横浜国立大学　大学院環境情報学府
博士学位論文

機械学習に基づく
医用画像からの生物学的特徴の抽出における
説明可能性に関する研究

A Study on Explainability in Extracting Biological Features from Medical Images Based on Machine Learning

情報環境専攻　情報学プログラム

小林　達明
Tatsuaki KOBAYASHI

請求学位　　博士(情報学)
責任指導教官　長尾　智晴　教授

提出年月日　令和5年1月5日
請求年度　令和4年度3月修了

あらまし

データサイエンスは，数学（統計，確率，機械学習など）を基本として，世の中の問題に挑戦していく分野の1つであり，医療にも広く応用されている．データサイエンスの医療応用では，問診，診察などの診療記録に関するテキストデータ，血圧，脈波，血液検査などの臨床検査データ，遺伝子検査などの分子診断データ，X線画像，X線Computed TomographyやMagnetic Resonance Imaging などの医用画像データなどがさまざまな解を推定するための変数として利用される．このうち，特に医用画像を用いるデータサイエンスの手段に，コンピュータ支援診断（ computer-aided diagnosis; CAD ）やRadiomicsがある．これらの研究は，良悪性や悪性度などの分類，イメージングバイオマーカーの開発などを目的として，従来から盛んに研究されてきた．

CADやRadiomicsから生み出される成果物は，診断の意思決定支援や治療のプロセスの効率化に貢献する可能性があり，さらなる応用が期待される．しかし，その一方で課題も残されている．CADやRadiomicsが目的のタスクを推定するための数理モデル（以下，モデルと略す）を主な成果物とする場合，モデルが行う推論に対する高い精度が重要とされる．加えて，これらの精度だけでなく，医療従事者が納得する根拠をモデルが提供する能力，すなわち，説明可能性が求められるようになってきている．CADやRadiomicsの成果が医療で広く利用されるようになるためには，処理の精度だけではなく，処理の可読性の根本的な改革やアルゴリズムの可視化が求められる．そこで，本論文は，医用画像の分類を対象として，より良い分類精度と説明可能性の追及のために，機械学習に基づく生物学的特徴の抽出に関する検証結果を報告する．

本論文ではまず，Radiomicsによるアプローチを実現するためのRadiomics特徴計算ライブラリの開発について報告する．この計算ライブラリは，記述統計的な特徴，ヒストグラム特徴，形態的特徴，テクスチャ特徴，フラクタル特徴などを含めた画像特徴を計算する能力を有する．

次に，従来の（表形式の学習データを用いる）機械学習プロセスにおける説明可能性の向上を検討するために，特徴選択に着目して検討を行う．分類のために重要な特徴を選択することは，分類精度を向上させるだけでなく，因子による客観的な説明を可能にする．本検討では，遺伝的アルゴリズムによる特徴選択に重要度を導入した新しい特徴選択手法を提案し，これをグレード2,3脳腫瘍の染色体変異（1p/19q共欠失）の推定を目的としたRadiomicsアプローチへ応用した結果を報告する．

そして，精度の高い分類器の作成において第一選択となりうる深層学習によるアプローチの事例として，マンモグラフィ上石灰化有無分類を対象として，EfficientNetによる分類精度およびGrad-CAMによる説明可能性を検証し，分類器としての深層学習モデルの有用性について考察する．

最後に，説明可能な深層学習モデルの一種であるGenerative Contribution Mapping（ GCM ）を，マンモグラフィ上乳腺石灰化有無分類を対象として検証した結果について報告する．GCMは高い分類精度を期待できるモデルであり，GCM以外の深層学習モデルが出力するGrad-CAMと比べ関心領域を正確に捉えることができる直観的な可視化マップを生成できる可能性がある．

Abstract

Data science is one of the fields based on mathematics (statistics, probability theory, machine learning, etc.) to challenge social issues, and is widely applied to medicine. In healthcare applications of data science, text data related to medical records (such as interviews and consultations), clinical data (such as blood pressure, pulse wave, and blood tests), molecular diagnostic data (such as genetic tests), and medical image data (such as X-ray images, X-ray Computed Tomography, and Magnetic Resonance Imaging), these are used as variables to estimate various solutions. Of these, computer-aided diagnosis (CAD) and radiomics are data science tools that use medical images in particular. These research have been actively conducted in the past for the purpose of classification such as distinguishing benign or malignant, malignancy grades, and development of imaging biomarkers.

Products produced from CAD and Radiomics have the potential to contribute to diagnostic decision support and improve the efficiency of the treatment process and are expected to have further applications. On the other hand, challenges remain. When CAD and Radiomics' main output is a mathematical model (a model in short) for estimating the desired task, high accuracy with respect to the inferences performed by the model is essential. In addition, there has been a growing demand not only for the accuracy of models but also for the ability of models to provide evidence acceptable to healthcare professionals and patients, i.e., explainability. For the products of CAD and Radiomics to become widely used in healthcare, not only processing accuracy but also fundamental reforms in processing readability and visualizing algorithms are required. Therefore, this paper reports the validation results of machine learning-based biological feature extraction for the classification of medical images in the pursuit of better classification accuracy and explainability.

First, we present the development of a radiomics feature computation library to implement a radiomics approach. This computational library has the ability to compute image features including statistical features, histogram features, morphological features, textural features, fractal features, etc.

Second, to investigate the improvement of explainability in the traditional (using tabular data) machine learning process, we focus on feature selection. Selecting important features for classification not only improves classification accuracy but also allows for an objective explanation by factors. In this study, we propose a new feature selection method that adopts the importance into feature selection by genetic algorithm and reports the results of applying this method to the radiomics approach for estimating chromosomal mutations (1p/19q co-deletion) in grade 2 and 3 gliomas.

Third, as an application of a deep learning approach that could be the first choice in creating a highly accurate classifier, we investigate the classification accuracy and the explainability with Grad-CAM using EfficientNet for classifying the presence or absence of calcifications on mammograms and validate the usefulness of deep learning models as classifiers.

Finally, we report the results of the validation of Generative Contribution Mapping (GCM), an explainable deep learning model, to classify the presence or absence of breast calcifications. GCM is a model that is expected to have high classification accuracy and could provide visualization maps that capture the region of interest more accurately than the Grad-CAM output from other convolutional neural networks.

目次

図目次

表目次

Algorithm

第1章　序論

1.1 研究背景

根拠に基づく医療は，データ駆動型の医療情報システムを成長させている．医療情報システムは，医療機関において利用される情報システムの総称であり，医療機関等の医事会計システム，電子カルテ，オーダリングシステム，Picture Archiving and Communication System（PACS）等の診療を支援するシステムだけでなく，何らかの形で患者の情報を保有するコンピュータ，患者情報の通信が行われる院内・院外ネットワークが含まれる．コンピュータ技術の進化とともに，医療情報システムの躯幹システムの計算能力，ストレージ，メモリなどの性能は向上しており，膨大な量の医療データの生成や保存・通信が可能となっている．従来は紙運用であった資料の大半は電子的に扱われるようになり，治療や診断の結果生成されるテキスト，波形，表，画像（静止画，動画），音声，イラストなどの電子的なデータによる運用が一般化されつつある．厚生労働省が公表している電子カルテシステム等の普及状況の推移によれば，電子カルテの普及率は一般病院において平成20年の14.2%から令和2年の57.2%まで上昇している [1]．また，Health Level 7（HL7）やDigital Imaging and Communications in Medicine（DICOM）などの国際的な医療データの標準化が産業へ浸透したことも後押しし，蓄積される医療データは，国際標準規格の上で整理・統合され，Analysis-readyなデータセットの資源にできるよう配慮されている．このように蓄積されたデータセットは，確率モデルに当てはめることができるタスクでは特に，データサイエンスへの応用が期待できる [2]．

このようなデータサイエンスの医療応用のうち，医用画像を用いた検討が盛んに行われてきた．従来，画像診断を行う放射線医学の分野において，訓練を受けた医師が医用画像を定性的に評価し，病気の検出，特徴づけ，経過観察を行ってきた．これに対して，人工知能（artificial intelligence; AI）は，画像データの複雑なパターンを自動認識し，医用画像の特徴を定性的ではなく定量的に評価することに優れている．AIの技術の1つである機械学習（machine learning; ML）は，計算機に人を模倣させることを研究するデータサイエンスに欠かせない技術であり，医療において医用画像や非構造化テキストなどを学習データとして用いた検討が活発に行われている [3, 4]．

このうち，主にX線Computed Tomography（CT），Magnetic Resonance Imaging（MRI）やPositron Emission Tomography（PET）などの医用画像から目的の推定や予測を試みる手段として，コンピュータ支援診断（computer-aided diagnosis; CAD）[5, 6]やRadiomics [7, 8]がある．CADは，画像診断において，良悪性，悪性度の分類予測，治療効果予測，イメージングバイオマーカー [9]の開発等を目的として行われる．Radiomicsは，医用画像から得られる特徴が分子生物学的な特徴を表現できるという仮説に基づき，画像から計算可能な信号強度やヒストグラム，テクスチャなどから遺伝子型や腫瘍染色体変異などの予測を試みる手段である．

しかし，機械学習に基づくCADやRadiomicsによるアプローチは，データセットの質や量が充分であること，モデルの推論精度や汎化性，人に理解可能であるかどうかに着目した説明可能性の検証など，機械学習を医療に応用するための条件が必要とされるが，これらは未だ解決できない課題として残っている．このような課題を解決するための研究が必要とされている．

1.2 研究目的

データサイエンスの医療応用は専門領域ごとに多岐に渡るが，医用画像を用いて何らかの分類パターンを捉えることを目的として行われる推論に関しては，枠組みの多くが共通している．これを踏まえ，本論文では次の2点に着目して検証を行うことを目的とする．

1. 分類精度の高い機械学習モデルの検証
2. 説明可能性が配慮された機械学習アプローチ

1.3 本論文の構成

第2章はコンピュータ支援診断とRadiomicsの先行研究，医用画像分類問題に共通している機械学習におけるプロセス，および人間が納得しうる根拠を示す技術について述べる．第3章は，充実した画像特徴を用いてRadiomicsを実践するためのRadiomics特徴計算ライブラリの開発について述べる．画像特徴は学習データとしてだけでなく，人が理解可能な特徴量としての役割を果たすことで説明可能性の向上に貢献する．第4章では，第3章の成果物であるRadiomics特徴計算ライブラリを用いて，脳腫瘍の染色体変異推定問題を対象として，分類精度および説明可能性の向上のための進化計算による特徴選択手法を提案する．第5章では，近年，第一選択肢となりつつある深層学習によるアプローチの検証のために，マンモグラフィ上乳腺石灰化有無推定を対象として，深層学習モデルから作成可能な可視化マップによる説明可能性の検討を行う．第6章では，よりよい説明可能性の探求のために，説明可能な深層学習モデルをマンモグラフィ上乳腺石灰化有無推定に応用することによって有用性を考察する．最後に，第7章で本論文のまとめと今後の課題について述べる．

第2章　本研究に関する研究

本章では，特に本研究と関わりのある研究領域について振り返る．まず，医用画像を対象とした機械学習の研究領域について整理した後，画像分類を対象とした機械学習アプローチの一般的なプロセスについて説明する．最後に，機械学習における説明可能性に焦点を当て，人間が納得しうる根拠を示す技術に関する先行研究について述べる．

2.1　研究領域

2.1.1　　医用画像を用いたコンピュータ支援診断（CAD）

CADは，主に画像診断を支援することを目的として，良悪性・悪性度の鑑別分類のための分類器やイメージングバイオマーカーの開発等を行う分野である(図　2-1)．CADには2つの意味がある．1つはコンピュータで病変を自動的に検出（detection）し，その位置を検査のオペレータや医師に提示することで見落としを減らすことであり，もう1つは病変の良悪性鑑別など，診断（diagnosis）を支援する情報を提供することで 医師の主観による診断のばらつきを減らすことである．前者は，CADe，後者はCADxと呼称される．

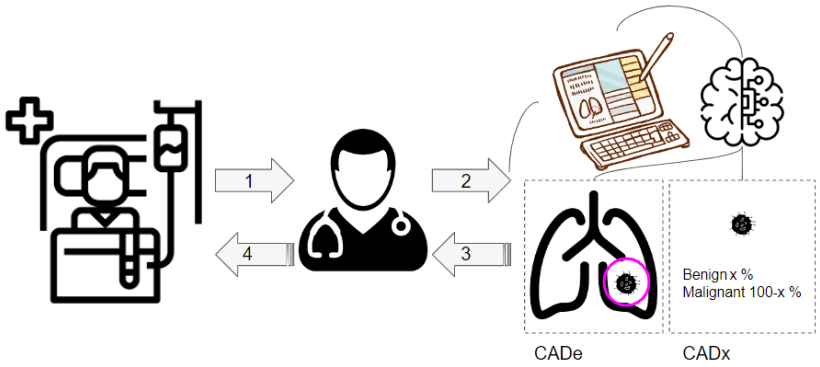


図　2-1 CADの概要

（左：患者，中央：医師・医療従事者，右：CADシステムが実装された医療情報システム．矢印１は診療情報，矢印２はCADシステムに必要な入力情報，矢印３はCAD処理結果，矢印４は医師や医療従事者がCAD処理結果を定量的に解釈した結果が反映された診断・治療に関する情報）

機械学習に基づくCADは，ある患者集団の過去の症例から得られた画像データや非画像データを解析し，抽出された情報と特定の疾患転帰を関連付ける「モデル」を開発する．ここでいうモデルとは，数理モデルであり，学習アルゴリズムそのもの，あるいは，学習アルゴリズムに従って作成された成果物としてのモデルを意味する．作成されたモデルは，未知の症例のデータが入力されると，なんらかの転帰を推定することが期待される．本論文では，便宜上，推定（estimation），推論（inference），予測（prediction）の定義について仮定しておく．推定は推し量って定めること，推論は既知の事柄を元にして未知の事柄について予想すること，予測は予め推測することを意味する．それぞれの用途は，モデルが未知の入力に対して出力を行う場合に推定あるいは推論を用い，人がモデルを利用して意思決定を行う場合は予測と表現する．

意思決定支援のために機械学習を用いて患者データを分析するアプローチは，疾患や病変の検出，特徴による層別化，がんのステージング，治療計画，治療効果判定，再発モニタリング，予後予測など，患者のケアのためのプロセスに適用可能である．これらのアプローチにおいて，医用画像の果たす役割の価値は大きい．

1980年代初頭，シカゴ大学のKurt Rossmann研究室において，様々な疾患に対するCAD手法の体系的な研究開発 [5]が始まった．Chanら [10]は，デジタル化されたマンモグラム上のクラスター状微小石灰化に対するコンピュータ支援検出システムを開発し，画像診断医のセカンドオピニオンとしてのCADeが診断のパフォーマンスを改善できる可能性を示すことを目的として最初の観察者研究 [11]が実施された．1998年，米国食品医薬品局は，マンモグラフィ検診における乳がん検出を支援する事を目的として，最初の商用CADeシステムを承認した．CADの研究の大部分は画像上の様々な種類の疾患の検出と特徴付けに向けられたものであるが，腫瘍の不均一性の定量的画像解析，画像の表現型と根底にある遺伝的・生物的プロセスとの相関，がんのサブタイプの分化，がんのステージング，治療計画および治療応答評価への CAD の適用に対する関心が高まってきている．これらの用途における画像特徴の定量的解析の研究領域はRadiomicsと呼ばれている．

