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

平成　３０　年　　５　月　３１　日現在

機関番号：　３２６０７
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ０８３８６

研究課題名（和文）「病理診断の客観化＝数値化モデル」構築と「ITによる病理診断支援システム」開発

研究課題名（英文）Application of text mining techniques and image analysis in pathology diagnosis

研究代表者
原　敦子（Hara, Atsuko）

北里大学・医学部・准教授

研究者番号：１０２７６１２３

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：病理診断は客観化＝数値化が難しく、ITによる診断支援システム開発は遅れている。一方、病理診断は組織像から特徴量を解析する思考過程でありこれを数値化出来れば客観化は可能という発想の元、①病理報告書テキスト及び②標本画像を材料に「客観的病理診断モデル」構築を行った。
①では病理報告書をテキストマイニング法等で解析、疾患と報告書内キーワードとの論理的関係を数値化し「インスペクションプログラム」を構築。また報告書内の医学的矛盾や記載ミスを提示する「病理診断支援装置」を開発した。②では乳腺標本を電子化した後、学習データを作成。次に深層学習法を用い新規画像の推定疾患を提示する「病理診断支援装置」を開発した。

研究成果の概要（英文）：We have developed a pathological information data base system and a diagnostic processing model with the use of pathology reports and images of digitized specimens. We first described an algorithm to enable the numeric transformation of pathology reports (breast, gastrointestinal, and esophageal disease) using both text mining and statistical analysis, and we attempted to develop an inspection program that point out the medical inconsistency and/or correct the erroneous description. Images of digitized specimen (breast disease) were divided into many small images, then Wavelet transformation and cluster analysis was performed. They were taken as training data, and identified by pattern recognition by the deep learning method or the K-nearest neighbor method. The result of this identification was used as a feature vector for a specimen image. Similar images were retrieved by comparing the feature vectors of the targeted images and the specimen images in the database.

研究分野：　病理学

キーワード：　病理診断　テキストマイニング解析　画像解析　深層学習法　診断支援システム

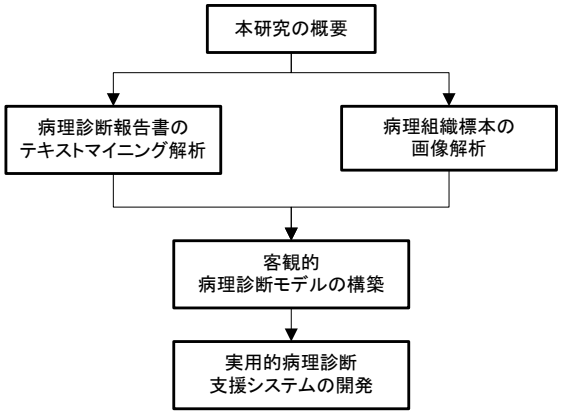
１．研究開始当初の背景

我国では癌患者数が増加傾向にある。それに伴い病理診断の重要性が高まり、大量の検体が病理部門に提出され最終診断が要求されている。一方、病理医数は慢性的に不足し、病理診断の現場は「量との格闘に追われ質の担保にまで手が届かない」のが現状である。この問題解決のためには病理診断を直接的にサポートする「診断支援システム」が不可欠であり、そのニーズはますます増加している。近年、IT 技術の進展で医療情報の電子化や創出されるビッグデータ利活用が活発化し、放射線画像分野ではコンピュータ診断支援システムの実用化が進んでいる。病理分野でも「①病理診断報告書」および「②組織ガラス標本」の電子化が可能となっているが、①②の電子化データを利活用した病理診断支援システムの開発は極めて遅れている。報告書からのテキスト情報を扱ったものは皆無、組織標本の画像解析は複数の施設で模索されるも悪性腫瘍の領域抽出のみにとどまり、実用診断レベルにまで達していない。その背景には、病理診断は全て人間の“目”で行われる作業であり時に主観的で再現性に乏しく、人工知能解析モデルによる数値化＝客観化が困難であることが挙げられる。

２．研究の目的

病理医は組織標本に含まれる視覚的情報から多くの特徴量（細胞異型など）を抽出・解析するという複雑な思考過程を介して病理診断を行う。例えば観察された特徴量を **x**、特徴量の重要度係数を **a**、その総和を **f** とすると、$f = a_1x_1 + a_2x_2 + ... + a_px_p$ という式が頭の中で形成され、**f** が一定以上なら悪性、それ以下なら良性等と判断し診断に至る。従ってこの思考過程を数値化出来れば、本来主観的な病理診断も原理的には客観化は可能であると考える。

