Identifying transcription patterns of histology and radiomics features in NSCLC with neural networks

Nova F. Smedley¹⁻³, Denise R. Aberle¹⁻⁴, and William Hsu¹⁻⁴

1. Medical & Imaging Informatics, University of California Los Angeles, Los Angeles, California, United States of America
2. David Geffen School of Medicine, University of California Los Angeles, Los Angeles, California, United States of America
3. Samueli School of Engineering, University of California Los Angeles, Los Angeles, California, United States of America
4. Bioinformatics IDP, University of California Los Angeles, Los Angeles, California, United States of America

Short Title:

Transcription patterns in NSCLC histology and radiomics

Keywords:

non-small cell lung cancer, radiomics, gene expression, neural networks, machine learning

Corresponding author:

William Hsu

whsu@mednet.ucla.edu

Conflicts: The authors declare no potential conflicts of interest.

Abstract

Purpose: To investigate the use of deep neural networks to learn associations between gene expression and radiomics or histology in non-small cell lung cancer (NSCLC).

Materials and Methods: Deep feedforward neural networks were used for radiogenomic mapping, where 21,766 gene expressions were inputs to individually predict histology and 101 CT radiomic features. Models were compared against logistic regression, support vector machines, random forests, and gradient boosted trees on 262 training and 89 testing patients. Neural networks were interpreted using gene masking to derive the learned associations between subsets of gene expressions to a radiomic feature or histology type.

Results: Neural networks outperformed other classifiers except in five radiomic features, where training differences were <0.026 AUC. In testing, neural networks classified histology with AUCs of 0.86 (adenocarcinoma), 0.91 (squamous), and 0.71 (other); and 14 radiomics features with ≥ 0.70 AUC. Gene masking of the models showed new and previously reported histology-gene or radiogenomic associations. For example, hypoxia genes could predict histology with >0.90 test AUC and published gene signatures for histology prediction were also predictive in our models (>0.80 test AUC). Gene sets related to the immune or cardiac systems and cell development processes were predictive (>0.70 test AUC) of several different radiomic features while AKT signaling, TNF, and Rho gene sets were each predictive of tumor textures.

Conclusion: We demonstrate the ability of neural networks to map gene expressions to radiomic features and histology in NSCLC and interpret the models to identify predictive genes associated with each feature or type.

Author Summary

Non-small-cell lung cancer (NSCLC) patients can have different presentations as seen in the CT scans, tumor gene expressions, or histology types. To improve the understanding of these complementary data types, this study attempts to map tumor gene expressions associated with a patient's CT radiomic features or a histology type. We explore a deep neural network approach to learn gene-radiomic associations (i.e., the subsets of co-expressed genes that are predictive of a value of an individual radiomic feature) and gene-histology associations in two separate public cohorts. Our modeling approach is capable of learning relevant information by showing the model can predict histology and that the learned relationships are consistent with prior works. The study provides evidence for coherent patterns between gene expressions and radiomic features and suggests such integrated associations could improve patient stratification.

1 Introduction

Personalized medicine has driven the study of high-throughput profiling of both molecular and medical imaging data to discern survival, treatment outcomes, and subtyping in cancer patients. In particular, radiogenomic studies attempt to integrate the two complementary data types to explain tumor imaging patterns using molecular information and vice versa. For example, radiogenomic studies have shown molecular patterns (e.g., gene expression, gene mutation, or molecular subtype) can be predicted from imaging features (e.g., computed tomography (CT) or magnetic resonance (MR) images) [1, 2, 3, 4, 5, 6]. Radiogenomic studies support the derivation of tumors' biological states from noninvasive imaging and the association mapping of molecular and imaging observations to better understand cancer heterogeneity. However, radiogenomic studies are often limited by the use of feature selection, linearity assumptions, and lack of testing datasets [7].

Deep learning techniques have been widely used on molecular and imaging datasets due to their ability to have high-dimensional input without the need for engineered features, and model nonlinear and hierarchical relationships. Several studies have used deep learning models such as convolutional neural networks, generative adversarial networks, and autoencoders for radiogenomic mapping [5, 8, 9, 6]. However, such works focus on accurate prediction of imaging features or imaging characteristics and often do not provide insight into the biological differences leading to accurate predictions by the models. While classification performance is a factor, it is also of interest to probe the trained models to validate the learned radiogenomic associations and to discover new hypotheses.

In our previous work, we addressed the model understanding challenge by presenting methods, such as gene masking to interpret trained neural networks [10]. We showed the models were capable of learning radiogenomic associations that were consistent with prior work while also generating new associations for further consideration. A limitation of [10] is that the analysis was performed using a single dataset from glioblastoma patients. We did not demonstrate the generalizability of our approach in other domains. Therefore, the purpose of this study was to investigate the ability of neural networks to model gene expression in different radiogenomic datasets, such as in a different cancer domain in which there are multiple different histologies or stage, there are computationally derived imaging features such as textures, and there is an external validation or testing dataset available.

Here, we present deep feedforward neural networks to model transcriptomes using two similarly derived radiogenomic datasets recently published in non-small cell lung cancer (NSCLC) [11]. As one of the few publicly released radiogenomic datasets available, the paper reported radiogenomic associations and provided a basis for comparison. First, we evaluate the ability of neural networks to separately predict each of two clinical traits and 101 radiomic features using a transcriptome consisting of 21,766 gene expressions in a training dataset of 262 patients. Next, we demonstrate the generalizability of our neural network models in the external dataset of 89 patients. Finally, we systematically probe the trained neural networks to define specific patterns of gene expression related to a clinical trait or radiomic feature. We compare the model's learned relationships to previously published associations in similar work.