2.1.2 Radiomics

医療は，一人ひとりの遺伝子，生活習慣，環境などの個人差を考慮した個別化医療へと向かっている．個別化医療の実現のために，遺伝学，プロテオミクス，メタボロミクスなど様々なオミクスデータを統合し，機械学習ベースの推論アルゴリズムを用いて生体システムの複雑な働きを解明するための横断的な研究が行われている [12]．しかし，個別化医療には，侵襲的な生検，高いコスト，遺伝子変異を調べるためのスループットの遅さなどの課題がある．例えば，固形がんの生検を行う場合，腫瘍は多くの場合不均一であるため，1回の生検で得られた腫瘍の一部では個別化医療の信頼性が低く，がん治療において個別化医療を行うことが困難な場合がある．そこで，これらの課題を解決するための手段としてRadiomicsが検討されている．

Radiomicsは，医用画像情報が人間の肉眼による分解能では捉えられない分子生物学的な特徴を表現できる可能性があるという仮説に基づき，画像情報から病理や分子診断，遺伝子型，予後予測などを試みる手法である．ここで，画像情報は，CT，PET，MRIなどで得られた医用画像から解析された定量的画像特徴や，評価クライテリアに基づく判断材料（例えば，腫瘍径の変化率や石灰化の有無など）を意味する．定量的画像特徴は，腫瘍などの病変の大きさ，形状，信号強度，ヒストグラム解析，テクスチャなどをもとに数理的な処理によって計算された定量値である．Radiomicsは，このような定量値を機械学習のための説明変数として用いてモデルを作成・検証する [13]．図　2-2に一般的なRadiomicsの概要を示す．

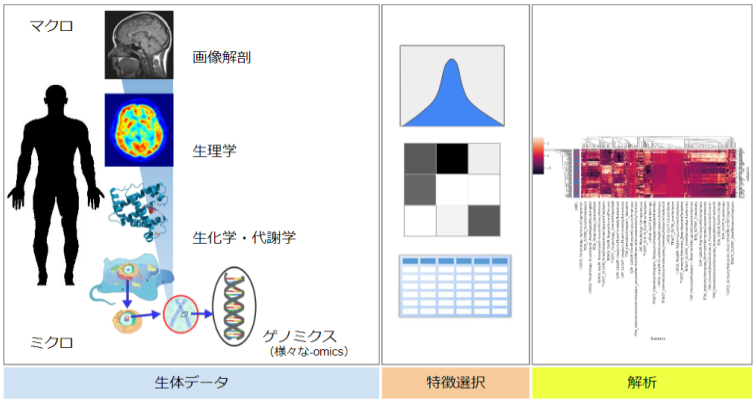


図　2-2 Radiomics概要

（生体データから取得された情報を説明変数として用いて目的変数（分子診断結果など）を推定する）

2.1.3　　PACSの機能として稼働する医用AI

CADおよびRadiomicsにより作成されたモデルは，医療従事者の単独の判断よりも，モデルの意思決定支援が組み合わさった判断の方が医療の質が向上すると考えられる場合に利用される．図　2-3にワークフローの例を示す．より具体的には，モデルを搭載した医用 Artificial Intelligence (医用AI) アプリケーションが医療情報システム内に実装される．例えば，画像診断に関わる医用AIアプリケーションは，マンモグラフィ，CT，MRI などのモダリティにパッケージされた画像処理機器に，それ以外の多くは，モデルのライフサイクルを管理可能な医用AI管理システムとして，PACSの一部に組み込まれることが考えられる．

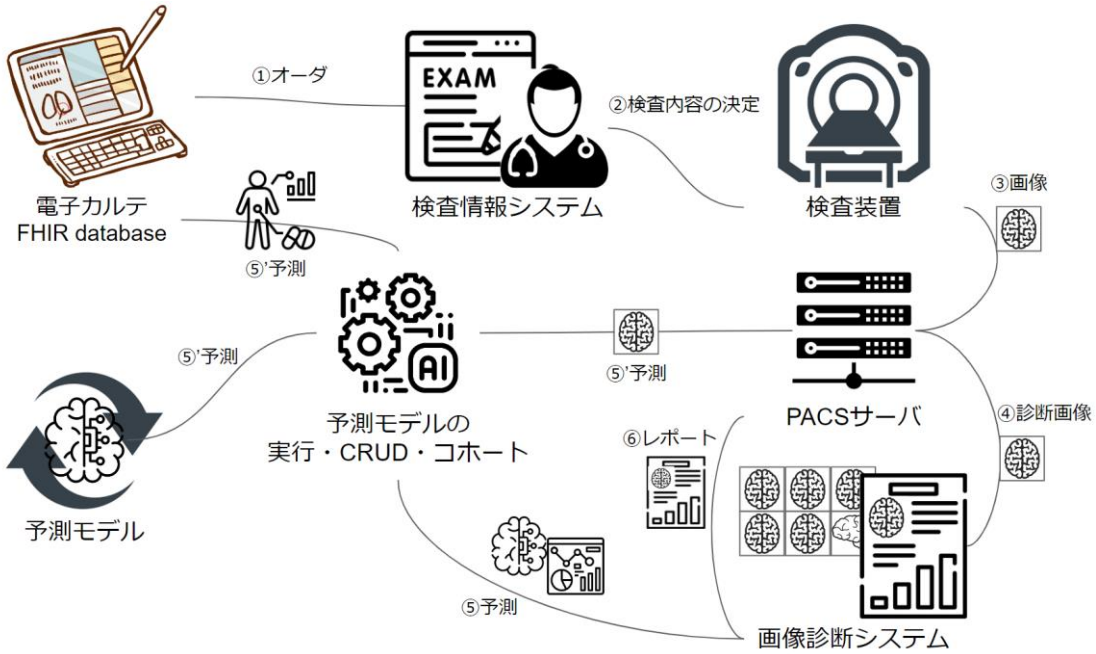


図　2-3 医用AIを利用する画像診断ワークフロー例

例えば，医用画像を入力に必要とするモデルを運用する場合，まず，主治医によって①検査オーダが作成された後，②放射線科医により検査内容が確定され，MRI検査が実施される．作成された画像は③PACSサーバに保管され，電子カルテ，画像ビューワ，および④画像診断システムからのクエリによる呼び出しに対応する．画像診断医は画像診断システムを通じて画像解釈のために⑤モデルを利用しながら診断の定量化や確信度の裏付けを試みることができる．画像解釈の結果作成された画像診断レポートは，⑥構造化レポートとしてPACSサーバで保管される．モデルの管理システム(医用AI管理システム) は，モデルの実行，CRUD (Create・Read・Update・Delete) 処理，利用状況のコホート提供を担う．モデルの管理パッケージでモデルの更新や再学習を行う際は，電子カルテのFHIR (fast healthcare interoperability resource) データベースやPACSサーバから学習データの提供を受ける．

2.1.4　　保健医療分野AIの産業

日本では，2022年度(令和4年)の診療報酬改定により，画像診断管理加算3に該当する特定機能病院を対象に，関連学会のガイドラインに則り医用AIが適切に管理される場合には，画像診断管理加算3を300点から340点に一律に引き上げられることになった．これにより，保健医療分野で医用AIが適切に運用できる施設では，患者一人につき，月に一回340点が算定できる．この画像診断補助のための医用AIが評価されたことによる40点の病院側の利益の一部は，医用AIアプリケーションの産業を支えることが期待される．また，画像診断管理加算3が取得可能な大規模な医療機関だけでなく，中規模病院でも医

用AIの利用が可能になるよう，診療報酬改定が行われる可能性がある．2022年9月時点では，医用AIによる医療行為の補助に関する診療報酬の適用は画像診断のみとなっているが，将来的には，画像診断だけでなくその他の臨床のワークフローを補助する医用AIの評価も行われる可能性がある．

2.2 画像分類を対象とした機械学習アプローチ

一般に，機械学習は大きく，教師あり学習（ supervised learning ），教師なし学習（ unsupervised learning ）に分けられる．この他，強化学習（ reinforcement learning ）や半教師あり学習を加えて分類されることもある．図 2-4に分類問題を前提とした教師あり学習と教師なし学習のイメージ例を示す．

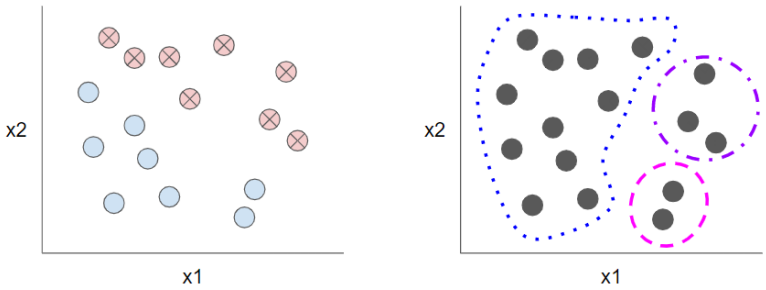


図 2-4 分類問題を前提とした教師あり学習と教師なし学習のイメージ
（左：教師あり学習，右：教師なし学習）

教師あり学習（図 2-4左）は，説明変数（Ｘ）から目的変数（Ｙ）を推定するモデルを求める手法をいう．目的変数は従属変数，説明変数は独立変数とも呼ばれる．目的変数は目標値（教師データ）であり，モデルは説明変数からこの目標値を推定するために学習データ内のパターンを学習する．図 2-4中，縦軸と横軸を成すx1，x2は説明変数を意味し，これらのデータから分類の推定が行われることを意図している．例えば，腫瘍のテクスチャ（説明変数）から，悪性グレード（目的変数）を推定するなどに利用される．目的変数となる教師データは，分類問題の場合，タグ付けやカテゴリ化などで，特定のグループやクラスを離散的な値（0,1,2など）で区分して表現される．

教師なし学習（図 2-4右）は，入力データそのものに着目し，データの性質のパターンを見つける，あるいは，データの縮約を行うための手法である．モデルの訓練には目的変数（教師データ）を必要としない．クラスタリング（入力データを類似グループに分ける）や，データ次元（説明変数の数）を元のデータの情報を失わないように少数に縮約する主成分分析（ principle component analysis; PCA ）などの手法がある．

本研究では主に教師あり学習に主眼を置いている．図 2-5に，画像分類のための機械学習（教師あり）プロセスを示す．本研究における機械学習（教師あり）のパイプラインは，本プロセスに基づき設計される．以降，プロセスの各論について述べる．

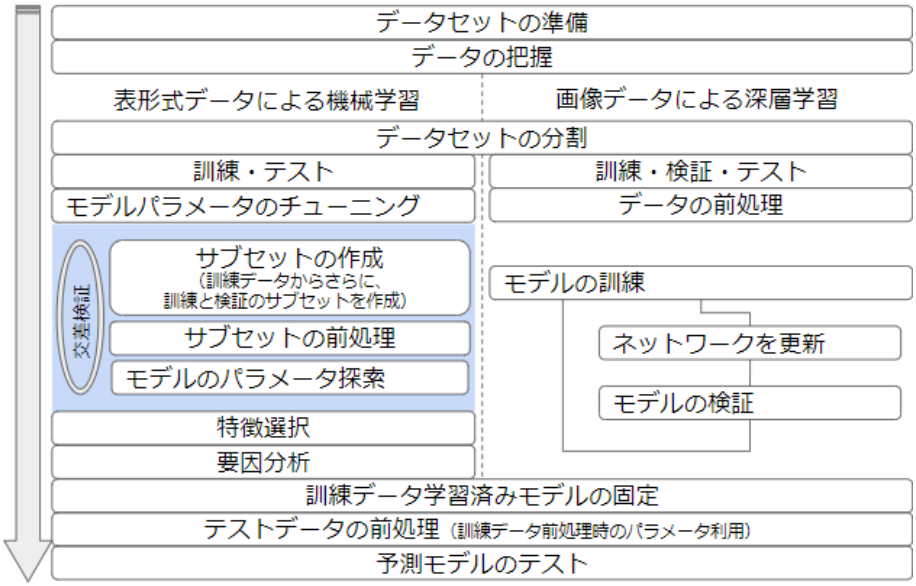


図　2-5 画像分類のための機械学習（教師あり）によるモデルの作成と評価

2.2.1　　データセットの準備と把握

機械学習のために医用画像を用いる場合，専門のドメイン知識が必要とされる．例として，Kumar Vら[14]はRadiomicsを行うためのデータセットの準備に関する留意点を，検査装置の違い，検査方法の違いなどのバイアスの観点も含め示唆している．

モデルを作成するための学習データセットは，説明変数と目的変数を担うデータが収集される．例えば，モデルが信号強度などの画像特徴を用いて画像上にある腫瘍が良性か悪性かを推定する場合，説明変数は画像特徴となり，目的変数は良悪性診断結果となる．

収集されるべき必要なデータセット量は，しばしば議論の対象となる．生物統計上の例数設定（サンプルサイズの設定）はデータセットの準備をする上で必要な観点の1つである．サンプルサイズは，被検体の個体数あるいは機械学習データセット量（例えば，画像データセットの場合は，画像枚数）を意味する．生物統計におけるサンプルサイズの計算方法は研究デザインによって異なるが，単純な計算方法は，確率論に基づき，ある要因（例えば，血圧など）について標識された母集団の標準偏差を基準として計算するなどがある．

生物統計と機械学習のサンプルサイズの考え方は，母集団の性質を捉えるという統計的観点からみて多量であるほど好ましいとされる．機械学習の場合は，作成されるモデルの精度や汎化性が向上することへの期待もある．一方で，臨床研究では，倫理的側面から必要最小限のサンプルサイズでの検討が求められる場合があり，モデルを作成するための十分なデータセットを収集することが難しいケースがある．Larracyらは，小さなサンプルサイズで機械学習モデルの検討を行う必要がある場合に限られたデータセットを逐次的に増加させることで得られた評価結果を用いて，目的の評価値を達成するために必要なサンプルサイズを非線形モデルによって推定する方法を提案している [15]．この他，2値分類の場合は，AUCからサンプルサイズを推定する手法 [16]もある．

収集されるデータは，統計学的な観点から，目的とする因子に対する内訳（クラスバランス）が均等であるほうが望ましい．データのクラスバランスは，例えば，良性データセットと悪性データセットがそれぞれ同数あるなど，均等に近いほどモデル学習に関するバイアスを減らすことができる．クラスバランスに大きな偏りがある場合は，モデル作成時に，アンダーサンプリング（多いクラスを少ないクラスのデータ数に合わせる），オーバーサンプリング（データオーグメンテーションによって少ないクラスのデータを増強する，あるいはクラスバランスが均等になるようにしつつデータ全体を増強する）を適用することができる．ただし，

