



OPEN

Fused feature signatures to probe tumour radiogenomics relationships

Tian Xia¹✉, Ashnil Kumar², Michael Fulham⁵, Dagan Feng¹, Yue Wang⁶, Eun Young Kim³, Younhyun Jung⁴ & Jinman Kim¹

Radiogenomics relationships (RRs) aims to identify statistically significant correlations between medical image features and molecular characteristics from analysing tissue samples. Previous radiogenomics studies mainly relied on a single category of image feature extraction techniques (ETs); these are (i) handcrafted ETs that encompass visual imaging characteristics, curated from knowledge of human experts and, (ii) deep ETs that quantify abstract-level imaging characteristics from large data. Prior studies therefore failed to leverage the complementary information that are accessible from fusing the ETs. In this study, we propose a fused feature signature (FF_{sig}): a selection of image features from handcrafted and deep ETs (e.g., transfer learning and fine-tuning of deep learning models). We evaluated the FF_{sig}'s ability to better represent RRs compared to individual ET approaches with two public datasets: the first dataset was used to build the FF_{sig} using 89 patients with non-small cell lung cancer (NSCLC) comprising of gene expression data and CT images of the thorax and the upper abdomen for each patient; the second NSCLC dataset comprising of 117 patients with CT images and RNA-Seq data and was used as the validation set. Our results show that our FF_{sig} encoded complementary imaging characteristics of tumours and identified more RRs with a broader range of genes that are related to important biological functions such as tumorigenesis. We suggest that the FF_{sig} has the potential to identify important RRs that may assist cancer diagnosis and treatment in the future.

Lung cancer is one of the leading causes of cancer death among men and women worldwide. Non-small cell lung cancer (NSCLC) accounts for approximately 85% of all cases of lung cancer¹. NSCLC diagnosed at an early stage has a 5-year survival rate of up to 80% for small and localised tumours (stage IA)². When compared to patients in advanced stage of NSCLC (stage IV), the 5-year survival rate is 2%².

Advances in the understanding of molecular characteristics of NSCLC have provided insights into the biology of NSCLC and assisted in more precise treatment^{3,4}. The usual approach for molecular characterisation is with large-scale gene expression profiling, a technique that determines the process by which information from a gene is converted into a functional gene product, such as proteins. Gene expression analysis at different levels of transcription can provide a global picture of different biological functions and can be identified using computational and statistical methods⁵. Gene expression analysis provided insights that facilitated the development of therapies that target specific biological pathways such as epidermal growth factor receptor (EGFR) in NSCLC that have improved clinical outcomes^{6,7}. Cetuximab⁸ is an example of target therapy medications that downregulate the EGFR. Specific types of EGFR mutations, such as the exon 19 deletions and the L858R point mutation are particularly responsive to gefitinib⁹ and erlotinib¹⁰. These medications are small-molecule tyrosine kinase inhibitors (TKIs) that restrict EGFR from transmitting cellular signals that are related to tumour progression¹¹.

Gene expression profiling, however, requires adequate tumour tissue samples that are obtained from core biopsies that sample only a part of the tumour and is invasive and expensive. In contrast, medical imaging is a non-invasive technique that plays a crucial role in routine clinical practice by capturing important imaging visual

¹School of Computer Science, Faculty of Engineering, The University of Sydney, Sydney, NSW 2006, Australia. ²School of Biomedical Engineering, Faculty of Engineering, The University of Sydney, Sydney, NSW 2006, Australia. ³Department of Radiology, Gil Medical Center, Gachon University College of Medicine, Incheon, Republic of Korea. ⁴School of Computing, Gachon University, Seongnam, Republic of Korea. ⁵Department of Molecular Imaging, Royal Prince Alfred Hospital, Camperdown, NSW 2050, Australia. ⁶Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA. ✉email: tian.xia@sydney.edu.au

characteristics. These characteristics are known as image features and can describe tumour's size, sphericity and location¹². Computerised medical image analysis has enabled the high-throughput and quantitative extraction of image features that can capture imaging features that are not quantifiable by visual assessment alone. Previous works showed that image features derived from tumours can predict tumour prognosis and treatment responses for NSCLC^{13,14}. These findings contributed to 'radiogenomics', a growing research field where the aim is to investigate the relationships between medical imaging features and molecular characteristics. Radiogenomics presents opportunities for the non-invasive assessment of important molecular characteristics that contribute to tumour development. Radiogenomics relationships (RRs) can be determined by identifying statistically significant correlations between image features and gene expressions^{15,16}. Another approach to determine RRs involves the use of functional enrichment analysis, such as gene sets enrichment analysis (GSEA)⁵, which uses statistical approaches to associate image features with functions of genetic products. These functions of genetic products are frequently described by using Gene Ontology (GO) terms, which is a formal representation that describes the biology domain with respect to three aspects: molecular functions, cellular components, and biological processes¹⁷. Published studies showed that RRs may predict the mutation status of key genetic biomarkers in NSCLC such as EGFR and KRAS^{18,19}. These biomarkers have been shown to have important implications for the treatment of NSCLC²⁰.

There have been attempts at using various image feature extraction techniques (ETs) to determine RRs, for example, between computed tomography (CT) images and tumour prognosis¹⁵. These radiogenomic studies typically employ a category of ETs that are based on the statistical analysis and medical knowledge of human experts. These ETs quantify 'handcrafted (HC)' features that quantify (i) "semantic features" that describe tumour's visual characteristics, including tumour shape, size, necrosis and contextual information, such as tumour's surrounding structures and; (ii) "agnostic features" that quantify statistical information of linear relationships about the pixels of the image that human observers consider important¹⁶ (e.g., colour histograms, Haralick textures and wavelet features^{21,22}). Handcrafted ETs have been extensively used to predict mutation status²³, model cancer outcomes²⁴, and response to therapy^{13,15}. However, handcrafted ETs are restricted to the human understanding of the disease and prespecified imaging representations. Furthermore, they may not be able to quantify the complex patterns of tumour imaging characteristics and may limit the potential of determining RRs.

Advances in machine learning algorithms are now enabling data-driven approaches for quantifying medical image's visual characteristics that can complement HC approaches. Deep learning is a method of machine learning that uses techniques such as Convolutional Neural Networks (CNNs) to learn sophisticated abstract and complex imaging characteristics directly from a large volume of labelled training image data²⁵. The use of CNNs has achieved state-of-the-art performances in a number of automated medical image analysis tasks that rely on visual characteristics, including tumour classification, detection and, segmentation^{26,27}. The data-driven nature of deep learning means that they may be less susceptible to the subjectivity of a human interpreter²⁵. Deep learning gave rise to other category of ETs, termed deep ETs, that can quantify intricate complex information from large training data sets, allowing the possibility to detect subtle variations in images of different diseases. The employment of deep ETs in radiogenomics, however, requires large volumes of labelled training data²⁸. The quantity of labelled data in medical domains are limited and present a challenge for CNNs to learn comprehensively. Transfer learning (TL) is often used in these circumstances for its ability to leverage CNNs that were pre-trained on large well-labelled natural image (photography) datasets. TL allows the pre-trained CNNs to learn image's visual characteristics that encode generic visual representations from the natural image datasets²⁹; these representations can then be employed in the medical domain to extract image features from smaller medical dataset for radiogenomics. Image features that are quantified using TL-based deep ETs are termed TL features. Since TL does not learn from any medical dataset, they may not necessarily encode specific imaging characteristics of medical images and the different diseases represented within, and hence may be suboptimal for constructing radiogenomic associations. Fine-tuning (FT) is a type of TL technique that uses backpropagation to refine the TL CNN weights through further training on smaller, directly relevant, dataset. This adapts the pre-trained TL CNN to the small relevant dataset without requiring a large well-labelled dataset. Image features that are quantified using FT-based deep ETs are termed FT features. FT has been applied in a range of different medical image analysis tasks such as ultrasound anatomy identification³⁰ and lung abnormality detection³¹, and has been used in a study to distinguish the molecular subtype in breast cancer³².

Image feature ensemble algorithms offer the opportunity to leverage handcrafted and deep ETs to extract complementary visual characteristics and provide additional information for medical image analysis. Feature fusion is a common ensemble technique that integrates both categories of ETs to produce a more comprehensive image representation of the problem³³. There have been applications of feature fusion to improve in a range of medical image analysis tasks. Kooi et al.³⁴ proposed a computer-aided detection system for mammography by using handcrafted and deep ETs to quantify image features. The deep features were found to be prone to misclassifying benign abnormalities as tumours because both share similar visual characteristics. In their study, handcrafted ETs complemented deep ETs by introducing information that is more difficult for deep features to learn, such as the location and surrounding structures of tumours, thereby increasing the detection performance when compared with using a single category of ETs. Hagerty et al.³⁵ demonstrated that using both categories of ETs to quantify image features improved melanoma classification with increased area under the curve (AUC) of receiver operator characteristics (ROC). The handcrafted ETs quantified medically meaningful image features such as lesion colour distribution and atypical pigment network and were complementary to deep ETs that quantified the low-level descriptive image features. Although these ensemble methods demonstrate notable advantages, to the best of our knowledge, the ensemble feature method has yet been investigated for radiogenomics analysis.

In this study, we propose a fused feature signature (FF_{Sig}), which is a selection of image features from both HC and deep ETs to encode complementary tumour imaging visual characteristics. We hypothesise that FF_{Sig} can identify more and exclusive RRs when compared to the use of a single category of ETs.