この発想の元、これまで①の「病理報告書」（乳腺疾患）の電子化テキストデータを材料に自然言語処理技術を用いて病理医思考過程を数値化し、報告書内の矛盾や記載ミス・推定疾患の候補等を提示する新しい「客観的診断モデル」の構築を行ってきた。（原ほか「病理診断におけるテキストマイニングの応用」計算機統計学 2014;28(1):1-12、「病理診断報告書作成支援装置」特許第 5816915 号）。またバーチャルスライド装置によって得られた②の「組織標本」(乳腺疾患)電子化データを材料に、機械学習方式で解析・数値化(＝客観化)・パターン認識し推定疾患や類似画像を提示する、画像解析による「客観的病理診断モデル」の構築を試みてきたが、(a)特徴小画像抽出が手作業のため非効率(b)学習データ量が少ない(c)用いたパターン認識法は疾患種が多いと認識精度が悪い等の課題が生じ、現手法では診断精度向上には限界があった。そこで今回は、①については対象臓器やチェック機能等の拡大、②については「大量画像データ」を材料に「新規アルゴリズム・機械学習方式（深層学習法など）」を用いて自動でパターン認識する高効率・高精度な画像解析方法の構築を試みた。図－1 に研究概要を示す。



図－1

３．研究の方法

A. 病理診断報告書テキストマイニング解析

材料：北里大学病院で蓄積保管されている電子化された病理診断報告書(2010 年～2016 年、乳腺・食道・胃・大腸疾患、各 1500 症

例)。疾患分類は各癌取扱い規約（金原出版）に準じた。報告書内容は、患者情報・臨床診断名・受付材料・採取法・病理診断名・病理所見など多岐にわたるが、ここでは病理診断名・病理所見をテキストデータとした。簡略化のため疾患名は記号化した（例： 乳腺硬癌→IB2a3 など）

方法（図－2 に概要を示す）：

(1)テキスト内のミススペルチェックおよび文字を統一（英数字を半角・小文字等）。

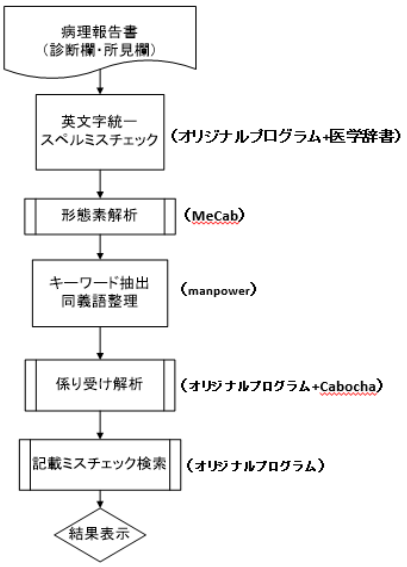
(2)MeCab を用い、テキストデータを形態素解析（言語で意味を持つ最小単位への分割と品詞の判別）する。

(3)同義語を整理し、診断に関連する数百語のキーワードを抽出し辞書を作成。

(4)Cabocha 及びオリジナルプログラムを用いキーワード間の係り受け頻度解析を行う。

(5)以上から得られた疾患と報告書内キーワードとの論理的医学的関係の数値化情報を病理情報データーベース(DB)に格納。

(6)新規症例テキストが与えられた場合、DB 内の情報を基に、①報告書内容の医学的・論理的矛盾や記載ミスの提示や②ベイズの定理を用いた推定疾患や診断確率の提示を行う「病理診断書インスペクションプログラム」を構築。

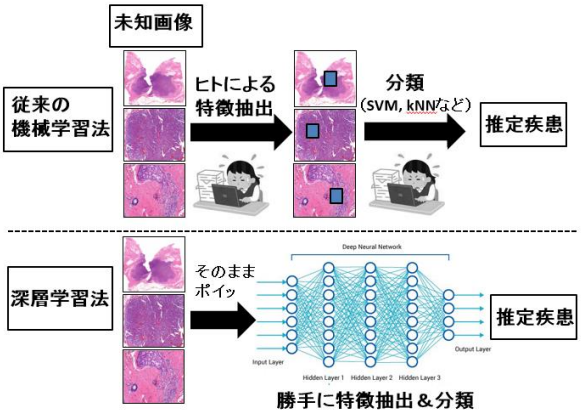


図－2

B. 組織標本電子データの画像解析

材料：北里大学病院で蓄積保管されている病理組織標本(2007 年～2014 年)で、報告書のテキスト解析を既に行った乳腺症例 1500 件。疾患分類は乳癌取り扱い規約(日本乳癌学会編第 17 版 2012 年発行)に準じた。

