

修士論文

畳み込みニューラルネットワークを用いた 肺ガンの形態特徴と遺伝子発現量の相関解析

出口 和樹

奈良先端科学技術大学院大学

先端科学技術研究科

情報理工学プログラム

主指導教員：金谷 重彦 教授

計算システムズ生物学研究室（情報科学領域）

令和 2 年 3 月 13 日提出

本論文は奈良先端科学技術大学院大学先端科学技術研究科に
修士(工学) 授与の要件として提出した修士論文である。

出口和樹

審査委員：	金谷 重彦	教授	(主指導教員，情報科学領域)
	佐藤 嘉伸	教授	(副指導教員，情報科学領域)
	小野 直亮	准教授	(副指導教員，情報科学領域)

畳み込みニューラルネットワークを用いた 肺ガンの形態特徴と遺伝子発現量の相関解析

出口和樹

内容梗概

ガンは細胞内の一部の遺伝子に変異することで細胞が異常に増殖し続けることで発生することが知られており,生物学的研究からそのメカニズムは明らかにされてきた。しかし,ガンの形状(細胞の凝集度,細胞質の形態,細胞核の形態)に影響を与える遺伝子はほとんど明らかにされていない。ガンの形状に影響を与える遺伝子を特定することで遺伝子発現情報から将来的なガン組織の形状を予測できるようになりガンの予防的治療が可能となる。本研究では教師ありの畳み込みニューラルネットワークと教師なしの畳み込みニューラルネットワークを用いて,高次元のガン組織学画像から低次元に空間的特徴や形態学的特徴を集約し,遺伝子発現量データとの相関を調べることでガン組織の形状に影響を与える遺伝子を特定することを目的とした。その結果,細胞増殖,細胞修復,免疫系で働く遺伝子の一部が主に組織形状に影響を与えていることが示唆された。

キーワード

畳み込みニューラルネットワーク,肺ガン,遺伝子発現量,

*奈良先端科学技術大学院大学先端科学技術研究科,修士論文,令和2年3月13日.

Correlation analysis between morphological features and gene expression levels of lung cancer using convolutional neural networks

Kazuki Deguchi

Abstract

It has been known that cancer occurs when cells in the cell continue to grow abnormally due to mutations in some genes in the cells, and biological studies have gradually revealed the mechanism. Genes that affect the shape of the cancer (cell cohesion, cytoplasmic morphology, cell nucleus morphology) are not well defined. Because the shape of cancer tissue can be predicted and preventive treatment of cancer is possible, in this study, we use supervised convolutional neural networks and unsupervised convolutional neural networks to obtain high-dimensional cancer histology images. By integrating spatial and morphological features in a low dimension and examining the correlation with gene expression data, we identified genes that affected the shape of cancer tissue. This suggests that some of the genes that work in cell proliferation, cell repair, and the immune system affect tissue morphology.

Keywords:

Convolutional neural network, lung cancer, gene expression

*Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology, March 13, 2020

目次

第 1 章 序論	1
1.1 研究背景	1
1.2 ガンのサブタイプ	2
1.3 Convolutional Neural Network (CNN)	3
1.3.1 CNN 構造	3
1.3.2 CNN における特徴集約部	4
1.4 遺伝子発現解析	5
1.5 病理組織画像解析	7
1.6 研究目的	8
第 2 章 データセット	9
第 3 章 提案手法	12
3.1 組織病理画像の特徴抽出法	12
3.1.1 AutoEncoder	12
3.1.1.1 概要	12
3.1.1.2 AutoEncoder 機構	13
3.1.1.3 事前学習(転移学習用の特徴抽出器の作成)	14
3.2 教師あり畳み込みニューラルネットワーク	15
3.2.1 概要	15
3.2.2 教師あり CNN による特徴抽出	15
3.3 教師なし畳み込みニューラルネットワーク	16
3.3.1 概要	16
3.3.2 Information Maximizing Self-Augmented Training(IMSAT)	16
3.3.3 Self-Augmented Training(SAT)	17
3.3.3.1 Virtual Adversarial Training(VAT)	18
3.3.4 IMSAT による特徴抽出	22
3.4 遺伝子発現量とガン形態特徴の相関探索	23
3.4.1 概要	23
3.4.2 相関係数計算手法	23
3.5 CNN 判断根拠の可視化	25

3.5.1 概要	25
3.5.2 Score-Weighted Class Activation Mapping (Score-CAM)	26
第 4 章 結果・考察	28
4.1 教師あり CNN	28
4.1.1 教師あり CNN 分類	28
4.1.2 教師あり CNN による遺伝子発現量と特徴量の相関探索	29
4.1.2.1 教師あり CNN による遺伝子発現量と特徴量の相関係数	29
4.1.2.2 正の相関を取る遺伝子と病理組織形態の関係	30
4.1.2.3 正の相関を取る遺伝子と病理組織形態の関係まとめ	39
4.1.2.4 負の相関を取る遺伝子と病理組織形態の関係	40
4.1.2.5 負の相関を取る遺伝子と病理組織形態の関係まとめ	45
4.1.2.6 教師あり CNN における相関探索まとめ	46
4.2 IMSAT による遺伝子発現量と特徴量の相関探索	46
4.2.1 IMSAT による遺伝子発現量と特徴量の相関係数	46
4.2.1.2 正の相関を取る遺伝子と病理組織形態の関係	48
4.2.1.3 正の相関を取る遺伝子と病理組織形態の関係まとめ	51
4.2.1.4 負の相関を取る遺伝子と病理組織形態の関係	52
4.2.1.5 負の相関を取る遺伝子と病理組織形態の関係まとめ	55
4.2.1.4 IMSAT における相関探索まとめ	55
4.3 まとめ	56
5 結論	57
6 謝辞	58
付録	65

図目次

図 1 肺ガンサブタイプ	2
図 2 CNN の構造	3
図 3 CNN における畳み込み像	3
図 4 CNN における特徴集約部	4
図 5 RNA シーケンシング	6
図 6 HE 染色組織画像	7
図 7 全体の組織病理画像 (HE 染色)	10
図 8 $5 \times 1 \times 2 \times 5 \times 1 \times 2$ ピクセルデータセット(HE 染色).....	10
図 9 Autoencoder の全体構造	13
図 10 CAE の Encoder を特徴抽出層として使用した CNN 構造	14
図 11 教師あり CNN の全体像	15
図 12 VAE における摂動 $\mathbf{r}$ の更新の模式図	21
図 13 IMSAT の全体像	22
図 14 Score-CAM による判断根拠の可視化	25
図 15 Score-CAM の概要図	27
図 16 遺伝子発現量と特徴量間の相関係数ヒートマップ	29
図 17 遺伝子発現量と特徴量間の相関係数の分布	30
図 18 UBE2C 遺伝子が影響を与える組織形状	32
図 19 CEP25 遺伝子が影響を与える組織形状	33
図 20 CDC20 遺伝子が影響を与える組織形状	34
図 21 CCDC109B 遺伝子が影響を与える組織形状	35
図 22 CDKN3 遺伝子が影響を与える組織形状	36
図 23 KIF11 遺伝子が影響を与える組織形状	37
図 24 KIF11 遺伝子が影響を与える組織形状	38
図 25 KIF11 遺伝子が影響を与える組織形状	41
図 26 HLA-DPB1, HLA-DPA1, CD74, HLA-DRA 遺伝子が影響を与える組織形状	42
図 27 ITGB2 遺伝子が影響を与える組織形状	43

図 28 HBA2 遺伝子が影響を与える組織形状	44
図 29 遺伝子発現量と特徴量間の相関係数ヒートマップ	47
図 30 遺伝子発現量と特徴量間の相関係数分布	47
図 31 TOB1 遺伝子が影響を与える組織形状	49
図 32 NEDD4L が影響を与える組織形状	50
図 33 FCGR2A 遺伝子が影響を与える組織形状	53
図 34 DSE 遺伝子が影響を与える組織形状	54

表目次

表 1 遺伝子発現量データセット(60864 遺伝子)	11
表 2 CAE 学習条件	14
表 3 CNN 学習条件	16
表 4 IMSAT 学習条件	22
表 5 遺伝子発現量データ(縦:遺伝子[505 種],横:被験者名[174 人])	23
表 6 肺ガン組織画像の特徴量データ(縦:特徴量[n]、横:被験者名[174 人]).....	24
表 7 遺伝子発現量と特徴量の相関係数(縦:特徴量[n],横:遺伝子名[505 種]).....	24
表 8 肺ガン病理組織画像の分類精度表	28
表 9 特徴量と正の相関をとる遺伝子(top7)	31
表 10 特徴量と負の相関をとる遺伝子(top7)	40
表 11 特徴量と正の相関をとる遺伝子(相関係数 0.35 以上)	48
表 12 特徴量と負の相関をとる遺伝子(相関係数-0.42 以下).....	52
表 13 特徴量と相関の高い正・負 TOP 25 の遺伝子リスト(教師あり CNN).....	65
表 14 特徴量と相関の高い正・負 TOP 25 の遺伝子リスト(IMSAT)	66

第1章 序論

1.1 研究背景

機械学習アルゴリズムは様々な分野で重要な革新を続けている。近年では、機械学習の一種である深層学習により画像認識分野で特に大幅な進歩が見られている。深層学習は脳の神経回路を模倣して提案された最も単純な神経発火モデルであるニューラルネットワークに基づいている[3]。ニューラルネットワークは入力に重みをかけて足し合わせたものに非線形な活性化関数を適用して出力する仮想的なニューロンを多数連結し、その結合の重みを最適化することで学習する機械学習のモデルである。近年その層を多数積み重ねることで複雑な図形などを学習できることが知られるようになり、深層学習モデルと呼ばれている。深層学習の画像認識への有益性は ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) における大規模画像分類で高い精度を示していることから証明されている[1]。ILSVRC において高い精度を出した深層学習手法は Convolutional Neural Network (CNN) を使用している[2]。CNN は画像の空間的特徴や形状的特徴を自動で導き出し、CNN 内の層に重要な特徴を集約することで画像認識を行っている。CNN はガン組織画像のサブタイプ分類のような病理画像の識別などでも高い成果を示している[4][8]。これらの研究は、CNN は高次元のガン組織学画像から低次元に空間的特徴や形態学的特徴を集約することが可能であることを示している。このように、機械学習の発展はガン組織画像などの病理画像において有用な特徴集約による次元削減を可能にした。このことから、CNN で集約されたガン組織形状情報と、従来の診断で得られている遺伝子発現量データや突然変異遺伝子データを組みあわせることで、未知の情報を探索できるようになった。未知の情報の探索によりガンの治療法に新規手法の開発が期待できる。

1.2 ガンのサブタイプ

ガンは通常、肺ガンや乳ガンのような発生部位ごとに大別される。それらの部位ごとのガンも通常、その症状によって「肺腺ガン」や「扁平上皮ガン」など様々なタイプに分類される。さらに近年では、それぞれのタイプを腫瘍細胞の中での突然変異や遺伝子発現のパターンの違いなどにもとづいてより細かなサブタイプに分類することが多くなってきている。このサブタイプによって効果的な薬品などの治療方法が異なる場合があることから、今後のガン診断ではサブタイプを特定することが重要になってくると考えられる。肺ガンを例にとり以下の図1に示す。本研究では、肺ガンのサブタイプで遺伝子的に分類される TRU (terminal respiratory unit)、PP (proximal-proliferative)、PI (proximal-inflammatory) の区別に着目し、その遺伝子発現パターンの違いと、細胞の形状などの表現型の特徴を抽出したものを比較し、サブタイプの違いと細胞の形状の特徴の関連を解析した[5]。

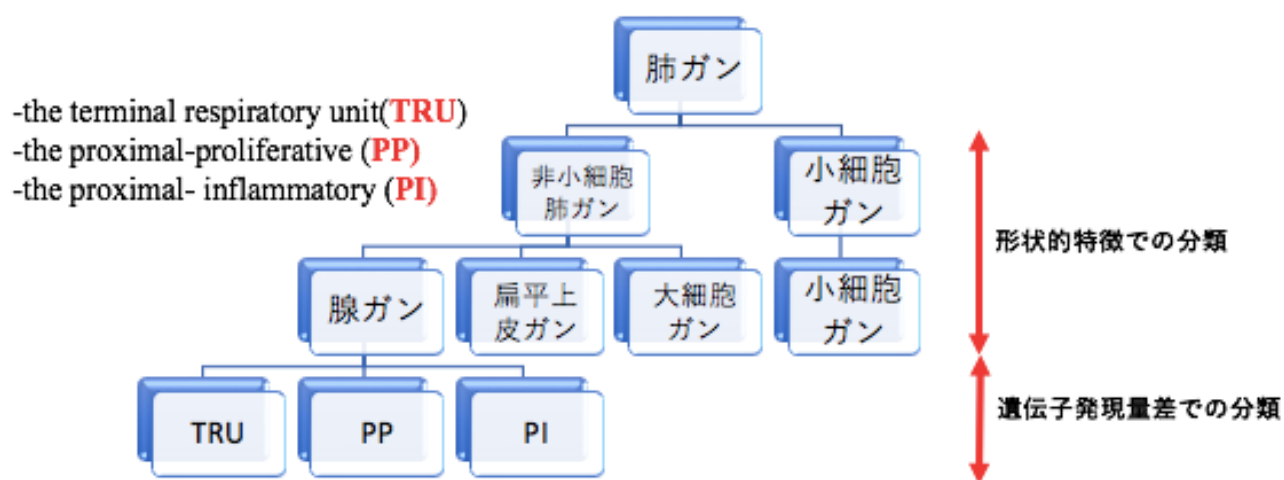


図 1 肺ガンサブタイプ

1.3 Convolutional Neural Network (CNN)

1.3.1 CNN構造

Convolutional Neural Network (CNN)は深層学習手法の一種で画像の分類などに使用されている。CNNは画像の空間的特徴や形状的特徴を自動で導き出し、CNN内の層に重要な特徴を集約することで画像認識を行なっている。一般的なCNNは以下の図2のような構造を取っている。

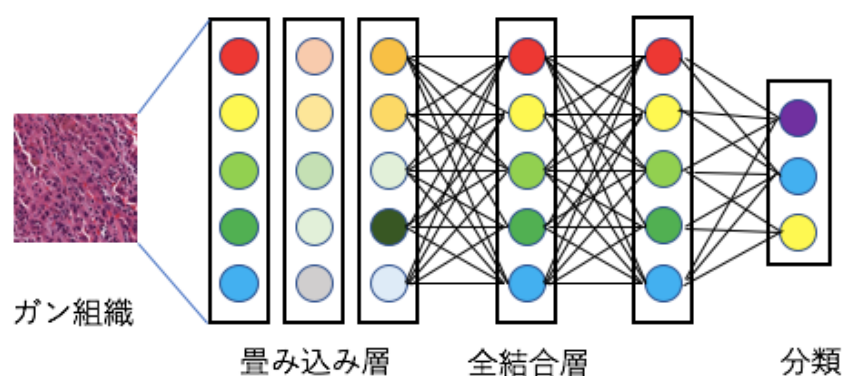


図 2 CNN の構造

CNNの畳み込み層は図3のような仕組みとなっておりフィルタを通して画像の空間的特徴や形状的特徴を求める特徴マップとして出力するプロセスを担っている。畳み込み層で使用するフィルタは画像の特徴をとる配列に最適化される。

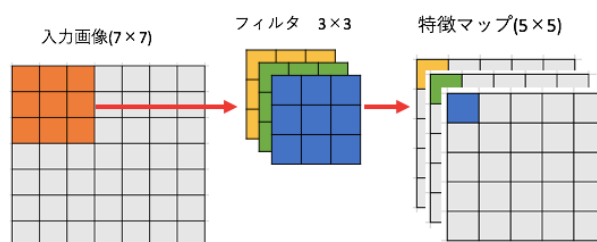


図 3 CNN における畳み込み像

1.3.2 CNNにおける特徴集約部

CNN は畳み込み層の段階で、画像の空間的特徴や形状的特徴を取ることが出来る [6]。さらに、全結合層を通して、特徴の集約が可能である。このことより、最終出力層の1つ手前の中間層において最も空間的特徴や形状的特徴情報を集約できる。手書き文字データセット Modified National Institute of Standards and Technology (MNIST)において CNN を特徴抽出器として使用し、抽出された特徴を従来の機械学習手法であるサポートベクターマシン (SVM) で分類することで 99.8%の精度を出すことができる研究が示す通り CNN は特徴抽出器として使用できることが裏付けられている [6]。本研究では、主に最終層の一つ手前の層がガン組織の空間的特徴や形状的特徴を集約しているものとして扱う。

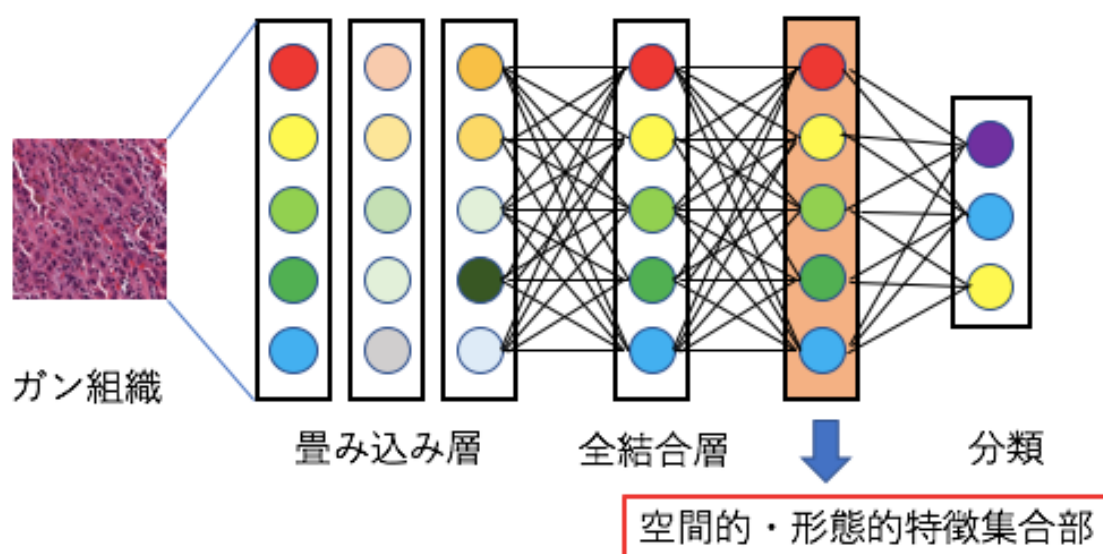


図 4 CNN における特徴集約部

1.4 遺伝子発現解析

遺伝子発現は、遺伝子がメッセンジャーRNA (mRNA) に転写され、タンパク質に翻訳される過程のことを言う。遺伝子発現量は細胞内でどれだけの mRNA が個々の遺伝子から転写されたかの本数に基づく指標であり、それぞれの遺伝子がタンパク質に翻訳された量にほぼ比例すると考えられている。そのため、遺伝子発現量の変化は細胞の機能や組織の形状に大きく影響を与えていると言える。遺伝子発現量の測定にはさまざまな方法があり、かつては細胞内の数千、数万種類の遺伝子の発現量を個別に定量するのはさまざまな困難が伴ったが、近年では次世代シーケンサーと呼ばれる実験技術の進歩により、細胞内の遺伝子の発現を網羅的に定量することが比較的安価かつ高速になった。この手法は RNA シーケンシングと呼ばれ、試料内の遺伝子から転写された mRNA から逆転写酵素によってクローン DNA (cDNA) を合成し、細胞内の全 mRNA から合成された多数の cDNA をまとめてシーケンシングすることで、それぞれの mRNA の配列とその本数を同時に解析する方法である [39]。次世代シーケンサーの手法ではシーケンシングの際に cDNA の分子は断片化されてしまうが、ヒトのように既に全ゲノムが解析されている生物の場合、既知のゲノム配列から最も相似な配列の箇所を探し出す（マッピングする）ことで、断片の配列からでも遺伝子の同定が可能である。簡単な RNA シーケンスの概要図（図 5）を以下に示す。