Methods

NSCLC–Radiomics–Genomics dataset. We used the public NSCLC Radiomics–Genomics dataset³⁶ from the Harvard University, and we refer to this dataset as the ‘NRG-H’. The dataset was sourced from the Cancer Imaging Archive (TCIA)³⁷. The NRG-H is a pre-processed and de-identified dataset. The creator of the dataset has indicated that the collection and processing of the dataset were conducted according to national laws and guidelines and approved by the appropriate local trial committee at Maastricht University Medical Center (MUMC1), Maastricht, The Netherlands. The dataset comprises 89 patients (29 W, 60 M; age range 37–85 years) with histologically confirmed NSCLC with T stage (T1–T4)³⁸. A detailed dataset description is presented in Supplementary Table S2.

All patients had a CT scan of the thorax/upper abdomen. CT scan slice thickness was between 1.5 and 5 mm. Gene expression information was acquired using the Rosetta/Merck human RSTA custom Affymetrix 2.0 microarray (Affymetrix HuRSTA-2a520709). Gene expression values were normalised using the RMA algorithm⁵ in Bioconductor. Gene expression information was accessed via the Gene Expression Omnibus (GEO)³⁹. The primary tumours were delineated by an experienced medical imaging specialist (M.F., more than 20 years of experience), slice-by-slice, on trans-axial image slices using open source software (Medical imaging Interaction Toolkit (MITK); version 2016.11⁴⁰). We excluded three patients (all men) because there were lung collapses distal to a proximal tumour and the extent of the tumour could not be reliably identified. Delineations were independently validated by a second clinician (E.K., 7 years of experience). Details of the delineation validation process are described in Supplementary Material Section 1. The annotation differences between the two clinicians are shown in Table S1 in the Supplementary Materials.

NSCLC-RADIOGENOMICS dataset. The NSCLC-Radiogenomics dataset reported by Bakr et al.⁴¹ from the Stanford University is a pre-processed and de-identified dataset, and we refer to this dataset as ‘NRG-S’. The creator of the dataset has indicated that the collection and processing of the dataset were conducted under IRB approval from Stanford University and the Veterans Administration Palo Alto Health Care System. The NRG-S dataset comprises CT images and RNA-Seq data from 117 subjects (29 W, 88 M; age range 46–85 years) with histologically confirmed NSCLC with T stage (Tis, T1–T4). A detailed dataset description is presented in Supplementary Table S3.

All patients had a CT scan from the apex of the lung to the adrenal gland in supine position. CT scan thickness was between 0.625 and 3 mm. Detailed scanning parameters, such as the manufacturer attributes are specified in the DICOM headers. Total RNA was extracted from the tissue and analysed with RNA sequencing technology. Gene expression information was processed using the STAR algorithm⁴² and Cufflinks version 2.0.2⁴³. Gene expression information was accessed via the Gene Expression Omnibus (GEO)³⁹. Primary tumours were segmented using an unpublished automatic segmentation algorithm on the axial image slices for all 117 subjects. Segmentations were viewed by a thoracic radiologist (M.K.) with more than 5 years of experience and edited as necessary using ePAD. An additional thoracic radiologist (A.N.L.) reviewed and approved the final segmentations.

Experimental overview. An overview of the experimental design is outlined in Fig. 1. HC and deep ETs are used to extract HC, TL and FT feature from delineated tumour ROIs from CT image volumes. HC features are extracted from the CT image volume directly. FT features are extracted from a 2.5D representation of the CT data around the tumour centroid³¹. The extracted HC, TL and FT features are fused into a feature matrix using concatenation. The FF_{Sig} is generated by applying a multi-step feature selection procedure involving median absolute deviation (MAD), minimum redundancy maximum relevance (mRMR), and least absolute shrinkage and selection operator (LASSO) generalised linear model. RRs are determined by using Spearman rank correlation between FF_{Sig} and the averaged gene expressions. RRs between image features signatures and GO terms are determined by using GSEA. For evaluation purposes, the same multi-step feature selection procedure is applied to HC, TL and FT features. The resulting collections of image features are denoted as HC_{Sig} , TL_{Sig} and FT_{Sig} , respectively. The training of the deep ETs was performed on the NRG-H dataset; the ETs were then used to extract image features and generate FF_{Sig} . We validated the robustness and generalisability of the FF_{Sig} by applying NRG-H trained deep ETs to the validation NRG-S dataset.

Image features. *HC and deep ETs.* We employed a set of standard HC ETs that are implemented in the pyradiomics framework to quantify HC features^{15,44}. For each patient, we extracted a well-documented set of 431 HC features from CT volumes^{45,46}. These 431 HC features comprised the following: (a) first-order statistics, describing the distribution of voxel intensities; (b) shape and size that are geometric descriptors of tumoural 3D characteristics such as compactness and surface area; (c) textural or co-occurrence matrix features to illustrate the spatial distribution of the voxel intensities and, (d) first order statistics and textural features of the wavelet decompositions of the raw imaging data. The detailed description to the 431 HC features is provided in Supplementary Material Section 4.

Deep ETs used a ResNet-101 backbone that was pre-trained on ImageNet ILSVRC challenge data. ResNet-101 is a well-established CNN architecture, which introduced the concept of ‘residual blocks’, a combination of skip connections and identity mapping to learn deeper features, and is robust to accuracy degradation⁴⁷. ResNet-101 is a robust and efficient CNN architecture that have been applied in a range of different medical image analysis tasks such as brain tumour classification⁴⁸ and segmentation for kidney and space-occupying lesion area⁴⁹. In comparison with other widely used pre-trained deep models, such as GoogLeNet⁵⁰, ResNet-101 has demonstrated superior performance in natural image detection across different datasets such as PASCAL VOC 2012 and ImageNet detection^{28,50}. The pre-training of ResNet-101 on ImageNet ILSVRC challenge data (millions of

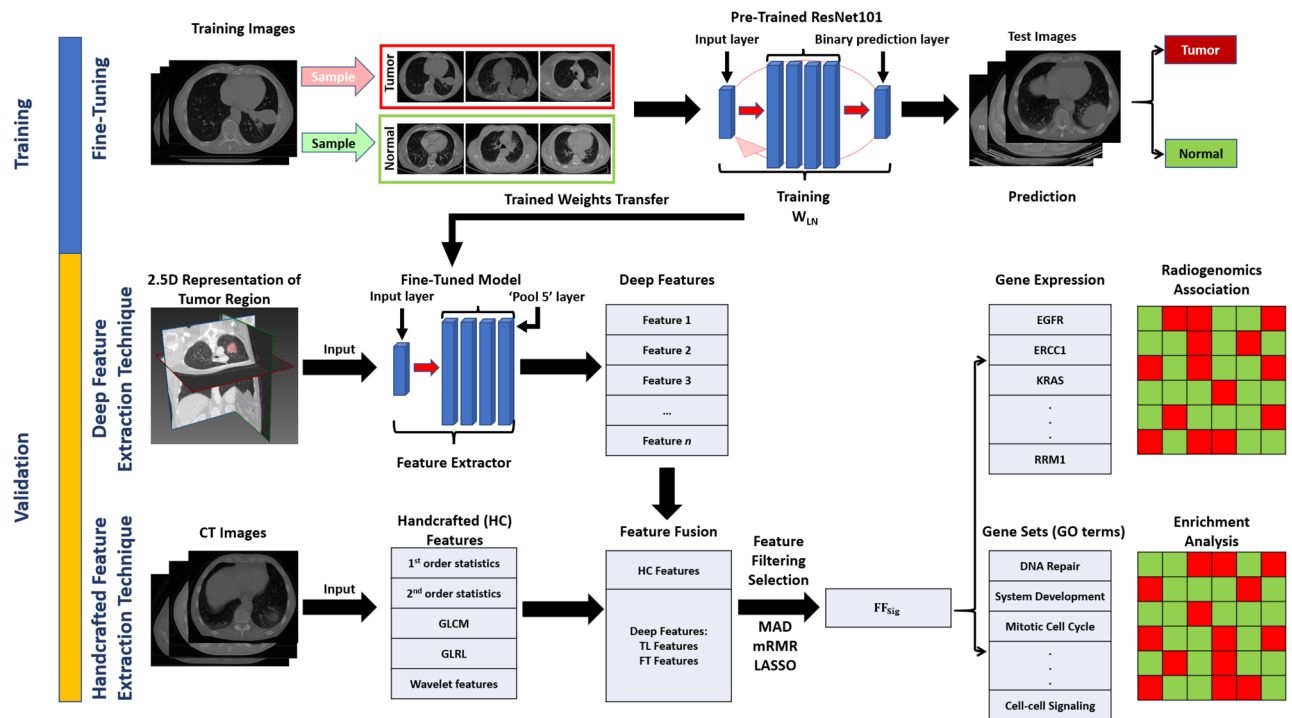


Figure 1. The workflow for generating the FF_{sig} and the identification of RRs with genes and GO terms. The workflow was implemented using the NRG-H dataset and validated on the NRG-S dataset.

well-annotated images that belongs to 1000 natural object classes) allowed it to extract deep features that represent generic image characteristics applicable to all images such as edge and texture²⁹, which has been demonstrated to be useful descriptors for medical images⁴⁷. We have compared the tumour classification performance between ResNet-101 and some of the most commonly used pre-trained deep models such as VGG-19⁵¹ on the testing set of the NRG-H. The detailed protocol for evaluating the tumour classification performance between the deep models is presented in Supplementary Materials Section 2.

To adopt the ResNet-101 model with pre-trained weight and to recognise the features in the NSCLC CT data, we fine-tuned it for the task of identifying CT images that contained tumours. The 86 subjects from the NRG-H dataset were divided into two groups: a training set that comprises imaging data from 69 patients and, a testing set that comprises imaging data from 17 patients. Subjects in the training and testing groups were randomly selected. We implemented a fivefold cross-validation strategy on the training set of 69 patients to fine-tune the ResNet-101 model. The testing set of 17 subjects was ‘held out’/‘unseen’ during the fine-tuning process and serves to assess the robustness and generalisability of those fine-tuned ResNet-101 models.

