-dimensional data. We used t-SNE to project the radiomics features in 2D to get intuitions about the data structure. It first calculates a similarity matrix for datapoints in the high-dimensional space and a similarity matrix for corresponding datapoints in the low-dimensional space, and then minimises the KL divergence between them. For a high-dimensional dataset X with N datapoints, it calculates a conditional probability matrix $p_{j|i}$ of shape $N \times N$:

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2/(2\sigma_i^2)\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2/(2\sigma_i^2)\right)}$$

where σ_i is a Gaussian distribution centred on x_i and $p_{i|i} = 0$. Each element in the matrix depends on the Euclidean distance between two datapoints and represents the similarity between them. Similarly, another probability matrix $q_{j|i}$ is defined for the low-dimensional representation:

$$q_{i,j} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq i} (1 + \|y_i - y_k\|^2)^{-1}}$$

where $q_{i|i} = 0$. Instead of using a Gaussian distribution, a Student t-distribution with one degree of freedom is employed. Student t-distribution has heavier tails than Gaussian distribution which assign non-negligible weights to points that are far apart. Then we minimise KL divergence between $p_{j|i}$ and $q_{j|i}$:

$$\sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{i,j}}$$

with respect to the low-dimensional dataset. The output is the low-dimensional representation after optimisation.

2.8 Evaluation metrics

Classification predictive modelling algorithms are evaluated based on their results. In this report, the evaluation metrics include accuracy, sensitivity, specificity and F1 score. All four scores are calculated based on the combinations of the true label and the predictive label, which refer to true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The formulae are described as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$sensitivity = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

$$F1\ score = \frac{2TP}{2TP + FP + FN}$$

Accuracy is the proportion of correctly predicted labels. Sensitivity evaluates the ratio of correctly identified positives, and specificity evaluates the percentage of correctly identified negatives.

The graphical pipeline of the whole process is shown in figure 2.10. It takes a CT scan as its input and the output is the predicted COVID-19 classification of the scan. Firstly, lung regions are segmented by "lungmask". Then imaging features are extracted from five lobes using Pyradiomics with the CT scan and its segmentation. The five extracted features are encoded from 93-dimension to 16-dimension by AE or VAE respectively. The URT layer is adapted to generate a weighted feature, which is used for the classification task by the random forest.

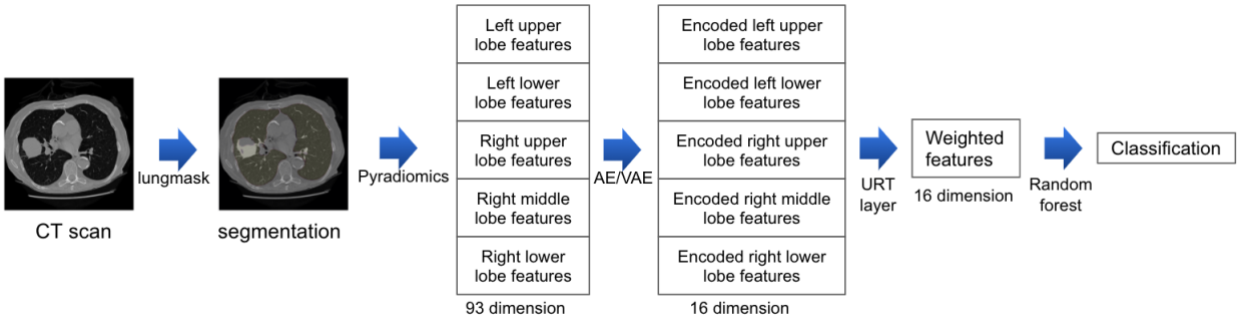

Figure 2.10. Graphical pipeline of the whole process

3 Results

3.1 Datasets

We compared the lobe volumes across each lobe between the two scan versions. Results are shown in Table 3.1. The p-value was calculated for paired t-test with a two-sided hypothesis. There was no significant difference in left lower lobe volumes ($p = 0.0962$). The average volume of left upper lobe and right upper lobe increased after the enhancement, while the mean volume of right middle lobe and right lower lobe on generated scans reduced comparing to original scans.

	Original (L)	Enhanced (L)	P value
Left upper lobe	1.1647±0.4064	1.1782±0.4077	0.0120
Left lower lobe	0.9921±0.3364	0.9834±0.3459	0.0962
Right upper lobe	0.8711±0.2801	0.8944±0.2602	0.0005
Right middle lobe	0.4017±0.1594	0.3943±0.1592	<0.0001
Right lower lobe	1.0394±0.3760	1.0268±0.3686	0.0490

Table 3.1. Lobe volume on original and enhanced scans. (n=59)

Then we measured the difference in lobe volumes of each individual between two repeated scans (Figure 3.1.1). The variance of right middle lobe volume was the smallest, which is natural because the volume of this lobe is also the smallest. For both types of scans, left upper lobes had the largest volumes, while variances of differences between right lower lobe volumes and left lower lobe volumes in pairs of scans are larger than the other three lobes.

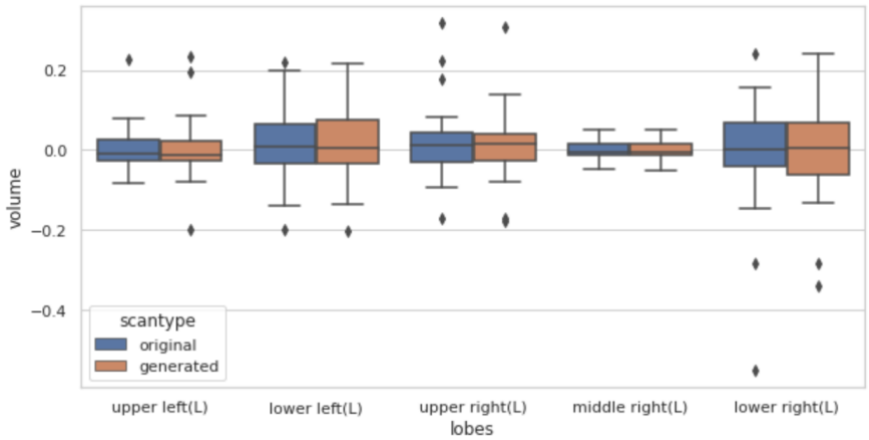

Figure 3.1.1. Difference of lobe volume between repeated scans of each individual.

Dice similarity coefficient (DSC) was used to measure the overlap between each pair of scan and rescan (Figure 3.1.2). Before calculating the coefficient, lung registration should be performed so that the difference caused by global positioning and respiratory motion was minimised and each pair of scans were aligned into the same coordinate system. Non-rigid registration was deployed which comprised an affine transform as initialization followed by B-spline transformation [52]. The dice coefficient on the lungs was larger than that on any single lobe. In terms of scan type, the overlaps of lungs and lobes on enhanced data were smaller than on original data. The occurrence of extreme values (DSC = 0) due to the lobe not being detected in some scans.

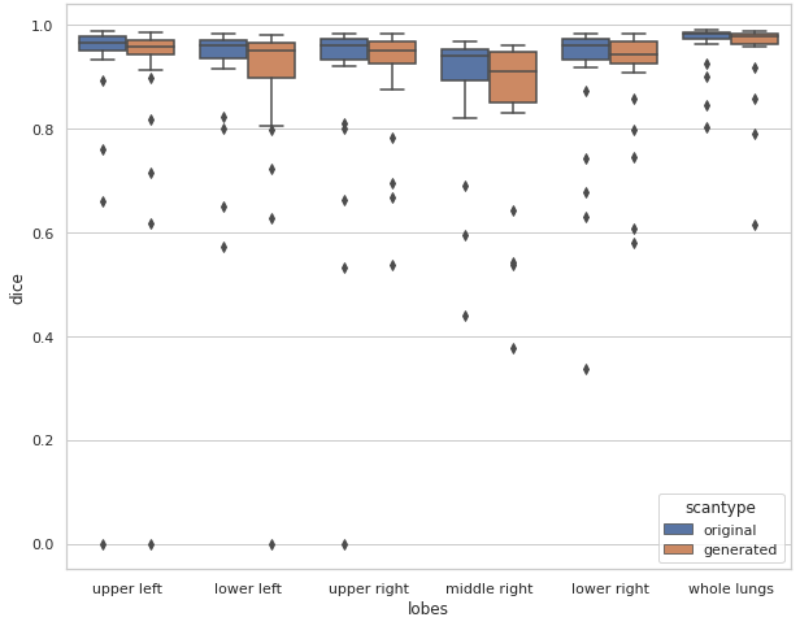

Figure 3.1.2. Dice coefficient of each lobe between repeated scans of each individual.

3.2 Feature Extraction

To get an insight into the feature space after the extraction, we generated a heatmap to show the extracted feature values of all 565 scans over 93 features with two lungs mask for feature extraction. The extracted features have a large range of values from 10^{-2} to 10^{12}. To be able to compare different features with each other, we normalised the feature values using z-score normalisation and figure 3.2.1 shows the resulting heatmap. The difference in the patterns between healthy scans and scans with

tumours and COVID-19 can be observed, but the difference between scans with COVID-19 and CAP is hard to distinguish.

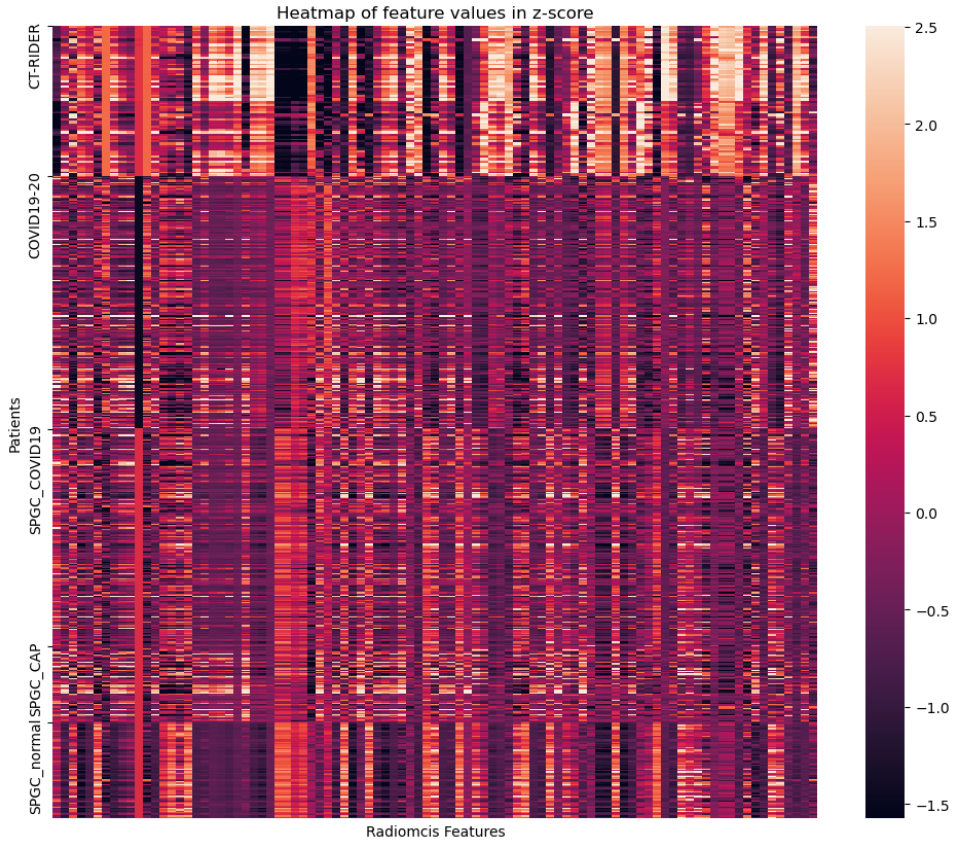

Figure 3.2.1. Heatmap of z-score normalised feature values for each patient.