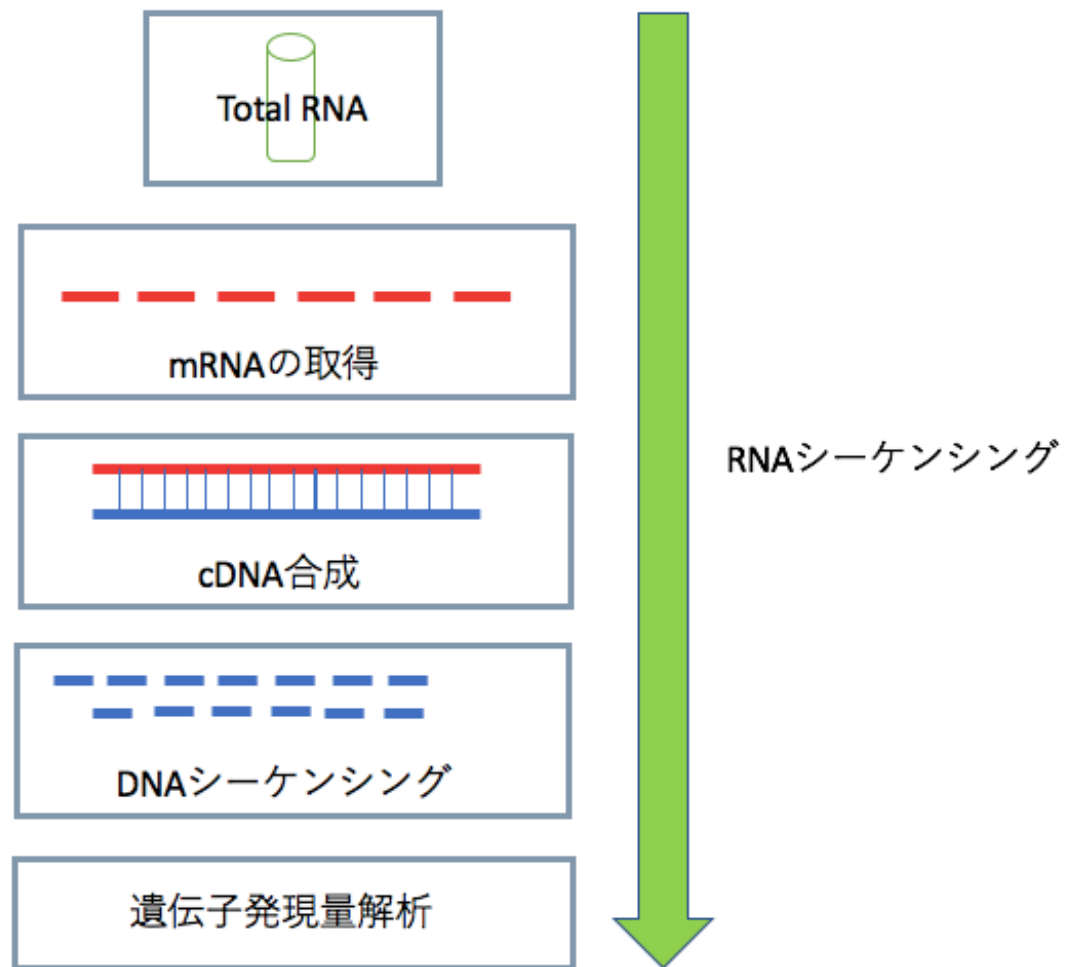


図 5 RNA シーケンシング

1.5 病理組織画像解析

組織から採取された細胞を顕微鏡で観察する際には、目的によって適切な色素によって染色する必要がある。ヘマトキシリン・エオジン (HE) 染色は病理組織画像の診断では最も一般的な染色法である。HE 染色 (図 6) はヘマトキシリンとエオジンの 2 種類の色素を用い、ヘマトキシリンで細胞核を青藍色に、エオジンで細胞質・線維を赤色に染めることにより細胞の形状や個数を把握しやすくする方法である。染色なしの病理組織は半透明であるため、染色は多くの情報を付加することができる。

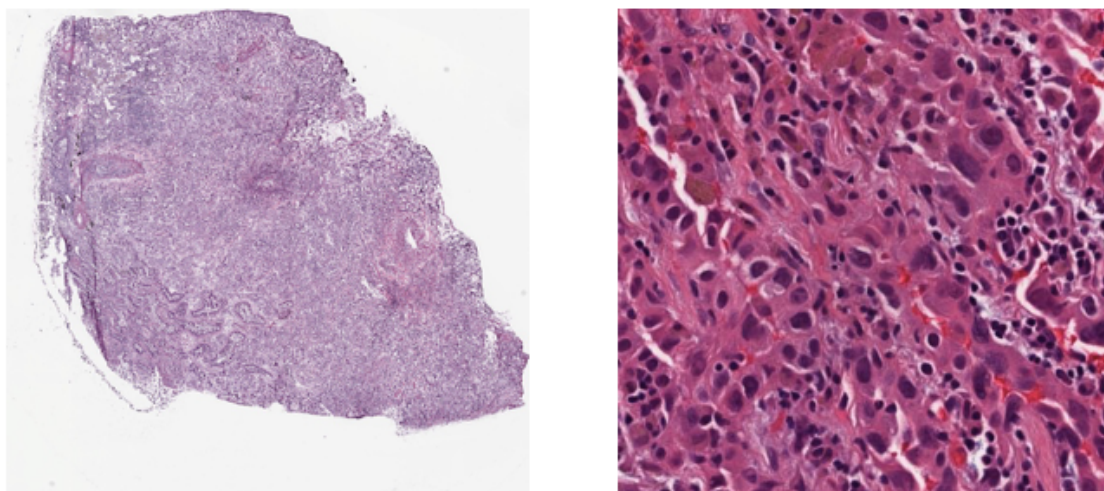


図 6 HE 染色組織画像(左：組織全体画像, 右：組織拡大画像)

1.6 研究目的

CNN は高次元のガン組織学画像から低次元に空間的特徴や形態学的特徴を集約することが可能である。機械学習の発展はガン組織画像などの病理画像において有用な特徴集約による次元削減を可能にした。CNN で集約された得られたガン組織情報と、従来の診断で得られている遺伝子発現量データや突然変異遺伝子データを組みあわせることで、未知の情報を探索できるようになった。

ガンは一部の遺伝子に変異することで発生することが知られており、生物学的研究からそのメカニズムは徐々に明らかにされてきた [7]。しかし、ガンの形状に影響を与える遺伝子はあまり明らかにされていない [8]。ガンの形状に影響を与える遺伝子を特定することで遺伝子発現情報から将来的なガン組織の形状を予測できるようになりガンの予防的治療が可能となる。そのため、CNN で得られた情報の集約されたガン組織情報と、従来の診断で得られている遺伝子発現量データの相関関係を解明することでガン組織の形状に影響を与える遺伝子を特定することが必要である。、本研究では畳み込みニューラルネットワークを用いガン組織形状に影響を与える遺伝子を調べることを目的とした。

第2章 データセット

本研究では The Cancer Genome Atlas Program (TCGA)より入手した被験者 174 人分の肺ガン病理画像、それに対応した遺伝子発現量データを用いた。本研究で使用した肺ガン組織は肺腺ガンを遺伝子発現量の差によりクラス分けした TRU(the terminal respiratory unit)、PP(the proximal-proliferative)、PI(the proximal- inflammatory) である[9]。

TCGA は米国ガン研究所 (National Cancer Institute: NCI)と米国ヒトゲノム研究所 (National Human Genome Research Institute: NHGRI)の共同プロジェクトであり 33 種のガン種においてゲノム変異・遺伝子発現変動・メチル化異常など 7 タイプのデータを網羅的に集めたオープンデータサイトである。

今回用いた病理画像は HE 染色を施されている。病理画像は高精細に取得されているため、サイズが非常に大きい。そのため、512×512 ピクセルサイズでランダムに切り出した画像をデータセットとして用いた(図 7、図 8)。後述する教師あり分類、教師なしクラスタリングにおける実験では、512×512 ピクセルの画像をトレーニングデータとして 250000 枚、テストデータとして 30000 枚使用した。

遺伝子発現量データ(表 1)は RNA シーケンス技術を用い各被験者の病理組織から取得されたもので、各被験者の病理組織と遺伝子発現量データは対応したものとなっている。遺伝子発現量データは後述する教師あり分類、教師なしクラスタリングにおける中間層(特徴抽出層)と遺伝子発現量に相関の有無の確認のため用いた。

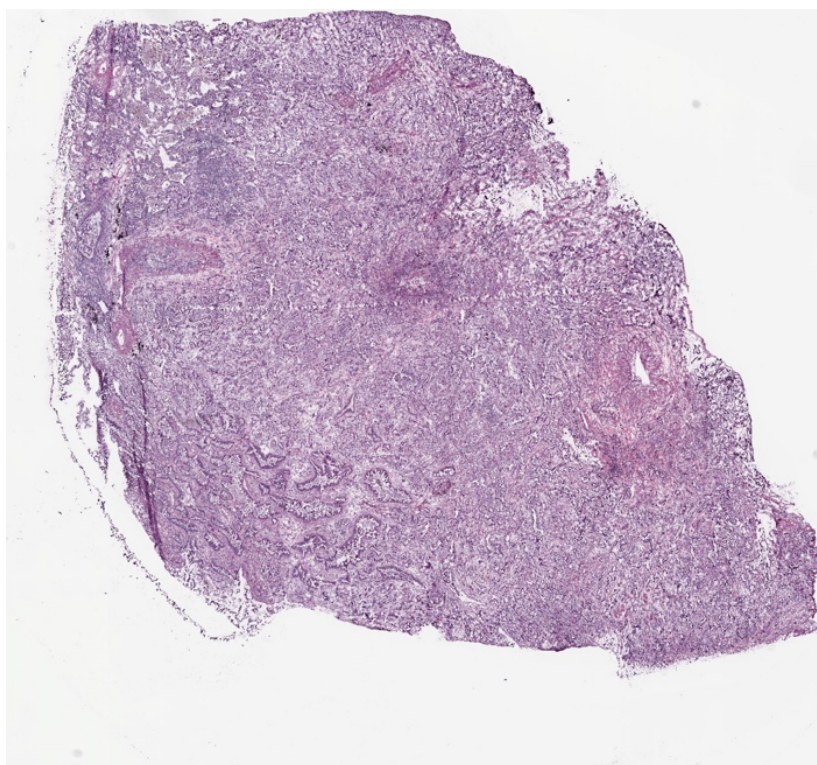


図 7 全体の組織病理画像 (HE 染色)

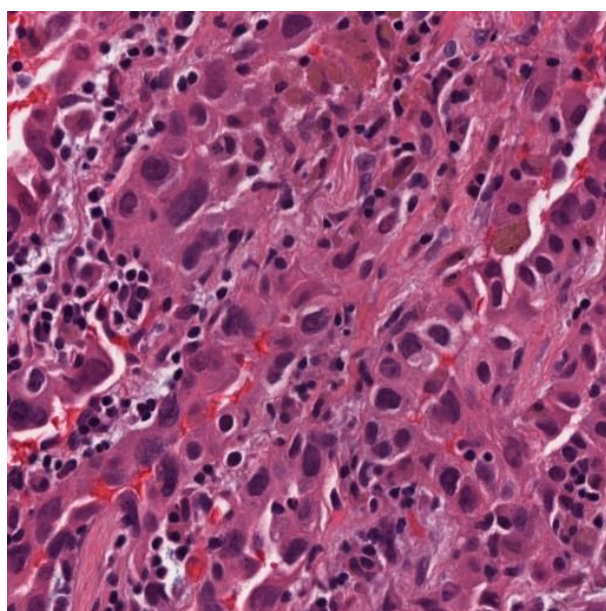


図 8 512×512ピクセルデータセット(HE 染色)

表 1 遺伝子発現量データセット(60864 遺伝子)

遺伝子名	遺伝子発現量
CYP4B1	0.05445011
TOP2A	0
MTHFD2	3.098068806
CCNB1	0
RRM2	3.734236455
:	:
:	:
TOP2A	0.07881799
MTHFD2	111.2763387
CCNB1	0.250710623
RRM2	0.053030971

第3章 提案手法

本章では、教師あり畳み込みニューラルネットワーク、教師なし畳み込みニューラルネットワークを用いた組織病理画像の特徴抽出手法を述べる。抽出された特徴と遺伝子発現量の相関探索手法についても説明し、抽出された特徴を可視化する技術を説明する。

3.1 組織病理画像の特徴抽出法

本研究は Convolutional AutoEncoder(CAE)をベースとした教師あり畳み込みニューラルネットワーク、教師なし畳み込みニューラルネットワークを用いて組織病理画像の特徴抽出を行った。以下にその手法の詳細を述べる。

3.1.1 AutoEncoder

3.1.1.1 概要

AutoEncoder(AE)は順伝播型のニューラルネットワークの一種で教師なし次元削減を行うことが出来る。代表的な線形次元削減手法として、主成分分析(PCA: Principal Component Analysis)や線形判別分析(LDA: Linear Discriminant Analysis)などがあるが、画像のような高次元データでは、うまく機能しないことが多い。そこで、画像データに対し高い精度で識別できるニューラルネットワークを次元削減に応用した **AutoEncoder** を用いることで、病理組織特徴を取る特徴抽出層を取得できると考えられる[10]。**AutoEncoder** で作られた特徴抽出層は教師あり畳み込みニューラルネットワーク、教師なし畳み込みニューラルネットワークへ転移学習(事前重み付加)する目的で作成した。

3.1.1.2 AutoEncoder機構

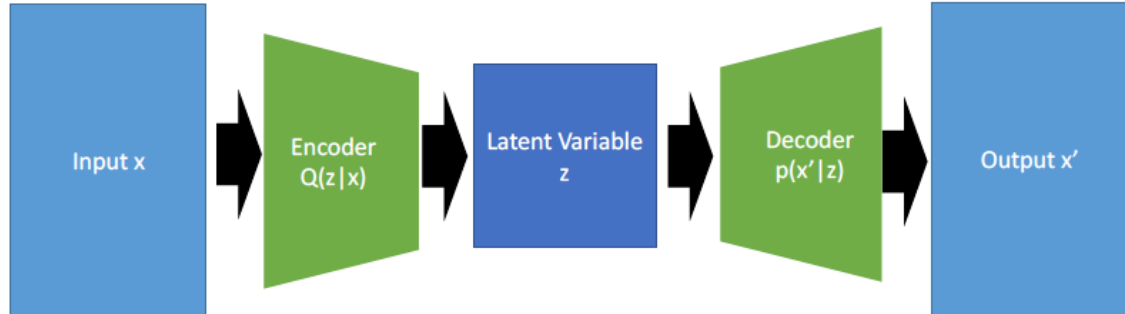


図 9 AutoEncoder の全体構造

図 9 のような入力層と出力層で同数のユニットを持つ順伝播型のネットワークを考える。この時、入力層より少数のユニットを持つ隠れ層 (Latent Variable) の出力 z へ次元削減する変換 (Encoder) は、重み W 、バイアス項 b 、活性化関数 f を用いて、

$$z = Q(z|x) = f(Wx + b) \quad (3.1.1)$$

と表現できる。出力 z を隠れ層で折り返し、入力 x と同じ空間 x' に戻す変換 (Decoder) は、

$$x' = p(x'|z) = f(\bar{W}z + \bar{b}) \quad (3.1.2)$$

と表現できる。これらの一連の計算は以下の式 (3.1.3) で示される入出力間の平均二乗誤差を最小化することで学習が行われる。

$$loss = \frac{1}{N} \sum_{i=1}^N \|x - x'\|^2 \quad (3.1.3)$$

z から元の入力 x を再構成できることから、 z は入力 x の特徴を抽出された低次元ベクトルと見なすことができる。

画像認識において、空間的特徴を捉える事のできる畳み込みニューラルネットワークが高い精度で識別できるのと同様に、畳み込みを適用した CAE (Convolutional Autoencoder) は重要な特徴を獲得できる。

3.1.1.3 事前学習(転移学習用の特徴抽出器の作成)

本研究は、初めに図 9 のような Encoder5 層,Decoder5 層の Convolutional AutoEncoder(CAE)を用いて病理画像再構成を行い、CAE の特徴抽出部(Encoder)とそれに対応する重みを取得する。

そして、図 10 のように、後に使用する教師あり畳み込みニューラルネットワーク、教師なし畳み込みニューラルネットワークの上層に CAE で得られた特徴抽出層を使用し、転移学習、fine tuning する事でそれぞれの学習を行う。