From the training set, a total of 2420 CT image slices were sampled for the fine-tuning task. The CT images from the training patients were augmented to avoid overfitting during the fine-tuning. Training data was augmented by randomly rotating images between 0 to 360 degrees and translating the rotated images between -5 to 5 pixels on both the x and y-axis. The last layer of the pre-trained ResNet-101 was replaced by a new fully connected layer to accommodate the classification task. The weight learn rate factor and bias learn rate factor were set to 20 for the new fully connected layer.

The fine-tuning process of the ResNet-101 model involved 300 epochs of training using stochastic gradient descent with a momentum of 0.9 and a batch size of 5. The Learning rate was set at 1×10^{-3} , with L2Regularization set at 0.001. For every 100 epochs, the learning rate decreased by the factor of 0.1. These hyperparameters were determined and tuned by using the widely adopted random search optimisation method⁵². This is achieved by finding the optimum model which consists of the combination of hyperparameters that give the best overall performance for the classification task. Fine-tuning was implemented using MATLAB 2019b on a machine running Ubuntu 18.04, with an 11 GB NVIDIA RTX 2080 Ti GPU and CUDA 10.1. The fine-tuned model with the best overall performance was selected to serve as deep ET for FT features (Supplementary Table S4). The performance of the selected deep ET was then assessed on the second NRG-S dataset (Supplementary Table S5).

Image feature extraction. HC features were extracted directly from the volumetric CT images using the pyradiomics framework. TL and FT features were extracted from the ‘pool5’ layer of the ResNet-101 model. We used the axial, sagittal and coronal views of the tumour ROI from the volumetric CT images as the input for deep ETs³¹. All views were centred on the physical centroid of the tumour ROI. Such an aggregated views display a 2.5-dimensional (2.5D) representation of the tumour ROI⁵³. The 2.5D representation for deep ETs contains richer spatial information of neighbouring pixels compared with traditional 2D images while demanding less computational power when compared with running ETs directly on 3D image volumes⁵⁴. For each view of the

2.5D representation, gray values were normalised from [0, 4096] to [0, 255] using a linear transformation. All three input slices were resized to 224×224 to fit the input size of ResNet-101 using nearest-neighbour interpolation and were padded with zeros to preserve the tumour aspect ratio. 6144 FT and 6144 TL were extracted from each CT image.

Image feature fusion and selection. We used a feature fusion strategy that concatenates the HC, TL and FT feature together to generate a feature matrix across the patients⁵⁵. The 431 HC, 6144 TL, and 6144 FT features were then fused into a single 12,719-dimensional feature matrix using concatenation. The resulting high-dimensional feature matrix presented challenges in performing statistically significant analyses as the number of features is much larger than observations^{56,57}. In such circumstances, small random fluctuations in individual features may be mistaken for important variance in the data and lead to the selection of features that are suboptimal for representing the observations. In addition, the concatenation may cause redundant features from individual extraction techniques to be contained within the matrix and add complexity during data interpretation. Feature selection is a technique to reduce the dimensionality and identify the subset of optimal and robust features that provide the best predictive power⁵⁸. We hence applied a multi-step image feature selection scheme that aims to: (i) reduce the dimensionality of the concatenated feature matrix; (ii) remove image features that are redundant or irrelevant to the histology classification of tumours; and (iii) identify a set of image features that are most relevant to the histology characteristics of patients.

The reduction of the dimensionality removed features that have poor variability and dispersion across patients. These features do not reflect the variances in tumour imaging characteristics and therefore unideal for identifying radiogenomics associations. We used the median absolute deviation (MAD) as an indication for these features as it measures the variability across features and is robust against outliers in the concatenated feature matrix.

The second stage reduced the dimensionality of the remaining features by removing those that are redundant or irrelevant to the histology characteristics of patients. The histology characterisation is a crucial parameter that indicates the subtypes of the disease and may also contain information that reflects distinct patterns of genetic alterations⁵⁹. The removal of biologically irrelevant and redundant features, therefore, prevents the discovery of meaningless radiogenomics associations. The histology characteristics of each patient were categorised into one of the following classes: (1) squamous cell carcinoma, (2) adenocarcinoma and (3) other types including Non-Small cell and Not otherwise specified (NOS).

We used mRMR, a widely-adopted approach for feature selection, to produce a subset of features with high biological relevance⁶⁰. The mRMR method selects features that have: (i) the maximal mutual information between the total feature set and the histology characterisation and (ii) the minimal mutual information between the selected features subset and the total feature set. A total of 100 features were selected using the mRMR method, taking into consideration of the number of patients as well as the original dimensionality of the feature matrix⁶¹.

The last stage of feature selection employed LASSO regularisation for generalised linear models to identify the set of remaining image features that are most relevant to the histology characteristics of patients. LASSO shrinks regression coefficients towards zero-based on regularisation weight λ ; features with non-coefficients are those that are related to predicting histological characteristics and hence are selected. We performed 10-fold cross-validation to identify the value of λ with the minimum cross-validation error. The outcome of this stage was the final FF_{sig} that was used for identifying radiogenomics associations with gene expressions and GO terms. We also applied the multi-stage image feature selection process to the HC, TL and FT features individually for comparison. The resulting image feature signatures are hereafter denoted as 'HC_{Sig}', 'TL_{Sig}' and 'FT_{Sig}', correspondingly.

Associating FF_{sig} with primary tumour T stages. The tumour, node, metastasis (TNM) staging is the most important clinical parameter to predict survival and establish treatment plans⁶². The T stage describes the size of the primary tumours and their involvement in the adjacent structures. We investigate investigated if the FF_{sig} is relevant to primary tumour T stages (T1–T4) prior to the radiogenomics analysis. We used unsupervised k-means clustering to the FF_{sig} to stratify the patients into distinct groups; the patient clusters were defined using 10 repeated new initial cluster centroid positions with a maximum of 1000 iterations. We compared the three patient clusters with the distribution of the T stage. We used the χ^2 test of independence to assess the ability of the FF_{sig} to encode tumour staging characteristics⁶³. For comparative evaluation, the HC_{Sig}, TL_{Sig} and FT_{Sig} were also validated for their relevance to the T stage.

Functional gene analysis. *Gene selection.* Probes that map to multiple unique gene symbols were discarded and the repeated total gene expression values of the same gene were averaged. Gene expression data may contain redundant genes that are irrelevant to the disease. We used the following process to remove genes that had low variance, entropy and absolute expression value because such genes showed poor variability and dispersion, and therefore may not reflect the differences in the underlying tumour biology. We firstly removed genes with a variance of less than one-quarter percentile, as such genes may not reflect changes in tumour biological behaviours. The averaged gene expression was filtered to remove the genes with a variance of less than one-quarter percentile across all patients. The remaining genes were then filtered to remove genes that have an absolute expression level in the lowest quarter percentile of the gene expression; genes with low absolute expression were removed because they are prone to errors due to large quantisation or spot hybridisation. Finally, gene expressions were filtered to remove the genes with an entropy value that is less than the quarter percentile; genes with low entropy are considered to be consistently expressed across patients and may not reflect the variance in tumour biological characteristics⁶⁴.

Radiogenomics analysis. We determined RRs between the FF_{Sig} with the averaged gene expressions using the Spearman rank correlation. We also employed functional enrichment analysis to enrich radiogenomics relationships with GO terms. We used 1046 gene sets from the C5 collection of MSigDB⁶⁵, which categorise the following GO terms: molecular function, cellular component and biological process. The gene list was generated by ranking the radiogenomics associations for each of the features from FF_{Sig} in descending order. Gene sets that include between 15 and 500 contributing genes were selected for the enrichment analysis as was the standard protocol in prior work¹⁵. The determined RRs were then assessed using a pre-ranked functional enrichment analysis. In this process, the radiogenomics relationships between FF_{Sig} and gene expressions were sorted to provide a ranked gene list based on the strength of the Spearman rank correlation.

We used the pre-ranked gene list to perform GSEA, which derives the association between the provided ranked gene list and GO terms by testing the enrichment of each annotated term iteratively in a linear model. The enriched radiogenomics relationships with GO terms can be quantified by calculating normalised enrichment scores (NES) based on the number of genes. NES indicates the degree to which a GO term is overrepresented by the radiogenomics relationships. To ensure that only significantly associated genes were used for functional enrichment analysis, RRs with p -value < 0.001 were selected and ranked and serve as input to the functional enrichment analysis with GO terms. The same procedure was applied to the HC_{Sig} , TL_{Sig} and FT_{Sig} for comparative experiments.

Evaluation strategy. We evaluated the performance of FF_{Sig} by: (i) determining if the proposed FF_{Sig} can encode complementary medical image visual characteristics when compared with other image feature signatures; (ii) determining if the proposed FF_{Sig} is relevant to the tumour T stage by using the χ^2 test of independence; (iii) assessing the distribution of RRs with genes; (iv) assessing the distribution of RRs with GO terms; (v) determining if the proposed FF_{Sig} can identify exclusive RRs with genetic biomarkers of NSCLC and GO terms that are related to NSCLC.