2 Materials and methods

Figure 1: An overview of this study's approaches to (a) training and (b) interpretation radio-genomic neural networks.

2.1 Data

All transcriptomes, radiomics, and clinical data were downloaded from [11], see Supp. Table S1. Dataset1 (262 patients from a North American institution) and Dataset2 (89 patients from a European institution) were used as the training and testing datasets. Briefly, each transcriptome was measured on the same Affymetrix microarray chip and listed 21,766 gene expressions. Clinical stage referred to pathologic TNM staging and was defined as I, II, III or other. Pathologic histology was defined as adenocarcinoma, squamous carcinoma, or other. Ra-

106 diomic features were extracted from three-dimensional tumor volumes in contrast-enhanced
107 presurgical CT scans to determine histogram statistics; morphology; textures, such as gray-
108 level co-occurrence matrix (GLCM), gray-level run-length matrix (GLRL, called RLGL), gray-
109 level size-zone matrix (GLSZM); Laplacian of Gaussian (LoG) transformations; and wavelet
110 decompositions.

111 In this study, each radiomic feature was a continuous variable that was transformed into
112 binary classes using k-means clustering in the training dataset. Dichotomizing patients into two
113 groups allowed us to focus on broad differences in the cohort for this initial work. For example,
114 surface area to volume ratio (one radiomic feature) was grouped into two clusters: cluster A
115 had a mean ratio of 0.26, and cluster B had a mean ratio of 0.52. Each cluster represented
116 one class. Clustering was performed via KMeans from sklearn. Radiomic features with class
117 frequency of $> 10\%$ were kept and resulted in 101 radiomic features. Clusters in the testing
118 were defined by the clusters created in training. Distributions and correlations for classes are
119 shown in Supp. Figs. S1–S2. For more details about the radiomic features, see Supp. Methods
120 and [11, 12, 13].

121 **2.2 Radiogenomic modeling**

122 Dense feed-forward neural networks were used to map transcriptomes (model inputs) to ra-
123 diomic features (model outputs). Gene expression was standardized by mean subtraction and
124 standard deviation division for each gene. Layers within the radiogenomic models were con-
125 structed with three or four hidden layers, where the first hidden layer had either 2000, 4000,
126 or 6000 nodes. The number of hidden nodes was halved with each subsequent hidden layer.
127 Neural networks were trained using cross-entropy loss, nonlinear activation functions, dropout,
128 batch normalization, and early stopping by monitoring loss.

129 Performance was measured with the receiver operating characteristic curve via the area
130 under the curve (AUC). Hyperparameters were grid searched and the hyperparameters with
131 the highest mean AUC in 10-fold cross-validation in the training dataset were selected. In
132 cross-validation, metrics for gene standardization were based on the training folds for each
133 split; in testing, metrics were based on the entire training dataset. Training performances were
134 averaged cross-validation folds. Accuracy was calculated at a threshold of 0.5 class probability.

135 Figure 1a depicts the overall procedure used to train radiogenomic neural networks. To

136 reduce the hyperparameter search for each radiomic feature, the grid search was applied in a
 137 neural network that predicted a *group* of radiomic features simultaneously, based on type of
 138 radiomic feature. A group was defined as a set of related features; the seven different groups
 139 are shown in Supp. Figs. S2. For example, the model would take a patient's transcriptome and
 140 predict all 22 GLCM features at once as a multi-label classification task. One neural network was
 141 trained for each radiomic group. The best performing hyperparameters for a group were then
 142 used in a neural network to predict a *single* radiomic feature within the group. Other classifiers,
 143 including logistic regression, support vector machines, random forest, and gradient boosted
 144 trees were also considered. Each alternate model was trained to predict a single radiomic
 145 feature and evaluated against a radiogenomic neural network. The best performing models for
 146 each radiomic feature classification were then retrained on the entire training dataset to obtain
 147 the final models. Final models were then evaluated with the testing dataset. Radiomic features
 148 with at least 0.70 test AUC were kept for further analysis.

149 More details on the architecture of radiogenomic neural networks are shown in Supp. Fig.
 150 S3. All models and their hyperparameters are listed in Supp. Table S2.

151 **2.3 Histology and stage modeling**

152 Transcriptomes were also mapped to histology and stage in two additional neural networks
 153 constructed with a similar architecture as the radiogenomic neural networks. Instead of pre-
 154 dicting radiomic feature classes, models were trained to predict history or stage. Histology and
 155 stage were each modeled as multiclass classifications (Table S1), where the neural networks
 156 used categorical entropy loss and softmax activation in the prediction layer. Train scores were
 157 based on micro-averages across all classes and folds in cross-validation. Test scores and model
 158 interpretation methods were based on one class versus all other classes unless otherwise noted.
 159 The models were also trained using an approach similar to the radiogenomic modeling, namely
 160 the methods for hyperparameter and model selection via grid search in the training set and the
 161 comparison with other classifiers.

162 **2.4 Extracting predictive gene expression patterns from neural networks**

163 Figure 1b illustrates the gene masking steps used to extract predictive gene expression patterns
 164 from trained neural networks. Predefined gene sets were used for gene masking. These in-

cluded the Hallmark and Gene Ontology (GO) biological processes from Molecular Signature Database (MSigDB) v7.0 [14]. Gene masking focused on a component of the model's input by using one predefined gene set at a time, a process related to sensitivity analysis [15]. For example, masking of genes related to hypoxia involved taking each patient's gene expression profile, keeping only the hypoxia genes' expressions and setting all other gene expressions to zero (i.e., the input is masked). The masked input was then pushed through the model and the model's prediction of a radiomic feature was recorded. This process was repeated for the entire cohort and classification performance was calculated. Gene masking measured the model's ability to predict radiomic features based on gene expressions of a particular gene set, where the higher the performance the stronger the association in the cohort, as initially used in [10]. Masking resulted in "radiogenomic associations" learned by the model. The strength of the associations was by test AUCs, which represented how well a gene set could retain the classification performance in a radiogenomic neural network. For simplicity, gene sets with a maximum of 500 genes were studied.

