

# 機械学習による胸部 CT 画像からの ドライバー遺伝子情報変異有無の識別法

## A Classification Method for Mutation of Driver Gene Information

### from Thoracic CT Images Based on Machine Learning

吉福 優汰<sup>1</sup>, 神谷 亨<sup>1</sup>, 寺澤 岳<sup>2</sup>, 青木 隆敏<sup>2</sup>

1) 九州工業大学

2) 産業医科大学

Yuta Yoshifuku<sup>1</sup>, Tohru Kamiya<sup>1</sup>, Takashi Terasawa<sup>2</sup> and Takatoshi Aoki<sup>2</sup>

1) *Kyushu Institute of Technology*

2) *University of Occupational and Environmental Health School of Medicine*

**Abstract:** Genetic testing confirms the mutation in driver gene information involved in cancer cell growth. If the mutation is identified, molecularly targeted drugs with significantly higher response rates and milder side effects are expected to play an active role, but due to the difficulty of identification by imaging findings, testing is performed by invasive biopsy. However, due to the difficulty of identification by image findings, the examination is performed by invasive biopsy. Therefore, CAD (Computer Aided Diagnosis) system is highly applicable to detect mutations non-invasively by applying the analysis results obtained from images. In recent years, the correlation between radiomics features and cancer has been confirmed, and the prediction, classification, and detection of lesions in unknown data by machine learning using these features have shown high performance. Therefore, with the goal of developing a non-invasive genetic testing CAD system, we propose a method for detecting driver gene information mutations from chest CT (Computed Tomography) images using Ensemble Learning. In this method, the radiomics features extracted from the chest CT images are used for supervised learning by ensemble learning. Then, the effectiveness of the proposed method is verified by classifying the images into different classes and performing evaluation experiments.

**Keywords:** Computer Aided Diagnosis, Lung Cancer, Radiomics, Machine Learning

---

Tohru Kamiya

1-1, Sensui, Tobata, Kitakyushu, Fukuoka 804-8550, Japan

Tel: 093-884-3185, E-mail: kamiya@cntl.kyutech.ac.jp

## 1. はじめに

近年, がんにおける死亡率は 1 位を占めており, その中でも肺がんの死亡率は 1 位であることから, 肺がんの早期治療は重要課題となっている. 肺がんの遺伝子検査では, がんの増殖に関わるドライバー遺伝子情報の変異を確認する. 変異が確認された場合, より奏効率の高い治療薬が選択できる. しかし, 現状として画像所見のみでの診断は困難であり, 侵襲的生検による同定法が行われているため, 非侵襲的に変異を検出できる画像処理システムの応用性は極めて高いと考えられる. 本

論文では, 機械学習による胸部 CT 画像からのドライバー遺伝子情報変異の検出法を提案する.

## 2. 手法

本論文では, ラジオミクス特徴量と患者の臨床情報を用いた機械学習による遺伝子変異の有無の識別を行う. 入力した CT 画像と臨床情報から特徴量を抽出し, それを説明変数とした機械学習法による教師あり学習を行う. その後, 未知データに対する遺伝子変異あり/なしの 2 クラス分類を行う.

### 2.1 ラジオミクス特徴量

ラジオミクス特徴量とは、CT、MRI、PETなどの放射線医用画像から抽出した、形状、濃度、テクスチャなどの特徴量群を指す[1]。大量の高度な定量的特徴を抽出・分析することにより、画像の特徴を遺伝子の表現型やシグネチャに関連付けるような予測モデルを構築することができる。

## 2.2 特徴量削減

特徴量抽出後、学習において冗長な特徴量を削減し、有意な特徴量を選択するため、RFE(Recursive Feature Elimination)による特徴量削減を行う。RFEは、機械学習モデルを用いた特徴量の評価手法であり、全ての特徴量を入力した状態から、重要度の低い特徴量を再帰的に削減していく[2]。

## 2.3 Ensemble 学習によるドライバー遺伝子情報変異有無の最終識別

Ensemble 学習は、複数の学習アルゴリズムを組み合わせるにより、単体の場合よりも高い精度を得るための学習手法である。本稿では、特に高識別精度が見込まれる SVM(Support Vector Machine) [3]、MLP(Multi-Layer Perceptron) [4]、RF(Random Forest) [5]の最大値投票によるアルゴリズムを構築する。

## 3. 実験と結果

データセットは、胸部 CT 像の 2 次元スライス画像または 3 次元画像と、臨床情報を用いる。学習には提案手法を含んだ、7 パターンの機械学習手法による識別と評価を行った。評価基準として、曲線の下部面積が 1 に近いほど判別能が高いことを示す、AUC(Area Under the Curve)を用いる。提案手法による識別の結果、AUC=0.90 となり、他のモデルと比較して最も良好な結果を得た。結果を図 1 に示す。比較に用いた図中の LR, KN, SVM, DC, RF, MLP はそれぞれ、Logistic Regression, K-neighbors Classifier, Support Vector Machine, Decision Tree, Random Forest Classifier, Multi-Layer Perceptron Classifier を示す。最後の Ensemble は本稿での Ensemble Learning による識別結果である。

## 4. おわりに

本稿では、機械学習による胸部 CT 画像からのドライバー遺伝子情報変異の検出手法を提案し、

概ね良好な分類結果を得た。今後の課題としては、データ拡張、腫瘍領域のみを抽出する画像処理手法の導入や、モデルの改良等が考えられる。また、他手法とのさらなる比較も必要である。

## 謝辞

本研究は、栢森情報科学振興財団 (K33 研 X X VI 第 579 号) の助成を受けて遂行された。

## 参考文献

- [1] Coroller TP, Grossmann P, Hou Y, et al., "CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma", *Radiotherapy & Oncology*, Vol.114, No.3, pp.345-350, 2015.
- [2] Guyon, J. Weston, et al., "Gene selection for cancer classification using support vector machines", *Mach. Learn.*, Vol.46, pp.389-422, 2002.
- [3] Ioannis Tsochantaridis, et al., "Large margin methods for structured and interdependent output variables", *The Journal of Machine Learning Research*, Vol.6, No.9, pp.1453-1484, 2005.
- [4] Geoffrey E.Hinton, "Connectionist learning procedures", *Artificial Intelligence*, Vol.40, No.1-3, pp.185-234, 1989.
- [5] L.Breiman, "Random Forests", *Machine Learning*, Vol.45, No.1, pp.5-32, 2001.

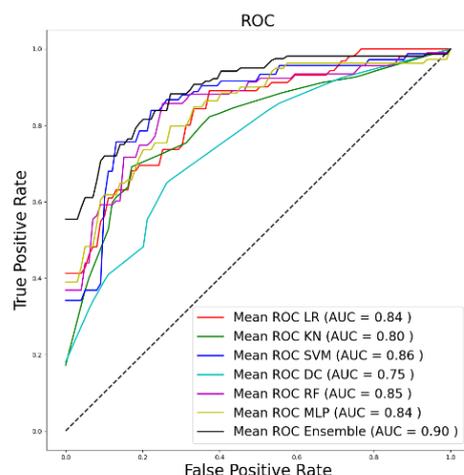


Fig.1 実験結果