様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月 １２ 日現在

機関番号：１７２０１
研究種目：若手研究
研究期間：2018～2023
課題番号：１８Ｋ１５５５３

研究課題名（和文）肺癌発症リスクの高い肺線維症CT画像を検出する解析基盤およびAI作成

研究課題名（英文）Creation of an analysis platform and AI to detect lung fibrosis CT images with high risk of developing lung cancer.

研究代表者

江頭　玲子（Egashira, Ryoko）

佐賀大学・医学部・助教

研究者番号：７０４５７４６４

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：工学系研究協力者との共同研究で「肺癌を発生した肺線維症のCT画像」と「長期経過観察をしても肺癌が発生しなかった肺線維症のCT画像」を人工知能（AI）に学習させ肺癌発生のリスクとなり得る肺線維症の画像所見を検出することを目的に研究した．ResNet18をSimSiamのEncoderの部分にを利用したモデルを構築し，潜在空間上に類似特徴量を有する画像が配置されるように設定した．経過で肺癌が発生した領域とその同一部位の過去画像（陽性画像），発癌しなかった領域の経過の画像（陰性画像）を検索画像として，類似画像検索から，発癌高リスク領域の検出は可能であることがわかった．

研究成果の学術的意義や社会的意義
肺線維症のCT画像において，肺癌が発生する領域には，発生しない領域と比べ何らかの特徴が存在し，それを把握しておくことにより，リスクの高い患者さんを早期発見し，厳重に経過観察することが可能となると考えられる．

研究成果の概要（英文）：The aim of the research was to detect image findings of lung fibrosis that may be a risk factor for lung cancer development by training an artificial intelligence (AI) on "CT images of lung fibrosis that developed lung cancer" and "CT images of lung fibrosis that did not develop lung cancer after long-term follow-up" in collaboration with an engineering research collaborator. The model was built using the Encoder part of SimSiam and set up so that images with similar features were placed on the latent space. Using the past images of the area where lung cancer occurred in the course of the disease and the same area (positive images) and the images of the course of the area that did not develop cancer (negative images) as search images, it was found that the detection of high-risk areas for carcinogenesis was possible from similar image search.

研究分野：間質性肺炎

キーワード：間質性肺炎における肺癌発生　肺線維症　人工知能

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

進行性かつ予後不良な疾患である肺線維症に対し，線維化の進行を抑制する薬剤が複数開発され，以前よりも肺線維症自体の予後改善が望まれるという中で，肺線維症診療に残る課題の一つが「肺線維症に併発しやすい肺癌を早期に発見加療すること」である．通常の目視による画像診断では，肺癌の発生しやすさを予測することは困難と言えるが，人工知能（AI）を用いることで，我々がこれまで見出し得なかった画像的特徴を利用して分離・同定する可能性があると考えた．また，画像情報の解析結果と網羅的な遺伝子情報との組み合わせを研究することで，画像情報を疾患に関連した遺伝子異常と結び付ける試み，Radiogenomics も注目され始めた．

２．研究の目的

「AI による識別結果を用いて，肺癌合併リスクの高い肺線維症の画像的特徴を見つける」，発展的には，更にその画像から遺伝子学的特徴を見つけ出す“ Radiogenomics ”を目指すことが目的である．

３．研究の方法

工学系研究協力者との共同研究にて，「肺癌を発生した肺線維症の CT 画像」と「長期経過観察をしても肺癌が発生しなかった肺線維症の CT 画像」を Convolutional Neural Network（CNN）を用いた人工知能（AI）に学習させる．さらに実臨床症例で識別を AI に行わせ、その判定結果および AI の識別過程の解析結果より肺癌発生のリスクとなり得る肺線維症の画像所見を検出する．

（1）材料

2007 年 1 月 1 日～2018 年 12 月 31 日までに当院にて臨床目的に胸部 CT を撮影された患者の胸部 CT 画像を対象とし，CT 画像アーカイブより，慢性線維化性間質性肺炎症例を抽出した．経過観察の画像も使用することとし，経過観察中に肺癌を合併した症例を，経過を追うことで確認した．また，正常例や肺気腫のみの症例も一定数用意することとした．

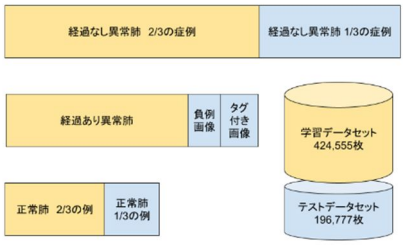
（2）方法

① 準備

間質性肺炎診断後の経過観察中に肺癌が発生した症例において，肺癌発生が確認された時点の CT 画像を起点に，その 1 年前，2 年前，3 年前…（X 年前）の画像の同じ場所を放射線科医が特定し，“ 将来的に肺癌が発生する領域（高リスク領域）”としてラベリングを行った画像を用意した．このラベリング情報と共に，　DICOM 画像（経過の画像も全て）を匿名化した状態で抜き出し，工学系研究者に提供した．間質性肺炎診断時の初回画像があるが，長期経過を追跡できなかった症例，長期経過観察をした上で肺癌が発生しなかった症例もラベルと共に用意した．