Relationships between gene expression and histology were also studied using gene masking. Gene expressions were previously used to predict histology in NSCLC. In one work, a 42-gene signature was used to distinguished adenocarcinomas (ADC) from squamous cell carcinoma (SCC) [16]. In another similar effort, a 75-probe set signature was found for ADC, SCC, and large cell carcinoma (LCC) [17]. These two signatures were defined as four additional gene sets for comparison in this work.

3 Results

3.1 Predicting histology or stage from the transcriptome

3.1.1 Neural networks outperform other models

In predicting histology and stage, most models achieved cross-validated AUC above 0.5; neural networks achieved the lowest error in the training dataset (Figure 2a). In testing, the neural network was unable to estimate stage with a test AUC equal to random guessing (0.5 AUC; Figure 2b). However, the histology model was able to generalize to the testing dataset and achieved the respective test scores of 0.86 AUC, 0.91 AUC, and 0.71 AUC in predicting adenocarcinomas, squamous cell carcinoma, or other histology, respectively; this resulted in an

Figure 2: The ability of models to predict NSCLC histology and stage in (a) training and (b) testing. Gene masking of the histology neural network using gene sets from (c) published gene signatures for histology [16, 17], (d) Hallmarks (top five out of 50) [18], and (e) Gene Ontology biological processes (GO.bp, top ten out of 7350) [14]. CV=cross-validation.

overall score of 0.77 test AUC when micro-averaged across classes. For this reason, the histology neural network was further analyzed.

3.1.2 Gene masking identities gene sets that predict histology

Gene masking of the histology neural network showed agreement with previously published gene signatures [16, 17] for predicting NSCLC adenocarcinomas (ADC) and squamous cell carcinoma (SCC). As shown in (Figure 2c), the aforementioned gene signatures were also found to be predictive in our histology neural network in the testing dataset. In particular, the gene signature from [16] resulted in 0.93 test AUC in both ADC and SCC.

Hallmark gene sets were also indicative of histology (Figure 2d). Gene expressions related to hypoxia, coagulation, and KRAS signaling could predict both ADC and SCC (>0.90 test AUC).

Similar to the overall test performances observed in Figure 2b, the histology neural network was driven by accurate predictions in ADC and SCC, where test performances were about 0.20 AUC higher than the "other histology" class. Subsequently, this behavior was reflected in gene masking, where the most predictive gene sets for estimating "other histology" in testing were inflammatory response (0.73 AUC) and angiogenesis (0.72 AUC). Notably, angiogenesis was more predictive of other histologies (0.73 AUC) than ADC (0.66 AUC) or SCC (0.63 AUC) classes in testing.

The most predictive GO biological processes (Figure 2e) were also associated with angiogenesis, epithelial mesenchymal transition, and hypoxia from the Hallmark gene sets. Of the gene sets considered in gene masking, negative regulation of DNA-binding transcription factor activity from GO had the best overall testing performance with the highest micro-averaged AUC of 0.79; the individual test AUCs were 0.93 in ADC, 0.91 in SCC, and 0.75 in other. The aforementioned gene set consisted of 170 genes, where 156 (91.2%) were in the gene expression profile. A summary of gene expression patterns is in Table 1. Notably, *KRAS* is a major gene studied in NSCLC and the *KRAS* Hallmark was found to be predictive of ACC and SCC.

3.2 Predicting CT imaging features from the transcriptome

3.2.1 Performance

Neural networks were overall better at classifying radiomic features than all other models within the training dataset (Figure 3a). The only exceptions were in gradient boosted trees that had better performance in four radiomic features (differences below 0.026 AUC) and random forest in one radiomic feature (0.012 AUC difference). In testing, neural networks had 0.42–0.89 AUC, 0.45–0.94 accuracy, and 0.09–0.98 average precision across all radiomic feature classifications (Supp. Fig. S4). A subset of 13 radiomic features had at least 0.70 testing AUC and was subsequently selected for interpretation. Figure 3b depicts the neural network's generalizability to classify the aforementioned 13 radiomic features in the testing dataset.

Table 1: Summary of predictive gene expression patterns in NSCLC histology.

histology	transcriptomic pattern			test	
	theme †	gene set	source*	AUC	AP
adenocarcinoma	synthesis	phosphatidylcholine biosynthetic process	G	0.94	0.93
		phosphatidic acid biosynthetic process	G	0.94	0.86
	transcription	neg. regulation of DNA-binding transcription factor activity	G	0.93	0.92
	vasculature	neg. regulation of sprouting angiogenesis	G	0.93	0.92
		coagulation	H	0.91	0.86
	cell development	retina development in camera-type eye	G	0.93	0.87
	cell death	hypoxia	H	0.91	0.88
	KRAS	KRAS signaling down	H	0.91	0.84
	UV	UV response down	H	0.92	0.90
	E2F	E2F targets	H	0.92	0.87
	cell development	metencephalon development	G	0.95	0.90
	differentiation	epidermal cell differentiation	G	0.94	0.94
		neuron fate commitment	G	0.92	0.87
	catabolism	neg. regulation of cellular catabolic process	G	0.94	0.88
squamous cell carcinoma	cell death	cornification	G	0.94	0.93
		hypoxia	H	0.91	0.90
	KRAS	KRAS signaling down	H	0.93	0.93
	hormone	estrogen response late	H	0.92	0.91
	cholesterol	cholesterol homeostasis	H	0.92	0.83
	transport	posttranslational protein targeting to membrane, translocation	G	0.86	0.35
		regulation of sodium ion transmembrane transporter activity	G	0.84	0.54
	AMPA receptor	regulation of AMPA receptor activity	G	0.85	0.44
	cell cycle	mitotic nuclear envelope reassembly	G	0.84	0.50
	ubiquitination	neg. regulation of protein K63-linked ubiquitination	G	0.84	0.38
other	immune system	inflammatory response	H	0.73	0.37
	vasculature	angiogenesis	H	0.72	0.35