検証やテストに用いるデータに対するデータオーグメンテーションは，現実に存在しないデータを扱うことになる場合は特に，配慮が必要となる．極端なクラスバランスの偏りがあるなど，十分に検証やテストにデータを割り当てることができない場合などは，実現可能性を確かめる意味でTest Time Augmentationなどの手法は許容されるが，この手法が第一選択かどうかは議論の余地がある．問題が極端なクラスバランスの偏りである場合は，教師なし学習による異常検知アプローチへの切り替えを検討することもできる．

データの量や内訳などの他，収集されたデータの傾向を事前に把握することは望ましいとされる．データの傾向を把握するための代表的な手段としては，各因子についてどの値をいくつ含むかを示す分布に関する傾向を示すヒストグラム，要約統計量（平均値，中央値，最頻値など），因子ごとのクラスの分布を示す散布図，因子ごとの外れ値の確認が容易に行える箱ひげ図などが用いられる．

2.2.2 学習データの分割

学習データは，モデルを訓練する前に訓練とテスト，あるいは訓練，検証，およびテストのデータセットに分割される（図 2-6）．訓練データはモデルの訓練に利用される．検証データは，モデルの学習アルゴリズムを最適化するパラメータの探索や特徴選択，深層学習モデルの学習経過観察などのために分割される．検証データの分割は，訓練データからさらにサブセットとして分割される場合と最初から分割される場合の2パターンがあり，一般に，従来の機械学習では前者，深層学習では後者が利用される．テストデータはモデルの評価に利用される．テストデータは多い方がよいとされるが，一般に，全データセットの2～4割がテストに用いられる．

データセットの分割時は，モデルが，モデルにとって未知であるはずのテストデータを誤って学習してしまう情報漏洩（information leakage，target leakage）を可能な限り排除するために，グループ化と層別化が考慮される．

グループ化は，例えば，同一被験者から連続して取得された画像などの対応のあるデータや，データを検査装置ごとにまとめておく必要がある場合など，分割時に混在することを防ぐなどの目的で行われる．例えば，同一被験者から，3枚の胸部レントゲン画像が治療前，治療後一年，治療後二年の経過観察で取得されたとする．これらの画像間には大局的に見たときに大きな差がないことが事前に想定可能と考えられる．これらの画像を意図せず，訓練と検証，あるいは，訓練とテストなどに混在するように振り分けた場合，モデルは訓練時に検証あるいはテストデータの内容に酷似したデータを学習できていることになり，情報漏洩を招く可能性が高くなる．このような状況は，同一被験者のデータをグループにして分割することで避けることができる．

層別化は，訓練，検証，テストに割り振られるクラス割合が，それぞれのデータセットでできるだけ均一に含まれるよう振り分ける操作をいう．例えば，10データのうち，陽性ラベルが6つあり，これを3つのデータセットに分ける場合，それぞれに2ずつ割り振る．層別化の手法は，クラス割合だけでなく，データの性質を基準に行われることもある．Khalid Mらは，データセットをクラスタリングしたのち，各クラスターで標準偏差による層別化を行い，各クラスターで平均的なデータと平均から離れたデータとを分け，これらを訓練，検証，テストに指定の割合で割り振るアルゴリズムを提案 [17]している．

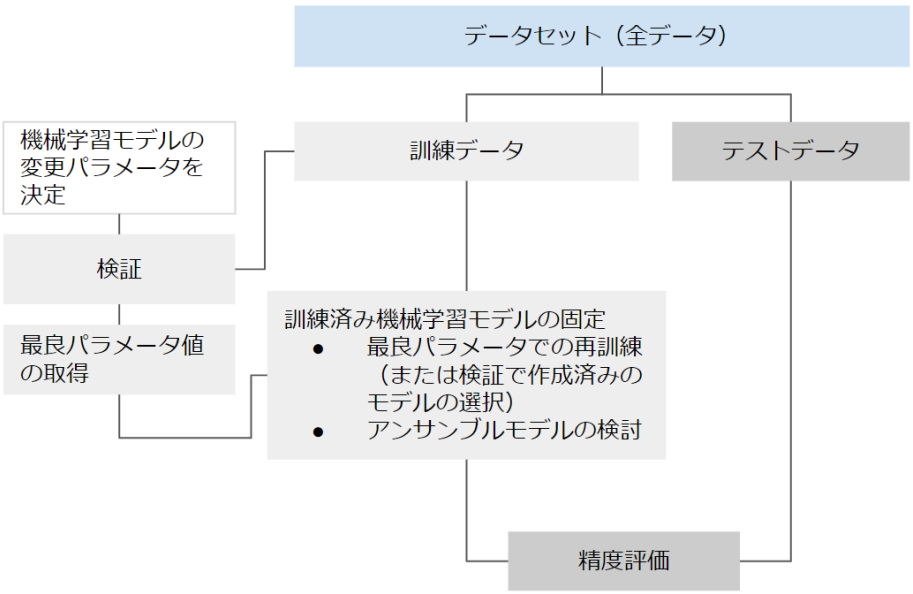


図　2-6 データセットの分割と学習のためのデータの流れ

データの分割方法は，検証方法によって異なる．ここから代表的な検証方法について触れる．

● Hold-out法

Hold-out法は，データセットを指定の割合で分割する最も基本的な検証データの分割方法であり，他の分割方法の基本となっている．図　2-7はHold-out法で分割した例を示す．図中の番号は連番，下付き文字は教師クラスを表す．

図　2-7 Hold-out法の例

（図中の番号は連番，下付き文字は教師クラスを表す．）

● 交差検証（Cross-validation）法

交差検証は，検証データを入れ替えながらHold-out法を繰り返し，複数の検証用サブセットを作成する手法である．繰り返す回数(k)を指定し，訓練［100－（100／k）％］と検証［（100／k）％］が割り振られたサブセットが作成される．例えば，3回繰り返す場合は，3分割交差検証と呼ばれ，66%を訓練，33％を検証に割り当てた3つのサブセットが作成される(図　2-8)．

図　2-8 単純なk-分割交差検証法

（3分割交差検証法の例）

交差検証はストラテジーによって分割の方法が区別される．図　2-9にストラテジーのタイプを示す．

リーブワンアウト交差検証法（図　2-9 No.1）は，分割の回数を訓練データセットに含まれるデータ数として，全訓練データを1つずつ検証データとして利用する．この手法は，データセットが少ない場合に有用であるとされる．次に，層別化交差検証法（図　2-9 No.2）は，指定した属性をサブセットに均等に割り振る．例えば，陽性ラベル（+）と陰性ラベル（-）を属性として指定した場合は，ラベルの割合が均等になるように各サブセットが作られる．そして，グループ交差検証法（図　2-9 No.3）は，データのグループを保持しながらサブセットを分割する．最後に，層別化グループ交差検証法（図　2-9 No.4）は，グループ化したデータをさらに層別化してサブセットに分割する手法である．

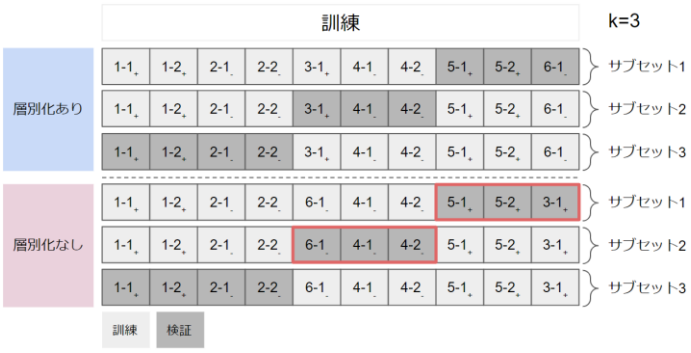
No.	ストラテジー	意味
1		リーブワンアウト交差検証法 全訓練データ数回サブセットを作成する．全訓練データは1つずつ検証データとして利用される．
2		層別化交差検証法 指定した属性をサブセットに均等に割り振る．層別化をしない場合，クラスラベルに偏りが生じる場合がある（左図中赤枠）．
3		グループ交差検証法 データのグループを保持しながらサブセットを分割する．
4		層別化グループ交差検証法 グループ化したデータをさらに層別化してサブセットに分割する．層別化をしない場合，クラスラベルに偏りが生じる場合がある（左図中赤枠）．

図　2-9 交差検証のストラテジー

ここで，情報漏洩について補足する．グループ化は類似したデータをグループ化することで情報漏洩を避けるために重要な意味を持つ．図　2-10にグループ化しないことにより起こる情報漏洩の例を示す．図中サブセット1，2は，5番に対応するデータが訓練と検証に混在している．サブセット3は，グループ化されているために，訓練と検証にデータが混在していないことがわかる．

一般に，CTやMRIなどの連続したZ軸方向の分解能が細かい画像（ thin slice画像 ）や，被写体が変わらない動画，同一被験者から取得された時系列の胸部X線画像などは，グループ化されなければ情報漏洩を招く可能性がある．

図　2-10 グループ化しないことにより起こる情報漏洩の例

（サブセット1，2は情報漏洩）

● ブートストラップ（Bootstrap）法

ブートストラップ法 [18]は，全訓練データから，サブセットの訓練データをランダムに重複を許してサンプリングし，検証データは訓練側にサンプリングされていないデータを一定の割合で割り振ることでサブセットを作成する方法である（図　2-11）．

図　2-11 ブートストラップ法の例（k=3）

2.2.3　　データ前処理

データの前処理として，カテゴリ変数，欠測値，外れ値，データのスケーリングの取扱いについて述べる．

● カテゴリ変数の取扱い

カテゴリ変数とは，身長や年齢のように数値で表せる変数ではなく，グループやタグ付け可能な属性（性別など）を分類した系列である．仮に，連番によってカテゴリ変数が設定されていた場合，連続変数のように数の大小に意味はないため，モデルを学習する前に，学習にバイアスを与えないようにカテゴリ変数を実数に変換する．表　2-1に代表的な手法を示す．

表 2-1 代表的なカテゴリ変数のエンコード方法

カテゴリ変数を実数へエンコード	操作
One-hot Encoding	ラベルごとに系列を作りTrue(1)/False(0)を割当てる.
Count Encoding	ラベルの出現回数を割当てる. ラベルの出現頻度の差に意味がある場合に利用する.
Label Count Rank Encoding	ラベルの出現回数ランクを割当てる. ラベルの出現頻度の順位差に意味がある場合に利用する.
Target Encoding	回帰問題の場合に, ラベルごとの目的変数に平均値を割り当てる.

● 外れ値の取扱い

統計学において, 外れ値は他の値から大きく外れた値をいう. 測定ミス・記録ミス等に起因する異常値とは概念的には異なる. 外れ値は, データセットから除外されるか, あるいは, モデル作成者が定めた許容範囲の最小値, 最大値に置き換えられる. 表 2-2に因子ごとの許容範囲を決めるための代表的な手法を示す.

表 2-2 代表的な外れ値の置換方法

外れ値の許容範囲設定方法	説明
パーセンタイル法	Minパーセンタイル以上, Maxパーセンタイル以下の範囲(例えば, 1パーセンタイル以上, 99パーセンタイル以下)にない値を外れ値とし, これらのパーセンタイル値で範囲外の値を置換する.
信頼区間による方法	平均値 ± (Z × SD)の範囲にない値を外れ値とし, この範囲の上限値と下限値で範囲外の値を置換する. ここで, ZはZスコア, SDは標準偏差を表す.

● 欠測値の取扱い

欠測値は, 取得できなかった計測値をいう. Not A Numberなども欠測値の対象となる. 欠測値は, 除外か補完で対応される. 除外は, 欠測値を含むサンプル(インスタンス)をデータセットから除外する. 補完は, 推定精度に過剰なバイアスを与えないことを前提として, 何らかの値を代入する. 補完の方法として, 単一補完法, 多重補完法などがある. 単一補完法は, ある系列データにあるすべての欠測に対して1つの値を補完する(例えば, Last Observation Carried Forward(時系列に取得していたなら, 最後に取得できた値で代用する方法), 同じ系列の値の中央値, 平均値, 最頻値などによる補完). 多重補完法は, 欠測値を回帰計算によって推定する(例えば, MICE [19]). Missing at randomとMissing completely at randomの一般的な条件下では, 単一補完と多重補完の両方で不偏推定となる. しかし, 単一補完では, 推定標準誤差が小さすぎる結果を招くことがあるため, 標準誤差と信頼区間を正しく推定する多重補完が望ましいことがある [20].

● データのスケーリング

画像特徴などのデータセットは, 各特徴によって値の取り得る範囲(スケール)が異なるため, すべての特徴を同じスケールに統一することで, モデル学習時のバイアスを小さくする. 表 2-3は代表的なスケーリングの手法を示す. スケーリングは基本的に系列(因子)ごとに行われるが, 深層学習に画像を用い

る場合など，訓練データセット全体を対象として処理する場合と，画像一枚ごとに行う場合など，スケーリングの対象範囲はモデル作成者の学習設計による．

表 2-3 代表的なスケーリングの種類

スケーリング手法	意味
Standardization	標準化．系列ごとに，平均値を各特徴量から差分し，標準偏差で除す．標準化された系列の配列は，平均が0，分散が1となる．事前に外れ値が除外されていない場合は，外れ値の性質が保持される．
Min Max Scaling	系列ごとの特徴量の最小値と最大値を指定された値の範囲（[Min 0，Max 1]など）にスケーリングする．外れ値の影響を受ける．
Maximum Absolute Value Scaling	系列ごとに，特徴量の絶対値の最大値が1.0となるように特徴量をスケーリングする．標準化やPercentile Scalingのように代表値によるデータのシフトやセンタリングは行わない．外れ値の影響を受ける．
Percentile Scaling（Robust Scaling）	系列ごとに，中央値を特徴量から差分し（中央値が0となる），分位範囲（q_min，q_max，ここで，0.0 < q_min < q_max < 100.0）に従って特徴量をスケーリングする．分位範囲を用いるため，外れ値に対して堅牢なスケーリングとなる．
Normalize	系列ごとに，特徴量を個々の単位ノルムに正規化する．各特徴は，その系列のノルム（l1（マンハッタン距離），l2（ユークリッド距離）または infinity（絶対値の最大値））が1になるように，他の特徴量とは独立してスケーリングされる．

ここまで，前処理について触れてきたが，分割された学習データセットに対して前処理を行う場合は，前処理に利用した各種パラメータを一貫して利用できるようにすることが望ましいと考えられる．例えば，訓練データに前処理を行った場合に，訓練データで設定された前処理パラメータを検証データとテストデータに利用する．このように処理することで，分割されたデータごとの前処理によるばらつきを防ぐことができる．

2.2.4 特徴選択

特徴選択は，主に，表形式データの学習データセットを用いたモデル作成のプロセスで利用される．特徴量（因子，説明変数）の選択は，情報量の少ない，あるいは，冗長な説明変数を学習データセットから取り除くことに主眼が置かれている．

表 2-4は，一般的な特徴選択方法である統計的手法，教師あり学習による手法，教師なし学習による手法を示す．

表 2-4 一般的な特徴選択手法

手法種別	特徴選択法	内容
統計的手法	分散しきい値法	分散がある閾値を満たさないすべての特徴を削除する．系列の値がすべて「1」であるなど，定数のみを値に持つ特徴量がある場合，分類問題の学習には情報量が少ないため，あらかじめ除外される（この操作は，分散しきい値を0にしたときの特徴選択と同義である）．