In CT-RIDER dataset, our aim is to extract features that are able to pair scans and rescans together. Metric for pairing was chosen to be the L2 distance on t-SNE projections of features in 2D. According to the matching performance mentioned above in section 3.1, the original scans were chosen for all following experiments. First, we used Pyradiomics to extract features from ROIs of the two lungs together. Feature normalisation was required since the range of values were different. We normalised the data so that every feature had zero mean with standard deviation of 1 across all samples. A typical value of perplexity in t-SNE is between 5 and 50, and it was set to 20 in our experiment. It is related to the number of nearest neighbours used and controls the trade-off between attention for global and local aspects. To confirm how many iterations are needed for t-SNE, we ran it multiple times with a different maximum number of iterations. Figure 3.2.2 shows the results of t-SNE projection of 28 pairs of repeated

original scans using normalised features with 500, 1000, 2000 and 3000 iterations. Since early stops usually happen before 3000 iterations, we concluded from the plots that 3000 iterations were enough for our task.

To evaluate the stability of the nearest neighbour for each scan, we encoded the nearest neighbour on the t-SNE plot into a heatmap. Since t-SNE has random components, we ran it 100 times and took the average. Figure 3.2.3 shows the stability of the nearest neighbour on the t-SNE projection before and after normalisation using the Euclidean distance metric on the original feature space. The (i, j) point in the figure encodes the probability of image j to be the nearest neighbour of image i, hence the sum of the value on each row should always be one. When using non-normalised features as the input of t-SNE, the average probability of pairing the data (i.e. the nearest neighbour of a scan being its re-scan) is 0.422. After calibration, the probability increased significantly to 0.816. Therefore, we decided to always work on normalised features, and extracted features refer to normalised features afterwards.

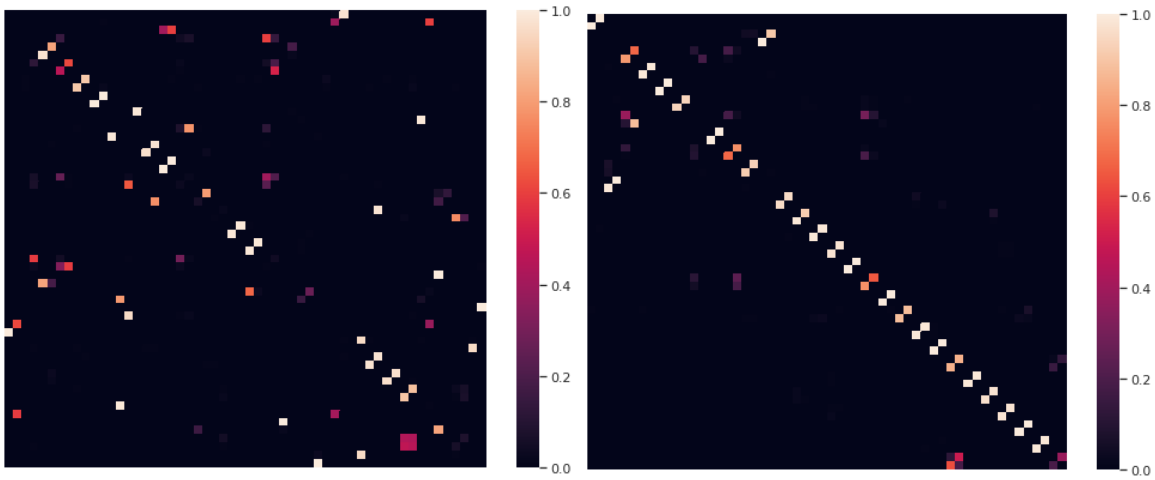

Figure 3.2.3. Probability of being the nearest neighbour on the t-SNE plot using non-normalised features (left) and normalised features (right).

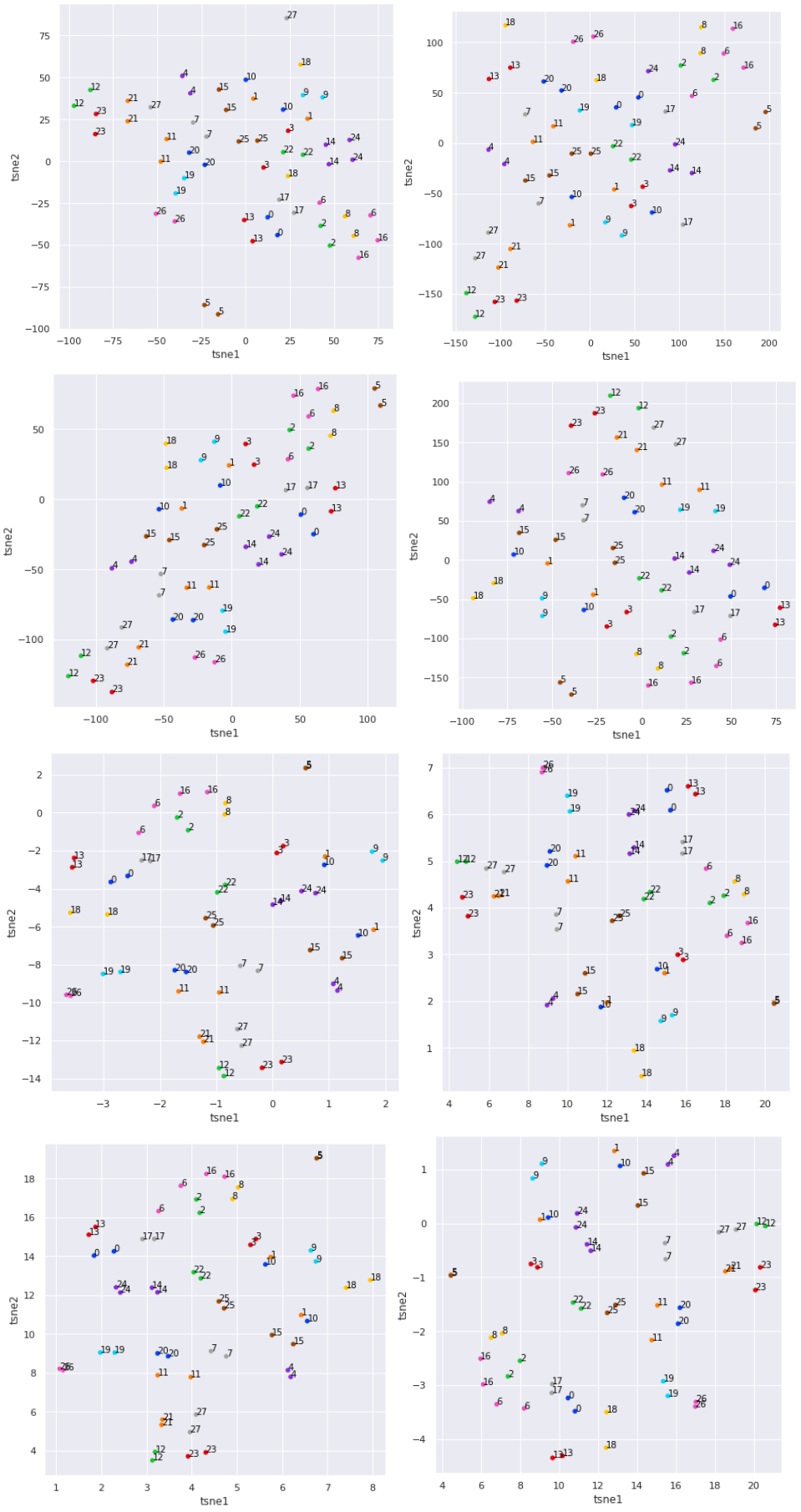

Figure 3.2.2. examples of the t-SNE plot of original scans with 500, 1000,2000 and 3000 iterations.

We also explore different normalisation methods to calibrate the feature values, including z-score, min-max and mean normalisation.

- Z-score normalisation is defined as $z_{ij} = \frac{x_{ij}-\mu_j}{\sigma_j}$ where x_{ij} is original feature value in training set/test set, z_{ij} is the normalised feature value in training set/test set, $\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$ is the mean value of the feature in training set, $\sigma_j = \sqrt{\frac{\sum_{i=1}^{N}(x_{ij}-\mu_j)^2}{N}}$ is the standard deviation of the feature in training set.
- Min-max normalisation is defined as $z_{ij} = \frac{x_{ij}-\min_i(x_{ij})}{\max_i(x_{ij})-\min_i(x_{ij})}$ where $\min_i(x_{ij})$ is the minimum value of the feature in training set, $\max_i(x_{ij})$ is the maximum value of the feature in training set.
- Mean normalisation is defined as $z_{ij} = \frac{x_{ij}-\mu_j}{\max_i(x_{ij})-\min_i(x_{ij})}$.

Figure 3.2.4 shows the classification results using three normalisation methods mentioned above. From the confusion matrices, four evaluation metric scores can be calculated for each method. Min-max normalisation has the highest accuracy (0.890), specificity (0.906) and F1 score (0.888), and z-score normalisation has the highest sensitivity (0.882). The COVID-19 is highly contagious and hence sensitivity is a more important metric than others when evaluating the results. Therefore, we chose z-score normalisation in our experiments, in spite of the accuracy of z-score normalisation (0.888) is slightly lower than that of min-max normalisation. Besides, the z-score normalisation is more robust to unseen datasets, because the impact of outliers is lower than that of min-max normalisation.

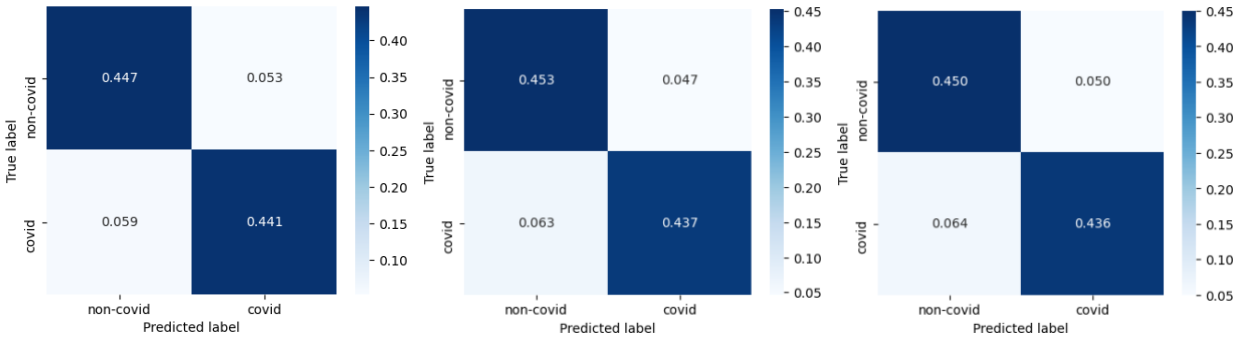

Figure 3.2.4. Confusion matrix for COVID-19 classification using z-score (left), min-max (middle) and mean (right) normalisation.