† Gene sets were categorized by comparing the top five gene sets in GO and Hallmark with at least 0.70 test AUC.

* from MSigDB v7.0, where H=hallmarks, G=Gene Ontology. neg.=negative.

3.2.2 Interpretation of radiogenomic associations and comparison with prior modules

Figure 4 illustrates the top GO gene sets associated with predicting each radiomic feature. Overall, gene masking suggested each radiomic feature is associated with a unique gene expression profile driven by different biological processes as none of the radiomic features had similar scores across all gene sets. Some gene sets were better for predicting one radiomic feature but not another. For example, the top two gene sets for predicting RLGL_longRunHighGrayLevEmpha, an imaging texture, were regulation of syncytium formation by plasma membrane fusion and pyrimidine nucleotide salvage had test AUCs above 0.75 but for all other 12 radiomic features these two gene sets were below 0.70 AUC and 0.65 AUC, respectively. Conversely, there were gene sets that could predict multiple radiomic features at once. For example, response to tumor necrosis factor was predictive of GLCM_entrop2 (0.78 AUC, 0.81 AP) and RLGL_runPercentage (0.76 AUC, 0.76 AP) in testing. Hallmark gene sets were also applied to the radiomic models in gene masking analysis (Supp. Fig. S6) but were not as predictive as the GO gene sets.

Radiogenomic associations were summarized for radiomic features related to histogram

Figure 3: Radiogenomic modeling performance (a) between neural networks and other models in training, and (b) neural network performance in training and testing datasets of the 13 radiomic features selected for further analysis.

statistics of either the tumor mask, the transformation of the mask (LoG features), and textures of tumors in Table 2. The radiomic feature, LoG_stats_std was completely predictable in the test cohort with the gene set post-translational protein modification and the gene sets related to the epidermis development. Processes involving immune system and cardiac system were top predictors for several radiomic features. Many gene sets were related to cell development, varying from muscle, liver, epidermis, fat cell, and renal gene sets. AKT signaling, a targeted pathway in NSCLC therapy [19] was somewhat predictive in RLGL_longRunHighGrayLevEmpha with 0.76 AUC but a lower 0.54 average precision in testing. TNF was associated with RLGL_runPercentage and GLCM_entrop2. Rho signaling, associated with tumor suppressor activity and another targeted pathway in NSCLC [20, 21] was associated with GLCM_diffEntro (0.75 AUC; 0.78 AP) and GLCM_invDiffmomnor (0.79 AUC; 0.55 AP). While GLCM_invDiffmomnor and invDiffnorm were correlated and clustered together (Supp. Fig. S1), the two radiomic features had differing gene sets.

Radiogenomic modules, defined as a set of correlated radiomic features and gene expressions were previously found in the same dataset [11]. Notably, not all the radiomic features were in those reported in their study's radiogenomic modules. Here, the radiomic models were

Figure 4: Gene masking of the radiogenomics models with biological processes from Gene ontology (GO). The top three ranked by test AUC for each radiomic feature are shown. There are 40 gene sets.

masked with the same Reactome gene sets from MSigDB v4.0 (Supp. Fig. S7). Supp. Table S3 summarizes the overlapping radiogenomic associations found in this study compared to [11]. The best agreement was between LoG_stats_entropy and module 13, where three of the pathways in the module were also among the top ten most predictive gene sets in our radiogenomic model. Other comparisons did not have overlapping associations. For example, the authors reported a radiogenomic association between GLCM_diffEntro and the five pathways in module 2, while we found the most predictive pathway in module 2 was ranked 197

268 out of the 664 Reactome pathways used in gene masking. Thus, our model suggested many
269 other pathways were more predictive of the radiomic feature than the five pathways in module
270 2.

271 4 Discussion

272 We demonstrate the ability of deep neural networks to learn associations between radiomic
273 features or clinical traits and gene expressions using two NSCLC cohorts. The datasets had a
274 relatively large training dataset of 262 patients, and another dataset of 89 patients that allowed
275 us to validate the generalizability of our neural network models. We show neural networks
276 outperformed other machine learning models and had test performances above 0.70 AUC in
277 predicting thirteen radiomic features and histology. We interpreted the models using gene
278 masking and identified specific sets of gene expressions that were indicative of a trait or feature.
279 Together, these results suggest potential biological associations exist to explain the differences
280 among histology classes and CT imaging characteristics of NSCLC patients.