CAE の学習に使用した具体的なパラメータは以下の表 2 に記載する

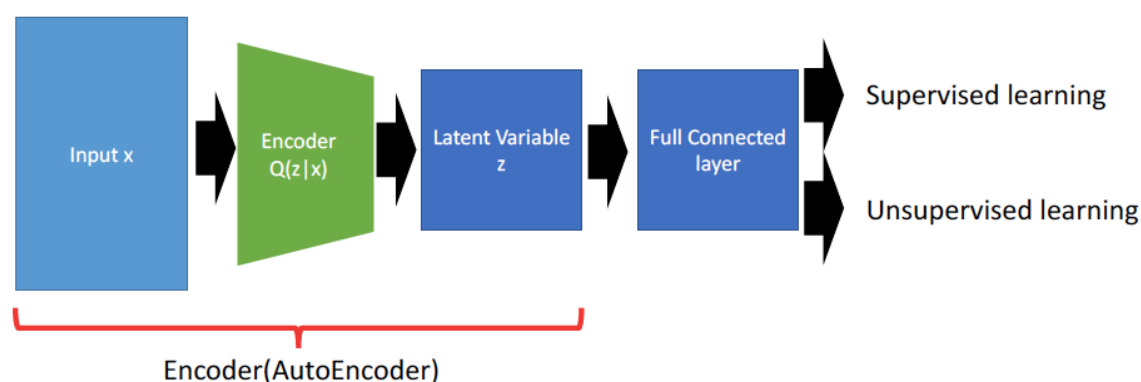


図 10 CAE の Encoder を特徴抽出層として使用した CNN 構造

表 2 CAE 学習条件

訓練データ数	250000
Epoch	20
Batch size	96
Optimizer	Adam

3.2 教師あり畳み込みニューラルネットワーク

3.2.1 概要

教師あり畳み込みニューラルネットワークは学習データに正解ラベルを対応させ学習する。教師あり畳み込みニューラルネットワークは学習データとラベルを対応づけて学習する特性上、学習が成功すれば中間層でラベルに対応した特徴をとる。

本実験では、遺伝子発現量と特徴量の相関を探索することを目標としているため、教師あり畳み込みニューラルネットワークを学習後、中間層(特徴量)を抽出する。

3.2.2 教師ありCNNによる特徴抽出

本研究の教師ありCNNによる特徴抽出では、Convolutional AutoEncoder(CAE)で事前に学習されたEncoder5層と未学習の全結合層3層を結合し作成したCNN(図11)をfine tuningし表3の条件で訓練を行った。モデル学習後、最終層の1つ前の中間層の出力を特徴量(Feature value:512 dimensions)として取得した。取得された特徴量は3.4で行う遺伝子発現量の相関探索で使用するため取得される。

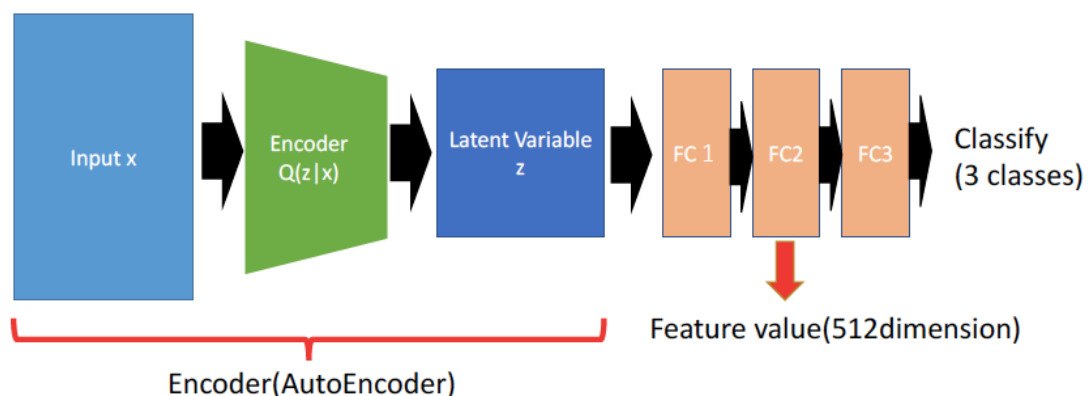


図 11 教師あり CNN の全体像

表 3 CNN 学習条件

訓練データ数	250000
テストデータ数	30000
Epoch	200
Batch size	96
Optimizer	Adam

3.3 教師なし畳み込みニューラルネットワーク

3.3.1 概要

教師なし畳み込みニューラルネットワークの 1 種に Information Maximizing Self-Augmented Training(IMSAT)がある。IMSAT は画像データであれば画像の形態、色的特徴を根拠にクラスタを作るように訓練される [11]。そのため、中間層で、画像の形態、色的特徴に基づいた特徴を得ることができる。

本実験では、遺伝子発現量と相関のある特徴量を探索することを目的とするため IMSAT 訓練後、中間層(特徴量)を抽出する。

3.3.2 Information Maximizing Self-Augmented Training(IMSAT)

Information Maximizing Self-Augmented Training(IMSAT)は教師なし学習であるのにも関わらず、手書き文字データセット MNIST で 98.4%の精度で分類に成功した手法である。IMSAT は Regularized Information(RIM)と Self-Augmented Training(SAT)に基づいて学習を行う。まず初めに RIM について述べる。RIM はデータ X と予測クラス Y の相互情報量を最大化することによりクラスタリングを行う手法である。相互情報量は 2 つの確率分布の相互依存の尺度を表すものである。RIM はエントロピー H 、 h で以下の式のように表せる。

$$\begin{aligned}
I(X:Y) &= H(Y) - H(Y|X) \\
&= h\left(\frac{1}{N} \sum_{i=1}^N p_{\theta}(y|x_i)\right) - \frac{1}{N} \sum_{i=1}^N h(p_{\theta}(y|x_i)) \quad (3.1.4)
\end{aligned}$$

$H(Y)$ はデータ全体における Y の予測分布のエントロピーである。 $H(Y)$ はデータ全体のクラスが一様分布になるように最大化する。一方、 $H(Y|X)$ はデータ X における Y の予測分布のエントロピーである。データ X が既知の時モデルが得る情報量は大きくなり、予測分布は尖った分布であることが望ましいため、 $H(Y|X)$ は最小化する。IMSAT は RIM における正則化項として後述する SAT を加えた手法であり式(3.1.5)で示される損失関数を最小化する事でクラスタリングを行う。

$$R_{IMSAT} = R_{SAT} - I(X:Y) \quad (3.1.5)$$

3.3.3 Self-Augmented Training(SAT)

SAT は data augmentation を利用してモデルのロバスト性を高める正則化方法である[12]。CNN は訓練データが少ない時、物体の位置、輝度などの変形に対して脆弱であることが知られている。そのため、入力データに回転や輝度の変化などをランダムに加えデータ数の水増しを行うことで、モデルの頑健性を高めることができる[13]。教師ありに基づいた識別モデルでは、教師ラベルが事前分布となるため、訓練データの水増しのみで精度が向上する。しかし、教師なしに基づいたクラスタリングモデルでは、水増しのみでの性能向上は難しい。これは、データ x と類似したデータ x' の予測分布間でエントロピーが大きくなるためである。よって、教師なしクラスタリングモデルでは予測分布に対して直接正則化を加えることで精度向上が可能である。IMSAT は教師なしクラスタリングであるため、予測分布に対して直接正則化を加えることのできる Virtual Adversarial Training(VAT)を SAT として使用した。

3.3.3.1 Virtual Adversarial Training(VAT)

Virtual Adversarial Training(VAT)は Virtual Adversarial example を変形後のデータ x として用いる事で、データ x の周辺で予測分布をなめらかになるように正則化を行う SAT の一つである[14]。CNN モデルにおいて、入力データに誤差関数の勾配方向へ小さな摂動を乗せると、モデルの識別が狂うことが知られている。このようにニューラルネットを大きく狂わせるデータを Adversarial Example という。このようなデータを取得し、学習時に用いる事で、ノイズに強いモデルを得る手法を Adversarial training と言う。Adversarial training における Adversarial example の摂動 r_{adv} はモデルの損失関数 D に対する勾配から作成される。教師ラベルに基づいた予測分布を $q(y|x)$ 、モデルから推定された予測分布を $p(y|x, \theta)$ とした時、摂動 r_{adv} は、

$$r_{adv} = \epsilon \text{sign}(\nabla_x D[q(y|x), p(y|x, \theta)]) \quad (3.1.6)$$

として生成される。Adversarial training における損失関数 D' は、データ x と Adversarial example $x + x_{adv}$ に対する損失関数の割合 α を用いて

$$D' = \alpha D[q(y|x), p(y|x, \theta)] + (1 - \alpha) D[q(y|x), p(y|x + r_{adv}, \theta)] \quad (3.1.7)$$

と表すことができる。このように教師あり分布に基づいた予測分布が必要のため Adversarial training は教師あり学習に限られていた。しかし、VAT では教師あり分布に基づいた予測分布を必要としないため、教師なし、半教師あり学習にも適用可能である。VAT はデータ x 周辺の予測分布から勾配と KL 距離から摂動 r_{adv} を求め、Virtual Adversarial example である $x + r_{adv}$ を作成する。

まず、データ x から求まる予測分布 $q(y|x, \theta)$ と、正規分布で初期化した r を加算したデータ $x+r$ から求まる予測分布 $q(y|x+r, \theta)$ 間の距離を求める。予測分布間の距離 LDS は以下の式(3.1.8)で表される。

$$LDS(x) = KL[p(y|x, \theta), p(y|x+r, \theta)] \quad (3.1.8)$$

この時、モデルの識別を最も狂わせる摂動 r_{adv} は式(3.1.9)のように $\|r\| \leq \epsilon$ を満たす時に最大値をとる $LDA(x)$ を求めれば良い。

$$r_{adv} = \arg \max_{r: \|r\| \leq \epsilon} KL[p(y|x, \theta), p(y|x+r, \theta)] \quad (3.1.9)$$

式(3.1.9)を線形近似で求めることは難しいため、 $r=0$ 周りでのテイラー展開を考える。式を簡略化するため、 $KL[p(y|x, \theta), p(y|x+r, \theta)]$ を $D_{kl}(r, x, \theta)$ と置き、二階導関数まで近似すると、式(3.1.10)

$$D_{kl}(r, x, \theta) \approx D_{kl}(0, x, \theta) + r^T \nabla_r D_{kl}(r, x, \theta)|_{r=0} + \frac{1}{2} r^T H(x) r \quad (3.1.10)$$

のようになる。但し、 H はヘッシアンである。 $r=0$ 付近では $p(y|x, \theta) = p(y|x+r, \theta)$ なので、 $D_{kl}(r, x, \theta) = 0$ となる。また、 $D_{kl}(r, x, \theta) \geq 0$ より、 $r=0$ 付近で最小値を取ると考えられるため、 $\nabla D_{kl}(r, x, \theta) = 0$ である。ゆえに、式(3.1.11)は、

$$D_{kl}(r, x, \theta) \approx \frac{1}{2} r^T H(x) r \quad (3.1.11)$$

として近似できる。 $r^T H(x) r$ を最大化する r は $H(x)$ の最大固有値に対応する固有ベクトルであるため、べき乗法を用いて求める。ランダムに初期化した単位ベクトルを初期値として、以下のようにして d を更新する。

$$d \leftarrow \overline{Hd} \quad (3.1.12)$$

$\overline{Hd}$ を用いて d を繰り返し更新することで $\overline{Hd}$ の最大固有値に対する固有ベクトルを得ることができる。式(3.1.12)のべき乗計算には $H(x)$ が必要であるが、直接計算することは困難である。そこで、差分法を用いて $H(x)d$ を以下の(3.1.13)のように近似する。

$$\begin{aligned} H(x)d &\approx \frac{\nabla_r D_{kl}(r, x, \theta)|_{r=\xi d} - \nabla_r D_{kl}(r, x, \theta)|_{r=0}}{\xi} \\ &= \frac{\nabla_r D_{kl}(r, x, \theta)|_{r=\xi d}}{\xi} \end{aligned} \quad (3.1.13)$$

ξ は非 0 の微少数である。そのため最終的な更新式は以下のように表せる。

$$d \leftarrow \frac{\nabla_r D_{kl}(r, x, \theta)|_{r=\xi d}}{\xi} \quad (3.1.14)$$

通常、べき乗法の反復回数は十分な回数必要であるが、論文において反復回数は 1 回で十分であることが報されているため、本研究でも同様に行った。べき乗法 1 回の時 r_{vadv} は

$$r_{vadv} \approx \epsilon \frac{g}{\|g\|_2} \quad (3.1.15)$$

$$\text{where } g = \nabla_r D_{kl}[p(y|x, \theta), p(y|x+r, \theta)] \quad (3.1.16)$$

として近似することが可能である。このように作成された r_{vadv} はデータ x に合計され、Virtual Adversarial example $x+r_{vadv}$ として Virtual Adversarial training に利用することでノイズに頑強なモデルを作成可能である。次ページに VAE における摂動 r の更新の模式図を置く(図 12)。

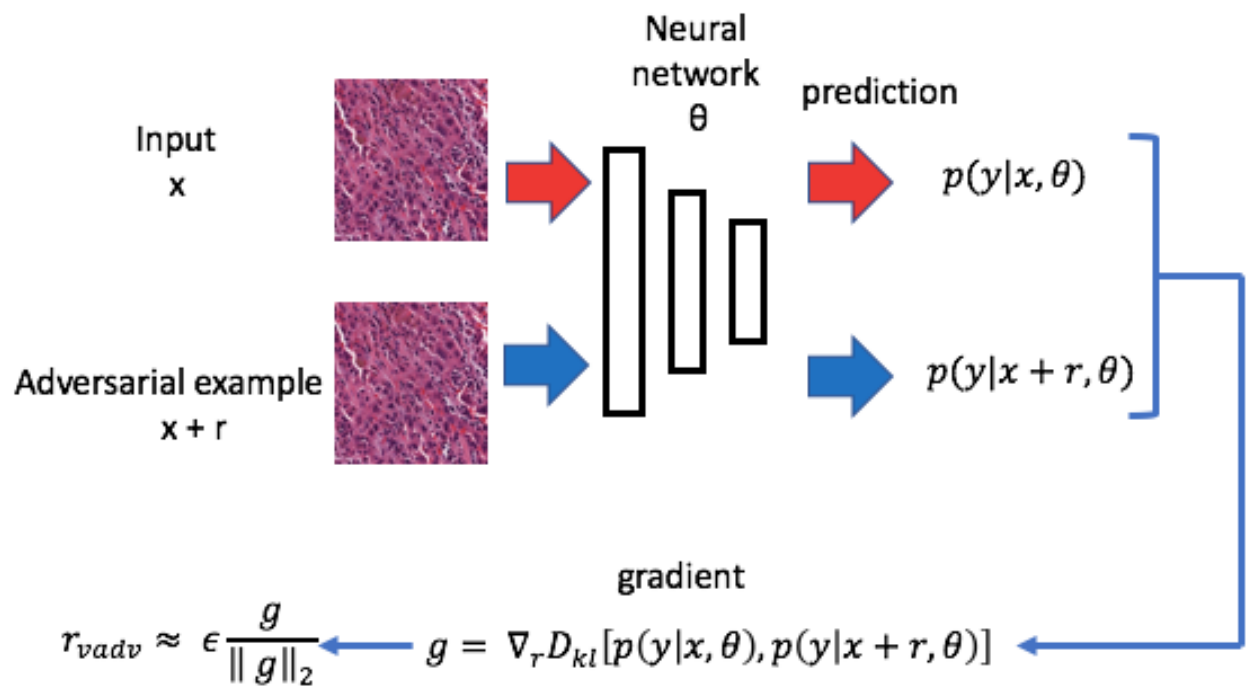


図 12 VAE における摂動 r の更新の模式図

3.3.4 IMSATによる特徴抽出

IMSAT に基づいたクラスタリングモデルは表 9 の条件で訓練を行った。メインのネットワークとして、Convolutional AutoEncoder(CAE)で事前に訓練された 5 層の Encoder 層を用い、3 層の未学習全結合層を結合したネットワークで fine tuning を行った(図 13)。また、出力は 8 クラスとした。モデル学習後、最終層の 1 つ前の中間層の出力を特徴量(Feature value:128 dimensions)として取得した。取得された特徴量は 3.4 で行う遺伝子発現量の相関探索で使用するため取得される。

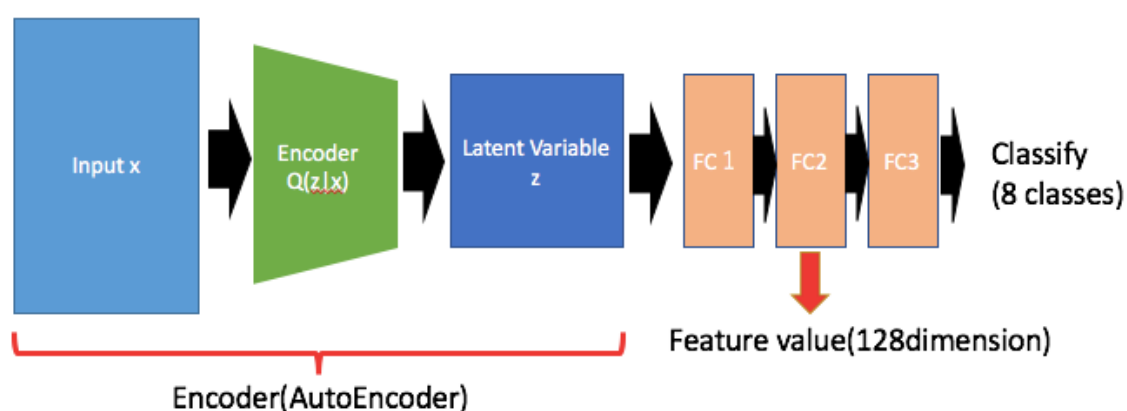


図 13 IMSAT の全体像

表 4 IMSAT 学習条件

訓練データ数	250000
テストデータ数	30000
Epoch	200
Batch size	96
Optimizer	Adam

3.4 遺伝子発現量とガン形態特徴の相関探索

3.4.1 概要

人間が持つ遺伝子数は 2020 年時点で 21206 個とされており、ガンに關係する遺伝子は一部と言われている。そのため、本研究で使用する遺伝子をガンと關係することが判明している転写因子 505 個のみに限定して、ガン形態特徴との相関を探索する。3.1、3.2 における教師あり CNN、教師なし CNN(IMSAT)で肺ガン病理組織画像の特徴を入手した。入手した特徴量と 505 個の遺伝子発現量との相関係数を 1 対 1 で対応させ計算することで相関係数の表を作成した。

3.4.2 相関係数計算手法

遺伝子発現量のデータセットとして、表 1 で示された遺伝子発現量データを被験者数(174 人)分、並べた表を作成する(表 5)。但し、遺伝子発現量が一部の遺伝子で偏っている場合があるので、全ての値に対し対数をとることとする。

表 5 遺伝子発現量データ(縦:遺伝子[505 種],横:被験者名[174 人])

	TCGA-05-4384	...	TCGA-05-4405	TCGA-05-4410
CYP4B1	32.77119628	...	1.918754633	1.114971556
TOP2A	2.79568537	...	7.351644946	17.51401224
...	...	...	...	...
STXBP1	3.836615103	...	5.114630143	12.4166835

次に、各被験者からランダムに 250 枚画像を取得し、教師あり CNN、教師なし CNN(IMSAT)に入力する。それぞれのモデルから出力された中間層の平均(/250)を取り、肺ガン病理組織画像の形状、空間的な情報が集約された特徴量(n 個)を入手する(表 6)。

表 6 肺ガン組織画像の特徴量データ(縦:特徴量[n]、横:被験者名[174 人])