方法：以下の(1)(2)の 2 種類の方法を試みた（図－3 に概要を示す）。

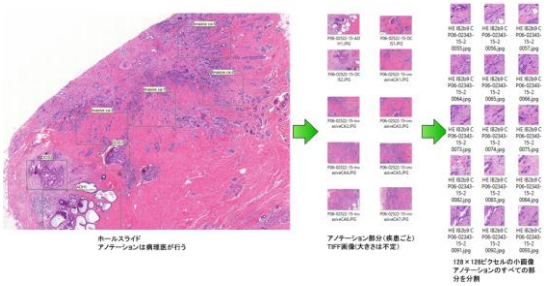


図－3

<u>(1)従来の数値化によるパターン認識</u>（大きな画像を細かく分割した後、ウェーブレット変換により数値化し、学習ベクトル量子化法・kNN 法でパターン認識を行う方法）

(1-1)　既に診断がついている組織標本をバーチャルスライドシステムにより電子化。

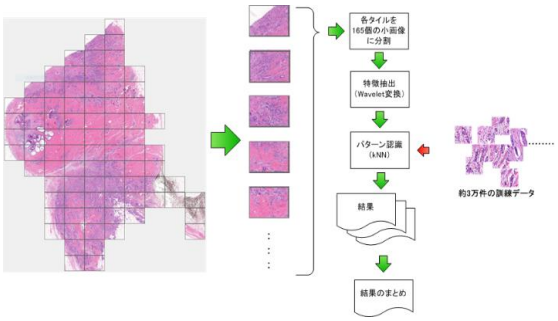
(1-2)　疾患部分を含む画像を抽出し 128x 128 ピクセルの小画像（疾患ごとに数万個）に分割（図－4）した後、各々をウェーブレット変換する。



図－4

(1-3)　さらにウェーブレット変換した結果をクラスタ分析により分類し、学習データとして適切なものだけを残す。本研究では約 12 万個の小画像データを分類整理し、約 3 万個の学習データを抽出。

(1-4)　新規画像（未診断症例）が与えられたら、全体を 1024x768 ピクセルの画像に分割（図－5）した後ウェーブレット変換する。



図－5

(1-5)　新規画像のウェーブレット変換したデータをテストデータとして、学習データに対して kNN 法（k - Nearest Neighbor 法）によるパターン認識を行い、推定疾患や類似画像を提示。

(1-6)　新規画像の分割やウェーブレット変換してパターン認識するには数時間かかる場合があるため、複数の計算機による並列処理を行い、n 台の計算機を用いることで計算時間を 1/n に短縮することができた。並列計算ソフトは、Argonne National Laboratory が開発した Windows 用 MPICH を用いた。

(2)深層学習によるパターン認識（学習データとして大量のデータを用い、深層学習法でパターン認識を行う方法）

(2-1) 前述の(1-3)で作成した小画像を深層学習法により学習させる。深層学習のフレームワークは Caffe（カリフォルニア大学バークレー校で開発）を用いた。