We also assessed the results of applying alternative distance metrics for t-SNE, such as Manhattan and cosine metrics in contrast to the traditionally Euclidean distance metric. Figure 3.2.5 shows the comparison between Euclidean, Manhattan and cosine distance metrics. The average probabilities of correct matching are 0.816, 0.832 and 0.755 respectively. The usage of Manhattan metric slightly raises the probability of successful pairing. Aggarwal [53] suggested that for a high dimensionality problem, a low value of p in Minkowski Distance is most preferable. This means that L1 (Manhattan) metric is preferred over Euclidean (L2) metric as the dimension of the data increases. Cosine metric is usually used when the magnitude of the data does not matter.

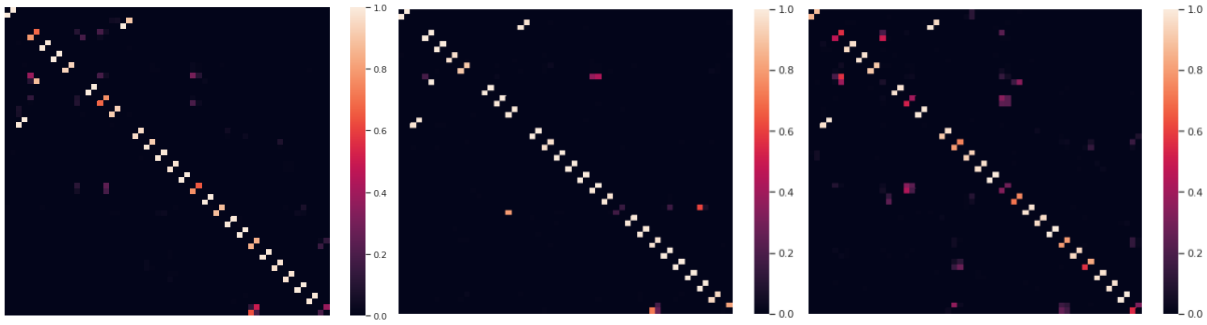

Figure 3.2.5. Heatmap of probabilities of being the nearest neighbour on t-SNE plot using Euclidean (left), Manhattan (middle) and cosine (right) distance metrics.

Furthermore, we explored the feature difference between scan/rescan across five lobes. We used lobe masks to extract features, and then compared these five groups of features. For each patient with scans 1 and 2, the difference of feature f was calculated as $|f_1 - f_2|$ to avoid negative values. The feature difference for each lobe is shown in Figure A.3.2.6 We concluded from the plots that there was no significant difference in

the stability of extracted features between lobes. In terms of feature groups, GLCM features have slightly larger variances across lobes.

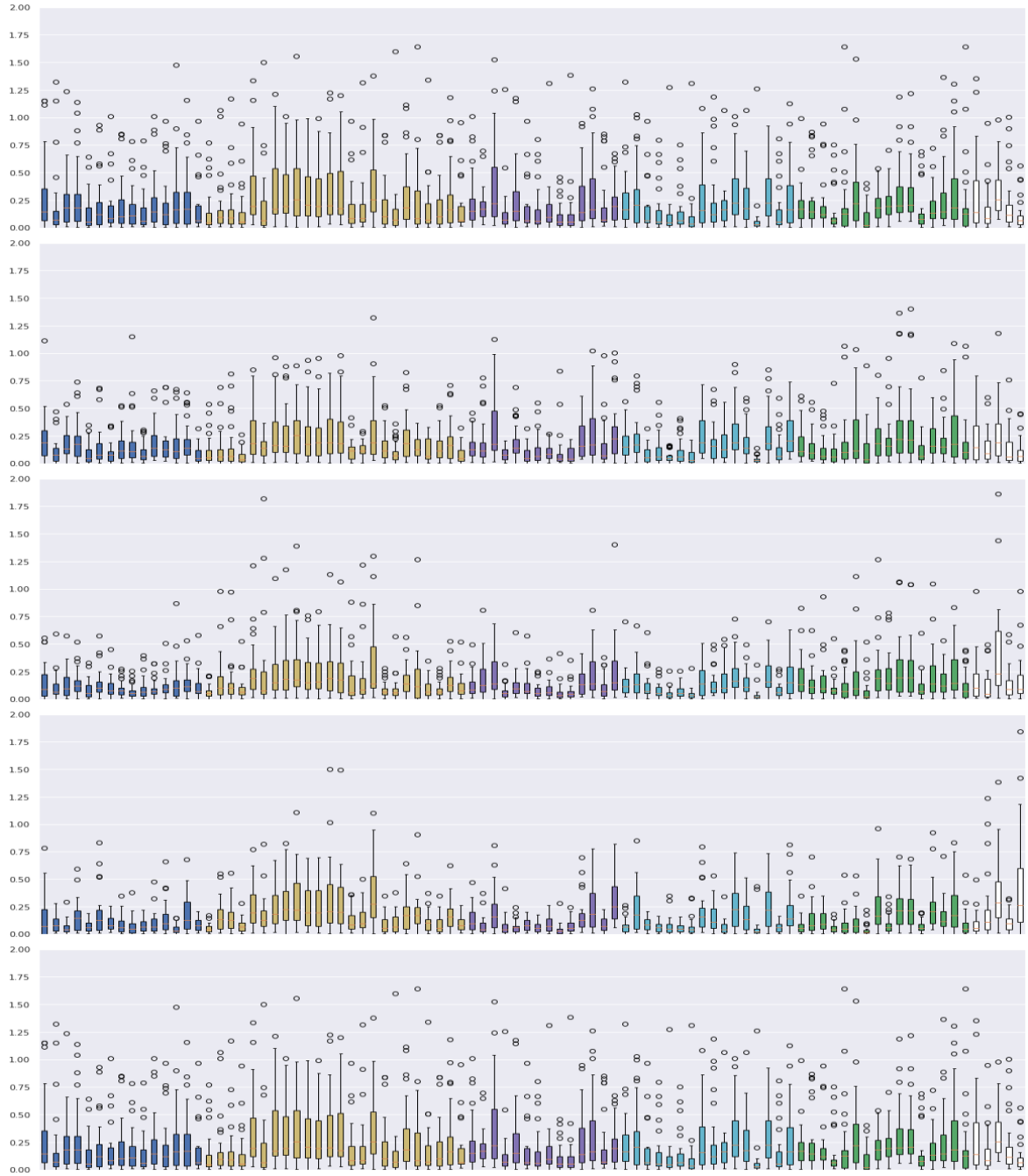

Figure A.3.2.6. Box plot of feature differences of left upper, left lower, right upper, right middle and right lower lobe respectively. Each group of features corresponds to a colour on the boxplot. (blue=first order, yellow=GLCM, magenta=GLDM, cyan=GLRLM, green=GLSZM, white=NGTDM)

Besides, we evaluated the feature difference between lobes. We calculated the root-mean-square error (RMSE) of all features between each pair of lobes and show them as a heatmap (Figure 3.2.7). All scans and re-scans were taken in the average. A higher score on the heatmap means a greater difference between the pair of lobes. This heatmap suggests that there is a large difference between the left lower lobe and the

right lower lobe in terms of normalised features, and a relatively small difference between the left upper lobe and the right upper lobe. These differences imply the inter-site lobe heterogeneity.

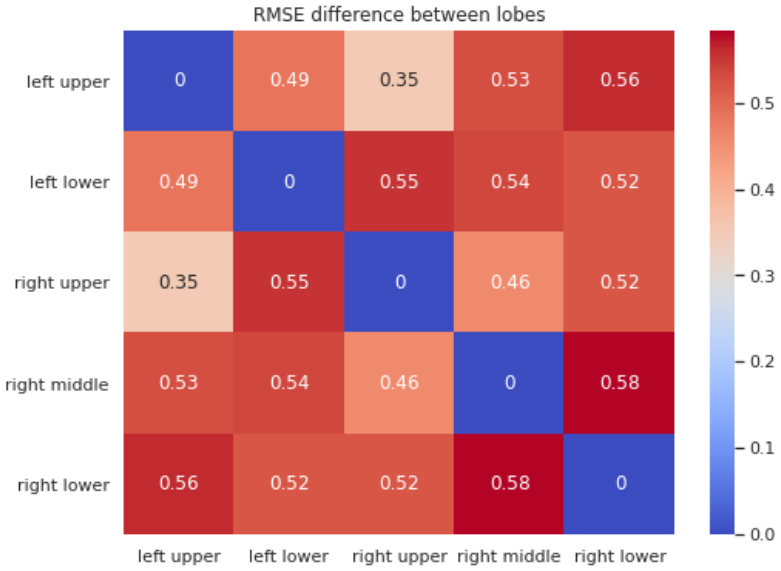

Figure 3.2.7. Heatmap of feature difference between lobes.

To explore the feature correlations, the Pearson correlation coefficients are calculated for each pair of features and showed as a heatmap in Figure 3.2.8. The values should be in the range of $[-1, +1]$ and 1 on the diagonal. A negative value represents a negative correlation of the two features and a positive value means that the two features are positively correlated. 1 and –1 imply a perfectly linear relationship between two features.

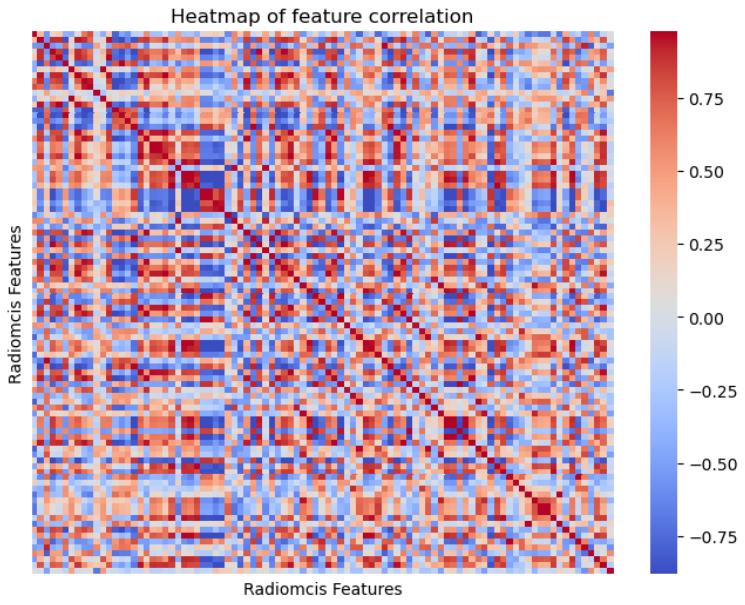

Figure 3.2.8. Feature correlation

3.3 Feature Selection

To evaluate the feature selection methods, we report the COVID-19 classification results as a confusion matrix, from which the accuracy, sensitivity, specificity and F1 score can be calculated. Figure 3.3 shows the classification result using all 93 extracted features with a random forest as the classifier. Since random forests consist of random components, the final results are averaged over 100 independent experiments. The accuracy, sensitivity, specificity and F1 score are 0.888, 0.882, 0.894 and 0.887 respectively. This result is set to be the baseline of our research.