281 Previous radiomic and radiogenomic efforts have recently been studied in NSCLC [12, 22,
282 23, 24, 11, 25, 9]. Molecular statuses are predicted using radiomic or other imaging features
283 in [12, 22, 24, 25]. A recent work used the same dataset of 89 to train models to predict
284 immune cell gene signatures of NSCLC tumors using CT radiomic features [26]. Compared to
285 these works, our model uses whole gene expression profiles to map to radiomic features. While
286 deep neural networks are used to map CT image patches to tumor gene expressions in [9], the
287 study does not report specific radiogenomic associations. Most related is the work from [11]
288 (the source of the two NSCLC datasets), who use Gene Set Enrichment Analysis and the Iter-
289 ative Signature Algorithm, a correlation and biclustering method to group radiomic features
290 with Reactome gene sets (i.e., pathways) into radiogenomic modules. The top ten most pre-
291 dictive Reactome pathways for a radiomic feature in our models have some overlap with the
292 radiogenomic modules created in [11] but this is often not the case. These results indicate the
293 studies identified different radiogenomic associations in the same data. Differences in radio-
294 genomic associations between our work and these authors [11] may relate to differences in
295 methodologies. Most importantly, our study assessed entire gene profiles, whereas their study
296 assessed correlations between sets of radiomics features and sets of genes. As well, we did

Table 2: Summary of predictive gene expression patterns in NSCLC radiomic features.

radiomic feature	transcriptomic pattern		test	
	theme †	gene set (from GO)	AUC	AP
Stats_skewness	cytoskeleton	reg. of actin filament-based process	0.82	0.82
		neg. reg. of cell adhesion	0.81	0.81
	immune system	neg. reg. of cell-cell adhesion	0.80	0.77
		reg. of hemopoiesis	0.81	0.78
		reg. of leukocyte differentiation	0.80	0.77
Stats_rms	transport	reg. of release of sequestered calcium ion into cytosol	0.95	0.65
		sequestering of calcium ion	0.93	0.30
	development	muscle organ development	0.93	0.46
		striated muscle cell differentiation	0.93	0.32
		actin filament-based movement	0.92	0.26
LoG_stats_std	post-translational	post-translational protein modification	1.00	1.00
	development	epidermis development ; epidermal cell differentiation	1.00	1.00
	DNA repair	DNA double-strand break processing involved in repair via single-strand annealing	0.99	0.79
LoG_stats_uniformity	cell cycle	neg. reg. of mitotic cell cycle	0.99	0.79
	cell development	liver regeneration	0.78	0.64
		epithelial tube morphogenesis	0.77	0.49
	transport	protein transmembrane transport	0.77	0.64
		intracellular protein transmembrane transport	0.76	0.61
LoG_stats_entropy	catabolism	organic acid catabolic process	0.75	0.54
	localization	establishment of organelle localization	0.82	0.47
		pos. reg. of protein localization to membrane	0.79	0.60
	heart rate	reg. of heart rate by cardiac conduction	0.77	0.40
	cell mobility	reg. of actin filament-based process	0.77	0.46
LoG_stats_kurtosis	external stimulus	cellular response to mechanical stimulus	0.77	0.44
	connective tissue	elastin metabolic process	0.73	0.84
		collagen metabolic process	0.72	0.85
	synthesis	pos. reg. of receptor biosynthetic process	0.73	0.81
		pos. reg. of hormone biosynthetic process	0.73	0.84
GLCM_diffEntro	immune system	response-regulating cell surface receptor signaling pathway	0.72	0.78
	muscle	muscle fiber development	0.76	0.82
		muscle cell differentiation	0.75	0.79
		cardiac ventricle formation	0.76	0.77
	bacteria	response to molecule of bacterial origin	0.76	0.73
GLCM_invDiffnorm	Rho	Rho protein signal transduction	0.75	0.78
	cell development	fat cell differentiation	0.83	0.63
		neg. reg. of cell development	0.80	0.60
	cell respiration	reg. of aerobic respiration	0.81	0.64
	immune system	neg. reg. of lymphocyte activation	0.80	0.63
GLCM_invDiffmomnor	nervous system	neuromuscular process controlling balance	0.79	0.70
	immune system	reg. of osteoclast differentiation	0.81	0.58
		osteoclast differentiation	0.80	0.58
	homeostasis	multicellular organismal homeostasis(G)	0.80	0.61
		tissue homeostasis	0.79	0.57
GLCM_entrop2	Rho	reg. of Rho protein signal transduction	0.79	0.55
	TNF	response to tumor necrosis factor	0.78	0.81
	muscle	muscle cell development	0.77	0.80
		ventricular septum morphogenesis	0.77	0.77
		striated muscle cell differentiation	0.76	0.78
RLGL_shortRunEmphasis	drug response	response to xenobiotic stimulus	0.76	0.84
	localization	pos. regulation of establishment of protein localization	0.79	0.77
	catabolism	lysine catabolic process	0.79	0.80
	cell death	pos. reg. of autophagy of mitochondrion	0.77	0.78
	hormone	pos. reg. of insulin secretion	0.77	0.77
RLGL_longRunHighGrayLevEmpha	cell mobility	neuron projection guidance	0.75	0.71
	syncytium	syncytium formation	0.78	0.54
		reg. of syncytium formation by plasma membrane fusion	0.77	0.47
	synthesis	pyrimidine nucleotide salvage	0.76	0.50
	AKT	protein kinase B signaling	0.76	0.54
RLGL_runPercentage	renal system	renal sodium excretion	0.76	0.44
	muscle	smooth muscle tissue development	0.76	0.75
		response to tumor necrosis factor	0.76	0.76
	TNF	tumor necrosis factor-mediated signaling pathway	0.75	0.76
		myeloid leukocyte differentiation	0.75	0.76
	immune system	myeloid leukocyte differentiation	0.75	0.76
	renal system	metanephros development	0.74	0.80

†Shown are the top five GO gene sets ranked by AUC and with at least 0.50 AP; reg.=regulation; pos.=positive; neg.=negative.

not predefine how many associations may be found prior to mapping or modeling; their study attempted to assess 20 radiogenomic modules. Finally, we used a classification (prediction) approach, whereas they used a correlation matrix to identify associations. Additionally, radiomic signatures are reportedly predictive of survival in [23, 11]. It is a part of future work to evaluate if radiogenomic relationships found in our neural network models can predict patient survival.