② AI 作成

・背景を削除した 512×512 マトリックスの CT 画像を 64×64 のパッチ画像に加工し，肺野が 10%以上含まれるパッチのみを抽出して使用．医師の作成したラベリングの位置情報をもとに，高リスク領域のパッチ画像を手動で作成し，それらを陽性画像として，経過をみて潜在（発癌前）と顕在（発癌後）に分ける．また，陽性画像とは逆に将来的に肺癌が発生しない間質性肺炎の領域を，これらの画像から陽性画像の領域を除くことで確保した．1 年間肺癌にならなかったことが担保されている画像と，3 年間肺癌にならなかったことが担保されている画像をそれぞれ陽性画像と同じ枚数作成し，1 年陰性 IP 画像，3 年陰性 IP 画像と命名した．Hold-out 法（評価を検証用と学習用のデータに分けることによって別々にモデルを評価する方法）を用い，経過あり，経過なし IP 肺と正常肺のパッチ画像を使用し，それぞれテストデータと，学習データに分けた（右図）．



②-1：ResNet を用いた潜在空間の作成

自己教師あり学習手法である SimSiam を用い，類似したパッチ画像が潜在空間上で近い位置に配置されるように設定した． ResNet18 を SimSiam の Encoder の部分にを利用したモデルを構築した．t-SNE を用いて 512 次元の特徴を 2 次元まで圧縮し，図にして可視化する．

パッチ画像一枚に対して，平面に１つの点がプロットされる．経過あり/なし IP 肺のクラスター，正常肺のクラスター，陽性画像のクラスター，それぞれについての重心もプロットした．

②-2：高リスク領域の自動検出手法

ResNet18 に先に作成したテストデータセットを入力する．これによって大量の経過あり/なし IP，正常肺のパッチ画像に対して特徴ベクトル（埋め込み）が作成される．ここで作られた埋め込み群をデータベースとする．
圧縮していない潜在空間上の L2 距離からクエリ画像と近い順に何枚かリトリーブすることで，類似画像検索を行えるが，この類似画像検索を応用して高リスク領域の自動検出を行った．（陽性画像同士が互いに同じ特徴を持つならば，陽性画像をクエリ画像としたときに，ほかの陽性画像が多く検索されると考えられる．陽性画像を検索できた枚数によって高リスク領域を検出）．

実験１：クエリ画像として陽性パッチ画像を ResNet18 に入力して埋め込みを取得し，テストデータセットに約 20 万枚のパッチ画像があり，L2 距離が近いものを類似画像としてその 0.8 パーセントにあたる 1600 枚を取り出す．実験は，リトリーブ画像の中に含まれる，陽性画像の枚数とそれぞれのメタデータを抽出する．クエリ画像のメタデータとリトリーブに含まれる陽性画像のメタデータを比較して，潜在空間の評価を行う．陽性画像は全部で 388 枚あるため，クエリ画像を除いた 387 枚を 20 万枚の中から検索する．ランダムに 20 万枚の中から 1600 枚を引いたとき平均的な陽性画像の検索枚数は 3.104 枚となる．実験１は，１枚のクエリ画像をもとにデータベースから 1600 枚の画像を検索し，取り出す．ただし，もとより偶然に陽性画像が検索結果に含まれる可能性があるため，ハイパージオメトリック分布を指標として使用する．この指標を用いて，検索システムの性能が偶然の結果ではないかを判断する．陽性画像をクエリ画像としたときのリトリーブに含まれるクエリ画像の症例以外の陽性画像の枚数と，陰性画像をクエリ画像としたときのリトリーブに含まれる陽性画像の枚数を比較する．検索枚数についてのヒストグラムを表し，そのヒストグラムから ROC 曲線を作成し，ROC 曲線下面積（AUC）を求める．

実験 2：陽性画像と陰性画像が，どの程度類似していないかについて分析する．具体的な方法として，陽性画像をクエリ画像としたときのリトリーブに含まれるクエリ画像の症例以外の陽性画像の枚数と，陰性画像をクエリ画像としたときのリトリーブに含まれる陽性画像の枚数を比較する．検索枚数についてのヒストグラムを表し，そのヒストグラムから ROC 曲線を作成し，ROC 曲線下面積（AUC）を求める．

４．研究成果

正常肺の画像は全 39 例，異常肺の画像は経過がある例が 57 例，経過がない症例群 C は 156 例含まれ，最終的にテストデータセットは 196777 枚，学習データセットは 423555 枚となった．
t-SNE を用いた可視化により，陽性画像が集中するところを確認すると，将来的に肺癌が発生する/すでに肺癌が発生した箇所がクラスターを形成しており，肺癌が将来的に発生する領域同士，肺癌が既にある領域同士，もしくは肺癌が将来的に発生する領域と肺癌がすでにある領域は，互いに類似した特徴を持っており，潜在空間上で近くに埋め込まれると示唆された．