	相関分析	共線性（コリニアリティ），多重共線性（マルチ・コリニアリティ）とは，説明変数の中に相関係数が高い組み合わせが存在することをいう．データセットが共線性を有する特徴を含んでいる場合，モデル学習バイアスの原因になる [21]．この対策として，例えば，AとBという説明変数の相関係数が極めて高い場合，両方を説明変数として使わずにどちらかを除外する方法や，分散拡大係数（variance inflation factor; VIF）と呼ばれる変数間の多重共線性の強さを定量化する指標を用いて，特定された特徴をデータセットから除外する方法などで対応する．
	単変量解析	単変量統計検定で得られる統計量の大小に基づいて，情報の少ない特徴の除外や有意差の大きい特徴量の絞り込みなどを行う．χ二乗検定，一元配置分散分析（ANOVA）など．
教師あり学習による手法	Filter法	Filter法は，各入力変数とターゲット変数の関係を統計的手法で評価し，そのスコア(例えばp値など)を基にモデルで使用する入力変数を選択（フィルター）する．
	Wrapper法	Wrapper法 [22]は，検証用モデルをサブセットデータで繰り返し評価する処理を内包（ラップ）しており，説明変数の追加または削除を行い，評価指標を最適化する特徴の組み合わせを見つける手法である．ステップワイズ法（変数増加法，変数減少法）や遺伝的アルゴリズムがあてはまる．
	Embedded法	訓練データで作成された一時的な検証モデルから特徴ごとの係数や重要度を得て，これらの値を基に特徴を選択（上位k個など）する． ロジスティック回帰モデルにペナルティ項を設けたLASSOモデルによる特徴選択が広く利用されている． 応用的な方法として，Permutation重要度がある．
教師なし学習による手法	次元圧縮など	次元圧縮を目的に利用されることが多い．PCA，tSNE [23]，UMAP [24]など．

特徴選択に画一的な手法はなく，表　2-4の手法を単独，あるいは，組み合わせて利用するなどの戦略が用いられる．例えば，先にFilter法で統計的に意味の少ないと考えられる特徴を除外し，次に，検証モデルから特徴の重要度を得た後，Wrapper法の内部ループとしてステップワイズ法で探索するなどが挙げられる（図　2-12）．

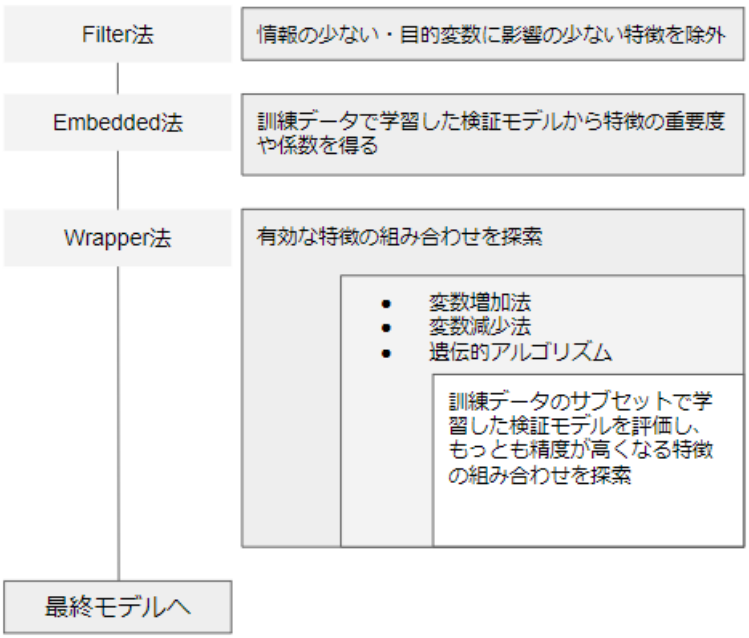


図　2-12 特徴選択のフレームワーク例

2.2.5　　要因分析：因果推論による学習に伴うバイアスの低減

特徴選択のフレームワークを用いた自動的な操作は，ドメイン知識から各特徴の因果関係を整理することは行われていない．選択された特徴が，目的変数との関わりの深いよく知られた概念であった場合，目的変数の情報漏洩を引き起こす可能性がある．また，推定タスクへの影響が大きいと考えられる因子（モデルによる推定にバイアスを生じさせる因子）をバックドア基準などによって特定し，傾向スコア分析により傾向の一致するデータセットにリサンプリングすることで，学習時のバイアスを減らすなどの対策を講じることができる [25].

2.2.6　　次元の圧縮と削減

特徴選択後，さらにデータセットの次元（説明変数の数）を減らすために次元圧縮を試みることができる．例えば，特徴選択の結果，i個の特徴が残った場合，主成分分析により次元圧縮を行い，第j成分（1 ≦ j ≦ i）までをデータセットとして選択し，学習に利用するデータセットの次元を削減（説明変数の数を減らす）することで，これらの特徴の総合的な情報のみを縮約した低次元のデータセットを作成することができる．次元削減によって，モデル学習や推定の計算負荷が低減される．

2.2.7　　機械学習モデル

機械学習モデルの発展は，医療研究に広く関わってきた [26]．表　2-5に代表的な機械学習モデル（教師あり学習アルゴリズム）を示す．また，これら以外の高い分類精度を期待できるモデルとして，XGBoost，勾配ブースティングモデル，複数の任意のモデルで構成するブースティングモデルなどがある．モデルの選択は作成者のドメイン知識やモデルのパフォーマンスを示す定量的評価指標の比較などを総合して行われる．

表 2-5 教師あり分類アルゴリズムに対応した代表的な機械学習モデル

モデル名	意味
kNN (k-Nearest Neighbors)	k 近傍法. 与えられた学習データをベクトル空間上にプロットしておき, 未知のデータが得られたら, そこから距離が近い順に任意のk個を取得し, その多数決でデータが属するクラスを推定するノンパラメトリックな教師あり学習法 [27].
Logistic Regression	ベルヌーイ分布に従う変数の統計的回帰モデルの一種である [28].
SVM	2つのカテゴリを分類する学習サンプルが与えられたとき, SVM [29]は, 2つのカテゴリ間のギャップの幅を最大化するように, 学習データを特徴空間上にマッピングする. 未知のサンプルが訓練済みの特徴空間にマップされることで, カテゴリ境界のどちら側に位置するかに基づいてカテゴリが推定される. 境界の計算には, 線形, 非線形の両方が設定できる.
Gaussian Process	ガウス過程に基づく分類器 [30].
Decision Tree	決定木による分類モデルはその分類に至る過程の解釈を容易にする. 決定木は, 葉が分類を表し, 枝がその分類に至るまでの特徴の集まりを表す木構造を示す [31].
Random Forest	ランダムフォレストは, データセットの様々なサブサンプルに多数の決定木を当てはめ, 推定精度の向上とオーバーフィッティングの抑制のために平均化を使用するアンサンブル学習器である [32].
Neural Net	多層パーセプトロン分類器 [33]である.
AdaBoost	データセットに対してアルゴリズムに用意された分類器を適用し, 逐次的に学習済みの分類器をコピーしながら, 同じデータセットを学習に適用する [34]. ただし, 誤って分類されたデータサンプルの重みは, 後続の分類器の学習効率を高めるよう調整される. 推定は逐次的に作成された学習モデルの出力を集約して算出される.
Naive Bayes	ナイーブベイズ分類器 [35]は, 特徴量間の強い(ナイーブな)独立性仮定を用いてベイズの定理を適用した, 確率的分類器の一種である.
Deep Learning	深層学習モデル [36]

なお, 場合によっては, 単一モデルではなく, 複数のモデルを作成し, 複数のモデルの個々の出力を集約(例えば, 平均や投票)することで得られた推定結果を採用することがある. このような, 複数モデルの出力を集約するようにデザインされたモデルはアンサンブルモデル [37]と呼ばれる. アンサンブルモデルの用途は, 例えば, k分割交差検証を用いて, 特徴セットを統一したk個のモデルを作成し, アンサンブルモデルを構成するなどが挙げられる. このようにすることで, それぞれ異なる訓練データで作成されたモデルの推論を利用できるため, 単一モデルに比べて過学習リスクが抑えられることや, 推定精度を補い合うなどが期待できる.

2.2.8 モデルの学習

機械学習モデルの学習は, モデルに訓練データを与えて学習させるプロセスである. モデルは説明変数と目的変数を対応付けるパターンを学習データから獲得する.

モデルごとにパターンの捉え方は異なる. 図 2-13は, 2値分類シミュレーションデータを用いて, 複数の異なる学習アルゴリズムを持つモデルを学習させ, 弁別境界上の学習データのサンプルの位置を図示した結果である. シミュレーションデータは, 線形分布, 非線形な円形分布, 三日月形分布が用いられている. モデルが有する学習アルゴリズムごとにパターンの捉え方が異なることが分かる. モデル作成者は事前に学習データの性質を観察することで適切な弁別境界を提供する学習アルゴリズムを選定できる.

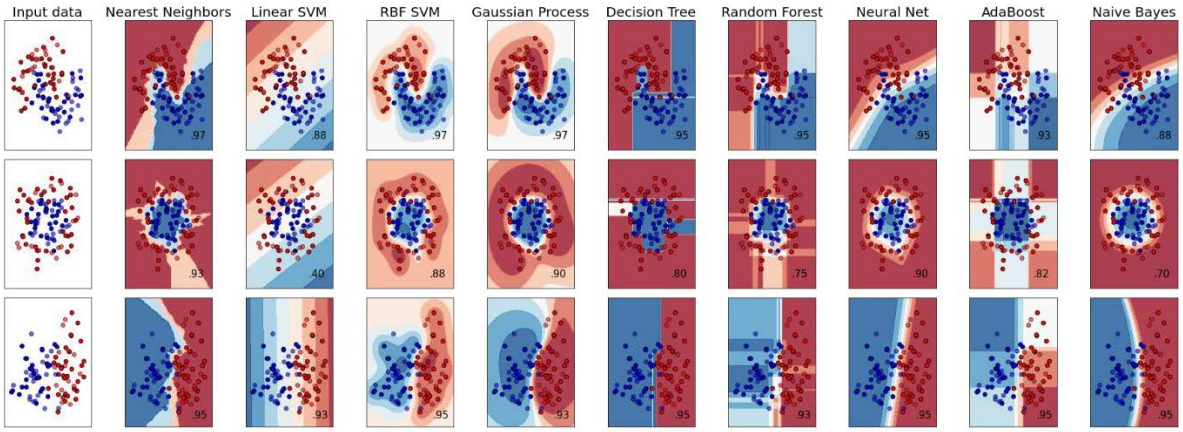


図　2-13 シミュレーションデータを用いた学習アルゴリズムごとの2値分類弁別境界の例

モデルの学習アルゴリズムは学習の設定条件が必須であることが多く，このような設定条件はハイパーパラメータと呼ばれる．ハイパーパラメータは学習アルゴリズムごとに異なる．モデルの最適なハイパーパラメータは，検証のサイクルで繰り返し探索され，検証の結果，最終的に求まったハイパーパラメータがモデルへ適用される．

2.2.9　　モデルの分類性能テスト

一般に，モデルの分類性能は，推定の精度（正解率など），推定値と教師データとの一致度，相関および損失から評価される．表　2-6に主な評価指標を示す．一致度，相関，損失は，推定値とクラスラベルの乖離の大きさを示すと考えることもできるため，モデルが学習したパターンとの当てはまりの良さの評価，過学習の評価に利用される．例えば，AUCが同じ2つのモデルがあった場合に，Log-lossが少ないほうがモデルの当てはまりのよさが良く，Log-lossが大きい状態は，過学習が起こっていることを判断する指標となる．

分類問題の評価指標は2値分類が基本とされる．多クラス分類の場合は，One vs One（一対一）総当りでモデルを作成し，各モデルの結果の平均値を精度として用いる方法や，One vs Rest（一対その他すべて）で精度が計算される．

これらの評価指標は，検証とテストの過程で算出される．検証時の評価指標算出の目的は，従来の機械学習モデルでは，モデルのハイパーパラメータ探索や特徴選択のためであり，深層学習モデルでは，学習の経過観察のために計算される．

最終的にテストデータにてテストが行われ，予めデザインされた主要評価指標によってモデル性能が比較される．

表　2-6 主な評価指標

評価指標	意味
Accuracy	正解率
Recall（Sensitivity）	感度
Specificity	特異度
Precision（ positive predictive value; PPV）	陽性的中率

評価指標	意味
NPV (negaitive predictive value)	陰性的中率
AUC (ROC-AUC)	ROC曲線下面積 (area under the receiver operating characteristic curve)
Balanced accuracy	アンバランスなデータセットに対する評価に用いられる. 各クラスで得られたリコールの平均値として定義される.
Average precision	Precision-Recall Curveを各閾値で達成されたPrecisionの加重平均としてまとめる. このとき, 前の閾値からのrecallの増加は重みとして使用される.
F1 score	F1スコアはPrecisionとRecallの調和平均として解釈することができ, F1スコアは1で最高値, 0で最低値となる. balanced F-score, F-measureとも表現される.
Fbeta score	F-betaスコアはPrecisionとRecallの加重調和平均であり, 1が最適値, 0が不適値を意味する. 算出時, リコールの重みを決定するbetaパラメータを指定する. beta < 1はPrecisionをより重視し, beta > 1はrecallを重視する(betaが0のとき, Precisionのみを考慮し, betaが大きくなるほどリコールのみを考慮する).
Cohen kappa	分類問題における2者間の分類評価の一致度を表すスコアである.
DCG (discounted cumulative gain) score	対数変換された推定スコアの順序値によって補正された正解に対する真のスコアの合計. 真のラベルが上位にランクされている場合に高い値を得る.
Matthews corrcoef	Matthews相関係数 (matthews correlation coefficient; MCC) は, 機械学習において, 2クラスおよび多クラス分類の精度の質を測る尺度として用いられる. これは, 真陽性と偽陽性を考慮し, クラスが非常に異なるサイズであっても使用できるバランスの取れた指標と一般にみなされている. MCCは-1から+1までの値をとる相関係数である. 1は完全に教師ラベルと一致した場合の予測を表し, 0は平均的なランダム予測, -1は逆予測を表す. この統計量は, φ係数(phi係数)としても知られている.
Jaccard score	Jaccard index(Jaccard類似度係数)は, サンプルの推定ラベルと正解ラベルの2つのラベルセットの共通部分の大きさを, これらの和集合の大きさで除した値として定義される.
Zero one loss	誤判定の割合や誤判定の数の値. 最良の性能は 0 である.
Brier score loss	Brierスコアは, 推定された確率(0≦p≦1)と正解ラベル(0か1の値をとりうる)との差の平均を2乗して得られる値. 0と1の間の値をとる. 最良の性能は 0 である.
Log loss (logistic loss, cross-entropy loss)	ロジスティック回帰やその拡張であるニューラルネットワークなどで用いられる損失関数である. ラベルが2つ以上の場合のみ利用できる. 損失が0に近いほどよい.
Hinge loss	2値クラスの場合で, 正解ラベル(y_true)が +1 と -1 でエンコードされていると仮定すると, あるサンプルの決定関数(decision function)で推論にミスが発生した場合, 分類器からの決定関数の出力(pred_decision)は, margin = y_true * pred_decision は符号が一致しないため常に負となり, 1 - margin は常に 1 よりも大きくなる. 正解した場合は0となる. これらを各サンプルで累積した結果は, 分類器の決定境界に対する推定ミスの上限と解釈できる. SVMモデルの評価などに利用される.
Hamming loss	誤った推定ラベルの割合.