Compared to our prior radiogenomic work in glioblastoma (a high-grade brain tumor), NSCLC is a cancer with multiple stages and histology types. We further explore the capability of neural networks to map transcriptomes to other relevant patient imaging features by training models to predict stage and histology. While neural networks are better at predicting stage and histology in the training dataset compared to other classifiers, stage is poorly estimated in the testing dataset. However, the histology neural network has 0.77 test AUC when averaged across each histology type.

Several prior works for predicting NSCLC histology using radiomics have been based on differentiating adenocarcinoma from squamous cell carcinoma. For example, using five radiomic features a study trained a Naive Bayes model to distinguish adenocarcinoma from squamous cell carcinomas and achieved a 0.72 AUC in a test set of 152 patients [27]. In another similar study, a radiomic signature was able to distinguish adenocarcinoma from squamous carcinomas using a logistic regression model and 129 patients; the authors reported a 0.893 AUC in a test subset of 48 patients [28]. More recent work reported logistic models that achieved 0.694, 0.780, 0.800, and 0.923 AUC for clinical, standard CT features, radiomics, or all three to respectively predict adenocarcinoma versus squamous cell carcinomas [29]. The AUC scores were observed in a cohort of 93 patients but are likely overly optimistic given there was no validation or test set to evaluate their models. In contrast, our neural network model uses gene expression profiles to predict histology. In a test set of 89 patients, our model achieves a 0.86 test AUC when estimating adenocarcinoma versus all other (squamous and otherwise) histology types and a 0.91 test AUC when estimating squamous versus all other types (adenocarcinoma and otherwise).

Additionally, there is a notable difference in our models' performances when predicting histology compared to stage. The difference is likely because staging is based on factors such as tumor size, location, and spread to lymph nodes or metastasis in other sites, which is infor-

328 mation that may not be readily seen in gene expressions of the collected tumor tissue. On the
329 other hand, histology classification is based on the molecular and physical characterization of
330 a collected sample, which is likely more related to the transcriptome profiling of the tumor.
331 Subsequently, we extract the learned associations between gene expression and histology types
332 with gene masking and compare against two previous studies who report gene expression sig-
333 natures to predict NSCLC histology [16, 17]. The studies' gene signatures are predictive in our
334 models as well and indicate neural networks have found similar associations in prior work.
335 Other gene sets such as hypoxia (200 genes) and angiogenesis (36 genes) Hallmark genes sets
336 and the neuronal (156 genes) GO gene set are associated with histology in our model and may
337 have the potential to automate/standardize the histology determination process.

338 This study has several notable limitations. The retrospective datasets used in this analy-
339 sis were from two different sources with varying imaging protocols, which can add variance
340 to computationally extracted imaging features such as radiomics. A majority of patients were
341 early stage cancers and thus late stage patients were left out of the radiogenomic associations.
342 While sample size is an inherent concern of radiogenomic analysis, the dataset is one of the
343 few publicly available. Still, radiomic features are difficult to interpret, and researchers are cur-
344 rently attempting to understand their correlation with tumor biology and other clinical traits.
345 Radiomic features were binarized using clustering, and radiomic features with a minority class
346 below 10% were removed. The tumor tissue samples used for transcriptome profiling are lim-
347 ited in that only one sample was acquired per patient and may not fully capture tumor het-
348 erogeneity. These factors make it challenging to validate the relationships in the radiogenomic
349 neural networks. The reported findings are hypotheses about radiogenomic associations, are
350 not proven here, and would require cell or animal studies.

351 In future work, the issues may be addressed by the standardization of imaging protocols to
352 allow consistent comparison of imaging features and maps such as genomic atlases to better
353 characterize whole tumors. A larger sample size could result in a more complete representation
354 of the general population and allow modeling of radiomic features as continuous outputs. The
355 ranked gene sets are interpreted based on GO descriptions. A more quantitative approach to
356 compare the ranked gene sets between radiomic features, such as semantic similarity of GO
357 terms [30], and a sensitivity analysis of resultant radiogenomic associations are needed. Out-
358 side of the transcriptome, there are likely other contributing factors to tumor imaging features,

such as other molecular data (e.g., gene mutations and methylation) and patient covariates (e.g., smoking status). These factors could either be incorporated into the modeling process or used to stratify analyses. Additionally, the impact of knowing such radiogenomic associations at the time of tumor biopsy in relation to survival or treatment projection would be highly beneficial.

5 Conclusion

In this study, we present a machine learning approach based on neural networks for mapping gene expressions to radiomic features or clinical traits for patients with non-small cell lung cancer. Our models are evaluated through previously published training and testing datasets to show such neural networks are capable of modeling high-dimensional gene expression to predict tumor imaging features and lung cancer histology. We further interpret the models through gene masking and report the learned relationships between gene expressions and a radiomic feature or gene expressions and a histology type, which include both new and previously reported findings in prior works. The predictive relationships found in this work could be further studied to develop automated classification of histology, non-invasive imaging surrogates for gene expressions, or individualized patient prognosis or treatment therapies.