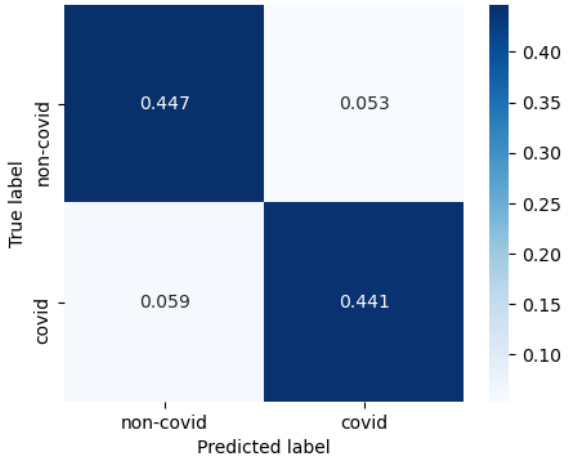

Figure 3.3. Confusion matrix of classification results using all 93 Pyradiomics features.

3.3.1 Fisher Score

For the Fisher score feature selection method [54], the input data to the model is normalised extracted features by Pyradiomics. The dimension of the input is 93 and the output of the model is one score for each feature. Then the input features are ranked from high to low according to the calculated fisher score. We select the first K (K<93) features and report the results in classification accuracy by a random forest model. Figure 3.3.1.1 shows the COVID-19 classification 10-fold accuracy, sensitivity and specificity respectively with a different number of selected features. Figure 3.3.1.2 shows the 4-class accuracy with tumour, COVID-19, CAP and healthy as the classes. The performance of the classification task has the trend to increase with the number of selected features overall, but the results improve much more slowly after around K=20.

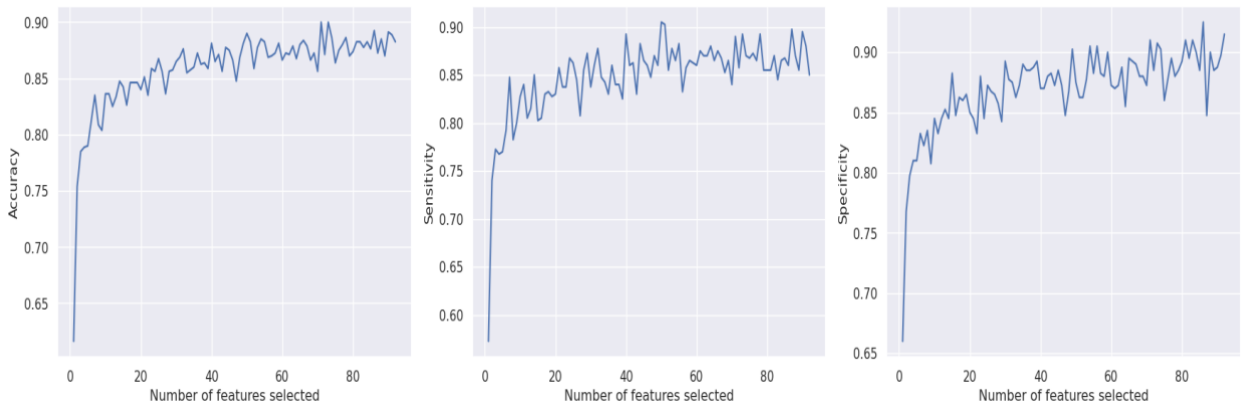

Figure 3.3.1.1. COVID-19 classification accuracy, sensitivity and specificity using Fisher score selected features.

Figure 3.3.1.2. Fisher score 4-class accuracy.

The model performance is also reported in a confusion matrix of classification using 16 features selected by Fisher score (Figure 3.3.1.3). The accuracy, sensitivity, specificity and F1 score are 0.846, 0.846, 0.846 and 0.846 respectively. The performance in all evaluation metrics is inferior to baseline, which means 16 features selected by Fisher score cannot capture all useful information for classification.

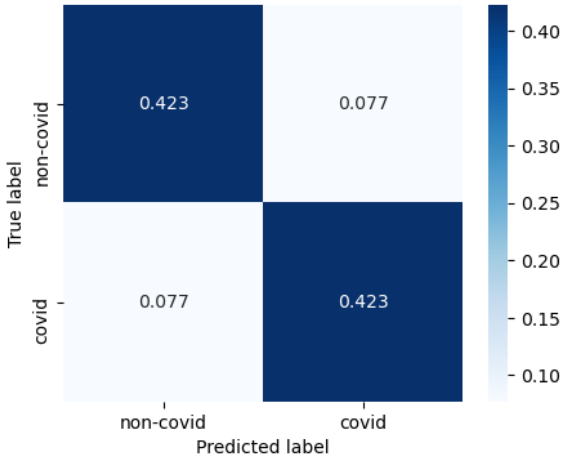

Figure 3.3.1.3 Confusion matrix of classification results using 16 features selected by Fisher score.

3.3.2 AFS

The code used for AFS is adapted from [55]. The input of the model is normalised extracted features by Pyradiomics. The output of the AFS model is the probabilities to be selected for each feature. The features are sorted according to these values and the top K features are selected. The feature selection performance is reported in COVID-19 classification accuracy by a neural network model (Figure 3.3.2). The accuracy increases dramatically as the number of selected features increasing from 1 to 4. After K=40, the accuracy shows a slight decrease, which proves the necessity for the feature selection.

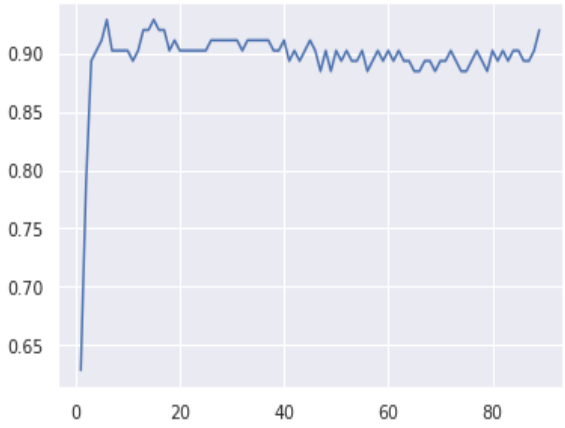

Figure 3.3.2. AFS COVID-19 classification accuracy.

3.3.3 AE and VAE

The input to both AE and VAE models is normalised extracted features of 93 dimensions. The output of the AE model is the encoded features with pre-defined dimension K. The performance of AE is described by COVID-19 classification results in four evaluation metrics with respect to the number of dimensions of latent feature space (figures 3.3.3.1). The model performance improves rapidly as the latent dimension increasing from 1 to 10 and remains a similar value after K=10.

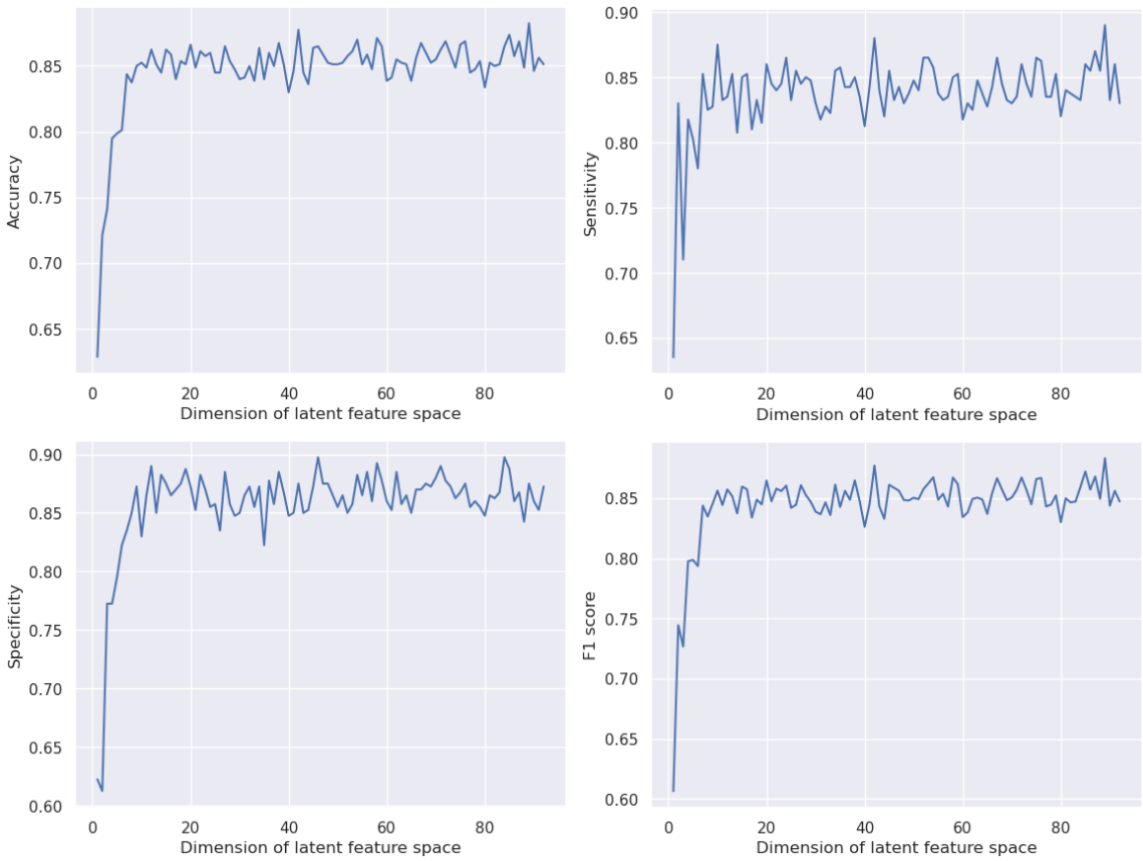

Figure 3.3.3.1. COVID-19 classification accuracy (top left), sensitivity (top right), specificity (bottom left) and f1 score (bottom right) using AE encoded features.

Figure 3.3.3.2 shows the confusion matrix of classification results using 16-dimensional AE encoded features. The accuracy, sensitivity, specificity and F1 score are 0.842, 0.828, 0.856 and 0.840 respectively for AE features. The performance of AE reconstructed features is also presented to verify the stability of the AE structure. The scores on the same four evaluation metrics are 0.857, 0.842, 0.872 and 0.855 respectively.