2.2.10　　モデルの性能比較

評価指標による評価は分類器の性能を客観的に特徴づけるために役立つが，分類器間の性能差を十分に評価できないことがある．より正確には，特定のデータセットにおいて分類器の性能に評価指標の差が示されたとしても，その違いが単なる偶然ではなく，統計的に有意であるかどうかを確認することで，性能差に対する評価の質を強化できると考えられる．

分類器の性能に対する統計的な検定は，ベンチマークデータセットを用いて，一般に帰無仮説（null-hypothesis statistical testing; NHST）によって行われる．例えば，同一のテストデータを対象として，同じ分類タスクを実行できる複数の分類器を比較する場合，① t 検定を用いてそれぞれの分類器から取得された推定値を比較する，②分類結果から得られた混同行列を χ 二乗検定にて比較する，③交差検証で繰り返し得られた評価指標について統計的な有意差があるかどうかを比較するなどがある．J Nathalieらは，2クラス分類，多クラス分類の統計的な検定の具体的手法について述べている [38]．この他の具体的な検定手法として，AUCから統計的有意差を検定するDeLong Test [39]などがある．

2.2.11　　機械学習プロセスの誤りへの対応

医用画像を用いた機械学習に関する研究は，専門的なデータを取り扱うことに伴う留意点の他に，モデルを作成する上での注意点がある．Varoquauxら [40]は，これらの注意点についてまとめ，学習データセットの量，モデルの作成手順や評価方法の観点から提言を行っている．

2.3　画像分類を対象とした深層学習アプローチ

機械学習の技術の1つに深層学習がある．本研究では，深層ニューラルネットワークを学習させるための技術やノウハウ全般を総称して深層学習と呼ぶ．深層学習に基づき作成されるモデルは，その精度の高さ，入力と出力の設定の柔軟さなどを理由に，近年注目を集めている．その基礎となる計算モデルは人工ニューラルネットワーク（あるいは単にニューラルネットワーク，neural networks; NN ）である．

1958年にRosenblattによってパーセプトロン（ perceptron ）が提案された [41]．パーセプトロンは，ニューラルネットワークのアルゴリズムの中で最も単純な数理モデルである．単一のパーセプトロンは，単純パーセプトロンと呼ばれ，入力データを受け取る入力層と推定結果を出力する出力層の2層のみで構成される．単純パーセプトロンの分類性能は，線形分類器と同等に扱うことができる．しかし，排他的論理和（XOR）のように，単純であっても非線形な演算は表現できなかった．そこで，非線形な演算にも応用できるように改良された階層型ニューラルネットワークが提案された．階層型ニューラルネットワークは，単純パーセプトロンをまるで神経ネットワークの伝達のための1つのシナプスのように信号の入出力を行うユニットとして捉えることで，複雑な演算に対応する．階層型ニューラルネットワークは複数の単純パーセプトロンを並べた列を層として表現し，さらに，入力層と出力層の間に隠れ層（あるいは中間層）を多数配置して構成される．多層型パーセプトロン（ multi-layer perceptron; MLP ）とも呼ばれる．一般に階層型ニューラルネットワークは，勾配の逆伝播 [42]に基づく確率的勾配降下法を用いて内部パラメータ，特に内積演算のための重みを最適化する．

階層型ニューラルネットワークは，理論的には，層数を増やすことで表現力を高めることができるが，単純な層数の増加では，層数を増やすに連れて逆伝播時に勾配が減衰し，学習が進まなくなる問題（勾配消失）があった．そのため実用的には層数を増やすことができず，層数の少ない（浅い）ネットワークしか構築できなかった．これを解決する方法の1つとして，より入力の特徴を効率的に学習できるネットワーク内部の仕組みが必要とされていた．

Hintonらは，レイヤーワイズトレーニングという技術によって勾配消失の問題に対処可能なディープビリーフネットを提案した [43]．これは，各層ごとに学習をするというシンプルな考えに基づく技術である．階

層型ネットワークの課題は，誤差を正しく各層へフィードバックしながら，ネットワーク全体としてのパラメータを適切に調整できなかった点にある．レイヤーワイズトレーニングはネットワーク全体を適切に最適化する仕組みとして有効に働く．

画像分類の研究領域では，レイヤーワイズトレーニングが提唱される以前に，LeCunら [44]によって提案された畳み込みニューラルネットワーク（ convolutional neural networks; CNN ）による進歩もあった．CNNは，階層型ニューラルネットワークの一種である． CNN ではフィルタの畳み込み演算（ convolution ），すなわち局所的な入力を用いた内積演算を空間的に反復することを特徴とする．LeCunらは，手書き数字画像分類でよく用いられた現在のCNNの原型となる LeNetを提案している．加えて，畳み込み演算を用いた画像認識系の着想は福島 [45]のNeocognitron に端を発していることも特筆すべきことである．

このような進歩を経て，現在のさまざまな深層ニューラルネットワーク（ deep neural networks; DNN ）があらゆる分野で利用されるようになった．従来の機械学習モデルと比べ，深層学習モデルの適用範囲は広く，画像を扱うテーマにおいては，画像分類のみならず，画像から自然言語を推定する画像キャプショニング，時系列画像からの予後予測などを試みることができるRecurrent Neural NetworkやLong Short Term Memory Network，画像から画像を推定するGenerative Adversarial NetworkやAuto Encoderなど応用は多岐にわたる．

2.3.1 従来の機械学習との違い

画像分類を対象とした場合，深層学習モデルの作成プロセスは，概ね従来の機械学習モデル作成プロセスと共通しているが，いくつか，訓練や検証のプロセスで異なる点がある．

深層学習モデルを訓練するプロセスにおける違いとして，事前学習，転移学習が挙げられる．事前学習は，訓練前に訓練データで教師なし学習を行っておき，学習データの特徴パターンを先に獲得させておく手法である．転移学習は，ある推論タスクを解決するために学習済みのモデルを，関連する別の推論タスクのために再学習して利用する手法である．例えば，一般物体を対象とした大規模な画像データセットであるImageNet [46]で学習済みの深層学習モデルは，医用画像を用いた研究のための転移学習モデルとしても応用されている．CTやMRI画像などの医用画像は，グレースケール画像であることが多く，画像解剖などの放射線学的情報をもったピクセルから構成される画像であり，花や車と言った一般物体とは性質が異なる．しかし，ImageNetのような多量のデータセットで学習されたモデルは，さまざまな特徴パターンを表現できるネットワークを獲得していると考えることができるため，医用画像を用いる推論タスクにおいても，モデルが学習する際の特徴のパターン化を助ける．どのようなときにより有効な転移学習が可能になるかを考察する必要がある場合は，設定した課題のために用意されたデータセットの性質と転移学習に用いる訓練済みモデルのデータセットの性質の近さが観点の1つとなる [47]．

検証のプロセスにおける違いとして，従来の（表形式データを用いる）機械学習の検証は，あらかじめ分割された訓練データを交差検証法などでサブセット化し，モデルパラメータのチューニングや特徴選択を目的として行われるのに対し，深層学習における検証は，繰り返し行われるモデルの学習を観察するために利用される．例えば，モデルの学習を観察することで，検証で最も評価指標がよくなった時点のネットワーク重みを最終的なモデルの重みとして採用するなどが一般に行われている．

この他，深層学習モデルを特徴抽出器として利用することもできる．

2.3.2 特徴抽出器としての深層学習モデル

ある特定のタスクに対して特徴パターンを学習済みの深層学習モデルは特徴抽出器としての役割を果たすこともできる（図　2－14）．例えば，CNNは，理論上，深い層になるほど特徴パターンの表現力が増すという意味で，浅い層からの中間出力は低次元なデータの特徴パターンを，深い層からの出力は高次

元なデータの特徴パターンを有している．このような特性を利用し，学習済み深層学習モデルの任意の中間出力を，機械学習のための説明変数として用いることが可能になる．このような機械学習モデルへの入力とすることを意図して中間層から得られる出力は，深層特徴（deep features）とも呼ばれる．深層特徴は，任意の層またはよく特徴を表現するとされるボトルネック層（bottleneck layer）から取得されることが多い．

「データセットの準備と把握」で述べたように，臨床研究では倫理的な観点から深層学習にとって理想的なサンプルサイズを用意することが難しいことがある．サンプルサイズが小さい場合，深層学習アプローチが第一選択とならない場合がある．このような場合，従来の機械学習アプローチが選択されるが，深層特徴を学習データとして用いることで，学習の補助に応用できる場合がある．

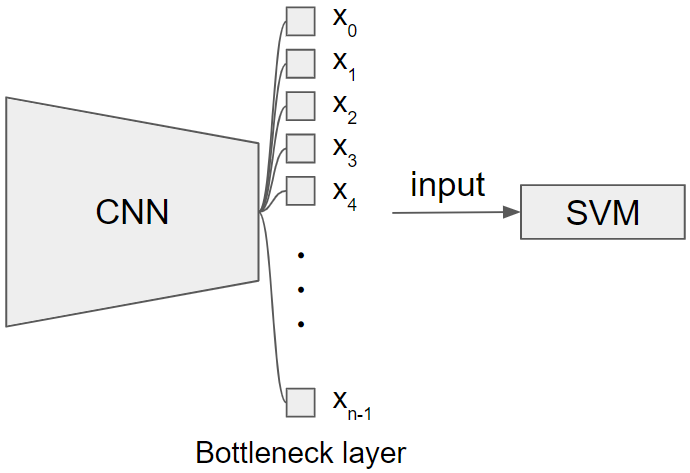


図　2-14 学習済みの深層学習モデルからの特徴抽出イメージ

（CNNを特徴抽出器として利用することで得られる深層特徴を従来の機械学習モデル(例えば，SVM)の学習に用いるアプローチの例）

2.4　人間が納得しうる根拠を示す技術

機械学習のアルゴリズムの改良と蓄積された学習データへのアクセス可用性が高まったことにより，医用AIが医療従事者のワークフローの一部を補強または代替することが期待されている．しかし，医用AIは広く採用されているとはいえない．医療への活用が進まない主な理由の1つは，特定の機械学習アルゴリズム，特にブラックボックスアルゴリズムに関連する透明性が乏しいことが挙げられる [48, 49]．根拠に基づく医療は，意思決定の透明性が重視される．このような医用AIにおける透明性の問題に対処するため，医療分野においても説明可能なAI（explainable artificial intelligence; XAI）が研究されている．AB Arrietaら [50]は，XAIの主な特性（表　2-7）を5つ挙げている．

表　2-7 XAIの主な特性

特性	意味
Understandability（理解しやすさ）	モデルの内部構造やモデルが内部でデータを処理するアルゴリズム的な手段を説明する必要がなく，その機能，つまりモデルがどのように動作するかを人間に理解させるモデルの特性を示す．

Comprehensibility（了解性）	学習アルゴリズムが学習した知識を人間が理解しやすい形で表現する能力を指す．定量化が困難なため，通常，モデルの複雑さの評価と関連付けられる．
Interpretability（解釈可能性）	人間にとって理解しやすい言葉で意味を説明する，またはそれを提供する能力として定義される．
Explainability（説明可能性）	人が意思決定を行うためのインターフェースとしての説明という概念に関連した機能であり，人に理解可能であること．
Transparency（透明性）	モデルは，それ自体が理解可能である場合，透明であるとみなされる．モデルには様々な理解しやすさの度合いがあるため，モデルの透明性は，シミュレーション可能なモデル，分解可能なモデル，アルゴリズム的に透明なモデルの3つの側面について説明される．

これらの特性は，例えば，入力データがモデルに与えられたときの，最終的な推定結果について，データのどの部分が重視されたかを示すこと，予測において判断基準になる要素はなにかを示すこと，あるいは，モデルの入出力の関係を人が見てわかりやすい仕組みに置き換え，推論の過程を示すことができることなどのために考慮される．

このような特性を実現するために，これまでに複数の可視化手法が提案されている．表 2-8に代表的なXAI技術を列挙する．

表 2-8 一般的なXAI技術

領域区分	手法名	特徴
従来型	LIME [51]	任意の機械学習モデルの推定を線形近似によって説明する．
	SHAP [52]	特徴量の貢献度をゲーム理論的な指標を用いて按分して説明する．
	Permutation Importance	特徴量の値をランダムにシャッフルし，推定誤差の増加を測定する．この操作を全特徴で繰り返し，シャッフルされたことで推定誤差が大きくなる特徴を見つける．誤差が大きくなる特徴量が推定に対する重要な特徴であると示すことで説明を可能にする．
	Partial Dependence Plot/Individual Conditional Expectation	特徴量の変化が機械学習モデルの推論結果に与える影響を可視化する．
	Tree Surrogate	決定木モデルにより，代理的に説明する．
深層学習型	Gradient	深層学習ネットワークの勾配を可視化する．
	SmoothGrad [53]	複数のノイズを加えた入力を用いて，入力の数だけ勾配を平均化する．
	Integrated Gradients [54]	入出力の勾配を積分し，入力と出力（参照層）の差との積を得る．
	DeepTaylor [55]	Relevance propergationにより計算された勾配を基にした可視化マップをえる．

	DeConvNet [56]	入力に対する出力を得たのち，出力側からの逆伝播を行い，その結果を得る．逆伝播の際は，Activationを減衰させるようなマイナスの値を取り除くために，値がマイナスになるような箇所を0にして伝播(逆伝播にReLUを用いることと同等)させる．
	Guided Backpropagation [57]	入力に対する出力を得たのち，出力側からの逆伝播を行い，その結果を得る．逆伝播の際は，順伝播時にReLUが適用された勾配に対してReLUを適用する．
	Layer-wise relevance propagation (LRP) [58]	Layer-wise relevance propagationによって，層ごとに勾配を逆伝播して計算された可視化マップを生成する．
	CAM/Grad-CAM [59]	CNNのネットワークを用いて，CNNの畳み込み層の勾配を利用して，画像内で重要な領域を強調したマップを生成する．
	Attention [60]	入力データに対する着眼点を学習するよう設定するAttention機構を利用して注意領域を可視化する．

表　2-8に挙げた技術の他，モデルの推論根拠の説明に限らず，統計や確率によって，データセット全体中のサンプルの位置を，標準正規分布やガウス分布から各クラスに属する確率として示すなどの俯瞰的な可視化も説明に役立つ．

XAIの技術は大きく，従来の(表形式データを対象とした)機械学習モデルと，深層学習によるものに分けられる．ここから，代表的な技術について触れる．

2.4.1　機械学習を対象とした推定根拠の説明

図　2-15，図　2-16に，説明可能なモデル例と解釈可能なモデル例を示す．図　2-15は，ワインの品質を推定するモデルから得られた推定結果を，LIMEを用いて説明した例を示している．ワインの品質を推定するために重要である各特徴が，ワインの三段階の品質の推定にどのように寄与したかが客観的にわかる．