	TCGA-05-4384	...	TCGA-05-4405	TCGA-05-4410
1	5.67889	...	0.37123	10.58596
2	8.53229	...	2.45671	9.65428
...	...	...	...	...
n	3.57892	...	15.23411	13.3459

表 6 と図 11,13 で示した、遺伝子発現量データと CNN の中間層から得られた肺ガン組織画像の特徴量データを利用して、単一の遺伝子における遺伝子発現量と単一の特徴量の相関係数を計算した。相関係数の計算にはピアソンの相関係数を用いた。計算により各遺伝子における遺伝子発現量と特徴量の相関係数を入手出来る(表 7)。

表 7 遺伝子発現量と特徴量の相関係数(縦:特徴量[n],横:遺伝子名[505 種])

	CYP4B1	...	CCL18	STXBP1
1	0.0324	...	0.54231	0.21345
2	0.1256	...	0.23981	0.32156
...	...	...	...	...
n	0.1287	...	0.0456	0.49872

各遺伝子における遺伝子発現量と特徴量の相関係数を求めることで、形態に影響を与える遺伝子を抽出することが可能となる。

3.5 CNN判断根拠の可視化

3.5.1 概要

深層学習は自動でデータの特徴を取捨選択し、学習することが可能である。しかし、自動で特徴を捉えるがゆえに、何を根拠に学習しているか不明である。単純なデータであれば、深層学習が何を根拠に学習したのかを推測可能であるが、複雑なデータであれば推測は難しい。深層学習の判断根拠は自動運転やコンピュータによる病理診断の補助など重大な決定を行う際、信頼性を保証するため非常に重要視されている。このような背景から近年、深層学習の判断根拠を提示できるアルゴリズムが開発され始めた。一例として、CNN モデルで判断根拠を可視化することができる **Score-Weighted Class Activation Mapping(Score-CAM)** などがある。**Score-CAM** は以下の図 14 のように CNN が犬と判断した時、犬の顔をみて判断した根拠を可視化できる[15][16]。本研究では、遺伝子とガン組織形状の相関を調べるために **Score-CAM** を用いた。

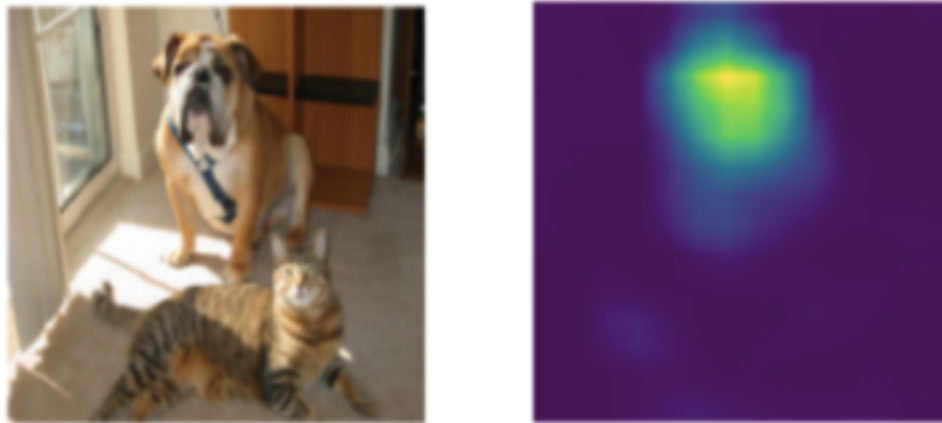


図 14 Score-CAM による判断根拠の可視化

3.5.2 Score-Weighted Class Activation Mapping (Score-CAM)

Score-Weighted Class Activation Mapping (Score-CAM) は CNN が判断した根拠を可視化できる手法の一種である。CNN の最後の畳み込み層に流れる勾配情報を使用して、ターゲットクラスの予測スコアの各活性化マップの重要性を表す。そして、CNN の最後の畳み込み層に流れる勾配情報を使用して、ターゲットクラスの予測スコアの各活性化マップの重要性を表すことができる。

Score-CAM は CNN の最終層で出力される特徴マップ $A_{m,h}^k$ を入力画像サイズまでバイリニア補完を用いて $A_{i,j}^k$ に拡大する。拡大された特徴マップを区間 $[0,1]$ に正規化する。(3.1.17)

$$A_{i,j}^k = \frac{A_{i,j}^k}{\max A^k - \min A^k} \quad (3.1.17)$$

拡大された特徴マップを正規化することで、入力画像 I に対し、チャンネルごとに積を取ることが可能となる。マスクされた画像 M を特徴マップのチャンネル数分準備する。(3.1.18)

$$M^K = A^K \cdot I \quad (3.1.18)$$

特徴マップがマスクされた入力画像 M を CNN モデル $F(x)$ に入力し、演算後の配列 S を取得する。(3.1.19)

$$S_k = F(M^k) \quad (3.1.19)$$

各チャンネルの重要度 w_k^c を、配列 S_k のスコア c で定義する。(3.1.20)

$$w_k^c = s_k^c \quad (3.1.20)$$

特徴マップ A^k に各チャンネルの重要度 w_k^c をかけ合わせ Relu 演算をおこなったものを。最終的なスコアマップ L とする 式(3.1.21). Relu 演算を行うのは、特徴マップには負値の要素を含んでいないからである。

$$L_{Score-CAM}^c = ReLU(\sum_k w_k^c A^k) \quad (3.1.21)$$

これにより、それぞれの特徴量を求める際プラスに働く領域を可視化することが可能となる。

Score-CAM の概要図(図 15)を以下に示す。

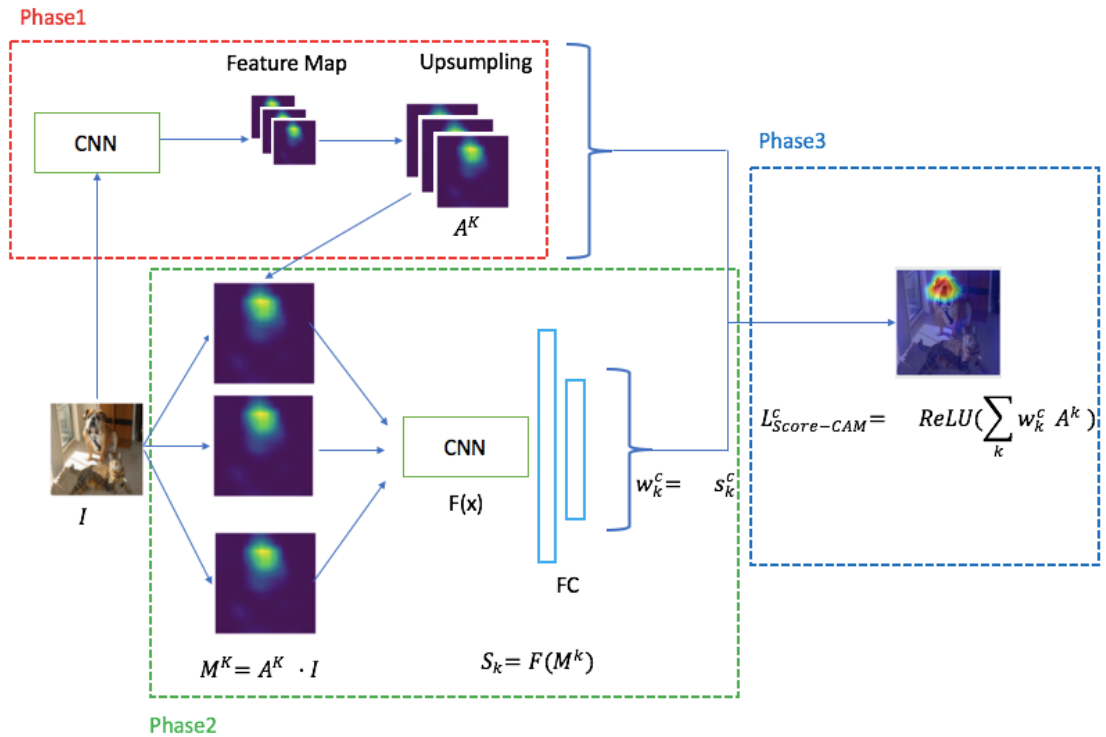


図 15 Score-CAM の概要図

第4章 結果・考察

4.1 教師ありCNN

本節では、遺伝子発現量の差でクラス分けされた3種の肺ガン組織サブタイプ画像、TRU(the terminal respiratory unit)、PP(the proximal-proliferative)、PI(the proximal-inflammatory)をCNNにより分類した結果を述べる。また、CNNの中間層から得られた特徴量と遺伝子発現量データの相関を調べることで、肺ガン組織形状に影響を与える遺伝子を特定する。

4.1.1 教師ありCNN分類

肺ガン組織のサブタイプTRU、PP、PIをCNNで分類した結果、精度93.4%の結果が得られた。(表8)

表8 肺ガン病理組織画像の分類精度表

		prediction			Total	Accuracy(%)
		PP	PI	TRU		
diagnosis	PP	9410	300	290	10000	94.1
	PI	150	9213	637	10000	92.1
	TRU	312	345	9343	10000	93.4
	Total	9872	9858	10270	30000	93.2

このことから、CNNは遺伝子発現量の分布により分類されたサブタイプを高精度で分類できる能力があることが示唆された。

4.1.2 教師ありCNNによる遺伝子発現量と特徴量の相関探索

4.1.2.1 教師ありCNNによる遺伝子発現量と特徴量の相関係数

4.1-1 で学習された CNN は、遺伝子発現量の分布よりクラス分けされた肺ガン病理組織サブタイプを、高精度で分類できる能力があることが示唆された。そのため、CNN の中間出力(以下特徴量とする)は、遺伝子発現量に影響を受けた組織形状の特徴が集約されていると予想される。遺伝子発現量(505genes)と特徴量(512dimension)1 対 1 の相関をとることで、どの遺伝子が組織形状に影響を及ぼしているかを確認した。遺伝子と特徴量の相関係数を求めたところ、258560 個の相関係数を求めることが出来た。相関係数の分布をヒートマップで示した図(図 9)を以下に示す。

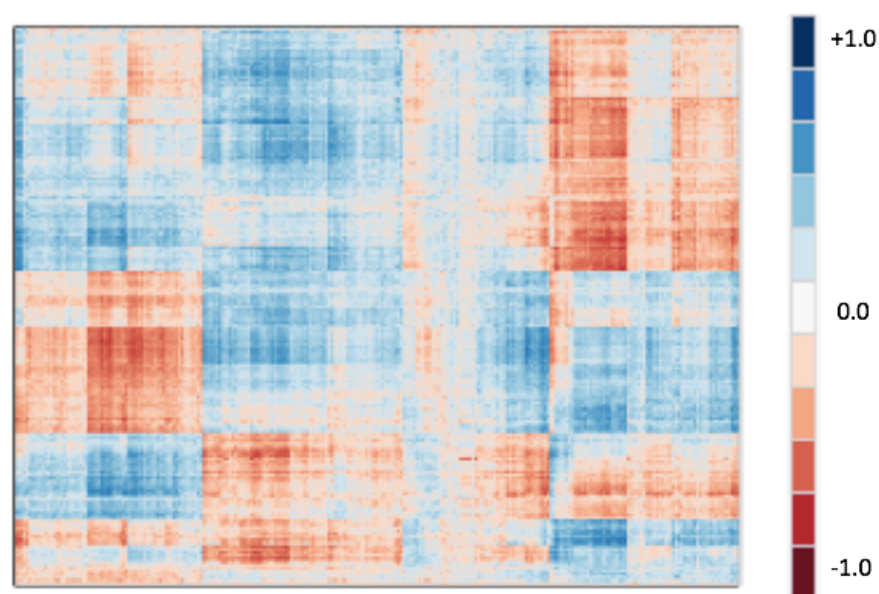


図 16 遺伝子発現量と特徴量間の相関係数ヒートマップ
(横:特徴量(512 dimension),縦:遺伝子発現量(505 genes))

ヒートマップから大まかな相関係数の分布を見ることができる。.

次に、相関係数の分布と最小値、最大値を調べるためにヒストグラムで描画した。(図 14)

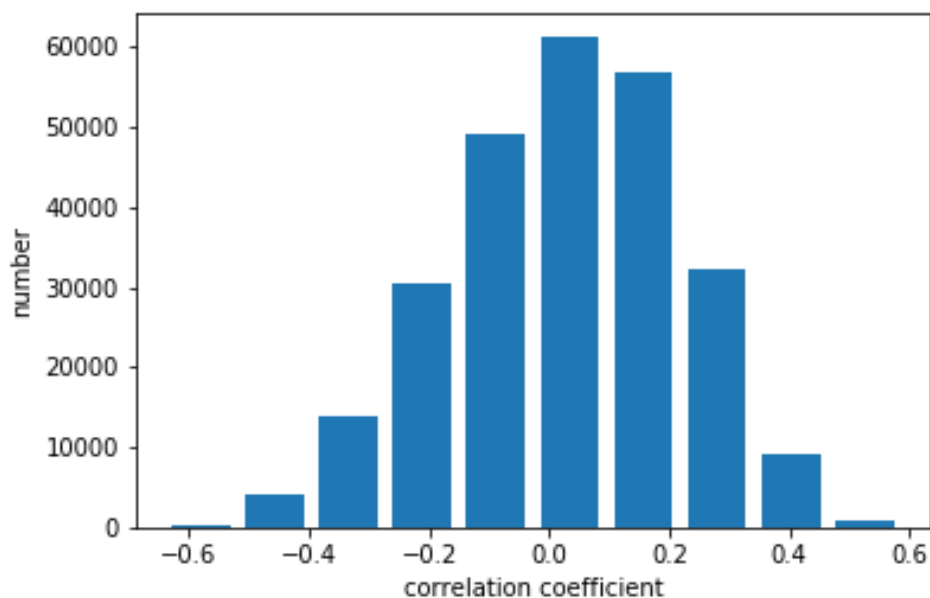


図 17 遺伝子発現量と特徴量間の相関係数の分布

相関係数の分布は正規分布で、正の相関係数の最大値は 0.58 で負の相関係数の最小値は -0.64 であった。

4.1.2.2 正の相関を取る遺伝子と病理組織形態の関係

遺伝子発現量と特徴量の相関係数の値は、遺伝子が肺ガン組織特徴に与える影響の高さの尺度と考えることができる。特徴量と正の相関係数を取った遺伝子は遺伝子発現量の増幅に比例して、細胞や組織形状を変化させることを示すと予測される。正の相関係数 top7 の値をとった遺伝子を肺ガン組織に影響を与えやすい遺伝子であると考ええる。表 13 で正の相関係数で top7 の値をとった遺伝子と、遺伝子の持つ役割を簡単に示す。

表 9 特徴量と正の相関をとる遺伝子(top7)

UBE2C	ユビキチン結合酵素
CEP55	中心体タンパク質
CDC20	細胞分裂周期タンパク質
CCDC109B	コイルドコイルドメインタンパク質
CDKN3	サイクリン依存性キナーゼ阻害タンパク質
KIF11	モータータンパク質
KPNA2	RAG コホートタンパク質

肺ガン病理画像の空間的、位置的情報を集約した特徴量と高い正の相関を取る遺伝子は UBE2C, CEP55, CDC20, CCDC109B, CDKN3, KIF11, KPNA 2 の 7 つの遺伝子であった。特徴量との相関を取る遺伝子が組織上でどのような特徴を示すかを Score-CAM で可視化する。Score-CAM ではそれぞれの遺伝子と相関が高かった特徴量を 3 つ用い可視化をおこなった。可視化した画像は左端が元々のガン組織画像である。また、左から 2,3,4 番の順に相関係数が高い特徴量で可視化した Score-CAM 画像である

7 つの遺伝子のうち UBE2C 遺伝子が影響を与える組織形状を Score-CAM で図 18 に可視化すると、組織密度が薄い肺ガン病理組織画像において、細胞密度が高く組織境界部位を示す傾向にあった。

UBE2C 遺伝子はユビキチン結合酵素タンパク質を作り出す。ユビキチン結合酵素はユビキチンに必要不可欠なものである。ユビキチンはタンパク質を修飾しタンパク質分解、DNA 修復、翻訳調節などさまざまな生命現象に関わる[17]。

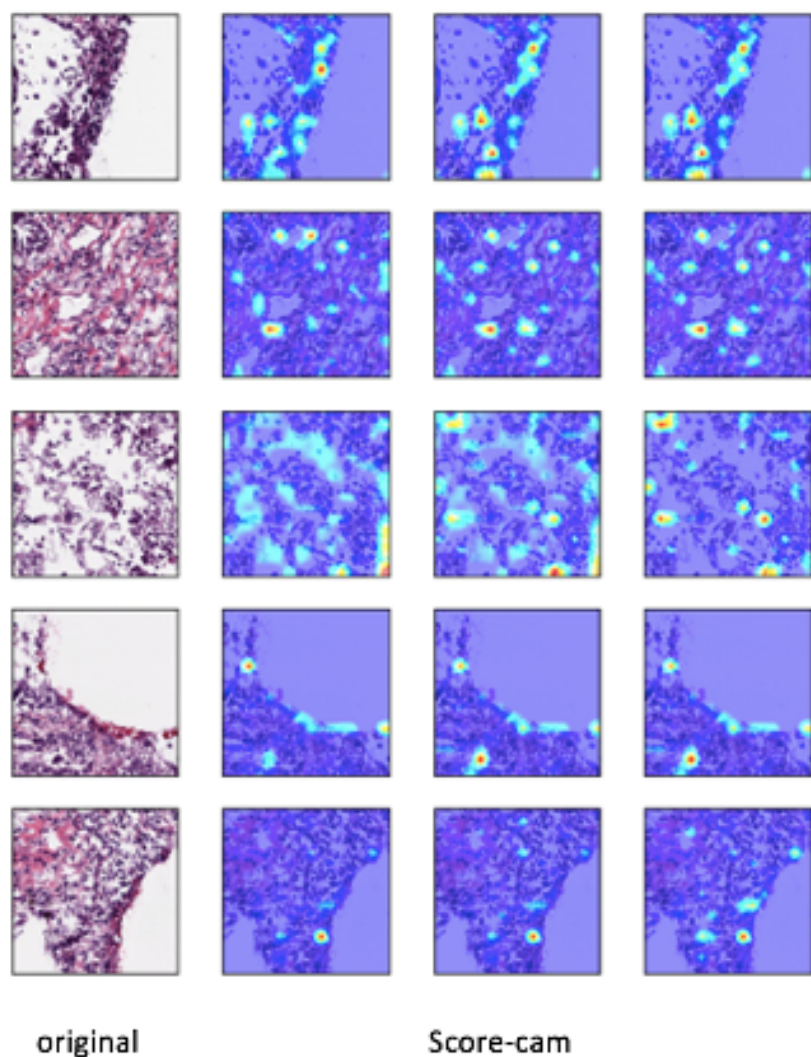


図 18 UBE2C 遺伝子が影響を与える組織形状

CEP55 遺伝子が影響を与える組織形状を Score-CAM で図 19 に可視化すると、組織密度が薄い肺ガン病理組織画像において、細胞密度が高く組織境界部位を示す傾向にあった。