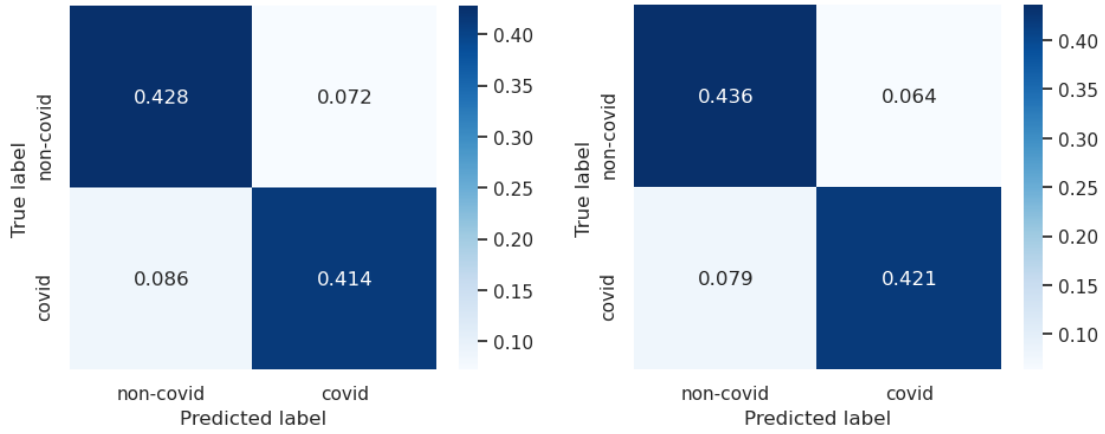

Figure 3.3.3.2. Confusion matrix of classification results using 8 AE encoded features (left), and 93 AE reconstructed features (right).

To check the stability of the 16-dimensional latent feature space of AE, we project the encoded features of 28 pairs of scans in the CT-RIDER cohort into a 2-dimensional space using t-SNE. The nearest neighbours for each scan are encoded as a probability on the heatmap averaged over 100 independent t-SNE projections. The average probability of correct pairing on the t-SNE plot is 0.161. Figure 3.3.3.3 shows an example of t-SNE projection and the nearest neighbour heatmap. For most of the scans, the nearest neighbour on the t-SNE plot is not its paired scan.

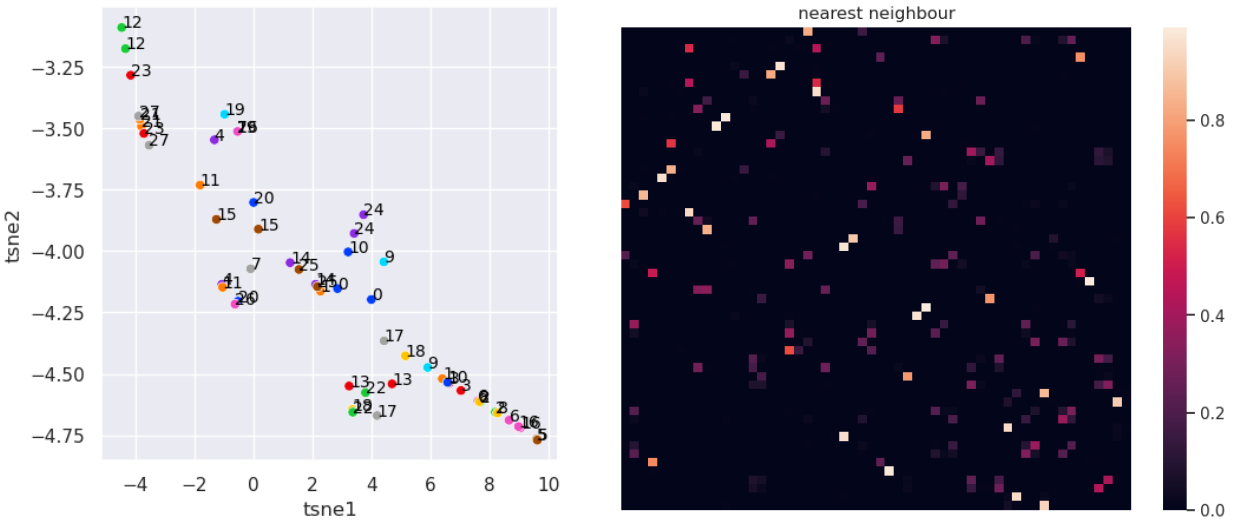

Figure 3.3.3.3. T-SNE projection of AE encoded features. (left)
Probability of being the nearest neighbour on t-SNE plot using 16 AE encoded features. (right)

Figure 3.3.3.4 shows the classification performances of VAE encoded features and VAE reconstructed features. The overall performance of VAE is better than AE. The scores on the same four evaluation metrics are 0.863, 0.856, 0.870 and 0.863 respectively. Since the VAE model reconstructs a variational version of the original input, the classification results are inferior to the reconstruction of AE. The reconstruction results confirmed the stability of the VAE model.

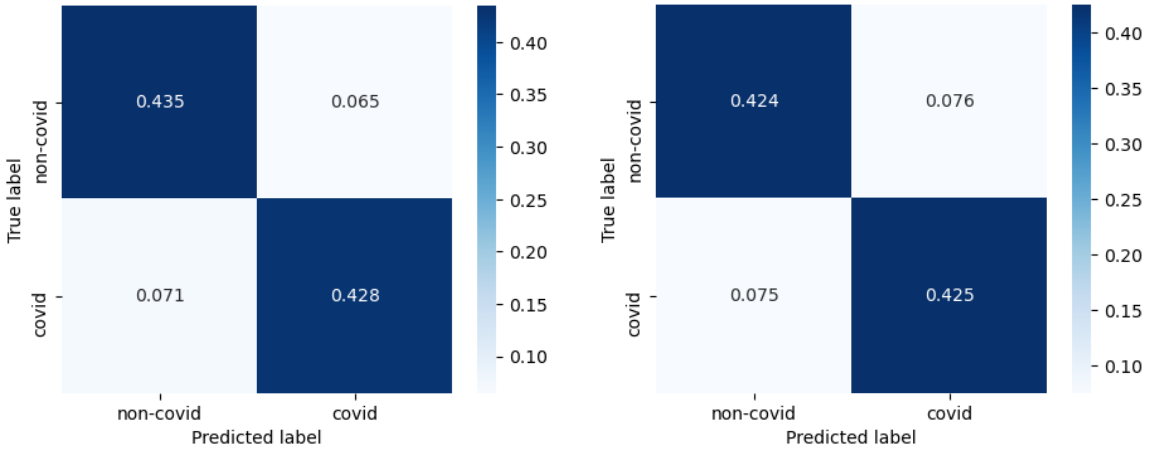

Figure 3.3.3.4. Confusion matrix of classification results using 16 VAE encoded features(left), and 93 VAE reconstructed features (right).

The VAE encoded features of 28 pairs of scans in the CT-RIDER cohort are projected into 2-dimensional space using t-SNE. Figure 3.3.3.5 shows an example of t-SNE projection and the nearest neighbour heatmap. Most pairs of scans are closed to each other on the 2D projection. The average probability of correct pairing on the t-SNE plot increased from 0.161 to 0.621 by changing the model from AE to VAE. The problem of discontinuous latent feature space can be solved by replacing AE with VAE.

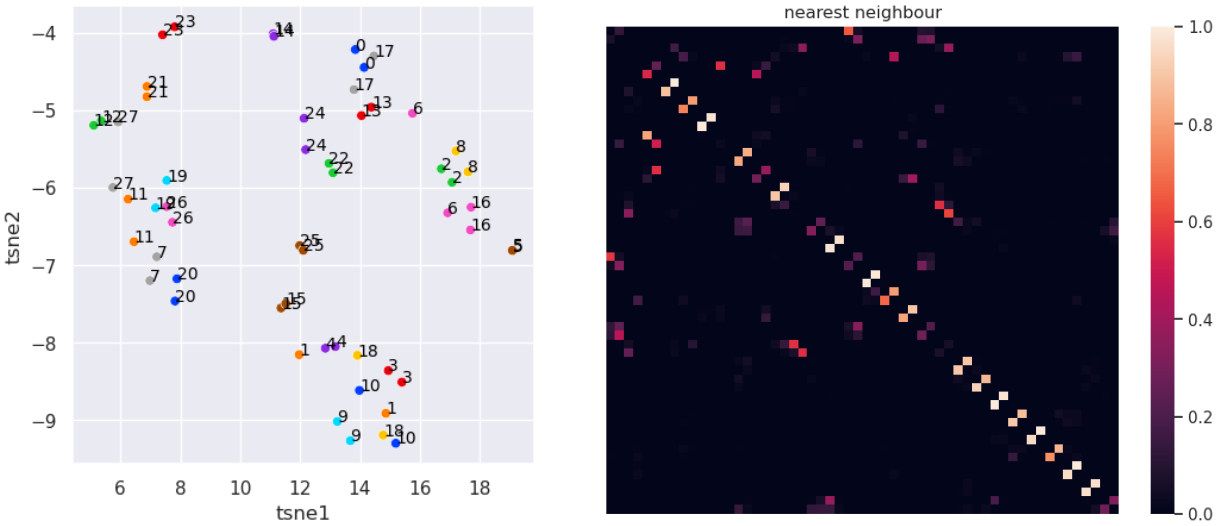

Figure3.3.3.5. T-SNE projection of VAE features. (left)
Probability of being the nearest neighbour on t-SNE plot using 16 VAE encoded features. (right)

3.3.4 URT layer

The input of the URT layer is five groups of normalised features extracted from each lobe using Pyradiomics. The backbones of the URT layer are chosen to be AE or

VAE. For each group of features, we first train a backbone to obtain encoded features of 16 dimensions to reduce the noise and the impact of outliers. Then five groups of encoded features are fed into the URT layer to get a weight for each group. The five groups of features are aggregated by the assigned weight, and the classification model takes the weighted features as its input. The classification results shown in Figure 3.3.4 imply the performance of the URT layer. The accuracy, sensitivity, specificity and F1 score are 0.917, 0.908, 0.926 and 0.916 respectively. The scores in all evaluation metrics increased by more than 0.02 compared to the baseline results. The training and validation model loss can be found in figure A.1, which verifies the stability of the URT layer.

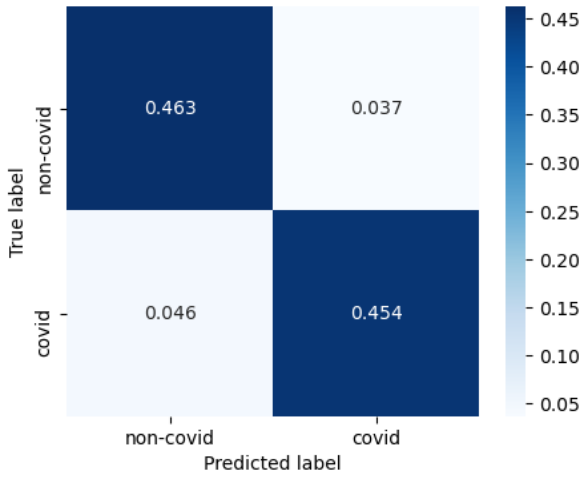

Figure 3.3.4. Confusion matrix of classification results using URT weighted AE encoded features.