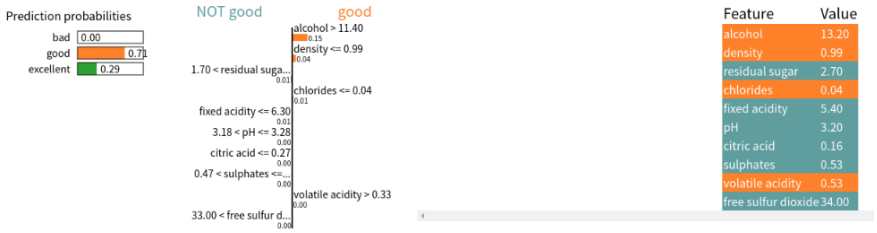


図　2-15 LIMEによる説明変数の推論への寄与度による説明例

(LIMEを用いてワインの品質'bad'，'good'，'excellent'を分類した例．各説変数が分類にどのように影響したかが客観的に説明されていることがわかる．ワインデータ [61])

図　2-16は，決定木モデル(J48)を学習し，決定木グラフを示した例である．決定木をグラフに示すことでモデルがどのように判断を行っているかを客観的に示すことができる．

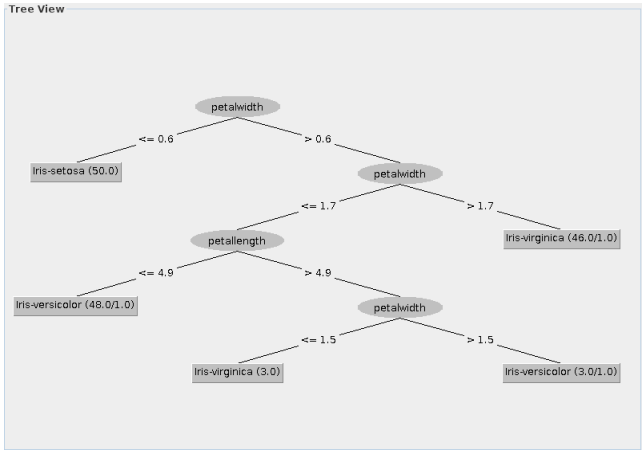


図　2-16 決定木による解釈可能なモデル例

(WEKAにてIrisデータセットをJ48が学習し，学習結果として得られた決定木を可視化．決定木モデルは，推論過程を辿ることができるため，解釈可能なAIの一種と言える．)

2.4.2　　深層学習モデルの推定根拠の説明

深層学習の推定根拠の代表的な説明手法には，大きく3つのアプローチがある．ユニットの反応マップを生成する手法，中間層の出力をそのまま可視化する手法，物体の概略位置を示すマップを生成する手法である [62]．これらのアプローチのうち，物体の概略位置を示すマップを生成する手法は直観的にわかりやすいという長所があるため，画像分類モデルによく利用されている．Zhou ら [63]は，ある画像を学習済みの画像分類ネットワークに入力した場合の畳み込み層の出力(特徴量マップ)を，注目クラスに対応する全結合層の重みを用いて重み付き加算することで注目クラスに関する物体の概略位置を示すマップを生成し可視化する手法(class activation mapping; CAM)を提案している．Selvaraju ら [59]は Zhou らの考え方を発展させた手法(Grad-CAM)を提案している．図　2-17にGrad-CAMの例を示す．これは全結合層における注目クラスの出力を特徴量マップで偏微分することで，注目クラスに対する特徴量マップの各チャンネルの重要度を求め，さらにこの重要度を重みとして重み付き加算することで，CAMと同様に注目クラスに関する物体の概略位置を示すマップを生成し可視化するものである．

図　2-17 ImageNet訓練済みXceptionを用いたGrad-CAMの例

(block14_sepconv2_act層の出力を利用)

第3章　Radiomics特徴計算ライブラリ：RadiomicsJ

3.1　はじめに

Radiomicsは，医用画像からの情報を分子生物学的特徴として捉え，解剖学的構造レベルから遺伝子まで，放射線医学とOmicsに基づく医学との関係を探る分野あるいは手段である．Radiomicsアプローチは，放射線医学や放射線技術の発展にとって重要な要素と考えられる．これまでに，主にがんの領域でRadiomicsアプローチの利用が報告されている [7]．Fanら [64]は，1p/19q共欠失（脳腫瘍の染色体変異）の有無を予測する可能性を示した．Wangら [65]は，Radiomics特徴が，腫瘍の免疫生物学と免疫療法への反応に関する新たなバイオマーカーになりうると報告した．

Radiomicsは，データ収集とキュレーション，関心領域（ region of interest; ROI ）マスク作成，Radiomics特徴量計算，モデル作成と評価というステップで行われる．一般に，医療データは電子カルテやPACSから収集・整理されるが，The Cancer Imaging Archive（ TCIA ）[66]などのオープンデータも広く利用されている．ROI マスクは，画像中の解析対象領域を定義するために作成される．これらの操作は，医用画像解析ソフトウェアやDICOMワークステーションを用いて行うことができる．

Radiomics特徴量の計算には専用のプログラムが必要である．現在までに，PyRadiomics [67]，Radiomics Calculator (RaCat) [68]，Medical Imaging Interaction Toolkit (MITK) [69]などの優れたオープンソースプログラムがリリースされている．例えば，PyRadiomicsは，計算量の多いテクスチャの計算をC言語のプログラムで実装し，Pythonのラッパーで操作を可能にすることで，高速処理と使いやすい操作性を両立している．また，3D Slicer [70]のプラグインとして動作し，グラフィカルユーザーインターフェース上での処理も可能である．しかし，計算できる画像の特徴量の種類や計算結果のばらつき，実行できるシステム環境に制限があるなどの課題が残されている．このような課題に対応可能な新たな計算ライブラリは，Radiomicsの基盤を支援するための要素になる．

本研究では，このような課題を解決する一助として，オープンソースのRadiomics特徴計算ライブラリであるRadiomicsJを開発する．

3.2　Radiomics特徴

Radiomics特徴は，ピクセル信号強度に基づく特徴，ヒストグラムに基づく特徴，形態的な特徴，テクスチャに基づく特徴などに分類される．これらは単に画像特徴とも表現される．これらの特徴は，人が画像を見て判断する特徴のパターン，あるいは，人の目では分かりづらい特徴のパターンを示すための人が理解可能な客観的なデータとして扱われる．

Radiomics特徴は多種多様であり，特徴同士で互いに足りない情報を補い合うことができる．例えば，記述統計的な特徴やヒストグラムに基づく特徴は，人が目で見て異なるパターンを認識できるようなタスクであってもパターンを分離することが難しい場合がある．このような場合にテクスチャに基づく特徴を併用することで，機械が理解可能な見た目のパターン情報を提供できる．一方で，Radiomics特徴の中には，相関する特徴や数学的に定義は同一であるが名称が異なるものも一部含まれている．このような特徴は機械学習のパイプラインにて適切な前処理で除外される必要がある．

3.3　Radiomics特徴計算ライブラリ開発

これまでにいくつかのRadiomics特徴計算が可能な計算ライブラリが公開されてきた．しかし，画像特徴を計算するプロセスや，計算方法の定義の違いなどの理由によって，同じRadiomics特徴の結果の整合

性が担保できないことが課題として残されている. このようなRadiomics特徴の計算を標準化するための取り組みに, Imaging Biomarker Standardisation Initiative (IBSI) がある. IBSIは多くのRadiomics特徴の定義を明確にするとともに, 計算結果の整合性を担保するために, デジタルファントム画像を用いた参考値を公開している [71].

本研究では, IBSIへ準拠することを原則としてRadiomics特徴計算ライブラリを開発した. また, IBSIには定義されていないが, 歴史的にみて重要と考えられた特徴(Fractal特徴など)もIBSI登録済みの特徴とは別に追加することで機械学習へ提供可能な資源の強化を図った.

表 3-1にRadiomicsJで計算可能な特徴ファミリーと特徴数を示す. Radiomics特徴ファミリーはそれぞれ異なる属性を有しており, モデルがパターンを学習することを助ける. 画像データのパターンを捉えるために, 画像の信号強度ベースの統計やヒストグラムベースの解析が従来から用いられている. 形態的な特徴は, 乳がんや肺がんなどの重要な評価指標とされる形状の特徴を説明するのに有効である. また, テクスチャの特徴として, GLCMは, 記述統計量やヒストグラム解析では抽出できない視覚的なパターンを定量的に評価するのに有効である. さらに, GLRLM(線のようにつながる構造), GLSZM(濃淡に対応するBlobサイズ), GLDZM(濃淡に対応するBlob位置とサイズ), NGTDM(信号差パターン), NGLDM(信号変化パターン)などがある. そして, フラクタル次元は構造の複雑さを表す.

また, これらの特徴は, 画像処理フィルタによって前処理された画像から計算される場合, 画像フィルタリングに基づく特徴として扱うことができる.

表 3-1 RadiomicsJが計算可能なRadiomics特徴(IBSIが非推奨の特徴は除外)

Feature families	Num of features	Additional	Excluded
Morphological	25	-	VolumeDensity_OrientedMinimumBoundingBox AreaDensity_OrientedMinimumBoundingBox VolumeDensity_MinimumVolumeEnclosingEllipsoid AreaDensity_MinimumVolumeEnclosingEllipsoid
Local intensity	2	-	-
Intensity-based statistical	21	TotalEnergy StandardDeviation StandardError	-
Intensity histogram	23	-	-
Intensity-volume histogram	6	-	AreaUnderTheIVHCurve
GLCM	25	-	-
GLRLM	16	-	-
GLSZM	16	-	-
GLDZM	16	-	-
NGTDM	5	-	-
NGLDM	16	-	DependenceCountPercentage
Fractal	1	-	-

(GLCM: gray level co-occurrence matrix; GLRLM: gray level run length matrix; GLSZM: gray level size zone matrix; GLDZM: gray level distance zone matrix; NGTDM: neighborhood gray tone difference matrix; NGLDM: neighboring gray level dependence matrix)

図　3-1は，DICOMビューワと連動するRadiomicsJの動作イメージを示す．RadiomicsJはRadiomics特徴を計算するモジュールとして機能することで，DICOMビューワへモデルに必要な説明変数を供給することが想定されている．

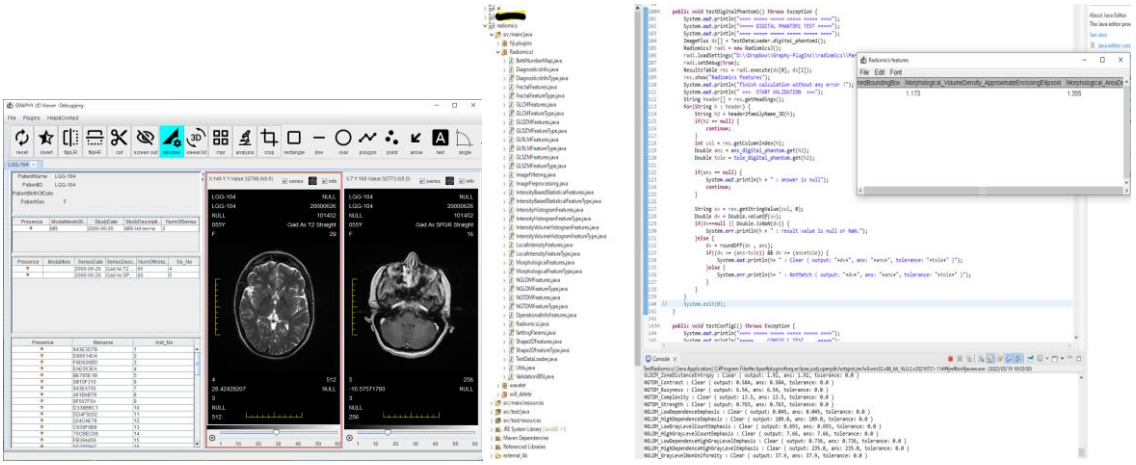


図　3-1 DICOMビューワと連動するRadiomicsJ

（左:DICOMビューワ，右:RadiomicsJ）

3.4　グレード2,3脳腫瘍MRI画像を用いた腫瘍染色体変異推定実験

Radiomicsは一般に，医用画像上に関心領域を定め，関心領域から特徴を抽出し，これらの特徴を他の変数（臨床変数など）と組み合わせてデータセットを構築し，このデータセットを用いてモデルを作成する．図　3-2に本実験手順の概要を示す．

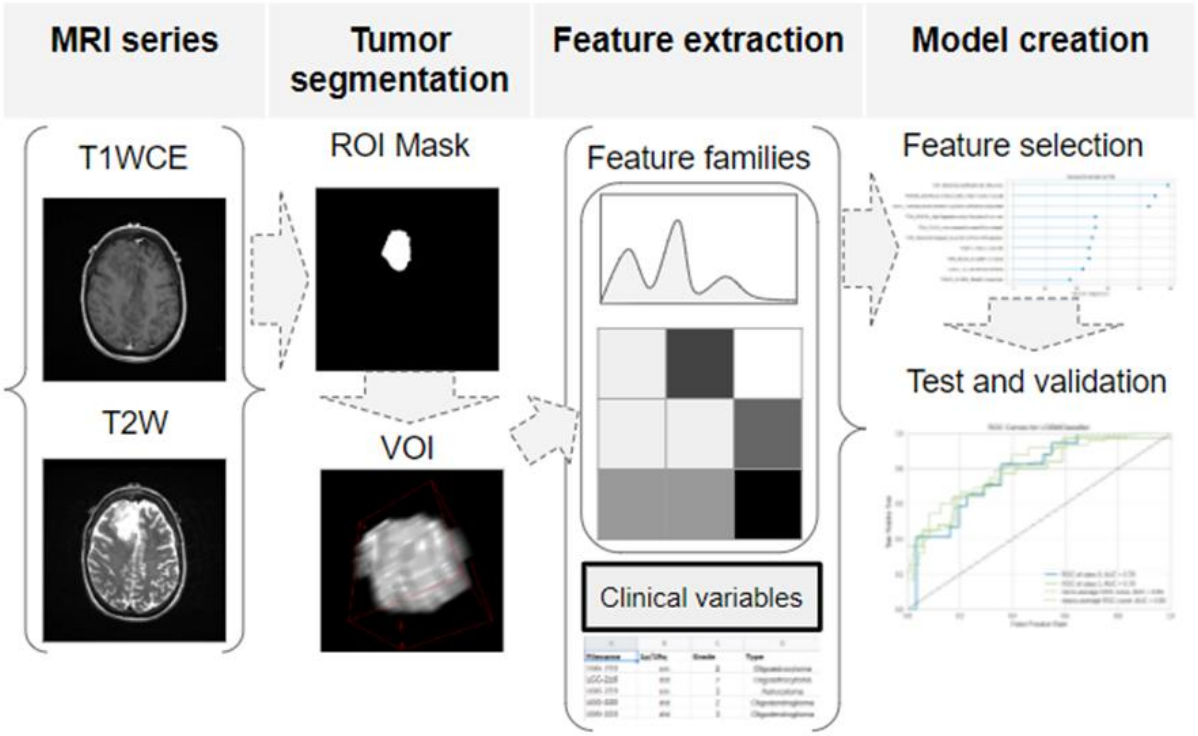


図　3-2 グレード2,3脳腫瘍MRI画像を用いた腫瘍染色体変異推定実験のためのRadiomics手順

3.4.1 データセット

本実験では，159人の被験者の術前MRIを含むCancer Imaging Archive 1p/19qデータセット[13]を使用した．このデータには，1p/19qのco-deleted armを持つ102人と，co-deleted armでない57人の被験者が含まれていた．病変のグレードはグレード2（n＝104）および3（n＝55）であった．LGGの種類は，乏突起星細胞腫（n = 97），乏突起膠腫（n = 45），および星細胞腫（n = 17）であった．年齢の中央値は42歳（範囲，13-84），このデータセットには女性76人，男性83人が含まれていた．このデータセットには，各被験者の造影T1強調画像，T2強調画像，および腫瘍ROIマスク画像が含まれている．