CEP55 遺伝子は中心体タンパク質を発現させる。中心体は微小管生成の中心として機能することが知られており細胞極性を作り出す。細胞極性は細胞に方向性を与える細胞の基本的な性質を担っている[18]。

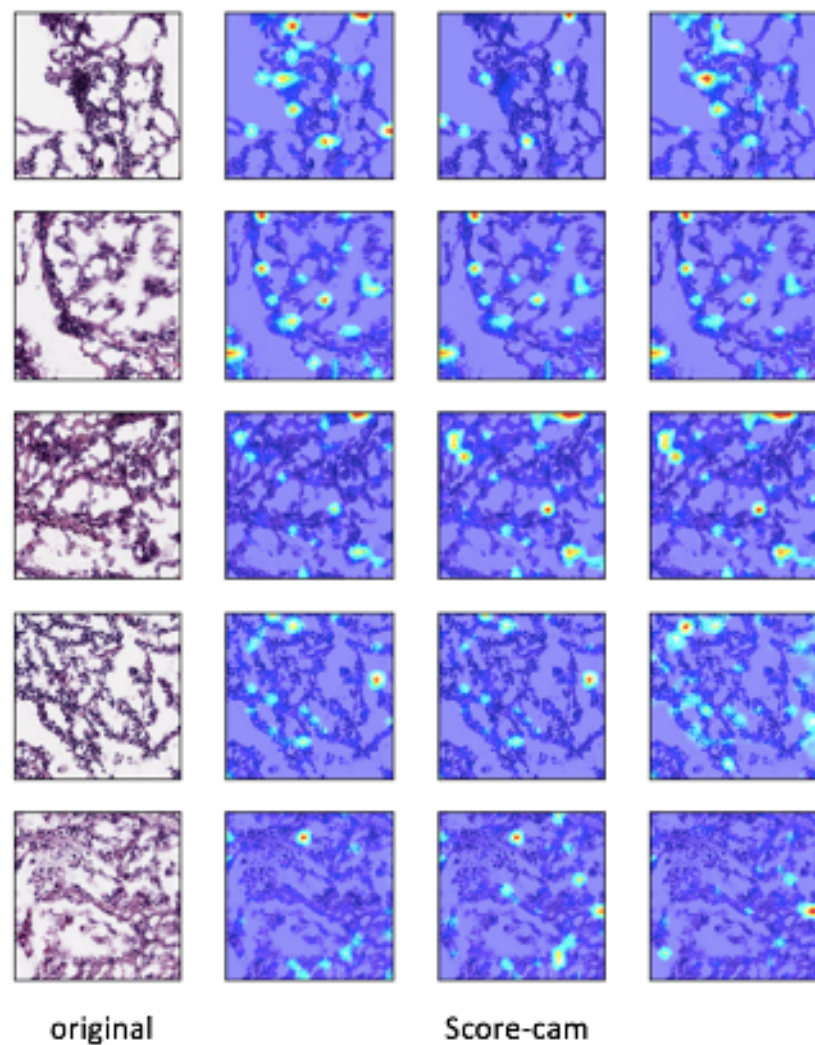


図 19 CEP55 遺伝子が影響を与える組織形状

CDC20 遺伝子が影響を与える組織形状を Score-CAM で図 20 に可視化すると、組織密度が薄い肺ガン病理組織画像において、細胞密度が高く組織境界部位を示す傾向にあった。

CDC20 は細胞分裂周期タンパク質を発現させ、細胞分裂の重要な調節因子である。主に染色体の分離と細胞分裂後期への進行を導く[19]。

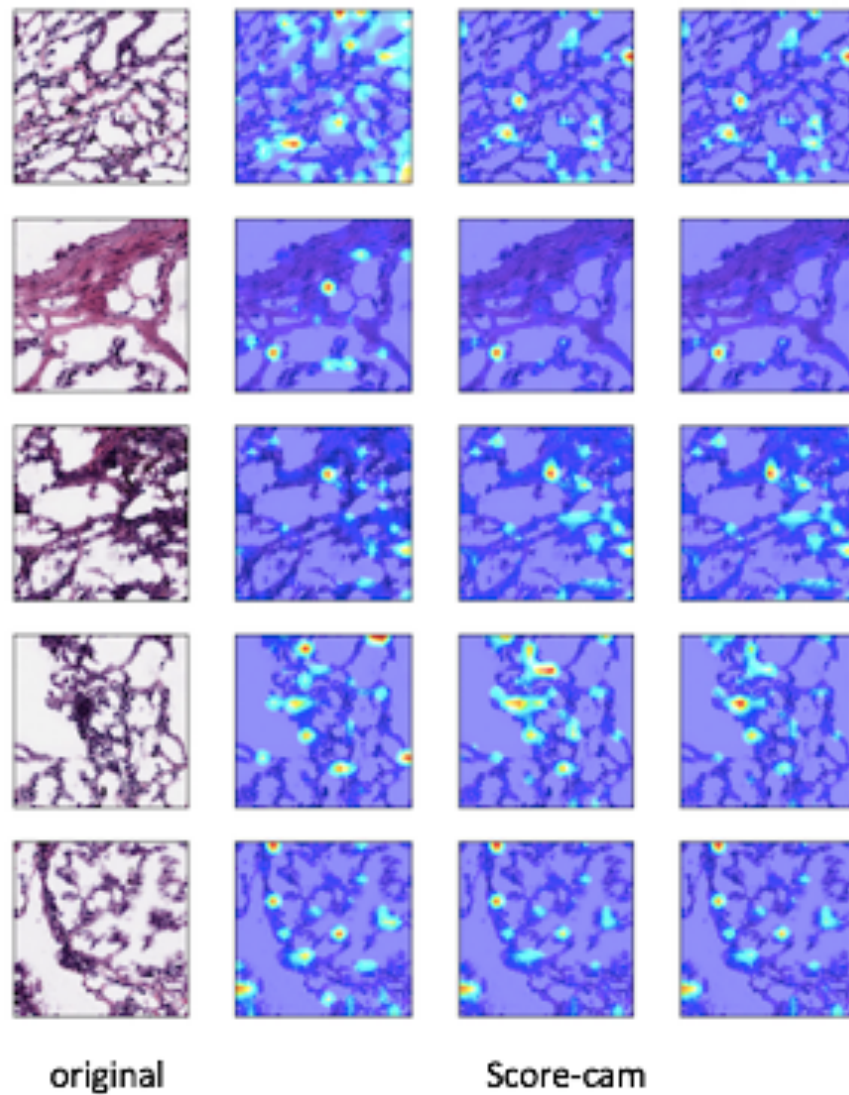


図 20 CDC20 遺伝子が影響を与える組織形状

CCDC109B 遺伝子が影響を与える組織形状を Score-CAM で図 21 に可視化すると、組織密度が高い肺ガン病理組織画像において、細胞密度が高い部位を示す傾向にあった。

CCDC109B 遺伝子はコイルドコイルドメインタンパク質を発現させる。コイルドコイルドメインタンパク質は細胞の膜に見られるイオンチャネルとして働く。イオンチャネルはイオンを透過させるタンパク質である。細胞の膜電位を維持変化させる他、細胞でのイオンの流出入を行う。細胞の膜電位は、細胞内外に大きな電位の差を作ることによって、電位の差を利用した細胞間の情報伝達が可能である[20]。また、悪性細胞の浸潤や移動に関連している。[40]

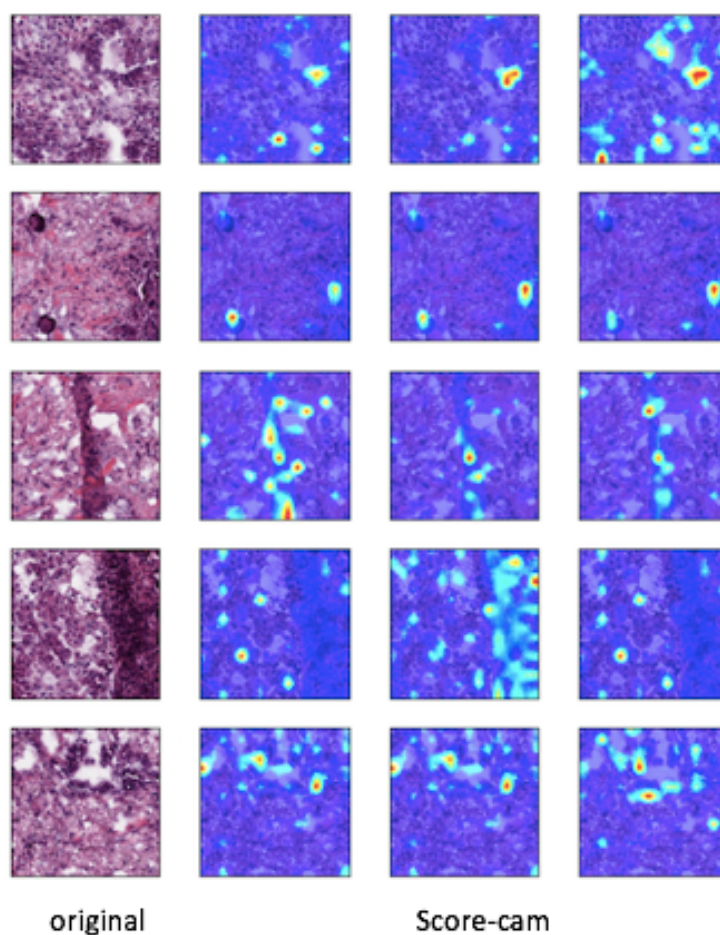


図 21 CCDC109B 遺伝子が影響を与える組織形状

CDKN3 遺伝子が影響を与える組織形状を Score-CAM で図 22 に可視化すると、組織密度が高い肺ガン病理組織画像において、細胞密度が高い部位を示す傾向にあった。

CDKN3 遺伝子は、サイクリン依存性キナーゼ阻害タンパク質を発現させる。サイクリン依存性キナーゼは細胞周期を調節することが知られており、サイクリン依存性キナーゼ阻害タンパク質は細胞周期を止めることで、細胞増殖を止める[21]。

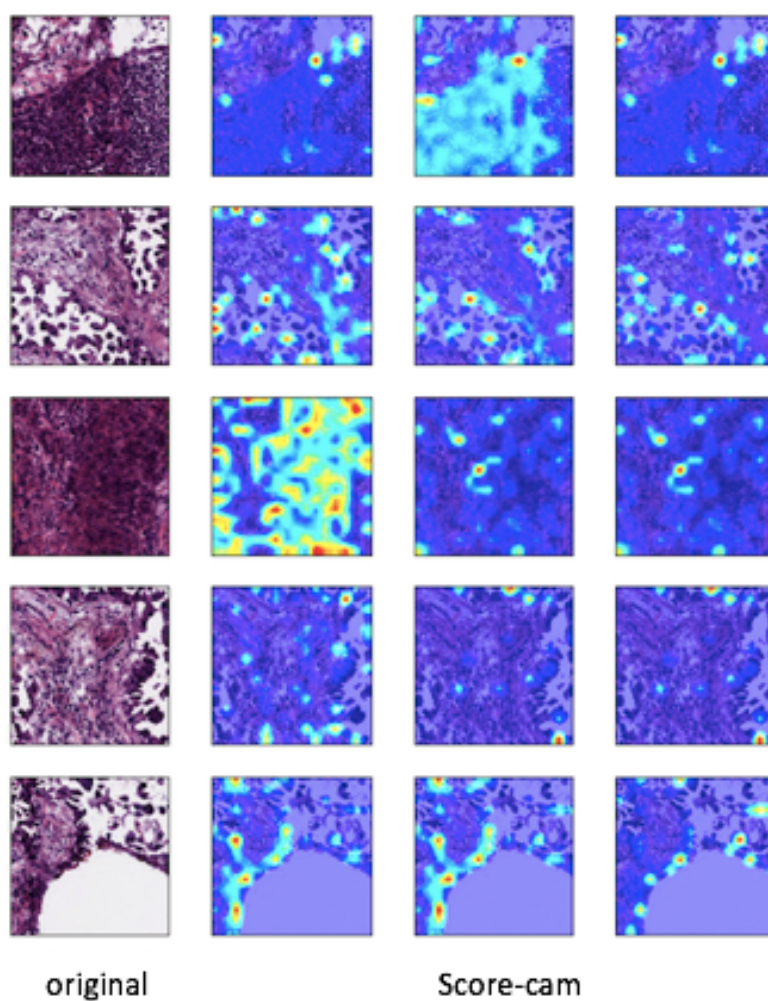


図 22 CDKN3 遺伝子が影響を与える組織形状

KIF11 遺伝子が影響を与える組織形状を Score-CAM で図 23 に可視化すると、組織密度が高い肺ガン病理組織画像において、細胞密度が高い部位と組織の境界部を示す傾向にあった。

KIF11 遺伝子はキネシタンパク質ファミリーに属するモータータンパク質を発現させる。このタンパク質は、紡錘体の形成に関与しており染色体の位置決め、中心体の分離、細胞分裂中の双極紡錘体を導く[22]。

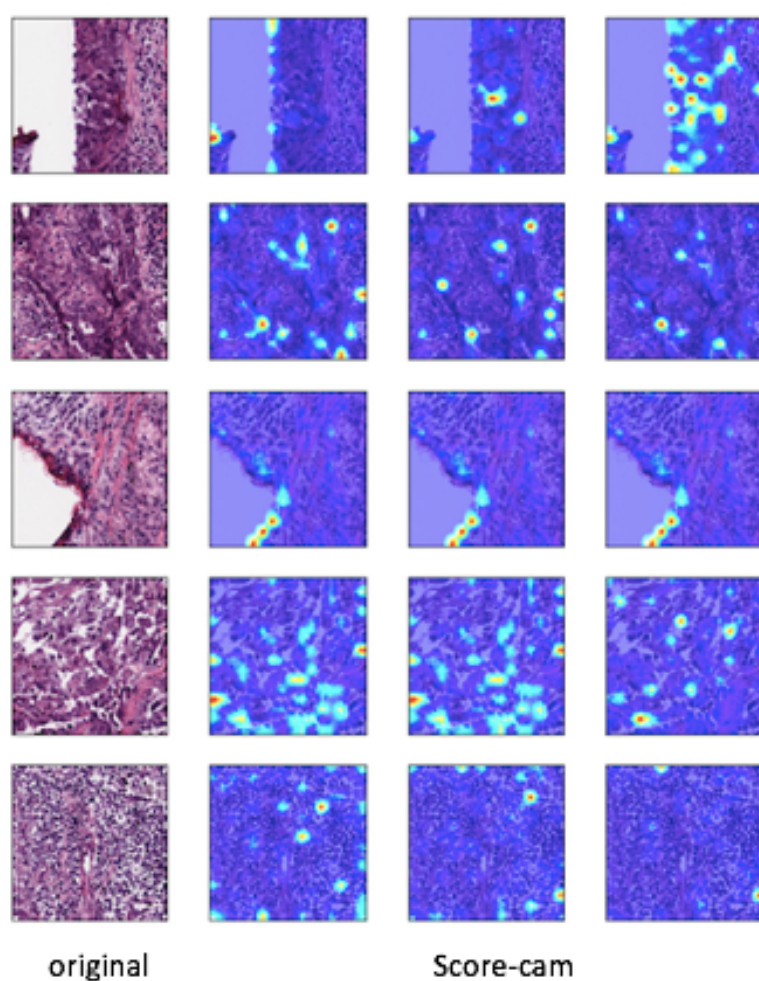


図 23 KIF11 遺伝子が影響を与える組織形状

KPNA2 遺伝子が影響を与える組織形状を Score-CAM で図 24 に可視化すると、組織密度が高い肺ガン病理組織画像において、組織の境界部を示す傾向にあった。

KPNA2 遺伝子は RAG コホートタンパク質を発現させる。RAG コホートタンパク質は DNA ヘリカーゼと相互作用することで、タンパク質の核輸送、細胞移動に関与すると言われている。また、KPNA2 遺伝子をノックダウンすることで、肺癌細胞の増殖、移動能力が抑制されることが判明している [23]。

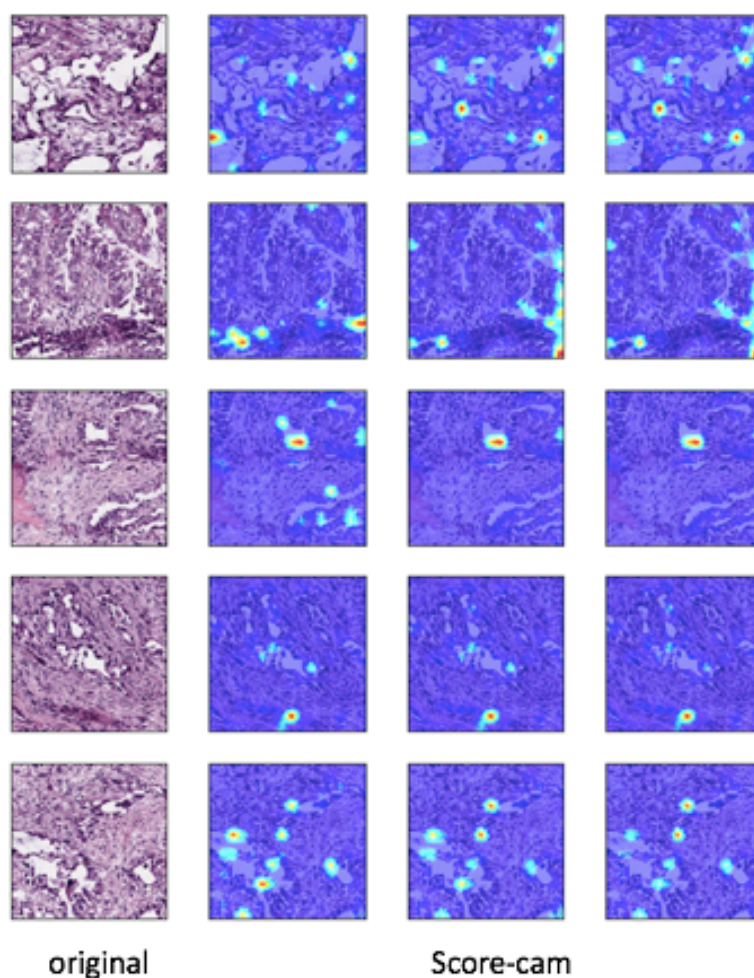


図 24 KIF11 遺伝子が影響を与える組織形状

4.1.2.3 正の相関を取る遺伝子と病理組織形態の関係まとめ

肺ガン病理画像の空間的・位置的情報を集約した特徴量と高い正の相関を取る遺伝子が、肺ガン特異的な特徴と関係しているかを調査するために、前節で UBE2C, CEP55, CDC20, CCDC109B, CDKN3, KIF11, KPNA2 遺伝子の機能を示し、遺伝子が影響を与えている病理組織部位を Score-CAM で可視化した。