We replace the backbone in the URT layer with a VAE model and remain other structure unchanged from the previous experiment. The classification results shown in Figure 3.3.4 imply the performance of the URT layer with VAE features. The accuracy, sensitivity, specificity and F1 score are 0.877, 0.888, 0.866 and 0.878 respectively. The performance decreased in all evaluation metrics comparing to the backbones of the AE model, which may be caused by overfitting.

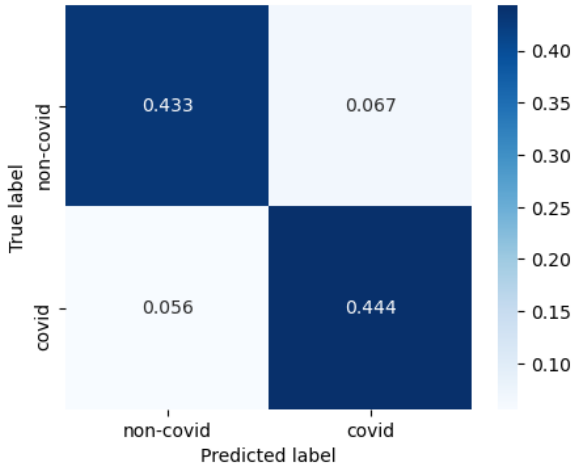

Figure 3.3.5. Confusion matrix of classification results using URT weighted VAE encoded features.

We project the training set onto 2D using t-SNE with 93-dimensional original features and 16-dimensional URT features respectively. Figure 3.3.6 shows the t-SNE projection. In the left plot, CT-RIDER, COVID19-20 and SPGC datasets are clustered into 3 classes. The two subsets of SPGC are mixed together despite one of them consist of healthy scans and the other consist of scans with COVID-19. The scans in COVID19-20 and SPGC(COVID) are closed to each other, but they still show two classes. On the right plot, the scans with COVID-19 from two datasets are mixed together, while the normal scans are more separated from the scans with COVID-19. Although the t-SNE plot does not generate three clear clusters for scans with tumour, COVID-19 or no disease, it still shows large improvements from the original features to the URT features.

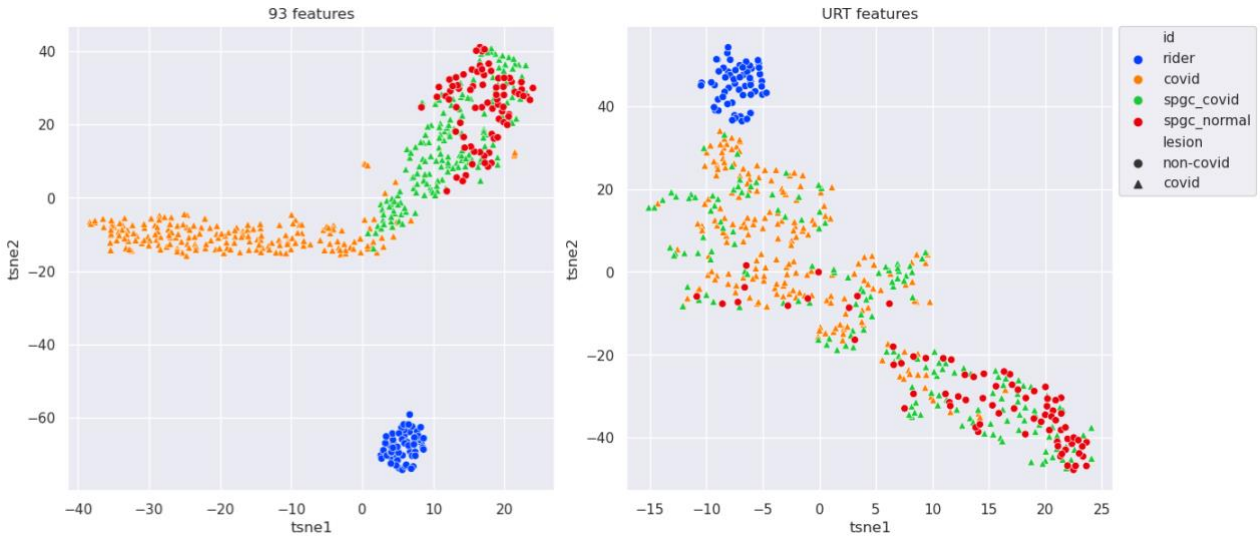

Figure 3.3.6. t-SNE plot using original features (left) and URT features (right).

4 Discussion

This work investigated feature extraction methods and feature selection methods for lung CT scans and attempted to provide a visualisation for clustering different subtype of lung diseases. Statistical analysis and classification modelling are employed to quantify the degree of stability of designed radiomics biomarkers with respect to various sources of variability. Our study demonstrated the value of whole lung radiomics and the feasibility of diagnosing COVID-19 with accuracy > 0.91 without any manual annotation.

4.1 Feature Extraction

In this work, our original plan was to score the radiological features for global lung CT scans in spite of different imaging acquisition techniques. We first attempted to segment the lungs and extract radiomics features using the original dataset, but the resulting feature values have large variabilities across three cohorts. This is because the feature matrices are largely determined by the resolution and slice thickness of the input scans. During the feature selection process, even if only one feature is selected to classify the scans, we will get a high classification accuracy. This implies that the differences in feature values are caused by technical influences rather than the intrinsic characteristics of lung disease. This technical influence can be eliminated by two methods. More datasets can be used in the study so that the unstable features, such as sensitive to imaging acquisition, will be eliminated during the feature selection process. Limited by the number of datasets, we have to employ another method, which is to rescale all the input scans to the same standard before feature extraction.

After rescaling, we explored three CT scan datasets and the extracted imaging signatures from them. From the patient/feature heatmap, we conclude that there are obvious differences in feature values for scans in different cohorts, which in line with the hypothesis that pre-defined imaging signatures can capture sufficient information

for disease prediction. However, the variation across different datasets is larger than the variation between different type of scans within a data cohort.

For the CT-RIDER cohort, the extracted features are capable of matching each pair of scans with an averaged probability of 0.816 in spite of various imaging acquisition protocols. This shows the fact that the feature extraction can retain the most essential characteristics of the CT scans.

We tried three normalisation methods for data calibration, and z-score normalisation was selected for our research because it has the highest sensitivity and also superior performances on other metrics. For highly contagious diseases like COVID-19, high sensitivity is important to help prevent the spread of the disease.

The quantitative evaluation of inter-lobe feature stability was conducted. The feature differences for each pair of scans are calculated and compared across lobes. There is no significant difference in feature stability across five lobes while the variances of some features are larger than others in all lobes, which shows that feature selection is necessary.

4.2 Feature Selection

The performance of feature selection methods was quantified through classification results and the probability of matching pairs of scans. Random forests are chosen to be the classifier of all feature selection methods, which can be easily trained and used with relatively low computation. We set the model which uses all 93 features as the baseline model. Our task is to select a small number of features to represent all the information and obtain results that are better than the baseline model.

For the Fisher score selection method, the model performance shows a trend of increase, and the increase rate decreases as the number of selected features increase. There is no significant improvement for any specific number of selected features comparing to using all features. This means the ability to identify the noises and outliers

is weak for the Fisher score feature selection method, or the dataset does not contain many noises or outliers.

For the AFS model, when 8 to 20 features are selected, the classification results are slightly better than the baseline. The accuracy decreases while choosing more than 40 features, this means features after 40 contains more noise and will disturb the results.

For the AE and the VAE models, after the dimensionality of the latent space is greater than 10, the classification results will no longer improve. The ability of VAE to match pairs of scans is much better than the AE because the latent feature space of VAE is forced to be continuous. The classification results of both the AE features and VAE features are slightly inferior to the baseline.

Finally, the encoded features from the AE and the VAE are fed into the URT layer. The 16-dimensional output features from the URT layer achieved a higher score in all performance evaluation metrics than the 93-dimensional original features. The t-SNE projection of the URT features shows more reasonable clusters. The results indicate that the backbone blending model improves the process of defining imaging biomarkers.

4.3 Limitations

There are only three sources of datasets are provided for this project, and there are obvious differences between these three datasets. CT-RIDER cohort is the single source of scans with tumours, and it only contains tumour scans. This result in a problem that the classification performance of tumour cannot be verified. The source of variability of the extracted features can be caused by both imaging protocols and the presence of a tumour. Although we have balanced the training set and test set, there could still be a probability that the results are impacted by the characteristics of datasets. We should use multiple datasets from different sources for healthy scans and each type

of disease scans, which can reduce the impact of the characteristics of the dataset on the results.

The t-SNE projection does not show clear clusters, but the classification results can be higher than 0.9. This may be due to the complex data structure and the intrinsic feature dimension is much greater than 2. Besides, the optimisation process of t-SNE needs to be run for every new input datapoint, which means it is expensive to use.

The Fisher score algorithm is an NP problem, and the use of the heuristic approach can lead to sub-optimal solutions because the feature scores are computed independently [41]. If two features are similar but both have high scores, they are both selected. If two features both have low scores, but the combination of these two features, which we do not compute, have a high score, they will not be selected.

4.4 Future Work

This research mainly focuses on feature building instead of segmentation and classification, and only uses a random forest model as the classifier. The results may improve if other lung segmentation and classification techniques are employed.

We used a pre-trained model for the lung segmentation task, which can be improved in the future. In our experiments, some lobes are not detected in some scans, and scans with severe diseases were not correctly segmented. However, lung segmentation is the basis of the whole process of radiomics, so the improvement of lung segmentation performance will lead to the improvement of the ultimate results.

The feature extraction method can be improved to be less sensitive to image scaling and slice thickness. The feature matrices are largely determined by the resolution of the input scans, and we need to rescale all scans to a uniform standard. By downsampling the high-resolution scans to match the low-resolution scans, we will lose some of the information contains in the high-resolution scans.

The random forest model that we are currently using is a supervised classification model. Some unsupervised clustering models, such as k-means and Gaussian mixture, are worthy to try since they can cluster scans with unseen diseases.

Other visualisation techniques can be used to provide the projection of encoded features. The UMAP method is worthy to try because it preserves more global structure in the data than t-SNE. This could provide an insight into the relationship between each subtype of diseases.

References

[1] H.J. Aerts, E.R. Velazquez, R.T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, and F. Hoebers, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature communications,* 5(1), pp.1-9, 2014.

[2] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G. Van Stiphout, P. Granton, C.M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H.J. Aerts, "Radiomics: extracting more information from medical images using advanced feature analysis," *European journal of cancer,* 48(4), pp.441-446, 2012.