3.4.2 他のRadiomicsライブラリとの比較

RadiomicsJとの比較のために，PyRadiomics（v.3.0.1）が利用された．各ツールからデフォルトで計算可能なすべての画像特徴を学習データセットとして利用した．学習に不要な幾何学的情報や被験者IDなどの管理情報はモデルの作成前に除去された．また，相関係数が0.9以上の特徴量はその一方が除外された．モデルの評価には，PyCaret（v.2.3.8）を使用した．データセットは70%をトレーニングデータとバリデーションデータに，残りはテストデータに被験者単位でグループ化して分割された．クラスラベルの不均衡は，Synthetic Minority Over-sampling Technique（SMOTE）法によって調整された．10分割交差検証法により，AUCが最も高くなる分類モデルが選択された．

比較の結果，RadiomicsJの画像特徴を基に作成されたモデルが，AUC 0.78，PyRadiomicsの画像特徴を基に作成されたモデルはAUC 0.76を達成し，RadiomicsJが提供できる画像特徴の方が分精精度の高いモデルをもたらした（図　3-3）．

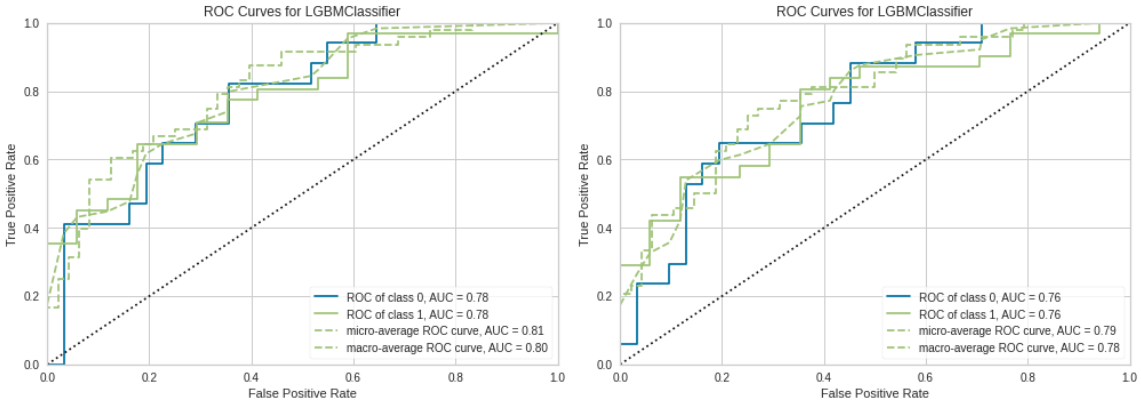


図　3-3 AUCによる比較

（左：RadiomicsJモデル，右：PyRadiomicsモデル）

RadiomicsJモデルのAUCが高くなった要因として，計算可能な特徴数が関係していると考えられた．PyRadiomicsは3 次元特徴量として 110 個（画像フィルタリングなし）の特徴量を算出できるのに対し，RadiomicsJ では 172個の特徴量を算出できるため，モデルの学習に有効な特徴を提供したと考えられる．例えば，RyRadiomicsは，Gray Level Distance Zone MatrixやFractal特徴を含んでいない．また，本検討では対象にならなかった特徴として，ホモロジー特徴 [72]などがある．このような特徴もRadiomicsアプローチに有効である可能性がある．

今回は，PyRadiomicsのみを比較に用いたが，この他にもRaCat，MITKなど，IBSIで定義されたRadiomics特徴を計算可能なツールもある．しかし，計算可能な特徴の種類や処理方法は各ツールに依存するため，今後，さらなる比較検討が必要である．

3.5 まとめ

本研究では，オープンソースのRadiomics特徴計算ライブラリであるRadiomicsJを開発した．RadiomicsJはIBSIリファレンスマニュアルで推奨されている特徴だけでなく，Fractal特徴を計算することができるなど，豊富な画像特徴資源を提供できる．RadiomicsJが提供する画像特徴は従来からの機械学習アプローチの成果物となるモデルの推定精度だけでなく，人が理解できる客観的データを提供するという意味での説明可能性の向上に貢献する．

RadiomicsJは，Radiomicsによる研究の効率化を図るための新たな研究ツールとなるだけでなく，オープンソースプログラムとして公開されることで，医用画像を対象とした特徴研究のための知識リソースとなることが期待される．

第4章　脳腫瘍染色体変異を対象とした分類精度および説明可能性の向上のための進化計算による特徴選択

4.1　はじめに

表形式データを学習データとして用いる従来の機械学習アプローチにおいて，特徴選択はそのパイプラインの構成に欠かせない手順となっている．特徴選択は，目的変数の推定に有効な特徴を取捨選択することによる推定精度向上，説明性や解釈性の向上，推定に影響の少ない特徴を削減（次元削減）することによる計算の高効率化などを目的として行われる．

これまでに多くの医用画像を対象とした画像分類のための特徴選択手法が各分野で検討されてきた．しかし，それでもなお画一的な手法はないことや，各分野でよりよい手法の開発が望まれていることから，特徴選択のアルゴリズムを検討する余地は残されている．

本検討では，遺伝的アルゴリズムを用いた新しい特徴選択手法を提案し，この手法を用いた検証事例を通じてその有用性について考察する．

4.2　遺伝的アルゴリズムから得られる選択個体カウント重要度

従来の特徴選択は，Filter法，Wrapper法，Embedded法，あるいはこれらの組み合わせにより行われることが一般的である．これらの手法による特徴選択は，モデルの推定精度を最大化させるために有用であるものの，特徴の組み合わせをグループとして捉えた場合に，グループ間の交互効果は加味されづらいという課題があった．例えば，モデルベースの特徴重要度を得て，重要度順に変数増加法によって精度を最大化する特徴の組み合わせを探索する際，重要度の高い特徴の組み合わせとしてのグループ比較は行われるが，重要度が低い特徴をグループに加えることによる影響は無視される．

これに対して，遺伝的アルゴリズムによる特徴選択は，特徴グループとしての交互効果を推定精度の観点から評価することができる発見的解法の1つである．しかし，遺伝的アルゴリズムによる特徴選択は，モデルベースの特徴重要度のように，モデルの推論に影響を与える因子を順位づける指標がないため，適応度が最大になる特徴グループを探索した後，その中で推定への影響が大きい特徴を特定することが難しかった．そこで，本検討では遺伝的アルゴリズムの世代ごとに選択された優良な特徴の組み合わせ（Hall Of Fameな個体）として抽出された特徴（個体を構成する因子）の出現回数を積算（カウント）し，これを重要度として捉えた「選択個体カウント重要度」（アルゴリズム 1）を用いた遺伝的アルゴリズムベースのWrapper法を提案する．

Algorithm A genetic algorithm that internalizes selected individuals count importance	
Procedure:	
1:	Population（個体群:特徴の組み合わせ群）を初期化する
2:	Hall Of Fame（適応度が最大の特徴組み合わせ）を初期化する※組み合わせる特徴数はランダム
3:	for number of generations do
4:	子孫(次世代の個体群)をトーナメント方式で選択する
5:	子孫を指定の交叉確率と突然変異確率で変化させる
6:	不適正な適応度を持つ子孫内の個体（特徴の組み合わせ）を評価する
7:	Hall Of Fameを子孫に戻す
8:	子孫内の個体でHall Of Fameを更新する
9:	Hall Of Fame（この世代で最高の精度を達成した特徴の組み合わせ）を保存する

10:	Populationを子孫に置き換える
11:	end for
12:	各世代のHall Of Fameに含まれる特徴の出現回数(選択個体カウント重要度)を算出
13:	最大精度のHall Of Fameの特徴セット, 選択個体カウント重要度を返す

アルゴリズム 1 選択個体カウント重要度を内挿した遺伝的アルゴリズム

4.3 実験:グレード2,3 脳腫瘍1p/19q共欠失分類を対象とした特徴選択手法の比較

4.3.1 データセット

3章のデータセットに準ずる. 学習データとして, Radiomics特徴, 深層特徴, 臨床変数(年齢)を使用する. このうち, 臨床変数である年齢は, 神経膠腫のグレードを推定する上で重要な因子であることが知られている [73].

4.3.2 実験設定

Radiomics特徴は, RadiomicsJを用いて172の特徴が計算された. 深層特徴の抽出は, モデルのスケーリングが定式化されているEfficientNet [74]を用いて行われた. 特徴抽出に用いられたEfficientNetのタイプは, 予備実験により最も分類に有効な特徴抽出器を調査した結果選ばれたEfficientNet-B3が用いられた. 深層特徴はEfficientNetのボトルネック層であるAverage Pooling Layerから抽出された. 深層特徴量はランダム初期化, ImageNet(転移学習なし), 転移学習モデルのそれぞれから抽出された特徴を用いた.

特徴選択法として, 分散閾値法, 単変量特徴選択法として F 統計量による分散分析(ANOVA-F), モデルベース手法としてランダム フォレストモデルの特徴重要度, モデルベースで説明力の高い手法 としてPermutation重要度, および本提案手法の5種類が用いられた. 表 4-1に各特徴選択法の設定パラメータを示す.

表 4-1 各特徴選択法の設定パラメータ

Feature Selection	Parameters
Variance threshold	Threshold: (.8 * (1 - .8))
ANOVA with F-statistics (ANOVA-F)	-
Random forest feature importance	Number of estimators: 300, class weight: 'balanced'
Permutation importance	Classifier: random forest (number of estimators: 300, class weight: balanced), scoring: roc_auc, number of repeats: 3
Genetic algorithm	Estimator: random forest (number of estimators: 300, class weight: balanced), cross-validation: StratifiedGroupKFoldCV(number of repeats: 3), scoring: roc_auc, max_features: 100, number of population: 50, crossover probability: 0.5, mutation_probability: 0.2, number of generations: 30, crossover independent probability: 0.5, mutation independent probability: 0.05, tournament size: 3, number of generations no change: 10

図 4-1にモデルの評価プロセスを示す. 特徴抽出のための前処理として, 標準化, 定数の除去, 相関係数の閾値(0.9 以上)を適用した. 学習データセットは, 70%を訓練データ, 30%をテストデータに利用した. 特徴選択は訓練データを用いて行われた. 分類性能の評価指標として, AUC, 正解率(Acc), 感度, 特異度, f1-score, Log-lossが用いられた. このうち, AUCは主要評価指標として用いられた. モデルの分類性能は, 各特徴選択法で得られる重要度や係数で降順に並べられた特徴を変数増加法によって組み合わせを変える都度算出された.

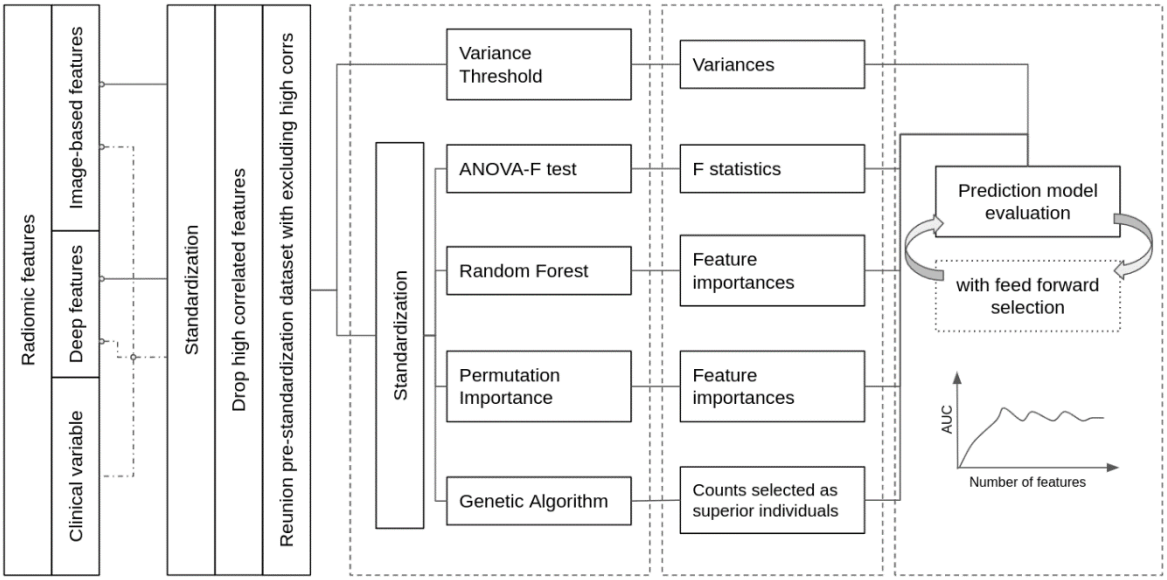


図 4-1 評価プロセス概要

そして，説明変数の変化による影響を検証するために，異なる学習データセットで学習されたType 0からType 2-cまでの7つのモデルが比較された．表 4-2に比較されたモデルの種類を示す．

表 4-2 比較されたモデルタイプ（それぞれのモデルで学習データセットの組み合わせが異なる）

Predictive model type (Type)	Combination of dataset					
	EffNet type	Image-based feature	Deep feature [RandomInit]	Deep feature [imagenet]	Deep feature [transfer]	Clinical variable (age)
Image-based feature (Type 0)	-	✔				
Deep with randominit (Type 1-a)	B3		✔			
Deep with imagenet (Type 1-b)	B3			✔		
Deep with transfer (Type 1-c)	B3				✔	
Img+Deep[randominit]+age (Type 2-a)	B3	✔	✔			✔
Img+Deep[imagenet]+age (Type 2-b)	B3	✔		✔		✔
Img+Deep[transfer]+age (Type 2-c)	B3	✔			✔	✔

4.3.3 分類精度

表 4-3は，各手法で最もAUCが高くなったモデルのタイプ，最大AUC時の特徴数，分類性能の各スコアを示したものである．すべての特徴選択法で，最も高いAUCを示したモデルのタイプはRadiomics特徴のみを学習データとしたType 0であった．Type 0モデルにおいて，提案手法とランダムフォレストによる特徴重要度によるアプローチでは，AUCが0.69，特異度が0.65と最も高く，Permutation重要度ではAccが0.64，感度0.64，f1-score 0.69と最も高かった．ANOVA-Fを用いたType 0モデルのAUCは0.64と最も低かった．