その結果、細胞周期・細胞増殖を制御する CDC20, CDKN3, KPNA2, CDKN3 遺伝子では、これらの遺伝子の発現と相関した特徴をとる領域で特に細胞密度の高い傾向があった。また、細胞内で細胞小器官を調節することで細胞形状・細胞周期に影響を与える CEP55, KIF11 遺伝子では、これらの遺伝子の発現と相関した特徴をとる領域で特に細胞密度が高く、細胞形状が異常な傾向があった。

イオンチャネルにより細胞でのイオンの流出入を行い細胞間の情報伝達を行う CCDC109B 遺伝子では、この遺伝子の発現と相関した特徴をとる領域で特に細胞組織間の境界部位を示す傾向があった。

これらの結果から、遺伝子の機能と特徴部位に対応があることが考えられるため、病理組織形態と高い正の相関を取る遺伝子は、肺ガン特異的な特徴に影響を与えていることが裏付けられる。また、他の正の相関を取る遺伝子も肺ガン特異的な特徴に影響を与えていることが予測される。

4.1.2.4 負の相関を取る遺伝子と病理組織形態の関係

遺伝子発現量と特徴量の相関係数の値は、遺伝子が肺ガン組織特徴に与える影響の高さの尺度と考えることができる。特徴量と負の相関係数を取った遺伝子は、遺伝子発現量の減少に比例して、細胞や組織形状を変化させることを示すと予測される。そのため、負の相関係数で top7 の値をとった遺伝子を肺ガン組織に影響を与えやすい遺伝子であると考え。表 14 で負の相関係数で top7 の値をとった遺伝子と、遺伝子が発現させるタンパク質を簡単に示す。

表 10 特徴量と負の相関をとる遺伝子 (top7)

TLR2	Toll 様受容体タンパク質
HLA-DPB1	HLA クラス II 組織適合性抗原タンパク質
HLA-DPA1	HLA クラス II 組織適合性抗原タンパク質
ITGB2	インテグリンサブユニットタンパク質
CD74	HLA クラス II 組織適合性抗原タンパク質
HLA-DRA	HLA クラス II 組織適合抗原タンパク質
HBA2	ヘモグロピンサブユニットタンパク質

肺ガン病理画像の空間的、位置的情報を集約した特徴量と高い負の相関を取る遺伝子は TLR2, HLA-DPB1, HLA-DPA1, ITGB2, CD74, HLA-DRA, HBA2 の 7 つの遺伝子である。7 つの遺伝子のうち TLR2, HLA-DPB1, HLA-DPA1, CD74, HLA-DRA, ITG2 の 6 つの遺伝子が免疫系と関係がある。特徴量との相関を取る遺伝子が組織上でどのような特徴を示すかを Score-CAM で可視化する。Score-CAM ではそれぞれの遺伝子と相関が高かった特徴量を 3 つ用い可視化をおこなった。可視化した画像は左端が元々のガン組織画像である。また、左から 2, 3, 4 番の順に相関係数が高い特徴で可視化した Score-CAM 画像である

TLR2 遺伝子が影響を与える組織形状を Score-CAM で図 25 に可視化すると、組織密度が低い肺ガン病理組織画像において、細胞密度が高く、組織の境界部を示す傾向にあった。

TLR2 は細胞表面に Toll 様受容体タンパク質を発現させる。Toll 様受容体タンパク質は病原体の認識と自然免疫の活性化において基本的な役割を果たし、細胞の炎症反応を調節することが知られている [24]。

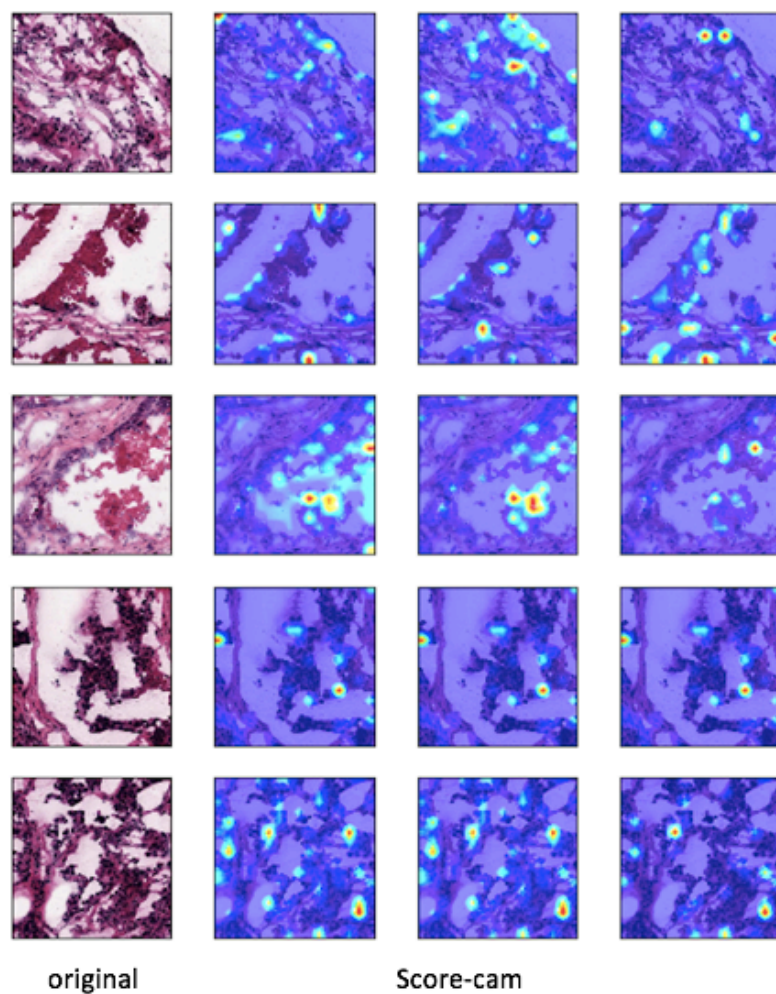


図 25 KIF11 遺伝子が影響を与える組織形状

HLA-DPB1, HLA-DPA1, CD74, HLA-DRA 遺伝子が影響を与える組織形状を Score-CAM で図 26 に可視化すると、組織密度が低い肺ガン病理組織画像において、細胞密度が高く、組織の境界部を示す傾向にあった

HLA-DPB1, HLA-DPA1, CD74, HLA-DRA は抗原提示細胞 (B リンパ球、樹状細胞、マクロファージ) で、HLA クラス II 組織適合性抗原タンパク質を発現させる。B リンパ球、樹状細胞、マクロファージはガン細胞の除去に関わっている細胞である HLA クラス II 組織適合性抗原タンパク質は細胞膜に固定されており、細胞外タンパク質に由来するペプチドを提示することにより、免疫系の中心的な役割を果たす [25][28]。また、CD74 遺伝子は血管新生と関連があることが判明している [27]。

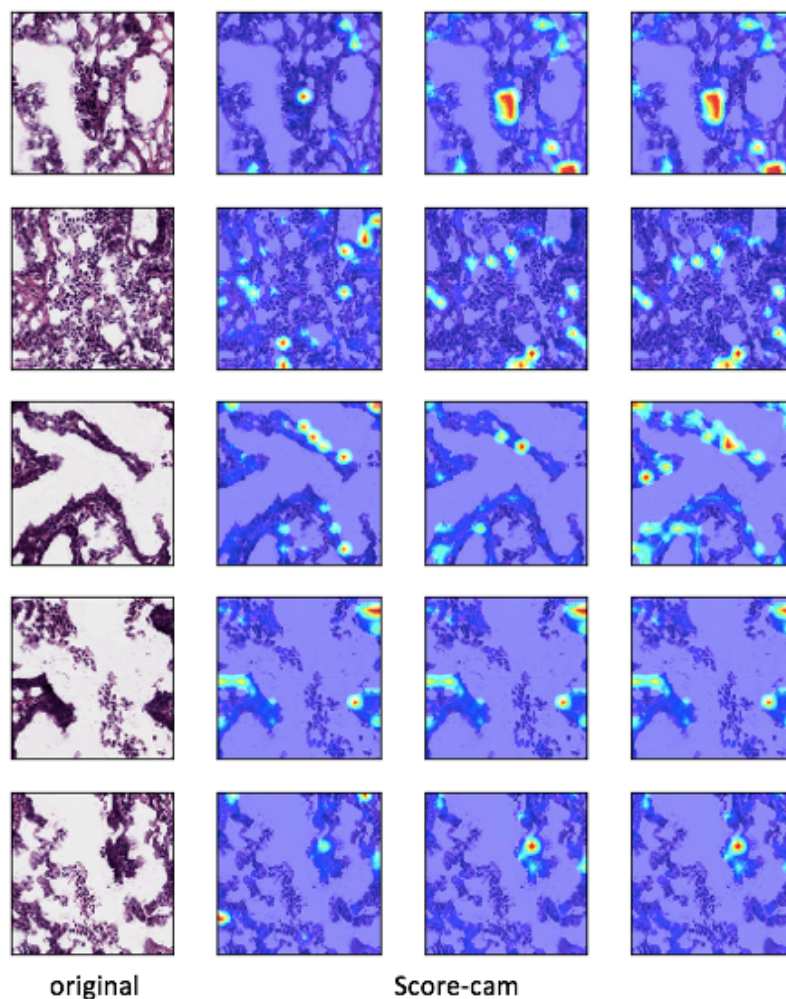


図 26 HLA-DPB1, HLA-DPA1, CD74, HLA-DRA 遺伝子が影響を与える組織形状

ITGB2 遺伝子が影響を与える組織形状を Score-CAM で図 27 に可視化すると、組織密度が低い肺ガン病理組織画像において、細胞密度が高く、組織の境界部を示す傾向にあった。

ITGB2 遺伝子はインテグリンサブユニットタンパク質を発現させる。インテグリンは、細胞接着および細胞表面媒介シグナル伝達に関与する細胞表面タンパク質である。ITGB2 に突然変異で欠陥を生じさせると、白血球接着不全を引き起こすことが知られている [26]。

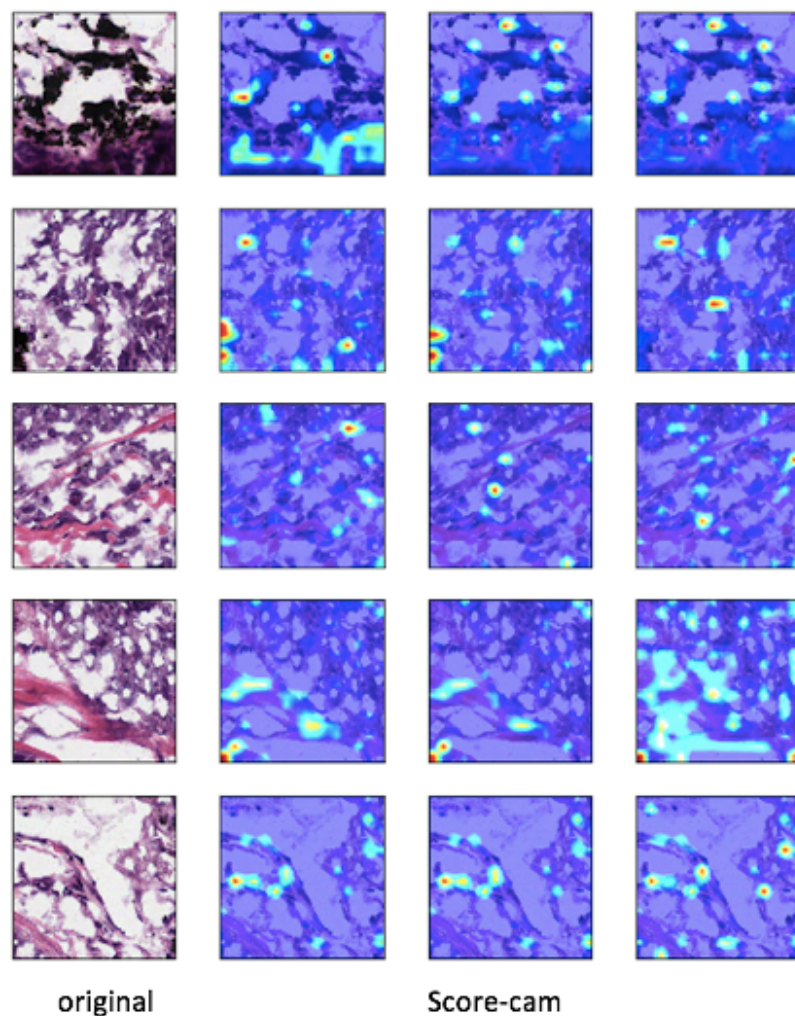


図 27 ITGB2 遺伝子が影響を与える組織形状

HBA2 遺伝子が影響を与える組織形状を Score-CAM で図 28 に可視化すると、組織密度が高い肺ガン病理組織画像において、赤色に染色された血管周辺組織を示す傾向にあった。

HBA2 遺伝子はヘモグロピンサブユニットタンパク質を発現させる。ヘモグロピンサブユニットタンパク質は他のサブユニットと結合しヘモグロピンとして働く、ヘモグロピンは血液中に含まれ、酸素と結合し肺から全身に酸素を行き渡らせる働きを持つ[29]。

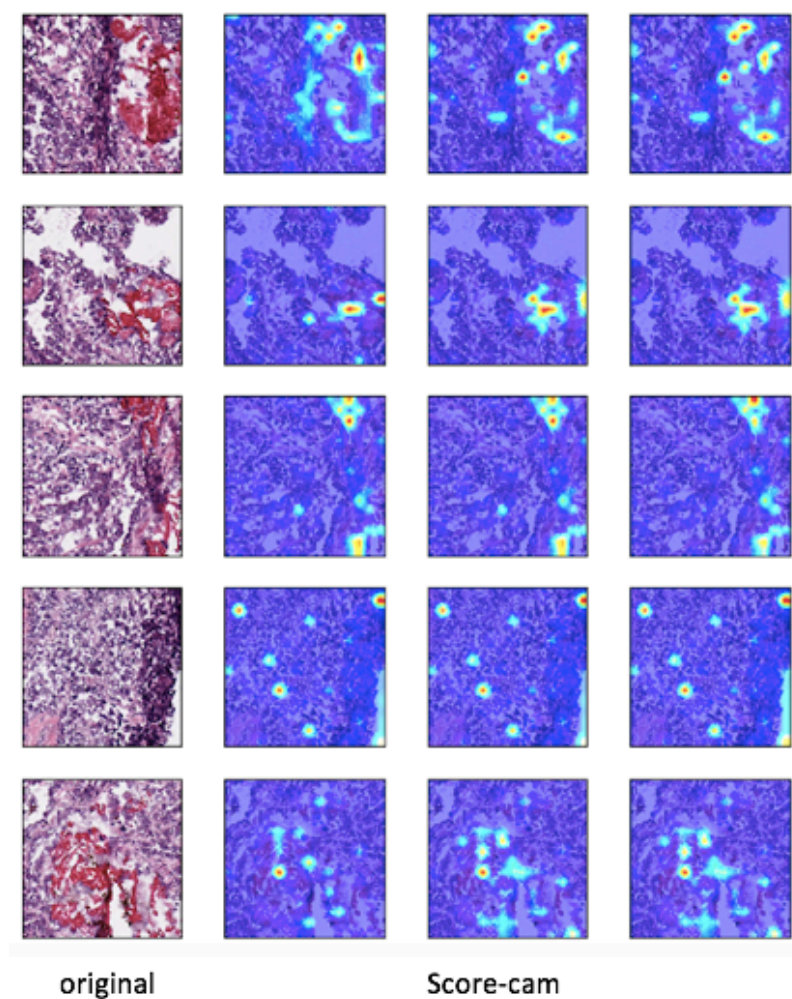


図 28 HBA2 遺伝子が影響を与える組織形状

4.1.2.5 負の相関を取る遺伝子と病理組織形態の関係まとめ

肺ガン病理画像の空間的、位置的情報を集約した特徴量と高い負の相関を取る遺伝子が、肺ガン特異的な特徴と関係しているかを調査するために、前節で遺伝子機能を示し、TLR2, HLA-DPB1, HLA-DPA1, ITGB2, CD74, HLA-DRA, HBA2 遺伝子が影響を与えている病理組織部位を Score-CAM で可視化した。

その結果、異常細胞で発現し異常細胞を破壊する働きを持つ TLR2, HLA-DPB1, HLA-DPA1, CD74, HLA-DRA 遺伝子では、これらの遺伝子の発現と相関した特徴をとる領域で特に細胞密度の高い傾向があった。また、これら6つの遺伝子の発現と相関した特徴をとる領域は同じであった。理由として、同じタンパク質ファミリーに所属することから同じ部位で遺伝子が発現していることが考えられる。

細胞表面シグナル伝達を行うことが知られている ITGB2 遺伝子では、この遺伝子の発現と相関した特徴をとる領域で特に細胞密度の高い部位と細胞質部位を示す傾向があった。

これらの結果から、遺伝子の機能と特徴部位に対応があることが考えられるため、病理組織形態と負の相関を取る遺伝子は、肺ガン特異的な特徴に影響を与えていることが考えられる。

一方、血液の構成要素として知られるヘムグロビンの一部を作成する HBA2 遺伝子では、この遺伝子の発現と相関した特徴をとる領域で特に血管付近に存在する細胞を示す傾向があった。HBA2 遺伝子は、これまでの研究において、ガンに関係しているか不明である。しかし、今回、強い負の相関を示した遺伝子7個のうち6つは、ガンに関わっているため、HBA2 遺伝子も何らかの形でガン組織の形状に関わっていることが考えられる。

4.1.2.6 教師ありCNNにおける相関探索まとめ

肺ガン病理画像の空間的、位置的情報を集約した特徴量と正・負の相関を取る遺伝子が肺ガン特異的な特徴と関係していることが考えられる。

正の相関を取る遺伝子には細胞増殖、細胞修復、イオンチャネルに関わるものが多く、負の相関を取る遺伝子には免疫系、シグナル伝達に関わるものが多かった

4.2 IMSATによる遺伝子発現量と特徴量の相関探索

本節では、遺伝子発現量の差でクラス分けされた3種の肺ガン病理組織サブタイプ画像、TRU(the terminal respiratory unit),PP(the proximal-proliferative),PI(the proximal-inflammatory)をIMSATによりクラスタリングした。クラスタリングは事前分布を指定しないため、純粹に組織形態の差でクラス分けされる。その結果、得られた肺ガン病理画像の空間的、位置的情報を集約した特徴量と遺伝子発現量データの相関を調べることで、組織形態に影響を与える遺伝子を探索する。また、教師なし学習でもガン組織に影響を与える遺伝子を探索可能であるかも同時に探索する。