[3] V. Kumar, Y. Gu, S. Basu, A. Berglund, S.A. Eschrich, M.B. Schabath, K. Forster, H.J. Aerts, A. Dekker, D. Fenstermacher, and D.B. Goldgof, "Radiomics: the process and the challenges," *Magnetic resonance imaging,* 30(9), pp.1234-1248, 2012.

[4] D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A.K. Jones, and L. Court, "Measuring CT scanner variability of radiomics features," *Investigative radiology,* 50(11), p.757, 2015.

[5] S.S. Yip, and H.J. Aerts, "Applications and limitations of radiomics," *Physics in Medicine & Biology,* 61(13), p.R150, 2016.

[6] R. Thawani, M. McLane, N. Beig, S. Ghose, P. Prasanna, V. Velcheti and A. Madabhushi, "Radiomics and radiogenomics in lung cancer: a review for the clinician," *Lung Cancer,* 115, pp.34-41, 2018.

[7] J.J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G. Beets-Tan, J.C. Fillion-Robin, S. Pieper, and H.J. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res*, 77(21), pp.e104-e107, 2017.

[8] V. Parekh, and M.A. Jacobs, "Radiomics: a new application from established techniques," *Expert review of precision medicine and drug development*, 1(2), pp.207-226, 2016.

[9] B. Zhao, L.P. James, C.S. Moskowitz, P. Guo, M.S. Ginsberg, R.A. Lefkowitz, Y. Qin, G.J. Riely, M.G. Kris, and L.H. Schwartz, "Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non–small cell lung cancer," *Radiology,* 252(1), pp.263-272, 2009.

[10] R.J. Gillies, P.E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology,* 278(2), pp.563-577, 2016.

[11] M. Kolossváry, M. Kellermayer, B. Merkely, and P. Maurovich-Horvat, "Cardiac computed tomography radiomics," *Journal of thoracic imaging,* 33(1), pp.26-34, 2018.

[12] G. Chassagnon, M. Vakalopoulou, E. Battistella, S. Christodoulidis, T.N. Hoang-Thi, S. Dangeard, E. Deutsch, F. Andre, E. Guillo, N. Halm, and S.E. Hajj, "AI-Driven CT-based quantification, staging and short-term outcome prediction of COVID-19 pneumonia," *arXiv preprint arXiv:2004.12852,* 2020.

[13] M. Fang, B. He, L.Li, D. Dong, X. Yang, C. Li, L. Meng, L., Zhong, H Li, H., Li, and J. Tian, "CT radiomics can help screen the coronavirus disease 2019 (COVID-19): a preliminary study," *Science China Information Sciences,* 63(7), 2020.

[14] C.D. Russell, J.E. Millar, and J.K. Baillie, "Clinical evidence does not support corticosteroid treatment for 2019-nCoV lung injury," *The Lancet,* 395(10223), pp.473-475, 2020.

[15] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, Y. Qiu, J. Wang, Y. Liu, Y. Wei, and T. Yu, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study," *The lancet,* 395(10223), pp.507-513, 2020.

[16] D. Dong, Z. Tang, S. Wang, H. Hui, L. Gong, Y. Lu, Z. Xue, H. Liao, F. Chen, F., Yang, and R. Jin, "The role of imaging in the detection and management of COVID-19: a review," *IEEE reviews in biomedical engineering*, 2020.

[17] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, and J. Liu, "Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: relationship to negative RT-PCR testing," *Radiology,* 296(2), pp.E41-E45, 2020.
[18] T. Ai, Z. Yang, H. Hou, C. Zhan, C., Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, "Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases" *Radiology,* 296(2), pp.E32-E40, 2020.
[19] Y. Li, and L. Xia, "Coronavirus disease 2019 (COVID-19): role of chest CT in diagnosis and management," *American Journal of Roentgenology*, 214(6), pp.1280-1286, 2020.
[20] J.P. Kanne, B.P. Little, J.H. Chung, B.M. Elicker, and L.H. Ketai, "Essentials for radiologists on COVID-19: an update—radiology scientific expert panel," 2020.
[21] S. Hu, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E.F. Fang, W. Menpes-Smith, J. Xia, and H. Ye, "Weakly supervised deep learning for covid-19 infection detection and classification from ct images," *IEEE Access,* 8, pp.118869-118883, 2020.
[22] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, Q. Chen, S. Huang, M. Yang, X. Yang, and S. Hu, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography," *Scientific reports,* 10(1), pp.1-11, 2020.
[23] C. Zheng, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and X. Wang, "Deep learning-based detection for COVID-19 from chest CT using weak label," *MedRxiv*, 2020.
[24] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, and K. Cao, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, 2020.
[25] B. Zhao, L.H. Schwartz, and M.G. Kris, "Data From RIDER_Lung CT," *The Cancer Imaging Archive*, 2015. doi: 10.7937/K9/TCIA.2015.U1X8A5NR
[26] P. An, S. Xu, S.A. Harmon, E.B. Turkbey, T.H. Sanford, A. Amalou, M. Kassin, N. Varble, M. Blain, V. Anderson, F. Patella, G. Carrafiello, B.T. Turkbe, B.J. Wood, "CT Images in Covid-19 [Data set]," The Cancer Imaging Archive, 2020. doi: https://doi.org/10.7937/tcia.2020.gqry-nc81
[27] 2021 IEEE ICASSP Signal Processing Grand Challenge (SPGC) "COVID-19 Radiomics", Mar, 2021 [Online] Available: http://i-sip.encs.concordia.ca/2021SPGC-COVID19/data.html
[28] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, 4(1), pp.1-13, 2020.
[29] J. Hofmanninger. "Automated lung segmentation in CT under presence of severe pathologies" github.com. https://github.com/JoHof/lungmask (accessed Mar. 13, 2021).
[30] R.T. Leijenaar, G. Nalbantov, S. Carvalho, W.J. Van Elmpt, E.G. Troost, R. Boellaard, H.J. Aerts, R.J. Gillies, and P. Lambin, "The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis," *Scientific reports,* 5(1), pp.1-10, 2015.
[31] F. Tixier, C.C. Le Rest, M. Hatt, N. Albarghach, O. Pradier, J.P. Metges, L. Corcos, and D. Visvikis, "Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer," *Journal of Nuclear Medicine,* 52(3), pp.369-378, 2011.
[32] R.M. Haralick, K. Shanmugam, and I.H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, (6), pp.610-621, 1973.
[33] C. Sun, and W.G. Wee, "Neighboring gray level dependence matrix for texture classification," *Computer vision, graphics, and image processing*, 23(3), pp.341-352, 1983.
[34] M.M. Galloway, "Texture analysis using grey level run lengths," *NASA STI/Recon Technical Report N,* 75, p.18555, 1974.

[35] A. Chu, C.M. Sehgal, and J.F. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," *Pattern Recognition Letters*, 11(6), pp.415-419, 1990.
[36] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P.Cau, N. Levy, J. Sequeira, and J. Mari, "Texture Indexes and Gray Level Size Zone Matrix. Application to Cell Nuclei Classification," *Pattern Recognition and Information Processing (PRIP)*: 140–145, 2009.
[37] M. Amadasun, and R. King, "Textural features corresponding to textural properties," *IEEE Transactions on systems, man, and Cybernetics*, 19(5), pp.1264-1274, 1989.
[38] A. Mirzaei, V. Pourahmadi, M.Soltani, and H. Sheikhzadeh, "Deep feature selection using a teacher-student network," *Neurocomputing,* 383, pp.396-408, 2020.
[39] D.G. Stork, R.O. Duda, P.E. Hart, and D. Stork, *Pattern classification*. A Wiley-Interscience Publication, 2001.
[40] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," *arXiv preprint arXiv:1202.3725*, 2012.
[41] N. Gui, D. Ge, and Z. Hu, "AFS: An attention-based mechanism for supervised feature selection," *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 3705-3713), 2019,
[42] D.P. Kingma, and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[43] D.J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *In International conference on machine learning* (pp. 1278-1286). PMLR, 2014,
[44] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence, 35*(8), pp.1798-1828, 2013.
[45] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.
[46] S. Shalev-Shwartz, and S. Ben-David, "Understanding machine learning: From theory to algorithms," *Cambridge university press*, 2014.
[47] L. Breiman, "Random forests," *Machine learning*, 45 (1), pp.5-32, 2001.
[48] H. Veeraraghavan, H.A. Vargas, A.J. Sanchez, M. Miccó, E. Mema, M. Capanu, J. Zheng, Y. Lakhman, M. Crispin-Ortuzar, E. Huang, D.A and Levine, "Computed tomography measures of inter-site tumor heterogeneity for classifying outcomes in high-grade serous ovarian carcinoma: a retrospective study," *bioRxiv, p.531046*, 2019.
[49] A. Vaswani, N. Shazeer, N., Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.
[50] L. Liu, W. Hamilton, G. Long, J. Jiang, and H. Larochelle, "A universal representation transformer layer for few-shot image classification," *arXiv preprint arXiv:2006.11702*, 2020.
[51] L. Van der Maaten, and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research, 9*(11), 2008.
[52] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "SimpleElastix: A user-friendly, multi-lingual library for medical image registration," *In Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 134-142), 2016.
[53] C.C. Aggarwal, A. Hinneburg, and D.A. Keim, "On the surprising behavior of distance metrics in high dimensional space," *In International conference on database theory* (pp. 420-434). Springer, Berlin, Heidelberg, 2001,
[54] J. Li. "scikit-feature" github.com. https://github.com/jundongl/scikit-feature (accessed Mar. 13, 2021).
[55] N. Gui. "AAAI-2019-AFS" github.com. https://github.com/upup123/AAAI-2019-AFS (accessed Mar. 13, 2021).

Appendix

A.1.Training curve

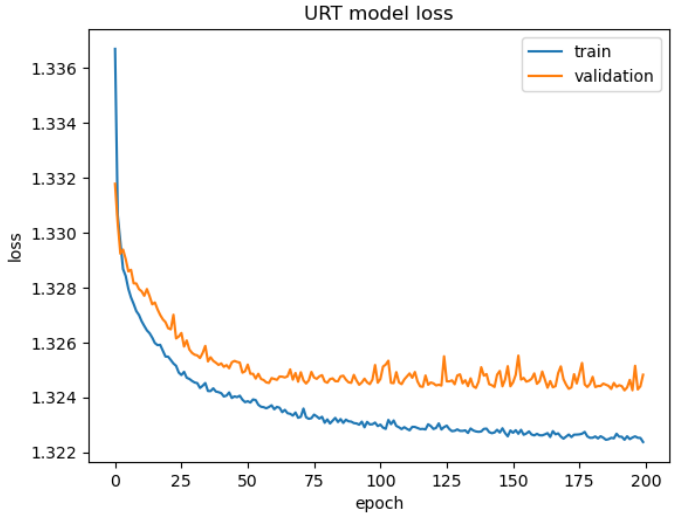

Figure A.1. Training and validation loss of URT layer.

A.2.t-SNE plot for different distance metrics/features

Figure A.2.1 t-SNE projection of 93-dimensional pyradiomics features using Euclidean distance metric.

Figure A.2.2 t-SNE projection of 93-dimensional pyradiomics features using cosine distance metric.