Results

Image feature signatures. After performing the multi-stage image feature fusion and selection on the NRG-H dataset, all four feature signatures were generated. FF_{Sig} is comprised of features that were all extracted from sagittal planes of the 2.5D presentation and has the highest number of features at 7. TL_{Sig} is also comprised of features that were all extracted from sagittal planes and has 6 features. FT_{Sig} is comprised of features that were extracted from 1 axial and 2 sagittal planes and has 3 features. HC_{Sig} is comprised of features that were extracted directly from image volumes and have 2 features. In contrast, our validation experiments on the NRG-S dataset show that only FF_{Sig} , FT_{Sig} and HC_{Sig} were generated after performing the multi-stage image feature fusion and selection. Our validation results from the NRG-S dataset show that the FF_{Sig} is comprised of features that were all extracted from 1 axial and 12 sagittal planes of the 2.5D presentation and has the highest number of features at 13. TL_{Sig} was not generated as none of the TL features were selected after the multi-step feature selection scheme. FT_{Sig} is comprised of 2 image features that were extracted from axial planes. HC_{Sig} is comprised of features that were extracted directly from image volumes and have 1 feature only.

Image signatures and T stage. In our experiment on the NRG-H dataset, the HC, TL and TF features were significantly associated with the T stage parameters (T1–T4) across patient clusters. The χ^2 test statistics for HC, TL and TF features with T stage parameters are $p < 2.9 \times 10^{-4}$, $p < 5.0 \times 10^{-3}$ and $p < 4.8 \times 10^{-2}$, respectively. For image signatures, FF_{Sig} was significantly associated with primary tumour T stages (χ^2 test, $p < 4.0 \times 10^{-2}$). None of the HC_{Sig} , TL_{Sig} or FT_{Sig} is found to be significantly associated with primary tumour T stages, their χ^2 test statistics are $p > 0.8$, $p > 6.0 \times 10^{-2}$ and $p > 0.5$, respectively. Figure 2 illustrate the relationships among FF_{Sig} , T stages and patient clusters from the NRG-H dataset. Each row of the heatmap represents one image feature that comprises the FF_{Sig} . Each column of the heatmap represents a single patient. Z-score is calculated for each radiomics feature across patients. Z-score shows the distinct distribution of T stage parameters across patient clusters. The association between FF_{Sig} and the T stage parameters is indicated by the grouped image features among the patient cluster II and III. The distinct pattern is represented using a z-score of image features that were extracted from each patient.

In our validation experiment on the NRG-S dataset, none of the HC, TL and TF features were significantly associated with the T stage parameters (Tis, T1–T4) across patient clusters. The χ^2 test statistics for HC, TL and TF features with T stage parameters are $p > 0.7$, $p > 0.7$ and $p > 0.8$, respectively. For image signatures, none of the FF_{Sig} , HC_{Sig} or FT_{Sig} was found to be significantly associated with primary tumour T stages, their χ^2 test statistics are $p > 0.5$, $p > 0.5$ and $p > 0.2$, respectively.

RRs between image feature signatures and genes. After gene expression filtering, a total of 11,318 gene expression remained from the NRG-H dataset to establish radiogenomics associations. Notably, two of the key biomarkers for NSCLC: KRAS and RRM1, were filtered due to low variance across the patients in the NRG-H dataset. Figure 3a represents the distribution of RRs that were determined between the averaged gene expression values of 11,318 individual genes and FF_{Sig} , HC_{Sig} , TL_{Sig} and FT_{Sig} . FF_{Sig} identified the highest number of RRs at 5039 and correlated with the highest number of genes at 3881. HC_{Sig} identified 1193 RRs with 886 genes. TL_{Sig} identified 3816 RRs with 3297 genes. FT_{Sig} identified 2089 RRs with 2008 genes. Figure 4a details the distribution of unique genes that were associated with FF_{Sig} , HC_{Sig} , TL_{Sig} and FT_{Sig} . Among the 3881 unique genes that were associated with the FF_{Sig} , 1896 unique genes cannot be associated with any of the HC_{Sig} , TL_{Sig} and FT_{Sig} . In contrast, a total number of 3269 unique genes were associated with one of the HC_{Sig} , TL_{Sig} and FT_{Sig} , but were not correlated with the FF_{Sig} .

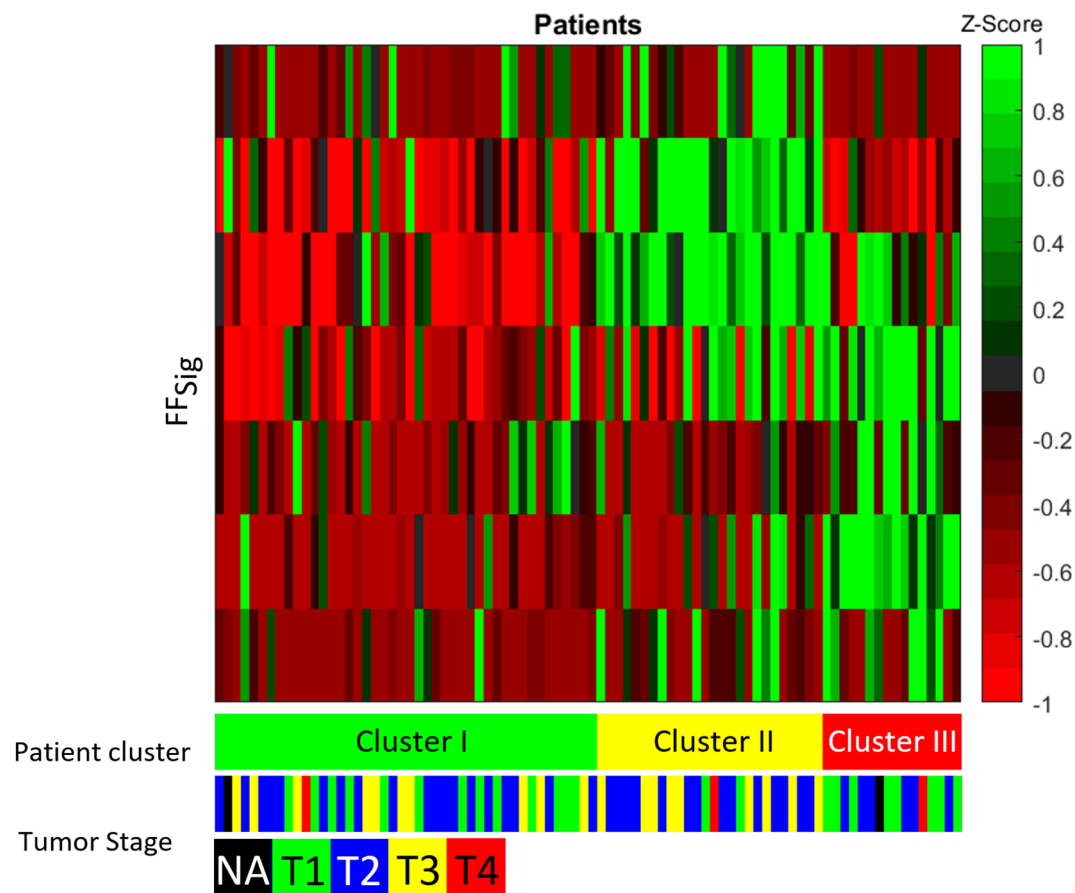


Figure 2. Heatmap of the FF_{Sig} across patient clusters with corresponding T stage from the NRG-H dataset. The heatmap was generated using MATLAB, version 2019b, URL: <https://www.mathworks.com/products/matlab.html>.

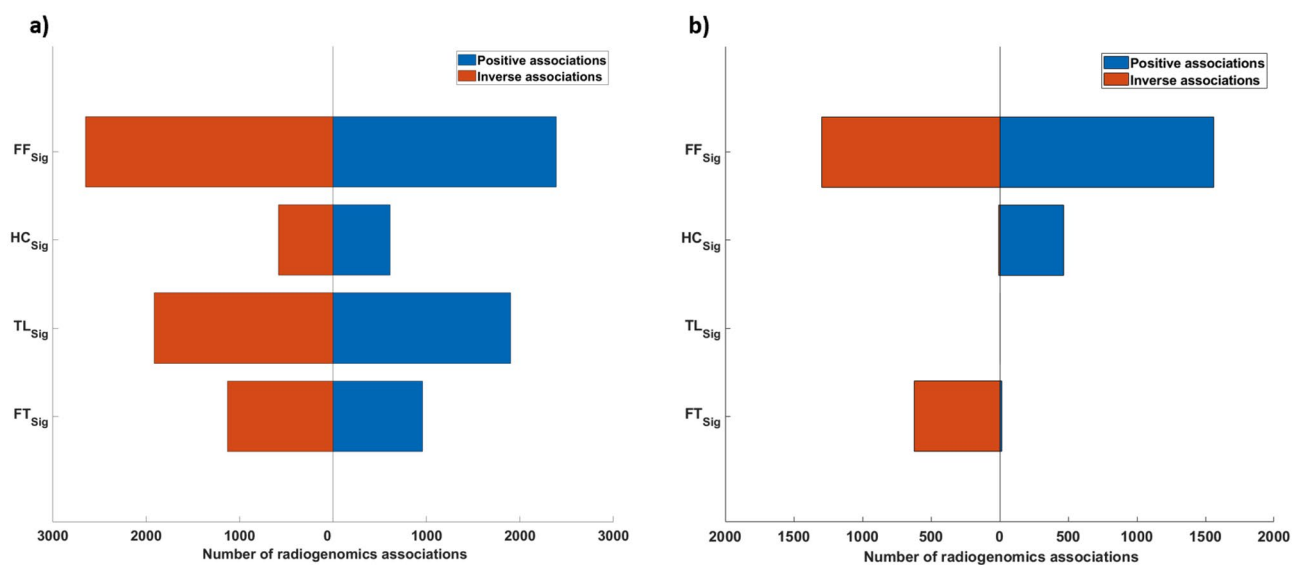


Figure 3. The distribution of RRs between feature signatures and: (a) gene expression value of the processed genes (n = 11,318) from the NRG-H dataset. (b) Gene expression value of the processed genes (n = 2993) from the NRG-S dataset.