4.2.1 IMSATによる遺伝子発現量と特徴量の相関係数

IMSATはデータの特徴情報を元にクラスタリングする手法である。そのため、IMSATの中間出力は肺ガン病理画像の空間的、位置的情報を集約した特徴量を集約したものと見做すことができる。そこで、遺伝子発現量(505genes)と特徴量(128dimension)1対1の相関をとることで、どの遺伝子が組織形状に影響を及ぼしているかを確認した。今回使用した特徴量はブランクなものであったので128個の特徴量を削減し38個に絞った。遺伝子と特徴量の相関係数を求めたところ、19190個の相関係数が得られた。

相関係数の分布をヒートマップで示した図(図 9)を以下に示す

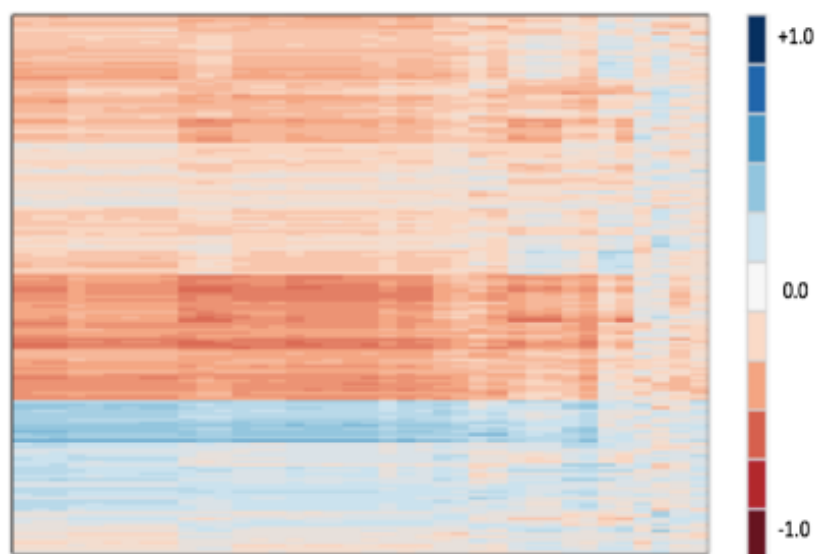


図 29 遺伝子発現量と特徴量間の相関係数ヒートマップ
(横:特徴量(38 dimension),縦:遺伝子発現量(505 genes))

次に、相関係数の分布と最大値、最小値を調べるためにヒストグラムを作成した。(図 27)

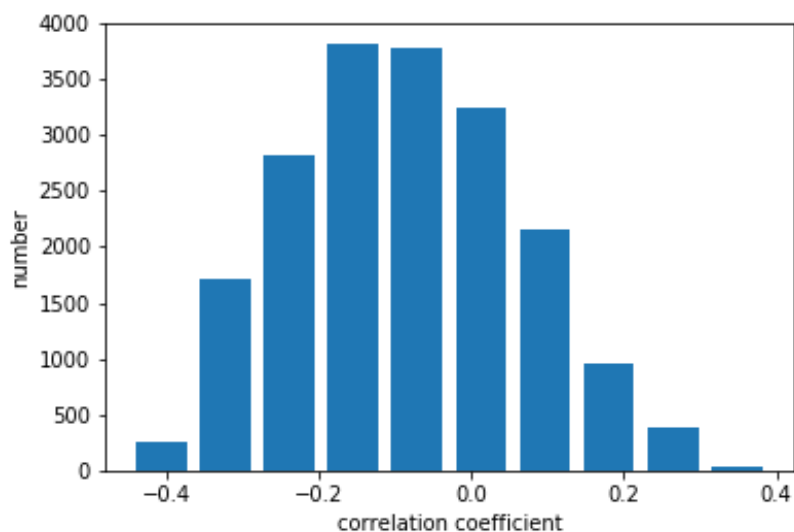


図 30 遺伝子発現量と特徴量間の相関係数分布

相関係数の分布は正規分布で、正の相関係数の最大値は 0.37 で負の相関係数の最小値は -0.43 であった。

4.2.1.2 正の相関を取る遺伝子と病理組織形態の関係

遺伝子発現量と特徴量の相関係数の値は、遺伝子が肺ガン組織特徴に与える影響の高さの尺度と考えることができる。特徴量と正の相関係数を取った遺伝子は遺伝子発現量の増幅に比例して、細胞や組織形状を変化させることを示すと予測される。相関係数が高い遺伝子を二つ例に挙げる(表 11)。

表 11 特徴量と正の相関をとる遺伝子(相関係数 0.35 以上)

TOB1	トランスデューサータンパク質
NEDD4L	ユビキチンプロテインリガーゼタンパク質

肺ガン病理画像の空間的、位置的情報を集約した特徴量と正の相関を取る遺伝子は TOB1, NEDD4L, 2 つの遺伝子であった。特徴量との相関を取る遺伝子が組織上でどのような特徴を示すかを Score-CAM で可視化する。Score-CAM ではそれぞれの遺伝子と相関が高かった特徴量を 3 つ用い可視化をおこなった。可視化した画像は左端が元々のガン組織画像である。また、左から 2, 3, 4 番の順に相関係数が高い特徴で可視化した Score-CAM 画像である

TOB1 遺伝子が影響を与える組織形状を Score-CAM で図 31 に可視化すると、組織密度が高い肺ガン病理組織画像において、細胞密度が高い部位を示す傾向にあった。

TOB1 遺伝子は細胞転座遺伝子タンパク質ファミリーをコントロールするタンパク質を発現させる。細胞転座遺伝子タンパク質ファミリーは細胞周期・細胞増殖を調節するタンパク質であり、TOB1 遺伝子は腫瘍抑制因子として機能すると考えられている[30]。

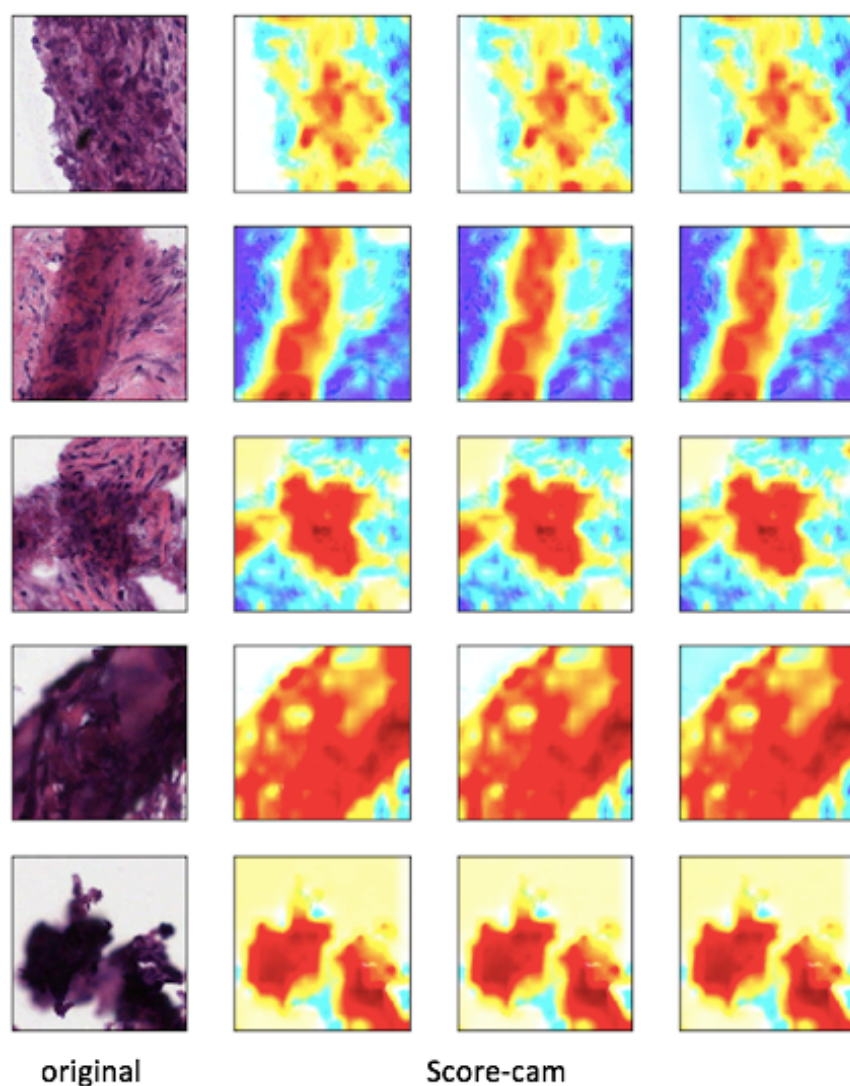


図 31 TOB1 遺伝子が影響を与える組織形状

NEDD4L 遺伝子が影響を与える組織形状を Score-CAM で図 32 に可視化すると、組織密度が比較的高い肺ガン病理組織画像において、細胞密度が高く、組織境界面を示す傾向にあった。

。NEDD4L 遺伝子はユビキチンリガーゼタンパク質をコードしており、細胞表面のイオンチャネルの発現量を調節することが知られている。また、細胞増殖調整に関わる TFG— β を制御していることが知られている [31]。

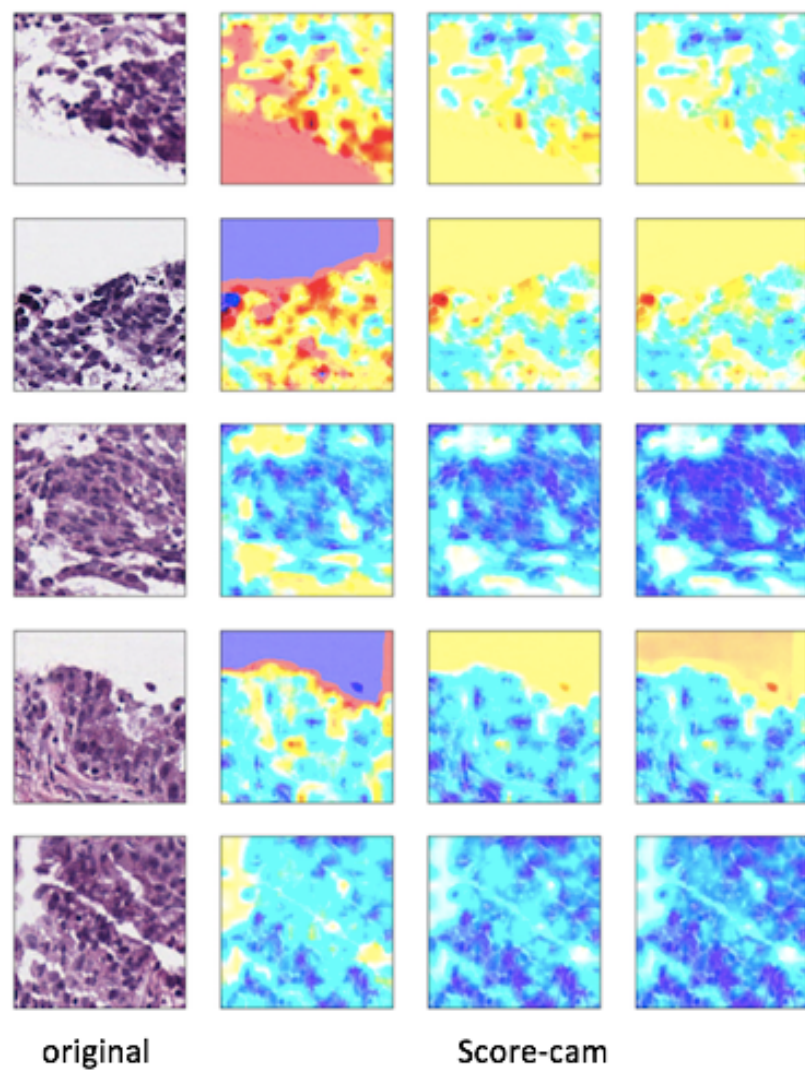


図 32 NEDD4L が影響を与える組織形状

4.2.1.3 正の相関を取る遺伝子と病理組織形態の関係まとめ

肺ガン病理画像の空間的、位置的情報を集約した特徴量と正の相関を取る遺伝子が、肺ガン特異的な特徴と関係しているかを調査するために、前節で TOB 1、MEDD 4 L 遺伝子の機能を示し、遺伝子が影響を与えている病理組織部位を Score-CAM で可視化した。

その結果、細胞周期・細胞増殖を調節する働きを持つ TOB 1 遺伝子では、この遺伝子の発現と相関した特徴をとる領域で特に細胞密度の高い傾向があった。

また、イオンチャネルの調節し細胞間と細胞内で情報伝達を行うことで、細胞周期を制御する MEDD 4 L 遺伝子では、この遺伝子の発現と相関した特徴をとる領域で特に細胞密度の高い傾向があった。

これらの結果から、遺伝子の機能と特徴部位に対応があることが考えられるため、特徴量と正の相関を取る遺伝子は、肺ガン特異的な特徴に影響を与えていることが考えられる。また、教師なし CNN(IMSAT)により次元削減された肺ガン組織と正の相関のある遺伝子は教師あり CNN で正の相関の高かった遺伝子と機能(細胞増殖、イオンチャネル)が類似していた。そのため IMSAT(教師なし CNN)を用いた遺伝子と肺ガン特異的な特徴の関係探索が可能であることが示唆された。

4.2.1.4 負の相関を取る遺伝子と病理組織形態の関係

特徴量と負の相関係数を取った遺伝子は遺伝子発現量の増幅に比例して、細胞や組織形状を変化させることを示すことが予測される。相関係数が高い遺伝子を二つ例にあげる (表 12)。

表 12 特徴量と負の相関をとる遺伝子(相関係数-0.42 以下)

FCGR2A	免疫グロブリン受容体タンパク質
DSE	腫瘍拒絶抗原タンパク質

肺ガン病理画像の空間的、位置的情報を集約した特徴量と高い負の相関を取る遺伝子は FCGR 2 A、DSE の 2 つの遺伝子であった。特徴量との相関を取る遺伝子が組織上でどのような特徴を示すかを Score-CAM で可視化する。Score-CAM ではそれぞれの遺伝子と相関が高かった特徴量を 3 つ用い可視化をおこなった。可視化した画像は左端が元々のガン組織画像である。また、左から 2,3,4 番の順に相関係数が高い特徴で可視化した Score-CAM 画像である

FCGR2A 遺伝子が影響を与える組織形状を Score-CAM で図 33 に可視化すると、組織密度が高い肺ガン病理組織画像において、細胞密度が高い部位と細胞質部位を示す傾向にあった。

FCGR2A 遺伝子は免疫グロブリン受容体タンパク質を発現させ、異常細胞や細菌の貪食作用や免疫複合体の除去に関与している[32]。

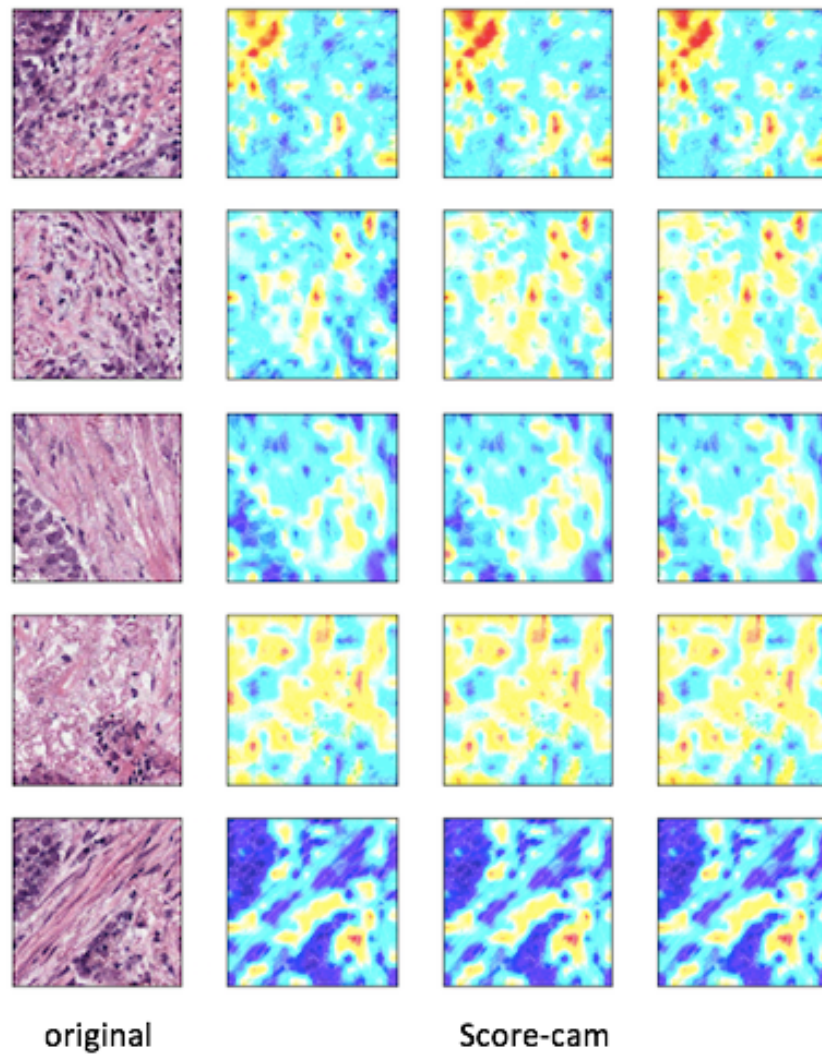


図 33 FCGR2A 遺伝子が影響を与える組織形状

DSE 遺伝子が影響を与える組織形状を Score-CAM で図 34 に可視化すると、組織密度が高い肺ガン病理組織画像において、細胞質部位が示す傾向にあった。

DSE 遺伝子は腫瘍拒絶抗原タンパク質を発現する。腫瘍拒絶抗原タンパク質は小胞体に局在しており、T リンパ球を誘導することで腫瘍を破壊する [33]。

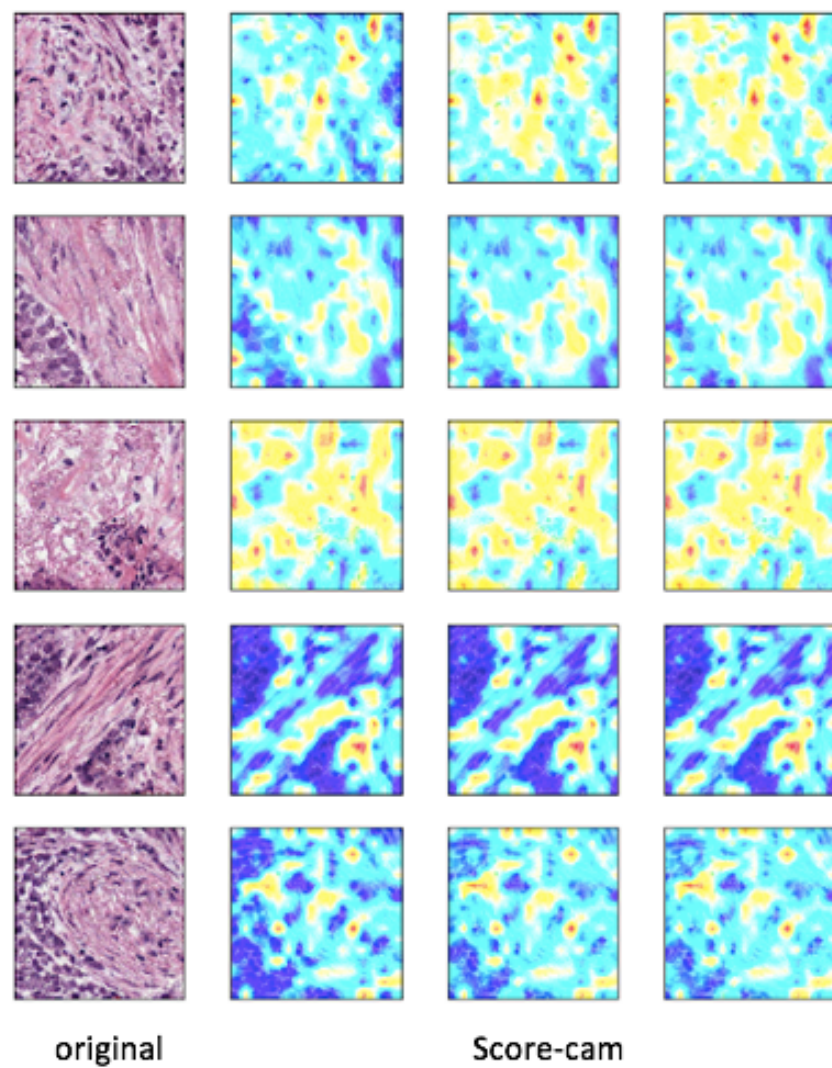


図 34 DSE 遺伝子が影響を与える組織形状

4.2.1.5 負の相関を取る遺伝子と病理組織形態の関係まとめ

肺ガン病理画像の空間的、位置的情報を集約した特徴量と負の相関を取る遺伝子が肺ガン特異的な特徴と関係しているかを調査するために、前節で FCGR2A、DSE 遺伝子の機能を示した。また、各遺伝子が影響を与えている病理組織部位を **Score-CAM** で可視化した。