Figure A.2.3 t-SNE projection of 93-dimensional pyradiomics features using city block distance metric.

Figure A.2.4 t-SNE projection of 16-dimensional Fisher score selected features using Euclidean distance metric.

Figure A.2.5 t-SNE projection of 16-dimensional Fisher score selected features using cosine distance metric.

Figure A.2.6 t-SNE projection of 16-dimensional Fisher score selected features using city block distance metric.

Figure A.2.7 t-SNE projection of 16-dimensional AE encoded features using Euclidean distance metric.

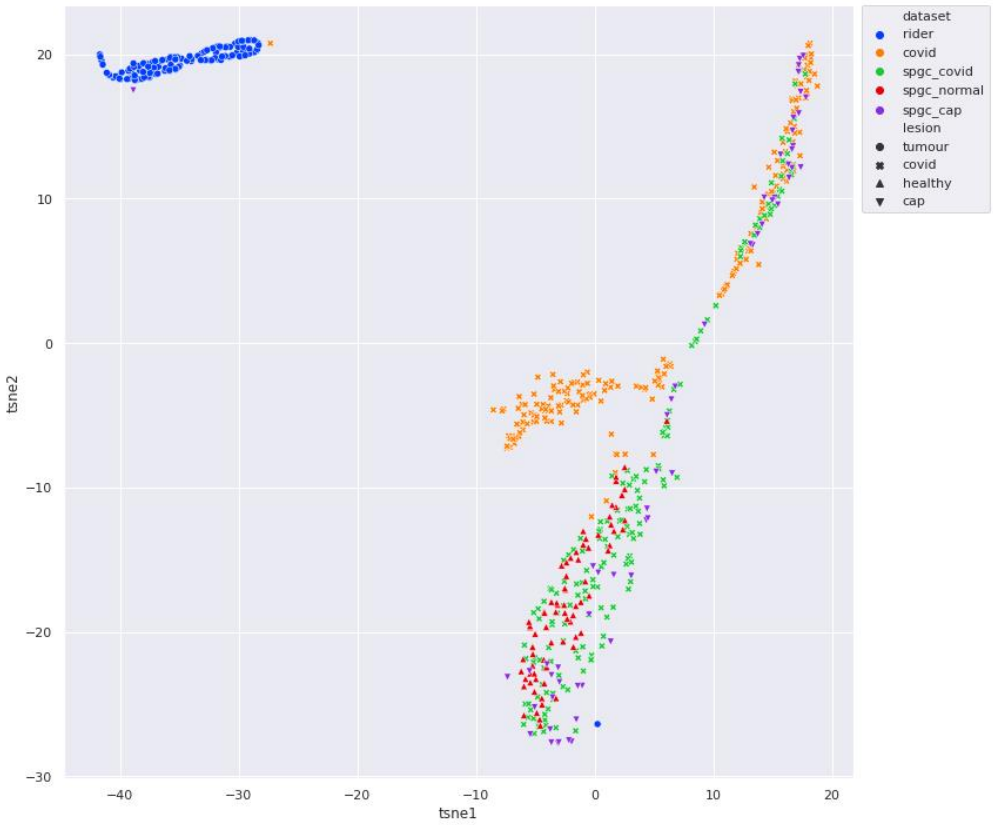

Figure A.2.8 t-SNE projection of 16-dimensional AE encoded features using cosine distance metric.

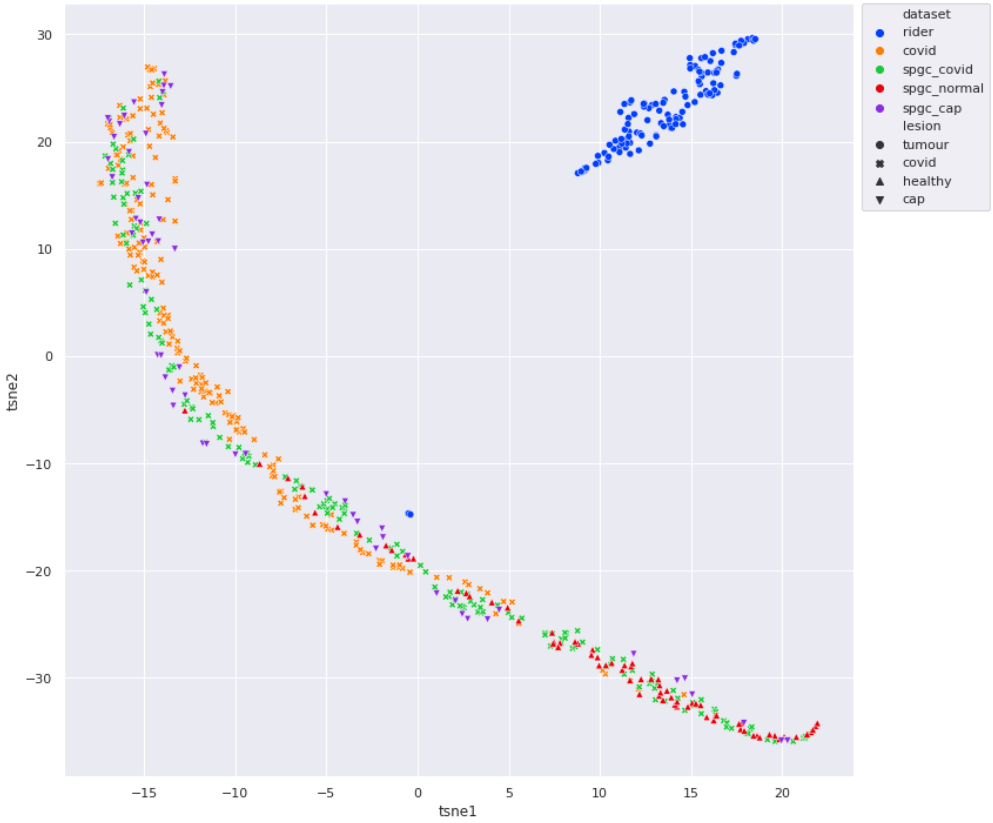

Figure A.2.9 t-SNE projection of 16-dimensional AE encoded features using city block distance metric.

Figure A.2.10 t-SNE projection of 16-dimensional VAE encoded features using Euclidean distance metric.

Figure A.2.11 t-SNE projection of 16-dimensional VAE encoded features using cosine distance metric.

Figure A.2.12 t-SNE projection of 16-dimensional VAE encoded features using city block distance metric.

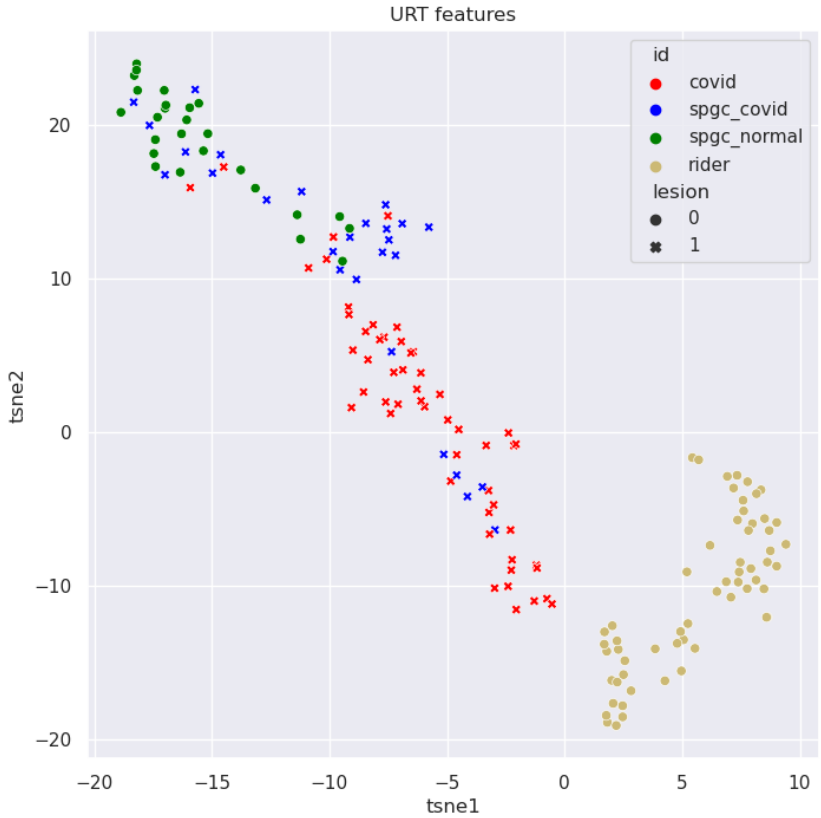

Figure A.2.13 t-SNE projection of 16-dimensional URT weighted features using Euclidean distance metric.

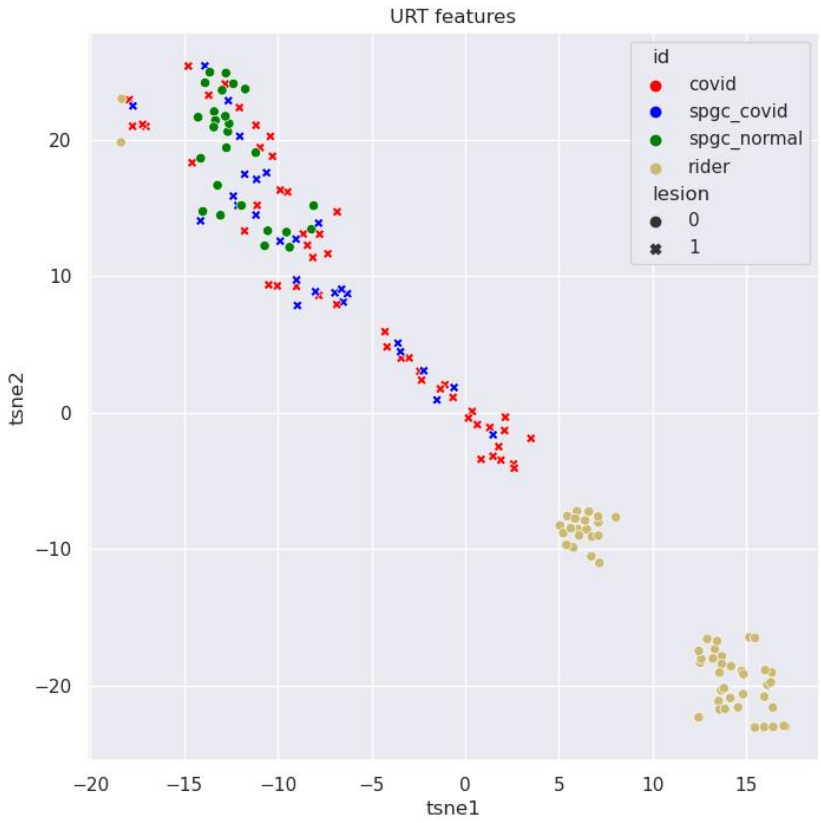

Figure A.2.14 t-SNE projection of 16-dimensional URT weighted features using cosine distance metric.