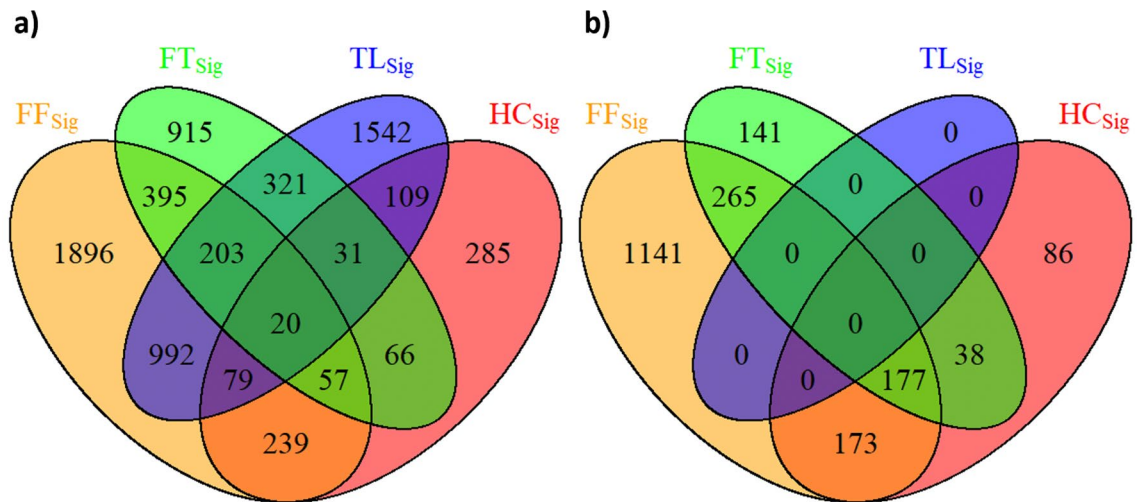


Figure 4. Venn diagram shows the distribution of unique genes that were associated with FF_{Sig}, HC_{Sig}, TL_{Sig}, and FT_{Sig}: (a) generated using the NRG-H dataset. (b) generated using the NRG-S dataset.

Feature signature	HC _{Sig}	TL _{Sig}	FT _{Sig}
Strength of positive RRs (two-sample <i>t</i> test)			
FF _{Sig}	$p > 0.2$	$p > 0.7$	$p > 0.3$
Strength of inverse RRs (two-sample <i>t</i> test)			
FF _{Sig}	$p < 1 \times 10^{-3}$	$p < 1 \times 10^{-2}$	$p > 0.6$

Table 1. Two-sample *t* tests that assess the strengths of all RRs constructed using the FF_{Sig} with HC_{Sig}, TL_{Sig}, and FT_{Sig} in both statistical directions on the NRG-H dataset.

Table 1 compares the strengths of all RRs that were determined using the FF_{Sig} against those determined using HC_{Sig}, TL_{Sig}, and FT_{Sig}. Our results show stronger RRs are identified between the FF_{Sig} and genes, when compared with HC_{Sig} and TL_{Sig}, in the inverse direction. The FF_{Sig}, however, did not show stronger inverse RRs when compared with FT_{Sig}. On the other hand, the FF_{Sig} did not show stronger positive RRs when compared with HC_{Sig}, TL_{Sig}, nor FT_{Sig}. Figure 5 illustrates the distribution of RRs that were determined between image feature signatures of the FF_{Sig}, HC_{Sig}, TL_{Sig}, FT_{Sig} with the gene expression value from the key genetic biomarkers of EGFR for NSCLC²⁰. Our result shows that the FF_{Sig} and FT_{Sig} were inversely correlated with EGFR expression. In contrast, HC_{Sig} is shown to be the only positive RRs with EGFR. Notably, FT_{Sig} shows to derive more and stronger inverse RRs with EGFR when compared with the FF_{Sig}. In addition, our result shows that ERCC1, a key genetic biomarker for NSCLC, is exclusively correlated with a single feature from the FF_{Sig}, where the same feature showed inverse RRs with EGFR previously.

The gene selection process was repeated in our validation experiments on the NRG-S dataset. A total of 22,126 unique genes were identified for each patient from the NRG-S dataset. After gene selection, 2993 gene expression remained from the NRG-S dataset to establish radiogenomics associations. In comparison to NRG-H, three of the key biomarkers for NSCLC: EGFR, KRAS and ERCC1, were filtered due to low variance across the patients in the NRG-S dataset. Figure 3b represents the distribution of RRs that were determined between the averaged gene expression values of 2993 individual genes and FF_{Sig}, HC_{Sig}, and FT_{Sig}. Radiogenomics analysis show that FF_{Sig} identified the highest number of RRs at 2856 and correlated with the highest number of genes at 1756. HC_{Sig} identified 474 RRs with 474 genes. FT_{Sig} identified 642 RRs with 621 genes. In addition, our result shows that RRM1, a key genetic biomarker for NSCLC, is exclusively correlated with a single feature from the FF_{Sig}. Figure 4b details the distribution of unique genes that were associated with the FF_{Sig}, HC_{Sig}, TL_{Sig}, and FT_{Sig}. Among the 1756 unique genes that were associated with the FF_{Sig}, 1141 unique genes cannot be associated with any of the HC_{Sig} and FT_{Sig}. In contrast, a total number of 265 unique genes were associated with one of the HC_{Sig} and FT_{Sig} but were not correlated with the FF_{Sig}. Table 2 compares the strengths of all RRs that were determined using the FF_{Sig} against those determined using HC_{Sig} and FT_{Sig}. Our validation results show that the FF_{Sig} did not identify stronger RRs with genes, when compared with HC_{Sig} and TL_{Sig}, in both statistical directions.

RRs between image feature signatures and GO terms. From our experiments using the NRG-H dataset, FF_{Sig} determined RRs with the highest number of GO terms at 244. HC_{Sig} determined RRs with 62 GO terms. TL_{Sig} determined RRs with 246 GO terms. FT_{Sig} determined RRs with 129 GO terms. Figure 6a details the distribution of GO terms that were associated with image feature signatures of FF_{Sig}, HC_{Sig}, TL_{Sig}, and FT_{Sig}.

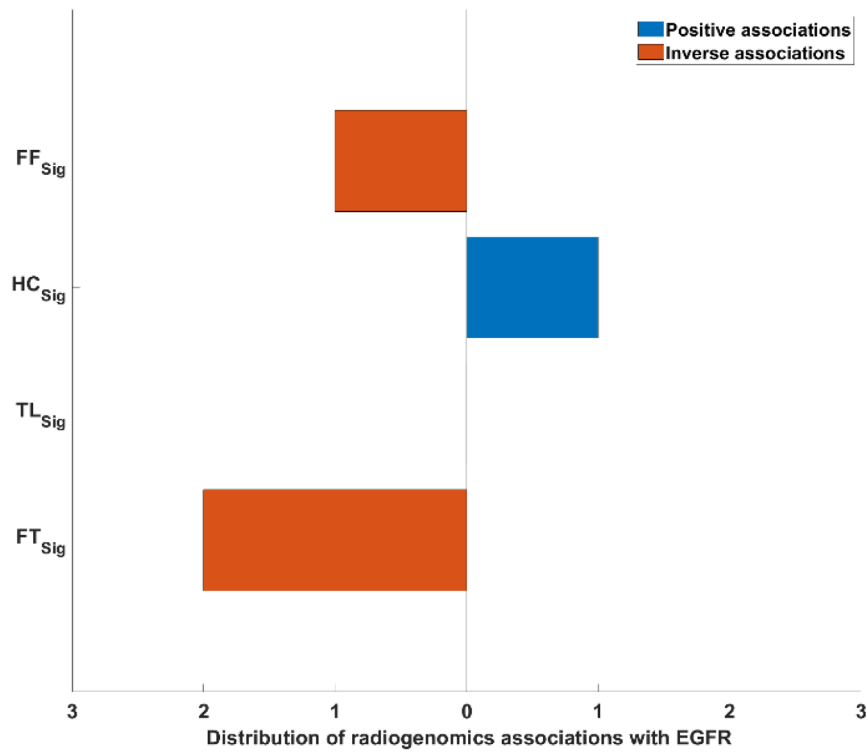


Figure 5. The distribution of RRs between the FF_{Sig} with the key genetic biomarker of EGFR from the NRG-H dataset, in comparison to HC_{Sig}, TL_{Sig} and FT_{Sig}.

Feature signature	HC _{Sig}	FT _{Sig}
Strength of positive RRs (two-sample <i>t</i> test)		
FF _{Sig}	p > 0.8	p > 0.2
Strength of inverse RRs (two-sample <i>t</i> test)		
FF _{Sig}	p > 0.3	p > 0.08

Table 2. Two-sample *t* tests that assess the strengths of all RRs constructed using the FF_{Sig} with HC_{Sig} and FT_{Sig}, in both statistical directions on the NRG-S dataset.