その結果、細胞表面で免疫受容体を作り出すことで異常細胞を破壊する機能を持つ FCGR2A 遺伝子では、この遺伝子の発現と相関した特徴をとる領域で特に細胞密度の高い部位と細胞質部位を示す傾向があった

また、腫瘍拒絶抗原タンパク質を細胞内の小胞体で発現させる働きを持つ DSE 遺伝子では、この遺伝子の発現と相関した特徴をとる領域で特に細胞質(細胞全体)部位を示す傾向があった

これらの結果から、遺伝子の機能と特徴部位に対応があることが考えられるため、特徴量と負の相関を取る遺伝子は、肺ガン特異的な特徴に影響を与えていることが考えられる。また、教師なし CNN(IMSAT)により次元削減された肺ガン組織特徴と負の相関のある遺伝子は、教師あり CNN で負の相関の高かった遺伝子と機能(免疫作用)が類似していた。そのため IMSAT(教師なし CNN)を用いた遺伝子と肺ガン特異的な特徴の関係探索が可能であることが示唆された。

4.2.1.4 IMSATにおける相関探索まとめ

肺ガン病理画像の空間的、位置的情報を集約した特徴量と正・負の相関を取る遺伝子が肺ガン特異的な特徴と関係していることが考えられる。

正の相関を取る遺伝子には細胞増殖、細胞修復に関わるものが多く、負の相関を取る遺伝子には免疫系、シグナル伝達に関わるものが多かった

4.3 まとめ

教師あり CNN で組織形態に影響を与えている遺伝子を探索したところ、細胞分裂、細胞分解、イオンチャネル、免疫系に関わる一部の遺伝子がガン組織の形態に影響を与えていることが示唆された。また教師なし CNN(IMSAT)でも細胞分裂、イオンチャネル、免疫系に関わる一部の遺伝子がガン組織の形態に影響を与えていることが示唆された。教師あり CNN と教師なし CNN(IMSAT)それぞれで関係する遺伝子の働きが類似しているため、教師なし CNN(IMSAT)でもガン組織の形態に影響を与える遺伝子を探索可能であることが考えられる。また、IMSAT ではクラスタ数を任意に変更して学習可能であるため、クラス数を変更することでより多様な特徴と遺伝子の関係を得ることが出来ると予想される。

5 結論

本研究では、畳み込みニューラルネットワーク (CNN) を用いて肺ガン組織の形態に影響を与える遺伝子の探索を行った。CNN は画像の形態的、空間的情報を抽出し低次元空間に落とし込むことができると知られている。そのため、教師あり CNN により肺癌病理組織を遺伝子発現量の差により分類し、遺伝子発現量に基づいて集められた病理組織の特徴を得た。得られた特徴と遺伝子発現量の相関を求めることで、組織形態を変化させる可能性のある遺伝子を提案することが出来た。また、深層学習に基づいた教師なしクラスタリング手法として知られている IMSAT を用い、遺伝子発現情報を事前に与えない CNN においても組織形態を変化させる可能性のある遺伝子を提案することができるかを確認した。IMSAT を用い得られた肺癌組織画像特徴と遺伝子発現量の相関を求めると、教師あり学習で得られた遺伝子と似た様な機能を持つ遺伝子を提案することができた。このことより、教師なし学習でも組織形態に影響を与える遺伝子を探索できることが示唆された。本研究では、組織形状に影響を与える遺伝子を特定した。これらの遺伝子を生物学的研究で理解を深めることにより、ガンに対する知見を深められることが期待される。

6 謝辞

本研究は多くの方々からの御指導、御協力をいただき、進めることができました。本研究を行うにあたり、丁寧な御指導と研究環境を与えてくださった奈良先端科学技術大学院大学情報科学研究科計算システムズ生物学研究室 金谷 重彦 教授に深く感謝致します。他大学の学部生であった私に快く研究室にて勉強する機会を与えてくださったことが、私のこれまでの人生の中での大きな転機となりました。また、研究室ゼミでは的確な御意見を頂きました情報科学研究科 Md. Altaf-Ul-Amin 准教授、情報科学研究科 黄 銘 助教に深く感謝致します。

研究発表の際、貴重な御意見及び助言を賜りました奈良先端科学技術大学院大学情報科学研究科 佐藤 嘉伸 教授に厚く御礼申し上げます。

日頃から研究のご助言を頂きました情報科学研究科小野 直亮 准教授に深く感謝致します。

計算システムズ生物学研究室の皆様には、大学院生活全般にわたり様々な面で支えていただきました。皆様のおかげで充実した大学院生活を過ごすことができました。誠にありがとうございました。

最後に、大学院修了まで多大なる支援を頂いた家族に深く感謝し、ここに謝辞とさせていただきます。

参考文献

[1]	YanmingGuo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu ,Michael S.Lew, “Deep learning for visual understanding,” <i>Neurocomputing</i> , 2016. 187, 27-48
[2]	Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, “Going Deeper with Convolutions,” <i>arXiv</i> , 2014.1409.4842
[3]	Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Andrew H. Beck, “Deep Learning for Identifying Metastatic Breast Cancer”, <i>arXiv</i> ,2016,1606.05718
[4]	Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel “Backpropagation applied to handwritten zip code recog- nition”. <i>Neural Computation</i> ,1989,1,541-551,
[5]	Ronglai Shen, Adam B. Olshen,Marc Ladanyi, “Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis”, <i>Bioinformatics</i> ,2009,25,22,2906-2912
[6]	Xiao-Xiao Niu,Ching Y.Suen, “A novel hybrid CNN–SVM classifier for recognizing handwritten digits”, <i>Pattern Recognition</i> , 2012,45,1318-1325
[7]	Patricia A. J. Muller & Karen H. Vousden,“p53 mutations in cancer”, <i>Nature Cell Biology</i> ,2013,15,2-8
[8]	Nicolas Coudray,Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyő, Andre L. Moreira ³ Narges Razavian and Aristotelis Tsirigos, “Classification and mutation prediction from non–small cell

	lung cancer histopathology images using deep learning”, <i>nature medicine</i> ,2018,24,1559-1567
[9]	Hawazin Faruki, Gregory M.Mayhew PH,Jonathan S.Serody ,D. Neil Hayes ,Charles M.Perou, Myla Lai-Goldman, “Lung Adenocarcinoma and Squamous Cell Carcinoma Gene Expression Subtypes Demonstrate Significant Differences in Tumor Immune Landscape”, <i>Journal of Thoracic Oncology</i> ,2017,12,943-953
[10]	Min Chen ,Xiaobo Shi, Yin Zhang,Di Wu,Mohsen Guizani, “Deep Features Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network”, <i>IEEE Transactions on Big Data</i> ,2017,1
[11]	Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, Masashi Sugiyama ,“Learning Discrete Representations via Information Maximizing Self-Augmented Training”,2017, <i>arXiv</i> ,1702.08720
[12]	Grant J. Scott,Matthew R.England, William A. Starms, Richa, “Training Deep Convolutional Neural Networks for Land–Cover Classification of High-Resolution Imagery”, <i>IEEE Geoscience and Remote</i> ,2017,549-553
[13]	Luis Perez, Jason Wang ,“he Effectiveness of Data Augmentation in Image Classification using Deep Learning”,2017, <i>arXiv</i> :1712.04621
[14]	Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, Shin Ishii ,“Distributional Smoothing with Virtual Adversarial Training”2016, <i>arXiv</i> :1507.006777
[15]	Haofan Wang, Mengnan Du, Fan Yang, Zijian Zhang,“Score-CAM:Improved Visual Explanations Via Score-Weighted Class Activation Mapping”,2019, <i>arXiv</i> :1910.01279

[16]	Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”, 2016, <i>arXiv</i> :1610.02391
[17]	Zhonglin Hao, Hui Zhang, John Cowell, “Ubiquitin-conjugating enzyme UBE2C: molecular biology, role in tumorigenesis, and potential as a biomarker”, <i>Tumor Biology</i> , 2012, 33, 723-730
[18]	Ligang Liu, Qi Mei, Jing, Zhao Yuhong Dai, Qiang Fu, “Suppression of CEP55 reduces cell viability and induces apoptosis in human lung cancer”, <i>SPANDIDOS PUBLICATIONS</i> , 2016, 1939-1945
[19]	Tatsuya Kato, Yataro Daigo, Masato Aragaki, Keidai Ishikawa, Masaaki Sato, Mitsuhito Kaji, “Overexpression of CDC20 predicts poor prognosis in primary non - small cell lung cancer patients”, <i>Journal of Surgical Oncology</i> , 106, 4
[20]	Weibo Liu, Haifeng Wei, Zhengyu Gao, Guanghui Chen, Yujie Liu, Xin Gao, Guangjian Bai, “COL5A1 may contribute the metastasis of lung adenocarcinoma”, <i>Gene</i> , 2018, 665, 57-66
[21]	Chao Fan, Lu Chen, Qingling Huang, Tao Shen, Eric A Welsh, Jamie K Teer, Jianfeng Cai, W Douglas Cress & Jie Wu, “Overexpression of major CDKN3 transcripts is associated with poor survival in lung adenocarcinoma”, <i>British Journal of Cancer</i> , 2015, 113, 1735-1743
[22]	Marc A. Schneider Petros Christopoulos Thomas Muley Arne Warth Ursula Klingmueller, “AURKA, DLGAP5, TPX2, KIF11 and CKAP5: Five specific mitosis-associated genes correlate with poor prognosis for non-small cell lung cancer patients”, <i>SPANDIDOS PUBLICATIONS</i> , 2017, 365-372

[23]	Chun-I Wang, Kun-Yi Chien, Chih-Liang Wang, Hao-Ping Liu, Chia-Chen Cheng, Yu-Sun Chang, Jau-Song Yu and Chia-Jung Yu, “Quantitative Proteomics Reveals Regulation of Karyopherin Subunit Alpha-2 (KPNA2) and Its Potential Novel Cargo Proteins in Nonsmall Cell Lung Cancer” <i>MOLECULAR & CELLULAR PROTEOMICS</i> ,2012,11,1105,1122
[24]	Aldo Pinto,SilvanaMorello,Rosalinda Sorrentino,“Lung cancer and Toll-like receptors”, <i>Cancer Immunology</i> ,2011,60,1211
[25]	Hiroshi Kamma,Takuya Yazawa,Takesaburo Ogata,Hisashi Horiguchi,Tatsuo Iijima,“Expression of MHC class II antigens in human lung cancer cells”, <i>Virchows Archiv B</i> ,1991,60,407
[26]	K Rubio, S Dobersch, G Barreto ,“Non-canonical ITGB2 signaling mediates resistance to TKI”, <i>Pnemologic</i> ,2019,73
[27]	Marc McClelland ,Liu Jian Zhao, Shannon Carskadon, Douglas Arenberg,“Expression of CD74, the Receptor for Macrophage Migration Inhibitory Factor, in Non-Small Cell Lung Cancer” <i>The American Journal of Pathology</i> ,2009,638-646
[28]	Aaron Osborne, Michael Tschickardt, George Blanch, “Retinoblastoma Protein Expression Facilitates Chromatin Remodeling at the HLA-DRA Promoter” <i>Nucleic Acid Research</i> , 1997,25,24
[29]	GeneCards:HBA2, http://web.fukuyama-u.ac.jp/pharm/htmls/Labo/labs/genome/link2/link.html (2020-01-23)
[30]	Yoshihiro Hiramatsu, Kyoko Kitagawa, Toru Suzuki, Chiharu Uchida, “Degradation of Tob1 Mediated by SCFSkp2-Dependent Ubiquitination”, <i>Cancer Research</i> ,2006,06-1603

[31]	Hiroiyuki Sakashita,Hiroshi Inoue,Shinji Akamine,Teruyoshi Ishida,Naohiko Inase,Kuniaki Shirao,Masaki Mori,Koshi Mimori, “Identification of the NEDD4L Gene as a Prognostic Marker by Integrated Microarray Analysis of Copy Number and Gene Expression Profiling in Non-small Cell Lung Cancer”, Annals of Surgical Oncology,2013,20,590-598
[31]	Mei-Hua,Qu,Chunhua,Han,Amit,Kumar,Srivastava,Tiantian Cui,Ning Zou,Zhi-Qin Gao,Qi-En Wang,“miR-93 promotes TGF- β -induced epithelial-to-mesenchymal transition through downregulation of NEDD4L in lung cancer cells”,Tumor Biology,2016,37,5645-5651
[32]	Janardan P. Pandey,“Prognostic Immune Markers in Non–Small Cell Lung Cancer—Letter”,2011,11-2103
[33]	GeneCards : DSE, https://www.genecards.org/cgi-bin/carddisp.pl?gene=DSE (2020-01-23)
[34]	Hoo-Chang Shin,Holger R. Roth,Mingchen Gao,“Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning”,IEEE,2016,35,5
[35]	Carl Doersch,“Tutorial on Variational Autoencoders”,arXiv,2016,1606.05908
[36]	Jonghoon Jin, Aysegul Dundar, Eugenio Culurciello,“Flattened Convolutional Neural Networks for Feedforward Acceleration”, arXive,2015,1412.5474
[37]	Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, Serge Belongie,“Kernel Pooling for Convolutional Neural Networks”,CVPR2017,2921-2939
[38]	Maxime Oquab, Leon Bottou, Ivan Laptev, Josef Sivic,“Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks”,CVPR2014,1717-1724

[39]	Fatih Ozsolak, Patrice M. Milos, “RNA sequencing: advances, challenges and opportunities”, Nature Reviews Genetics,12,87-98
[40]	Ran Xu, Mingzhi Han, Yangyang Xu, Xin Zhang, Chao Zhang, Di Zhang, Jianxiong Ji, Yuzhen Wei, “Coiled-coil domain containing 109B is a HIF1 α -regulated gene critical for progression of human gliomas”,Journal of Translational Medicine,2017,15,165

付録

教師あり CNN で 3 種のガン組織サブタイプを分類した際、特徴量と相関の高い正・負 TOP 24 の遺伝子リスト(表 14)

表 13 特徴量と相関の高い正・負 TOP 25 の遺伝子リスト(教師あり CNN)

	正の相関	負の相関
1	UBE2C	TLR2
2	CEP55	HLA-DPB1
3	CDC20	HLA-DPA1
4	CCDC109B	HLA-DRB1
5	CDKN3	ITGB2
6	KIF11	CD74
7	KPNA2	HLA-DRA
8	HSPD1	HLA-DMB
9	HSPE1	HBA2
10	AKR1C2	FAS
11	GPX2	DPYD
12	PLAU	AIM1
13	HSPD1	FGL1
14	HSPE1	DSE
15	AKR1C2	CCNB1
16	GPX2	SLC16A3
17	PLAU	RAP1GAP
18	MELK	SLC15A3
19	TPX2	LHFPL2
20	KIF4A	FLNA
21	DLG7	HLA-E
22	CCNB2	TRIM22
23	CCNA2	PBK
24	TTK	TGFB1

教師なし CNN(IMSAT)で 3 種のガン組織サブタイプを分類した際、
特徴量と相関の高い正・負 TOP 24 の遺伝子リスト

表 14 特徴量と相関の高い正・負 TOP 25 の遺伝子リスト(IMSAT)

	正の相関	負の相関
1	TOB1	FCGR2A
2	NEDD4L	DSE
3	VIPR1	GREM1
4	RAP1GAP	MMP12
5	TSPAN8	SEC23A
6	ERBB3	COL5A2
7	ATP1B1	HCK
8	GSTA1	CSF1R
9	SORBS2	LAPTM5
10	ATP2C2	MS4A6A
11	CLDN18	FAP
12	PLA2G10	AIM2
13	SLC44A4	OLFML2B
14	MRPL13	TUBB6
15	HSPE1	TNFAIP6
16	CACNA2D2	COL6A2
17	GPX2	CTSK
18	MCF2L	MAFB
19	CYP4B1	COL5A1
20	SELENBP1	CLEC2B
21	GJB1	SLAMF8
22	ALDH3A2	COL1A1
23	KIF20A	LCP2
24	CIT	GLIPR1