6 Supporting Information

References

- [1] Segal, E. *et al.* Decoding global gene expression programs in liver cancer by noninvasive imaging. *Nat. Biotechnol.* **25**, 675–80 (2007).
- [2] Diehn, M. *et al.* Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 5213–8 (2008).
- [3] Zinn, P. O. *et al.* Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme. *PLoS One* **6**, e25451 (2011).
- [4] Colen, R. R. *et al.* Imaging genomic mapping of an invasive MRI phenotype predicts patient outcome and metabolic dysfunction: A TCGA glioma phenotype research group project. *BMC Med. Genomics* **7**, 30 (2014).
- [5] Chang, K. *et al.* Residual convolutional neural network for the determination of IDH status in low- and high-grade gliomas from MR imaging. *Clin. Cancer Res.* **24**, 1073–1081 (2018).
- [6] Ha, R. *et al.* Predicting breast cancer molecular subtype with MRI dataset utilizing convolutional neural network algorithm. *J. Digit. Imaging* **32**, 276–282 (2019).

- 391 [7] Bai, H. X. *et al.* Imaging genomics in cancer research: limitations and promises. *Br. J.*
392 *Radiol.* **89**, 20151030 (2016).
- 393 [8] Korfiatis, P. *et al.* Residual deep convolutional neural network predicts MGMT methylation
394 status. *J. Digit. Imaging* **30**, 622–628 (2017).
- 395 [9] Li, S., Han, H., Sui, D., Hao, A. & Qin, H. A novel radiogenomics frame-
396 work for genomic and image feature correlation using deep learning. In
397 *2018 IEEE Int. Conf. Bioinforma. Biomed.*, 1, 1–8 (IEEE, 2018). URL
398 <https://doi.org/10.1109/BIBM.2018.8621432>.
- 399 [10] Smedley, N. F., El-Saden, S. & Hsu, W. Discovering and interpreting transcriptomic
400 drivers of imaging traits using neural networks. *arXiv:1912.05071 [q-bio.QM]* (2019).
401 1912.05071.
- 402 [11] Grossmann, P. *et al.* Defining the biological basis of radiomic phenotypes in lung cancer.
403 *Elife* **6** (2017).
- 404 [12] Aerts, H. J. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative
405 radiomics approach. *Nat. Commun.* **5**, 4006 (2014).
- 406 [13] Thibault, G., Angulo, J. & Meyer, F. Advanced statistical matrices for texture characteri-
407 zation: Application to cell classification. *IEEE Trans. Biomed. Eng.* (2014).
- 408 [14] Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–
409 1740 (2011).
- 410 [15] Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional
411 networks. In *Eur. Conf. Comput. Vis.*, 818–833 (Springer, 2014). URL
412 https://doi.org/10.1007/978-3-319-10590-1_53.
- 413 [16] Girard, L. *et al.* An expression signature as an aid to the histologic classification of non-
414 small cell lung cancer. *Clin. Cancer Res.* **22**, 4880–4889 (2016).
- 415 [17] Hou, J. *et al.* Gene expression-based classification of non-small cell lung Carcinomas and
416 survival prediction. *PLoS One* **5**, e10312 (2010).
- 417 [18] Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell*
418 *Syst.* **1**, 417–425 (2015).
- 419 [19] Heavey, S., O’Byrne, K. J. & Gately, K. Strategies for co-targeting the PI3K/AKT/mTOR
420 pathway in NSCLC. *Cancer Treat. Rev.* **40**, 445–456 (2014).
- 421 [20] Zhang, B. Rho GDP dissociation inhibitors as potential targets for anticancer treatment.
422 *Drug Resist. Updat.* **9**, 134–141 (2006).
- 423 [21] O’Hayre, M., Degese, M. S. & Gutkind, J. S. Novel insights into G protein and G protein-
424 coupled receptor signaling in cancer. *Curr. Opin. Cell Biol.* **27**, 126–135 (2014).
- 425 [22] Yamamoto, S. *et al.* ALK molecular phenotype in non-small cell lung cancer: CT radio-
426 genomic characterization. *Radiology* **272**, 568–576 (2014).
- 427 [23] Aerts, H. J. *et al.* Defining a radiomic response phenotype: A pilot study using targeted
428 therapy in NSCLC. *Sci. Rep.* **6**, 33860 (2016).
- 429 [24] Rizzo, S. *et al.* CT radiogenomic characterization of EGFR, K-RAS, and ALK mutations in
430 non-small cell lung cancer. *Eur. Radiol.* **26**, 32–42 (2016).

- 431 [25] Gevaert, O. *et al.* Predictive radiogenomics modeling of EGFR mutation status in lung
432 cancer. *Sci. Rep.* **7**, 41674 (2017).
- 433 [26] Yoon, H. J. *et al.* Deciphering the tumor microenvironment through radiomics in non-
434 small cell lung cancer: Correlation with immune profiles. *PLoS One* **15**, e0231227 (2020).
- 435 [27] Wu, W. *et al.* Exploratory study to identify radiomics classifiers for lung cancer histology.
436 *Front. Oncol.* **6**, 71 (2016).
- 437 [28] Zhu, X. *et al.* Radiomic signature as a diagnostic factor for histologic subtype classification
438 of non-small cell lung cancer. *Eur. Radiol.* **28**, 2772–2778 (2018).
- 439 [29] Digumarthy, S. R., Padole, A. M., Gullo, R. L., Sequist, L. V. & Kalra, M. K. Can CT radiomic
440 analysis in NSCLC predict histology and EGFR mutation status? *Medicine (Baltimore)*. **98**,
441 e13963 (2019).
- 442 [30] Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO
443 terms and gene products. *Bioinformatics* **26**, 976–978 (2010).
- 444 [31] Kotikalapudi, R. & Others. keras-vis (2017). URL
445 <https://github.com/raghakot/keras-vis>.