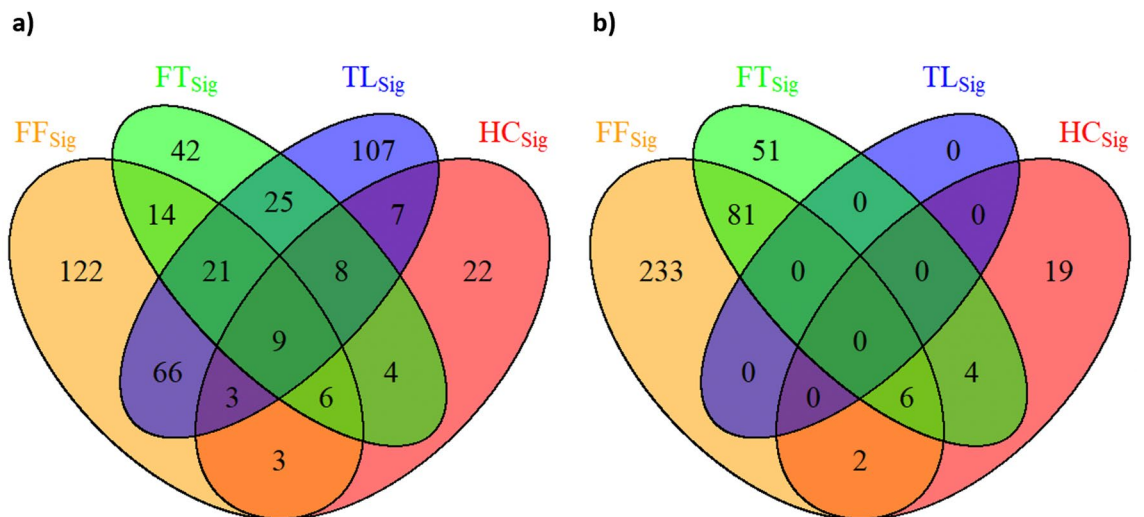


Figure 6. Venn diagram shows the distribution of GO terms that were associated with image feature signatures of FF_{Sig}, TL_{Sig}, FT_{Sig} and HC_{Sig}: (a) generated using the NRG-H dataset. (b) Generated using the NRG-S dataset.

	NES		NES
FF_{Sig}		HC_{Sig}	
Organelle lumen	2.43	Extracellular region	1.74
Nuclear lumen	2.22	Regulation of transferase activity	0.60
Membrane enclosed lumen	2.19	Transferase activity transferring phosphorus containing groups	0.58
Glycoprotein biosynthetic process	1.98	Protein kinase activity	0.58
Macromolecule biosynthetic process	1.94	Stress activated protein kinase signalling pathway	0.58
Response to virus	-1.98	Carbohydrate metabolic process	-0.99
Cell cell signaling	-1.98	Phosphoric monoester hydrolase activity	-0.99
Response to other organism	-2.00	Phosphoric ester hydrolase activity	-1.01
Anatomical structure morphogenesis	-2.01	Alcohol metabolic process	-1.02
Response to biotic stimulus	-2.01	Hydrolase activity acting on ester bonds	-1.02
TL_{Sig}		FT_{Sig}	
Cell fraction	2.17	Anatomical structure morphogenesis	1.85
Membrane fraction	2.03	Enzyme regulator activity	1.80
Phosphoric ester hydrolase activity	2.02	Enzyme activator activity	1.79
Soluble fraction	1.96	Enzyme linked receptor protein signalling pathway	1.77
Insoluble fraction	1.96	Membrane fraction	1.73
Homophilic cell adhesion	-1.66	Extracellular region part	-1.23
Sulfuric ester hydrolase activity	-1.67	Extracellular space	-1.23
Nervous system development	-1.68	Phosphorylation	-1.24
Regulation of anatomical structure morphogenesis	-1.68	Lipase activity	-1.25
Cell surface	-1.99	Female pregnancy	-1.27

Table 3. The GO terms that have RRs with FF_{Sig}, HC_{Sig}, TL_{Sig} and FT_{Sig} with positive and negative associations from the NRG-H dataset.

FF _{Sig} exclusive	NES	FF _{Sig} restricted	NES
Transmembrane receptor protein kinase activity	1.61	Soluble fraction	1.96
Protein tyrosine kinase activity	1.60	Insoluble fraction	1.96
Transmembrane receptor protein tyrosine kinase activity	1.53	Enzyme regulator activity	1.80
Generation of precursor metabolic and energy	1.47	Enzyme activator activity	1.79
Phospholipid metabolic process	1.42	Molecular adaptor activity	1.73
RNA processing	-1.85	Generation of neurons	-1.66
Organ morphogenesis	-1.95	Homophilic cell adhesion	-1.67
Response to virus	-1.98	Sulfuric ester hydrolase activity	-1.67
Response to other organism	-2.00	Regulation of anatomical structure morphogenesis	-1.68
Response to biotic stimulus	-2.01	Cell surface	-1.99

Table 4. The GO terms that have the highest NES and exclusively RRs with FF_{Sig} (left) and the GO terms that are restricted to have RRs with FF_{Sig} (right), experimented on the NRG-H dataset.

Among the 244 GO terms that have RRs with by FF_{Sig}, 122 GO terms were exclusively enriched; these GO terms account for 50% of the total enriched GO terms or 13.8% of the total 1046 GO terms.

Table 3 shows the GO terms with the highest NES. Notably, FF_{Sig} determined RRs with GO terms that exhibit distinct patterns relating to the biological functions and cellular behaviours: (i) 3 GO terms were related to lumen structures including organelle, nuclear and membrane; (ii) 2 GO terms were reflecting biosynthesis processes that involve glycoprotein or macromolecule; (iii) 3 GO terms were related to the response mechanism to viruses, other organism or biotic stimulus, and other types of stimulus processes. In comparison, our results also show that TL_{Sig} determined RRs with 4 GO terms that are associated with fraction activities. In addition, FT_{Sig} determined RRs with GO terms that are related to enzyme activities. In contrast, HC_{Sig} determined RRs with GO terms are shown to be without overlaps in their biological functionalities.

Table 4 shows the comparison between GO terms that have exclusive RRs with FF_{Sig} and those GO terms that are restricted to have RRs with FF_{Sig}. Among the GO terms with the highest NES, our result shows clusters of biological functions and cellular behaviours that have exclusive RRs with the FF_{Sig}: (i) 3 GO terms were related to kinase activities for transmembrane receptor protein and tyrosine kinase; (ii) 2 GO terms were related to metabolism activities; The identical 3 GO terms were most enriched by FF_{Sig} and related to the virus response

FF _{Sig}	NES	HC _{Sig}	NES	FT _{Sig}	NES
Perinuclear region of cytoplasm	2.62	Sensory perception	1.80	Intracellular protein transport	2.62
Nervous system development	2.58	Monooxygenase activity	1.78	Establishment of protein localisation	2.61
Membrane organisation and biogenesis	2.45	Oxygen binding	1.78	Macromolecule localisation	2.61
Intercellular junction	2.05	Electron transport (GO 0006118)	1.75	Protein localisation	2.54
Tight junction	1.96	Neurological system process	1.70	Protein transport	2.52
Kinase activity	-2.00	Second messenger mediated signalling	-0.77	Soluble fraction	-1.61
Endoplasmic reticulum	-2.11	Establishment and or maintenance of cell polarity	-0.77	Organelle lumen	-1.62
Nuclear lumen	-2.19	Regulation of catalytic activity	-0.77	Nucleolus	-1.65
Organelle lumen	-2.84	cAMP mediated signalling	-0.77	Nuclear lumen	-1.67
Membrane enclosed lumen	-3.06	G protein signalling adenylate cyclase activating pathway	-0.77	Membrane enclosed lumen	-1.71

Table 5. The GO terms that have RRs with FF_{Sig}, HC_{Sig} and FT_{Sig} with positive and negative associations from the NRG-S dataset.

FF _{Sig} exclusive	NES	FF _{Sig} restricted	NES
Perinuclear region of cytoplasm	2.62	Positive regulation of metabolic process	1.93
Membrane organisation and biogenesis	2.45	Positive regulation of cellular metabolic process	1.90
Intercellular junction	2.05	Neurite development	1.90
Tight junction	1.96	Steroid hormone receptor signalling pathway	1.89
Apical junction complex	1.94	Cellular lipid catabolic process	1.88
Serine type peptidase activity	-1.57	cAMP mediated signalling	-0.77
Serine hydrolase activity	-1.58	G Protein signalling adenylate cyclase activating pathway	-0.77
Serine type endopeptidase activity	-1.60	Intrinsic to Golgi membrane	-0.88
Peptidase activity	-1.75	Intrinsic to organelle membrane	-0.93
Endopeptidase activity	-1.76	Integral to organelle membrane	-0.93

Table 6. The GO terms that have the highest NES and exclusively RRs with FF_{Sig} (left) and the GO terms that are restricted to have RRs with FF_{Sig} (right), experimented on the NRG-S dataset.

mechanism. In contrast, our result shows 2 groups of related biological functions among the GO terms that were restricted to FF_{Sig}. Such GO terms are related to fraction processes and enzyme activities.

From our validation experiment on the NRG-S dataset, functional gene enrichment analysis reveals that FF_{Sig} determined RRs with the highest number of GO terms at 322. HC_{Sig} determined RRs with 31 GO terms. TL_{Sig} determined RRs with 0 GO terms. Figure 6b details the distribution of GO terms that were associated with image feature signatures of FF_{Sig}, HC_{Sig}, TL_{Sig}, and FT_{Sig}. FT_{Sig} determined RRs with 142 GO terms. Among the 322 GO terms that have RRs with by FF_{Sig}, 233 GO terms were exclusively enriched; these GO terms account for 72.4% of the total enriched GO terms or 22.3% of the total 1046 GO terms.

Table 5 shows the GO terms with the highest NES. Notably, FF_{Sig} determined RRs with GO terms that exhibit distinct patterns relating to the cellular structure: (i) 3 GO terms were related to lumen structures including organelle, nuclear and membrane; (ii) 2 GO terms that reflect the cell junction. In comparison, FT_{Sig} determined RRs with GO terms that are related to cellular structures, protein transportation and localisation. HC_{Sig} determined RRs with GO terms that are related to signalling pathways, such as cAMP mediated signalling and second messenger mediated signalling.