(2-2) 新規画像は前述の(1-4)で作成された画像によりパターン認識を行い、推定疾患や類似画像を提示。

４．研究成果

A.　病理診断報告書テキストマイニング解析

(1)　「病理診断支援システム」の開発：メニュー画面から簡便に機能表示可能な「病理診断支援システム」を開発。図－6 は矛盾や記載ミス検索の表示画面例である。

図－6

(2)検出可能項目

「病理診断書インスペクションプログラム」により、新規症例テキストに対し以下の項目の検出が可能となった。

・論理的矛盾の検出：病理診断名と病理所見内容の医学的矛盾（例：診断欄「線維腺腫（乳腺良性腫瘍）」 vs 所見欄「悪性所見を認めます」→両者の齟齬を検出・提示）

・記載ミスの検出：「左右」記載ミス・「臓器名」のミス・「診断名」のミス

「手術材料などで用いる記号」のミス（例：診断欄 ly(+)→所見欄 ly(-)）

「記号と所見文章」のミス（例：診断欄 ly(+)→所見欄　リンパ管侵襲陰性））

・英単語スペルミスチェック、日本語用語のチェック、正しい単語候補の表示

(3)検出可能臓器：乳腺・食道・胃・大腸

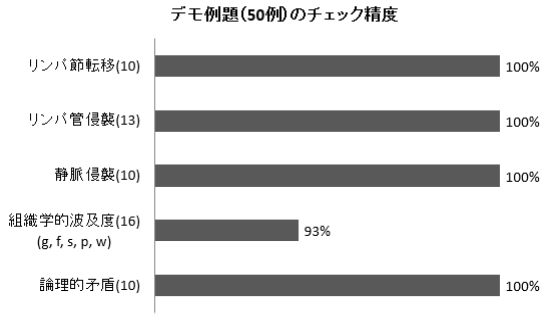
(4)その他

・一症例で複数検体がある場合（例：胃生検 5 個）は分割解析を可能とした。

・Excel ブックとのインターフェース機能を持たせ、多数の症例の一斉分析や既存システムとの間でのデータのやり取りを可能とした。

(5)検出精度：デモ用テキスト（乳癌手術材料）50 例を与えたところ、組織学的波及度項目を除いて記載内容や論理的矛盾のミスをほぼ 100%検出できた（図－6）。また既に報告さ

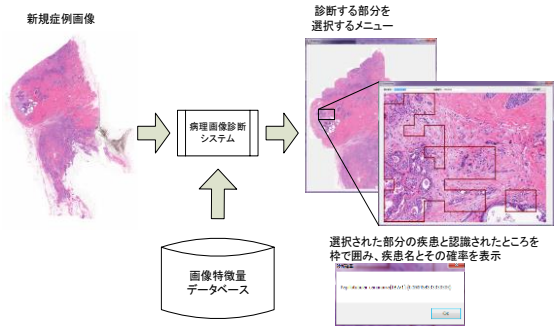
れた既存テキスト 1500 例を与えたところ、8 例で新たなミスを検出した（図－7）。



図－7

B. 組織標本電子データの画像解析

(1)「病理診断支援システム」の開発： メニュー画面から簡便に機能表示可能な「病理診断支援システム」を開発（図－8）。ホールスライドを分割し、各部分ごとの推定疾患名および確率の表示を可能とした。



図－8

(2) 検出精度：癌部と非癌部の鑑別や、浸潤癌と非浸潤癌の鑑別が可能となった。また浸潤性乳管癌の未知画像 500 個に対し、浸潤性乳管癌の三つのパターン（乳頭腺管癌・充実腺管癌・硬癌）を正しく識別できた accuracy は、kNN 法による従来の機械学習では 69.5%、深層学習法を用いた解析は 83.9%と後者で高い精度が得られた。さらに従来の機械学習法に比べ、深層学習法では画像特徴抽出や分類などの人手による作業が不要である点、解析時間が従来の 1/10 と早くなった点が利点として挙げられた。

５．主な発表論文等

〔学会発表〕（計 4 件）

(1) 石橋雄一、原　敦子：“病理診断インスペクションプログラムの全臓器への対応の方法　大規模医療データ科学に関する研究集会(20180207)．会場名　北海道大学(北海道札幌市)

(2) 石橋雄一、原　敦子：“病理画像データの空間的配置とディープラーニングによるパターン認識　科研費シンポジウム「空間データと災害の統計モデル」(20180127)．会場名　同志社大学(京都府京都市)

(3) 原　敦子、石橋雄一、三枝　信：”深層学習による病理自動診断　病理医不要の日は近い？　第 106 回日本病理学会総会(20170429)．会場名　京王プラザホテル(東京都新宿区)

(4) 原　敦子、石橋雄一、三枝　信：”電子化標本画像解析による客観的病理診断のモデル化　第 105 回日本病理学会総会(20160512)．会場名　国際センター(宮城県仙台市)

〔産業財産権〕

○取得状況（計 1 件）

名称：病理診断報告書作成支援装置
発明者：石橋雄一、原　敦子
権利者：同上
種類：特許
番号：特許第 5816915 号
取得年月日：2015 年 10 月 9 日
国内外の別：国内

６．研究組織

(1)研究代表者
原　敦子（Hara Atsuko）
北里大学・医学部・准教授
研究者番号：10276123

(2)研究分担者
三枝　信（Saegusa Makoto）
北里大学・医学部・教授
研究者番号：00265711

(3)研究協力者
石橋雄一（Ishibashi Yuichi）
（株）スタットラボ代表