Supplemental materials

1 Supplemental methods

A total of 101 radomic features were considered in this study out of the 636 from [11]. These included: 10 shape, 22 GLCM, 7 GLSZM, 10 RLGL, 12 stats, 9 LoG, and 31 wavelet features that had at least 10% of the training dataset with the least frequent class. LoG and wavelet features were selected from “LoG_sigma_0_5_mm_2D” and the “HHH” wavelet decomposition; the rest could be explored in future studies. Radiomic features not defined in [11, 12] were not considered for model interpretation and survival analysis.

Parameters for class saliency values were set to default in `keras-vis` 0.4.1 [31], except for the input range (determined by the gene expression range in the dataset) and `backprop_modifier` set to `guided`.

2 Supplemental tables and figures

2.1 Dataset

Table S1: Patient characteristics, estimated or replicated from [11].

		training n=262	testing (n=89)
source	country	USA	Netherlands
gender	male	100 (38%)	59 (66%)
	female	124 (47%)	28 (32%)
	n/a	38 (15%)	2 (2%)
histology	adenocarcinoma	129 (49%)	42 (48%)
	squamous	61 (23%)	33 (38%)
	other	34 (13%)	12 (14%)
	n/a	38 (15%)	0 (0%)
stage	I	123 (47%)	39 (44%)
	II	35 (13%)	26 (29%)
	III	46 (18%)	12 (14%)
	other	20 (8%)	10 (11%)
	n/a	38 (14%)	2 (2%)
smoking status	current	66 (25%)	n/a
	former	141 (54%)	n/a
	none	17 (7%)	n/a
	n/a	38 (14%)	n/a
tumor site	primary	224 (86%)	87 (98%)
	n/a	38 (14%)	2 (2%)
status	alive	134 (51%)	41 (46%)
	deceased	90 (35%)	46 (52%)
	n/a	38 (14%)	2 (2%)
follow-up	median months	32	31

Figure S1: Correlation between radiomic features in Dataset1, the training dataset. **(a)** The subset of radiomic features selected based on classification performance in testing. **(b)** All radiomic features used in this study (n=101). Features were clustered with pheatmap in R using the Spearman rank correlation and default parameters.

Figure S2: Class distributions of radiomic features after clustering into two classes.

2.2 Modeling

Figure S3: The architecture and hyperparameter tuning of a radiogenomic neural network.

Table S2: NSCLC radiogenomic models and hyperparameters. Grid search was used for selecting hyperparameters

model type	# models	hyperparameter	values
logistic regression	100	penalty type L1 ratio solver	elastic net [0–1] saga
support vector machines	400	kernel C penalty	linear, poly, rbf, sigmoid log(-6)-log(6)
random forest	120	trees split criterion max. features max. depth	[50: 100: 3050] Gini, entropy $\sqrt{G}$, $\log_2(G)$ None
gradient boosted trees	150	trees max. depth learning rate	[50: 100: 1000) [1-3] [0.01: 0.1: 0.50]
neural network	48	hidden nodes‡ hidden layers optimizer activation dropout loss epochs (patience) batch	[6000-250] 3, 4 Nadam, Adadelta sigmoid, relu 0.4, 0.6 binary cross-entropy 500 (200) 10

‡ First hidden layers were either 2000, 4000, or 6000 nodes, where the number of hidden nodes for a layer were halved with each subsequent layer.

G: number of genes in transcriptome

Figure S4: Classification performance of radiogenomic neural networks used to predict 101 radiomic features in training and testing. Not all radiomic features had both binary labels in testing; therefore some metrics are missing. Accuracy was defined at a 0.5 threshold of predictions.

2.3 Gene masking

Table S3: A comparison of the learned radiogenomic associations extracted from our neural networks and the modules identified in [11]. Each module consisted of a set of Reactome pathways and a set of imaging features. Shown are the modules that included the radiomic features used in this study. If any module's set of pathways was ranked among the top 100 in gene masking, the top three pathways were listed.

radiomic trait	Reactome pathway	this study		[11]	
		test AUC	rank†	module	# pathways
GLCM_diffEntro	cross presentation of soluble exogenous antigens endosomes	0.58	197	2	5
	phase II conjugation	0.69	9	12	35
	regulation of mitotic cell cycle	0.66	21		
	ABCA transporters in lipid homeostasis	0.65	28		
LoG_stats_std	regulation of ornithine decarboxylase (ODC)	0.94	15	2	5
	cross presentation of soluble exogenous antigens endosomes	0.82	89		
	antigen processing cross presentation	0.80	106		
	cholesterol biosynthesis	0.75	150	6	7
	signaling by TGF beta receptor complex	0.83	81	8	17
	elongation arrest and recovery	0.82	92		
	mRNA splicing	0.75	142		
LoG_stats_entropy	elongation arrest and recovery	0.73	1	7	8
	mitochondrial protein import	0.63	74		
	RNA pol III chain elongation	0.58	180		
	elongation arrest and recovery	0.73	1	13	26
	RNA pol II pre transcription events	0.69	7		
	formation of RNA pol II elongation complex	0.69	9		
Stats_skewness	antigen processing cross presentation	0.62	139	2	5

† All Reactome pathways were ranked by test AUC.

Figure S5: Hallmark gene set masking in the histology model. The top five ranked by AUC for each histology type were shown. There are twelve hallmarks.

Figure S6: Hallmark gene set masking in the histology model. The top three ranked by AUC for each radiomic trait were shown. There are 21 hallmarks.

Figure S7: Reactome gene set masking in the histology model. The gene sets are from MSigDB v4.0 to match the pathways previously used in [11] to allow a comparison. The top five ranked by AUC for each histology type were shown. There are 53 pathways shown. There were 674 pathways total and 669 pathways with less than 500 genes.