Table 6 shows the comparison between GO terms that have exclusive RRs with FF_{Sig} and those GO terms that are restricted to have RRs with FF_{Sig}. Among the GO terms with the highest NES, our validation results show a cluster of biological functions and cellular behaviours that have exclusive RRs with the FF_{Sig}: (i) 3 GO terms were related to peptidase activity; (ii) 2 GO terms that reflect the cell junction. In contrast, our result shows 2 groups of related biological functions among the GO terms that were restricted to FF_{Sig}. Such GO terms are related to the intrinsic components of organelle membranes and metabolic processes.

Discussion

Our main findings are that our FF_{Sig}: (i) encoded complementary medical image's visual characteristics when compared with other image feature signatures; (ii) determined a greater number of RRs with a greater number of genes; (iii) determined RRs with distinct GO terms; (iv) determined exclusive RRs with genetic biomarkers of NSCLC and GO terms that are related to NSCLC and (v) is robust and generalisable for determining RRs when validated on NRG-S.

From our experiments using the NRG-H dataset, the FF_{Sig} comprises 7 image features that are complementary to image features that were selected in the HC_{Sig} , TL_{Sig} , and FT_{Sig} . Image features that are included in the FF_{Sig} can be traced back to the 6144-dimensional TL features. This finding indicates that the multi-step feature selection scheme prioritised a set of complementary image features that are relevant to the histological characteristics while reducing the overall redundancy in the information captured. This finding suggests that the FF_{Sig} encodes unique medical imaging visual characteristics when compared with other image signatures. The FF_{Sig} was the only feature signature that produced a significant association ($p < 0.05$) with the T stage. The HC_{Sig} , TL_{Sig} , and FT_{Sig} did not have any association with the T stage, despite the fact that the FF_{Sig} was selected from the HC, FT, and TL features. Our results showed that the semantic information that is encoded in the HC features and the abstract-level information that are encoded in the TL and FT features contributed towards the selection of features in FF_{Sig} . This finding implies that the association between FF_{Sig} and T stage occurred because the FF_{Sig} leveraged complementary information using both HC and deep ETs.

The FF_{Sig} determined a greater number of RRs with a greater number of genes when compared with the other image feature signatures. The FF_{Sig} was also correlated with EGFR. One potential explanation for our finding is that the FF_{Sig} encodes the imaging characteristics of the tumour that can reflect the underlying molecular characteristics of NSCLC⁶⁶. The FF_{Sig} has also determined stronger inverse RRs with a range of genes when compared to HC_{Sig} and TL_{Sig} . There was no stronger positive RRs with genes when compared with the HC_{Sig} , TL_{Sig} and FT_{Sig} . The reason for this is because the FF_{Sig} did not incorporate any image feature that was learned from scratch from the raw data using deep ETs; the FT components were the closest and as stated previously were aligned with the non-medical TL features. We suggest that positive RRs may appear when deep ETs are directly trained from scratch on the NRG-H CT data.

In addition, from our experiments using the NRG-H dataset, the FF_{Sig} determined RRs with a distinctive collection of GO terms with higher NES when compared to the other image feature signatures. A higher NES of GO terms is typically the result of a stronger correlation between the image feature signatures and the affiliated genes that contribute to the GO term and, RRs with a greater number of affiliated genes that contribute to the GO term. Notably, GO terms with the highest NES consist of a range of biological functions that relate to cellular structures. It has been reported that abnormalities in cellular structures are related to the development of NSCLC⁶⁷. FF_{Sig} has shown to determine RRs with more GO terms when compared with HC_{Sig} and FT_{Sig} . A potential explanation for this finding is that the FF_{Sig} determined RRs with a greater number of unique genes. These genes may be affiliated with a greater range of biological functions and therefore provide opportunities for FF_{Sig} to determine RRs with more and unique GO terms. We note that while the TL_{Sig} determined RRs with a higher number of GO terms, these are generally related to normal human anatomical information rather than the subtle disease processes related to the primary tumour. This finding is evidenced by the most enriched GO terms, such as “Regulation of Anatomical Structure Morphogenesis”, as shown in Table 3.

From our experiments using the NRG-H dataset, FF_{Sig} determined exclusive RRs with a group of GO terms that consist of a range of biological functions that are related to protein kinase activities, such as “Transmembrane Receptor Protein Kinase Activity”. Atypical kinase and its activities have been reported previously as an oncogene in NSCLC⁶⁸, which play a crucial role in cell growth and tumorigenesis that may be observable in medical images⁶⁹. In contrast, GO terms that are restricted to have RRs with FF_{Sig} include, for example, “Soluble Fraction” and “Enzyme Regulator Activity”. A potential explanation is that the specific enzyme activities and fractions cannot be depicted by CT images and hence cannot be quantified by the FF_{Sig} .

Our validation experiments on the NRG-S dataset show that the FF_{Sig} comprises 13 image features that are complementary to image features that were selected in the HC_{Sig} and FT_{Sig} . Among the 13 image features, 12 can be traced back to the 6144-dimensional FT features and the other feature can be traced back to a HC feature. Our results using NRG-S demonstrated that the FF_{Sig} encoded complementary medical imaging visual characteristics. The validation results are consistent with our previous findings from the NRG-H dataset.

However, none of the FF_{Sig} , HC_{Sig} , nor FT_{Sig} from the NRG-S dataset produced a significant association with the T stage. We attribute our findings to the different scanning parameters used in the NRG-S dataset, for example, slice thickness that ranges from 0.625 to 3 mm. Such factors contribute to subtle imaging differences and have potential impacts on the feature extraction process.

In our validation study, FF_{Sig} has determined a greater number of RRs with a greater number of genes when compared with the other image feature signatures. This result validates that the FF_{Sig} is robust and generalisable in encoding the imaging characteristics of the tumour that can reflect the underlying molecular characteristics of NSCLC. However, in our validation study, using the NRG-S dataset, FF_{Sig} did not identify stronger RRs with a range of genes when compared with HC_{Sig} and FT_{Sig} . One potential explanation is that the FF_{Sig} did not incorporate any image feature that was fine-tuned on the NRG-S dataset. Despite NRG-S dataset has many similarities to the NRG-H dataset, such as the type of disease, the distribution of patients’ clinical parameters and their histopathology status are vastly different to the NRG-H dataset. We suggest that stronger RRs may appear when deep ETs are fine-tuned on the NRG-S dataset.

In our validation experiments, the FF_{Sig} has also shown to determine RRs with a distinctive collection of GO terms with higher NES when compared to the other image feature signatures. Notably, our validation results share a high degree of similarity with our previous findings from experiments using the NRG-H dataset. For example, from both experiments, the proposed FF_{Sig} determined RRs with GO terms such as ‘Membrane Enclosed Lumen’ and ‘Organelle Lumen’. Interestingly, such RRs with GO terms that relate to lumen structures are in opposite statistical direction. We attribute this finding to the differences between the NRG-H and NRG-S datasets where their distribution of T stage parameters and histology sub-types, as they played important roles in the multi-stage feature selection scheme. Such findings further demonstrate the robustness and generalisability of our proposed FF_{Sig} to determine RRs with GOs across different datasets.

Furthermore, in our validation experiments using the NRG-S dataset, FF_{Sig} determined exclusive RRs with a group of GO terms that consist of a range of biological functions that are related to peptidase activity such as 'Endopeptidase Activity'. Previous study has shown that bombesin-like peptides and other neuropeptides are autocrine growth factors for both small cell lung cancer (SCLC) and NSCLC⁷⁰. Our validation results demonstrate the robustness and generalisability of our proposed FF_{Sig} for determining GO terms that are related to NSCLC.

We recognise that a limitation of our study is the size of the dataset and that lack of knowledge about the patients' mutation status. This limits the ability to optimise deep ETs to quantify image features that are most relevant to the NSCLC. Another limitation of this study is the differences between the train dataset and the independent test dataset. The two datasets use different methods for gene expression profiling, and as such the NRG-H dataset has a greater amount of genetic information compared to the NRG-S dataset. The ideal situation would have been to utilise two datasets that use the same technology for gene expression profiling, but at the time of experimentation and to the best of our knowledge, no such public radiogenomics dataset existed. However, despite these differences we note that the NRG-S dataset shares similarity with the NRG-H dataset, such as the type of disease and histopathology subtypes, and these similarities mean that it is the closest dataset that can be used for independent validation.

The limited availability of the clinical parameters e.g., survival data in the datasets has restricted our study from designing a deep learning-based image feature selection scheme. We note that as more radiogenomics datasets becomes available in the future, a key area for radiogenomics studies is to investigate the feasibility for a data-driven method for image feature selection⁷¹. Another potential future direction for our study is to investigate deep learning-based gene expression level prediction. Such a deep model can encode imaging characteristics that are reflective towards changes in gene expression levels and therefore may provide more insights into RRs.

Conclusion

We used a selection of image features from handcrafted and deep ETs, which we named FF_{Sig} , to determine RRs. Our results show that the FF_{Sig} encoded complementary medical image visual characteristics when compared with other image feature signatures. The FF_{Sig} determined more RRs with genes and with a group of distinct GO terms. Our results show that FF_{Sig} is correlated with a key biomarker for NSCLC and GO terms that are related to tumour developments in NSCLC. Furthermore, our validation experiments demonstrate that the FF_{Sig} is robust and generalisable in different dataset. The FF_{Sig} has demonstrated its potentials to identify important RRs that may facilitate cancer diagnosis and treatment in the future.