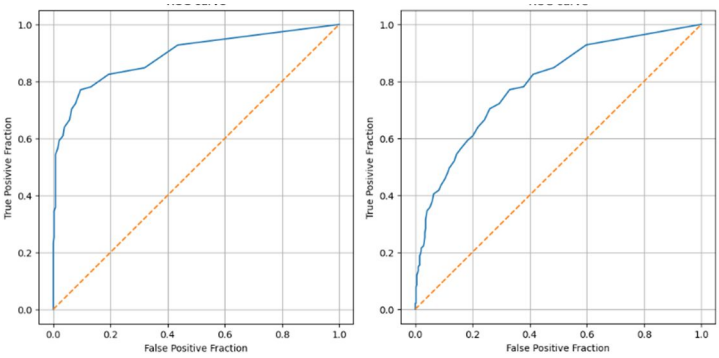
結果（実験 1）：ハイパージオメトリック分布を用い，検索性能を評価した結果，クエリ画像に対し抽出される画像が偶然である確率が極めて低くなる検索枚数は 7 枚以上とわかった．7 枚以上の例は 273 例であり，これはクエリ画像の数 388 例を全体として 70%にあたる．したがって，陽性画像の 70%は，他の症例の陽性画像と類似した特徴を画像内に持っており，偶然ではない結果と考察できる．潜在陽性画像をクエリ画像とすると，そのリトリーブに含まれる潜在陽性画像は 77.7%となる．顕在陽性画像をクエリ画像とすると，そのリトリーブに含まれる顕在陽性画像は 60%となる．よって，クエリ画像と同じ状態の画像が検索されやすいとわかった．しかし，顕在陽性画像がクエリ画像の場合は，潜在陽性画像も 40%検索される．これは，すでに肺癌が発生した領域は，まだ肺癌が発生していないが将来的に肺癌が発生する領域と類似した特徴を持つためであると考えられる．

結果（実験２）：1 年陰性画像をクエリ画像としたときのリトリーブ中の陽性画像枚数から作成したヒストグラムにおいて，最も頻度が高い検索枚数は 0 枚で，219 例，最大検索枚数は 27 枚で，1 例ある．検索枚数 7 枚以上の例は 25 例である．これはクエリ画像の数 388 例を全体として 6.4%にあたる．陰性画像であっても 6.4%は陽性画像と類似した特徴を有

すると言える．陽性画像をクエリ画像とした場合と，1 年陰性画像をクエリ画像とした場合で，検索の結果が異なるため，両者は別々の特徴を持つとわかる．陽性画像が他の IP 領域とは違う固有の特徴を持つと示唆される．3 年陰性画像をクエリ画像としたときのリトリーブ中の陽性画像枚数から作成したヒストグラムについて，最も頻度が高い検索枚数は 1 枚で，31 例，もっとも多い検索枚数は 73 枚で 1 例あり，検索枚数 7 枚以上の例は 101 例であった．これはクエリ画像の数 388 例を全体として 26%にあたる．これは陽性画像をクエリとした場合の 70%と比べると低い割合であり，陽性画像が他の IP 領域とは違う固有の特徴を持つと示唆する．しかし，1 年陰性画像の場合と比較すると，3 年陰性画像の方が高い割合で，陽性画像と類似した特徴を持つ．これは，1 年陰性画像の場合は，癌ができるまでの時間が少ないため，癌ができる場所と，そうでない場所がはっきりと分かれるためと考察できる．3 年陰性画像の場合は癌ができる場所が確定するまでに時間がかかるような肺野の状態であり，癌が発生するかの判断が難しいため，潜在陽性画像との差異が少なくなると考えられる．

陰性画像がクエリ画像の場合と，陽性画像がクエリ画像の場合のリトリーブ中の陽性画像検索枚数にもとづく 2 値分類の ROC 曲線で検証すると，1 年陰性画像がクエリ画像の場合と，陽性画像がクエリ画像の場合のリトリーブ中の陽性画像枚数にもとづく ROC 曲線下面積（AUC）は 0.890 であった．3 年陰性画像がクエリ画像の場合と，陽性画像がクエリ画像の場合では，AUC は 0.789 であった．したがって陽性画像とそれ以外の IP 領域は，それぞれの画像をクエリしたときの陽性画像の検索枚数で判別可能であると分かる．これは，IP 領域において肺癌が合併するリスクが高い領域と，低い領域では特徴に差異があると示す．したがって，深層学習を用いた潜在空間上での類似画像検索により，IP に合併する肺癌の高リスク領域を自動で検出できる可能性が示唆された．

図 左：1 年陰性画像をクエリ，右：3 年陰性画像をクエリとした場合



自己教師あり学習で作成した潜在空間上で，間質性肺炎に合併する高リスク領域は類似した特徴を有し，類似画像検索を用いた高リスク領域の検出は可能と考えられる．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 吉田直樹
２．発表標題 肺癌合併間質性肺炎患者のCT画像における機械学習を用いた非癌部画像情報を用いた発癌肺の判別法の検討
３．学会等名 生体医工学会大会
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	原　武史 (Hara Takeshi)		

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関