Received: 18 June 2021; Accepted: 14 January 2022

Published online: 09 February 2022

References

- Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E. & Adjei, A. A. in *Mayo Clinic Proceedings* 584–594 (Elsevier).
- Tsim, S., O'dowd, C., Milroy, R. & Davidson, S. Staging of non-small cell lung cancer (NSCLC): A review. *Respir. Med.* **104**, 1767–1774 (2010).
- Network, C. G. A. R. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
- Chen, Z., Fillmore, C. M., Hammerman, P. S., Kim, C. F. & Wong, K.-K. Non-small-cell lung cancers: A heterogeneous set of diseases. *Nat. Rev. Cancer* **14**, 535–546 (2014).
- Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
- Mitsudomi, T., Kosaka, T. & Yatabe, Y. Biological and clinical implications of EGFR mutations in lung cancer. *Int. J. Clin. Oncol.* **11**, 190–198 (2006).
- Herbst, R. S. & Shin, D. M. Monoclonal antibodies to target epidermal growth factor receptor-positive tumors: A new paradigm for cancer therapy. *Cancer* **94**, 1593–1611 (2002).
- Pirker, R. *et al.* Cetuximab plus chemotherapy in patients with advanced non-small-cell lung cancer (FLEX): An open-label randomised phase III trial. *Lancet* **373**, 1525–1531 (2009).
- Parra, H. *et al.* Analysis of epidermal growth factor receptor expression as a predictive factor for response to gefitinib ('Iressa', ZD1839) in non-small-cell lung cancer. *Br. J. Cancer* **91**, 208–212 (2004).
- Tsao, M.-S. *et al.* Erlotinib in lung cancer—Molecular and clinical predictors of outcome. *N. Engl. J. Med.* **353**, 133–144 (2005).
- Cataldo, V. D., Gibbons, D. L., Pérez-Soler, R. & Quintás-Cardama, A. Treatment of non-small-cell lung cancer with erlotinib or gefitinib. *N. Engl. J. Med.* **364**, 947–955 (2011).
- Kaus, M. R. *et al.* Automated segmentation of MR images of brain tumors. *Radiology* **218**, 586–591 (2001).
- Coroller, T. P. *et al.* Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiother. Oncol.* **119**, 480–486 (2016).
- Gevaert, O. *et al.* Non-small cell lung cancer: Identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—Methods and preliminary results. *Radiology* **264**, 387–396 (2012).
- Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).
- Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: Images are more than pictures, they are data. *Radiology* **278**, 563–577 (2015).
- GO Consortium. Creating the gene ontology resource: Design and implementation. *Genome Res.* **11**, 1425–1433 (2001).
- Shiri, I. *et al.* Next-generation radiogenomics sequencing for prediction of EGFR and KRAS mutation status in NSCLC patients using multimodal imaging and machine learning algorithms. *Mol. Imaging Biol.*, 1–17 (2020).
- Gevaert, O. *et al.* Predictive radiogenomics modeling of EGFR mutation status in lung cancer. *Sci. Rep.* **7**, 1–8 (2017).
- Ettinger, D. S. *et al.* Non-small cell lung cancer. *J. Natl. Compr. Cancer Netw.* **8**, 740–801 (2010).
- Castellano, G., Bonilha, L., Li, L. & Cendes, F. Texture analysis of medical images. *Clin. Radiol.* **59**, 1061–1069 (2004).
- Haralick, R. M., Shanmugam, K. & Dinstein, I. H. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **3**, 610–621 (1973).
- Le, N. Q. K. *et al.* Machine learning-based radiomics signatures for EGFR and KRAS mutations prediction in non-small-cell lung cancer. *Int. J. Mol. Sci.* **22**, 9254 (2021).

24. Le, V.-H., Kha, Q.-H., Hung, T. N. K. & Le, N. Q. K. Risk score generated from CT-based radiomics signatures for overall survival prediction in non-small cell lung cancer. *Cancers* **13**, 3616 (2021).
25. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
26. Kamnitsas, K. *et al.* Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **36**, 61–78 (2017).
27. Havaei, M. *et al.* Brain tumor segmentation with deep neural networks. *Med. Image Anal.* **35**, 18–31 (2017).
28. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
29. Donahue, J. *et al.* in *International Conference on Machine Learning* 647–655.
30. Kumar, A. *et al.* in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* 791–794 (IEEE).
31. Shin, H.-C. *et al.* Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).
32. Zhu, Z. *et al.* Deep learning for identifying radiogenomic associations in breast cancer. *Comput. Biol. Med.* **109**, 85–90 (2019).
33. Castanedo, F. A review of data fusion techniques. *Sci. World J.* **2013**, 704504 (2013).
34. Kooi, T. *et al.* Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303–312 (2017).
35. Hagerty, J. R. *et al.* Deep learning and handcrafted method fusion: Higher diagnostic accuracy for melanoma dermoscopy images. *IEEE J. Biomed. Health Inform.* **23**, 1385–1391 (2019).
36. Aerts, H. J. W. L., Velazquez, E. R., Leijenaar, R. T. H., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M. M., Leemans, C. R., Dekker, A., Quackenbush, J., Gillies, R. J. & Lambin, P. *The Cancer Imaging Archive* (2015).
37. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
38. Edge, S. B. & Compton, C. C. The American Joint Committee on Cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann. Surg. Oncol.* **17**, 1471–1474 (2010).
39. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2012).
40. Wolf, I. *et al.* The medical imaging interaction toolkit. *Med. Image Anal.* **9**, 594–604 (2005).
41. Bakr, S. *et al.* A radiogenomic dataset of non-small cell lung cancer. *Sci. Data* **5**, 1–9 (2018).
42. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
43. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
44. Van Griethuysen, J. J. *et al.* Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **77**, e104–e107 (2017).
45. Su, C. *et al.* Radiomics based on multicontrast MRI can precisely differentiate among glioma subtypes and predict tumour-proliferative behaviour. *Eur. Radiol.* **29**, 1986–1996 (2019).
46. Liu, X. *et al.* A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas. *NeuroImage Clin.* **20**, 1070–1077 (2018).
47. He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.
48. Ghosal, P. *et al.* in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)* 1–6 (IEEE).
49. Xia, K.-J., Yin, H.-S. & Zhang, Y.-D. Deep semantic segmentation of kidney and space-occupying lesion area based on SCNN and ResNet models combined with SIFT-flow algorithm. *J. Med. Syst.* **43**, 2 (2019).
50. Szegedy, C. *et al.* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1–9.
51. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014).
52. Bergstra, J. & Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **13**, 281–305 (2012).
53. Roth, H. R. *et al.* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* 520–527 (Springer).
54. Hesamian, M. H., Jia, W., He, X. & Kennedy, P. Deep learning techniques for medical image segmentation: Achievements and challenges. *J. Digit. Imaging* **32**, 582–596 (2019).
55. Fu, Y., Cao, L., Guo, G. & Huang, T. S. in *Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval* 127–134.
56. Johnstone, I. M. & Titterton, D. M. (The Royal Society Publishing, 2009).
57. Hira, Z. M. & Gillies, D. F. A review of feature selection and feature extraction methods applied on microarray data. *Adv. Bioinform.* **2015**, 1–13 (2015).
58. Tang, J., Alelyani, S. & Liu, H. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, 37 (2014).
59. Weigelt, B., Geyer, F. C. & Reis-Filho, J. S. Histological types of breast cancer: How special are they?. *Mol. Oncol.* **4**, 192–208 (2010).
60. Ding, C. & Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* **3**, 185–205 (2005).
61. Hua, J., Xiong, Z., Lowey, J., Suh, E. & Dougherty, E. R. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics* **21**, 1509–1515 (2005).
62. Mirsadraee, S., Oswal, D., Alizadeh, Y., Caulo, A. & van Beek, E. J. The 7th lung cancer TNM classification and staging system: Review of the changes and implications. *World J. Radiol.* **4**, 128 (2012).
63. Compton, C. C. *et al.* *AJCC Cancer Staging Atlas: A Companion to the Seventh Editions of the AJCC Cancer Staging Manual and Handbook* (Springer Science & Business Media, 2012).
64. Kohane, I. S., Butte, A. J. & Kho, A. *Microarrays for an Integrative Genomics* (MIT Press, 2002).
65. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
66. Jin, M. *et al.* Cyclin D1, p16 and retinoblastoma gene product expression as a predictor for prognosis in non-small cell lung cancer at stages I and II. *Lung Cancer* **34**, 207–218 (2001).
67. Zheng, C.-Y., Lam, S.-K., Li, Y.-Y. & Ho, J.C.-M. Arsenic trioxide-induced cytotoxicity in small cell lung cancer via altered redox homeostasis and mitochondrial integrity. *Int. J. Oncol.* **46**, 1067–1078 (2015).
68. Regala, R. P. *et al.* Atypical protein kinase C α is an oncogene in human non-small cell lung cancer. *Cancer Res.* **65**, 8905–8911 (2005).
69. Regala, R. P. *et al.* Atypical protein kinase C α plays a critical role in human lung cancer cell growth and tumorigenicity. *J. Biol. Chem.* **280**, 31109–31115 (2005).
70. Cohen, A. J. *et al.* Neutral endopeptidase: Variable expression in human lung, inactivation in lung cancer, and modulation of peptide-induced calcium flux. *Cancer Res.* **56**, 831–839 (1996).
71. Chatterjee, A. *et al.* An empirical approach for avoiding false discoveries when applying high-dimensional radiomics to small datasets. *IEEE Trans. Radiat. Plasma Med. Sci.* **3**, 201–209 (2018).

Acknowledgements

This study was funded, in part, by the Australian Research Council (ARC) DP170104304.

Author contributions

T.X. wrote the manuscript, data processing, data analysis and prepared figures and tables. A.K., Y.J., D.F. and Y.W. contributed to the design and formulation of the study. A.K., M.F. and J.K. supervised all steps of the study and reviewed the manuscript. M.F., E.K. annotated the dataset.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-06085-y>.

Correspondence and requests for materials should be addressed to T.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022