損失の評価では，提案手法のType 0モデルが最小のLog-loss（0.65）を示した．ANOVA-Fとランダムフォレストによる手法は，0.70の最大のLog-lossであった．

表 4-3 特徴選択別最大AUCモデルおよびその分類性能

Feature selection	Best model type	Num of selected	Scores of metrics					
			AUC	Acc	Sens	Spec	f1	Log-loss
Variance threshold	Type 0	24	0.66	0.60	0.58	0.62	0.65	0.68
ANOVA-F	Type 0	40	0.64	0.57	0.54	0.63	0.62	0.70
Random forest feature importance	Type 0	30	0.69	0.63	0.61	0.65	0.68	0.70
Permutation importance	Type 0	24	0.68	0.64	0.64	0.64	0.69	0.66
Genetic algorithm	Type 0	13	0.69	0.63	0.62	0.65	0.68	0.65

4.3.4 特徴選択法の比較に関する考察

提案手法は，最も少ない特徴数（すなわち，高い次元削減効果），最も高いAUCと特異度，および最小のLog-loss を持つモデルを提供したことから，比較した他の特徴選択法に比べ，総合的にもっともよいモデルを提供したと考えられる．他の手法が変数ごとに着目するのに対し，遺伝的アルゴ リズムは評価指標を最適化する特徴群を選択するため，Radiomics特徴，深層特徴，臨床変数のように変数間の意味的な繋がりを説明することが困難な状況下でも，分類性能を高める特徴群を選択できた可能性がある．

ただし，多すぎる説明変数による検討下では，適切な遺伝的アルゴリズムのハイパーパラメータが設定されない限り，本手法が有効に働かない場合があることも考えられるため，ハイパーパラメータの最適化が今後の課題であると考えられた．

選択個体カウント重要度は，遺伝的アルゴリズムで選択された各世代のHall Of Fameに含まれる特徴の出現回数を積算した値であるために，同じ重要度をもつ特徴が抽出されることがある．この課題に対処するために，世代の順位で重みづけを行うなどの重要度としての解像度を高める改良が今後必要であると考えられた．

統計的手法に基づく分散閾値とANOVA-Fは，他の特徴選択手法よりも分類性能が低かったが，統計的手法には統計値で説明できるメリットがあり，分散閾値やANOVA-Fなどの組み合わせで使用することもできるため，特徴選択手法として推奨されないわけではなく，目的に応じてモデル作成の前処理に適用されることが望ましいと考えられた．

4.4 まとめ

本検討では，遺伝的アルゴリズムによる特徴選択に選択個体カウント重要度を導入した．選択個体カウント重要度は，遺伝的アルゴリズムの世代ごとに選択された優良な特徴の組み合わせから得られる各特徴の出現回数の積算値である．この重要度は，抽出された特徴グループの中からさらに推定に影響の大きい因子を決定するために役立ち，従来の遺伝的アルゴリズムによる特徴選択では処理できなかった特徴グループ抽出後の重要度によるWrapper法を可能にする．脳MRI画像を用いた1p/19q共欠失分類を対象とした実験においては，提案手法は検証された他の特徴選択法と比べ，次元削減の効果が大きく，かつ，分類精度が高く，損失が小さいモデルを提供した．

第5章　EfficientNetを用いたマンモグラフィ上乳腺石灰化有無分類におけるGrad-CAMによる説明可能性の検討

5.1　はじめに

深層学習による画像分類のアプローチは，従来手法よりも分類精度を向上させることができる可能性があり，医用画像を用いたさまざまな研究に応用されている．また，CNN系の深層学習モデルは，Grad-CAMによって直観的でわかりやすい推定根拠の説明が可能となるなど，深層学習のブラックボックス問題に対する対応策としての説明可能性の改善も進められている．

このような背景のもと，本研究では，マンモグラフィ上乳腺石灰化の自動検出を視野に入れ，EfficientNet [74]を用いた石灰化有無の分類精度およびGrad-CAMを用いた説明可能性について実現可能性を検証した．乳がんを対象とした研究は歴史も古く，乳がんの検出に重要な所見となる乳腺石灰化を対象とした画像分類に関する研究はすでに高い分類精度が期待できる手法が報告 [75]されていることから，深層学習によるアプローチの初期検討を行うために適した研究対象であると考えられる．

マンモグラフィは，一般に超音波やMRIよりも石灰化を捉えることに優れている．石灰化（特に微小石灰化）の所見は，乳房病変の有無の診断や悪性度の判定に重要である [76]．石灰化の有無を自動判別するための単純な2値分類モデルは，コンピュータ支援診断システムにおける高性能なモジュールの一つとして機能する可能性がある．

5.2　EfficientNetとGrad-CAMによる推論根拠の提示

EfficientNetは，無秩序にニューラルネットモデルの層を深くし，中間パラメータを多くすることで高精度なモデルを得ようとするのではなく，深さ，広さ，解像度の観点からモデルのアーキテクチャを変更せずにモデルを最適化する定式化されたスケーリング手法で作成された深層学習モデルであり，ImageNetを対象とした検証で高い分類精度を記録している優れた分類器の1つである．B0からB7など，入力サイズに応じたネットワークパラメータでのスケールが可能となっており，モデルのアーキテクチャは変更されないため，入力サイズが変更されるようなタスクにおいても検証に用いやすいという長所がある．

Grad-CAMは，CNNのネットワークを用いて，CNNの畳み込み層の勾配を利用して，画像内で重要な領域を強調したマップを生成する技術であり，モデルの推論根拠を可視化し，人による意思決定を支援する．

5.3　乳腺石灰化を対象とした深層学習による石灰化有無推定の実験

5.3.1　　データセット

検診マンモグラフィのデジタルデータベース（ Curated Breast Imaging Subset of DDSM, CBIS-DDSM ）[77]に含まれる全石灰化病変1872例から高濃度乳房（乳房密度4以上）を除いた1471例を最終データセットとして使用した．図　5 -1に典型的な石灰化形状を示す．データセットには40種類の石灰化形状（"ROUND_AND_REGULAR "などの複数分類を含む）が含まれている．

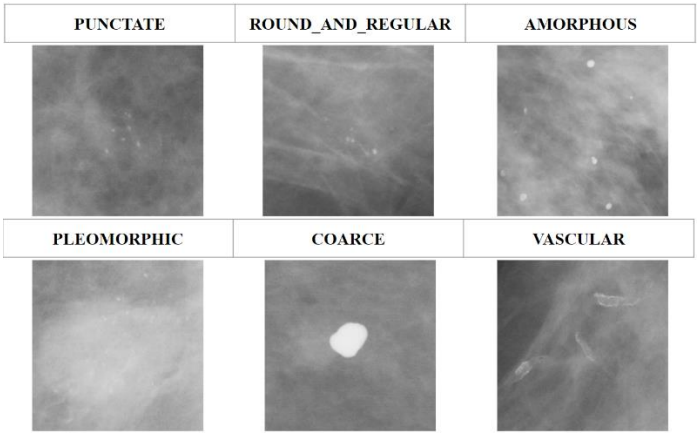


図　5-1 タイプ別の乳腺石灰化例

5.3.2　　実験設定

画像前処理として，Full Fields Digital Mammogram（FFDM）画像を8ビットグレースケールにダウンスケールし，k平均クラスタリングとモルフォロジー処理により乳房ラベル画像を算出した．乳房ラベル画像は，FFDM画像上のX線マーカや背景を消すため画像演算に使用された（ 図　5-2 ）．

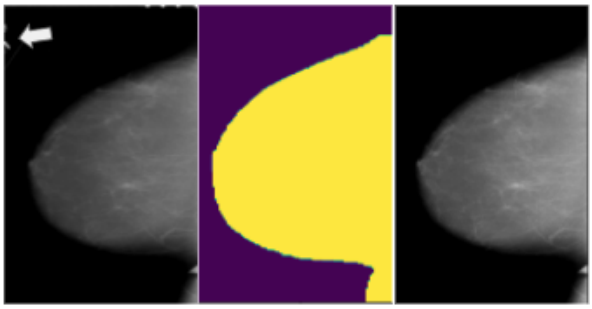


図　5-2 乳房マスク画像の作成

（左：オリジナル，中央：乳房ラベル画像，右：乳房マスク画像）

乳房マスク画像は，石灰化病変と非石灰化病変のパッチ画像（200×200）の作成に使用された．石灰化病変を含むパッチ画像は，石灰化病変のマスクが重なった領域からランダムに切り出すことにより作成され，石灰化を含まない非石灰化パッチ画像はsliding windowにより作成された．石灰化マスクは，Otsu法による2値化によって，CBIS-DDSMからランダムにサンプリングされた平均的な乳腺の信号強度を超える箇所を石灰化領域としてラベルされた．石灰化マスクの重なりや背景が0であるものは，非石灰化パッチ画像から除外された．

パッチ画像の分類器としてEfficientNet-B0が用いられた（ 図　5-3 ）．B0は既定のタイプの中で最も小さい入力サイズ（ 224×224×3 ）が設定されており，パッチベースの大量の計算処理を想定した場合に適したモデルであると考えられた．

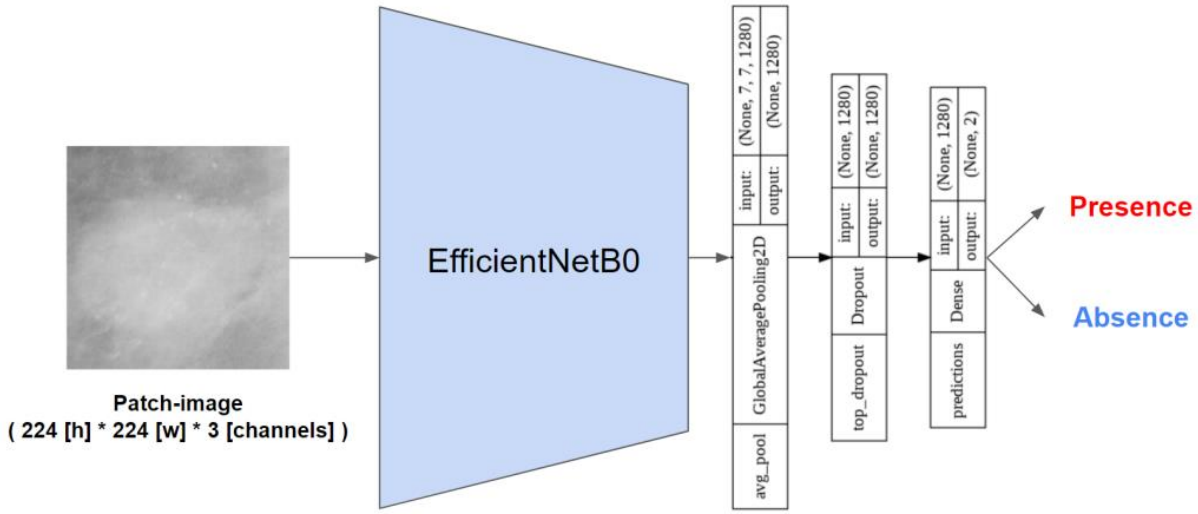


図　5-3 本研究で用いたEfficientNetモデル

モデルは，投影方向（CC，MLO）と左右の組み合わせに分けて計4つ作成された．各パッチ画像データセットは，70%を訓練，10%を検証，20%をテストに分割された．各モデルの重みは，検証AUCが最も高くなった時に保存された．

モデルの分類精度は，Acc，AUC，感度（Sensitivity），特異度（Specificity）にて評価された．モデルの説明可能性は，石灰化カテゴリ別にテストデータの中から選択された複数のパッチ画像を対象としてGrad-CAMによる可視化マップを用いて目視にて評価された．

5.3.3　　分類性能とGrad-CAMによる可視化マップ

表　5-1に各モデルの分類性能を示す．

表　5-1 各モデルの分類性能

	CC		MLO	
	R	L	R	L
AUC	0.86	0.83	0.88	0.85
Acc	0.77	0.75	0.79	0.76
Sensitivity	0.77	0.75	0.79	0.76
Specificity	0.77	0.75	0.79	0.76

すべてのモデルのAUCが0.8以上であったという結果は，一般に優れた分類器として判断できる精度の基準を達成しているため，本分類タスクの実現可能性を確認するためには十分であると考えられた．

次に，説明可能性の評価としてのGrad-CAMの結果は，石灰化ありパッチ画像の場合，石灰化のある領域を捉えて強調することが確かめられた（図　5-4）．しかし，その一方で，正しく推定が行われた場合であっても，石灰化でない領域を強調する傾向があることが確認された（図　5-5）．

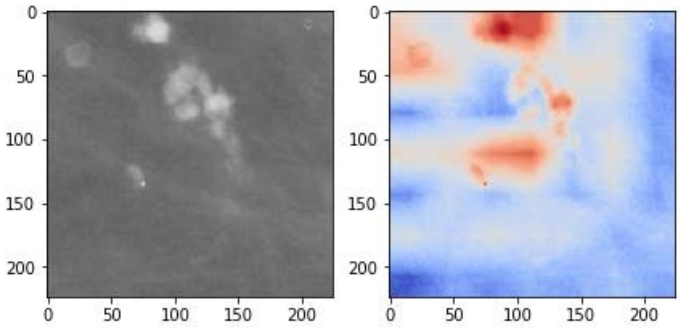


図 5-4 Grad-CAMが正確に関心領域を捉えていることがわかる真陽性例（石灰化推定確率：0.99）

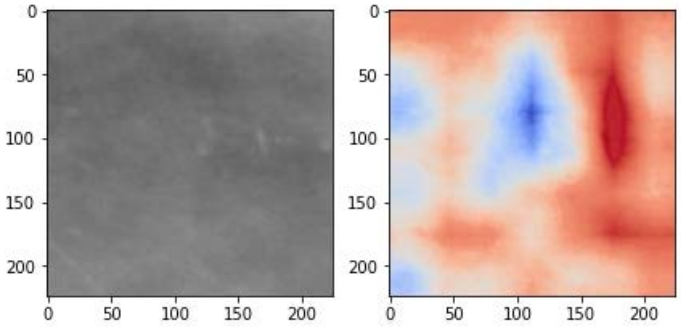


図 5-5 Grad-CAMが正確に関心領域を捉えなかった真陽性例（石灰化推定確率：0.82）

正しく推定が行われた場合であっても，石灰化でない領域を強調してしまった結果は，分類精度と説明可能性のミスマッチとも解釈できる．このようなミスマッチが改善されるためのアプローチの1つとして，推論精度の高さと，人が見て直観的に分かりやすい可視化マップを提供する技術が個々で独立して組み合わされるのではなく，推論と可視化マップとが連動することを前提に設計されたモデルによる対応が有用となる可能性が考えられた．

制限事項として，本研究実施時点で，CBIS-DDSMデータセットには，石灰化の教師ラベル画像（石灰化マスク画像）が含まれておらず，すべての画像を人の手でクラスに振り分けることが困難であったため，本検討では，石灰化を伴う腫瘤病変を対象として，乳腺濃度を考慮した2値化処理によって石灰化ラベルを定義した．この手順は，比較的高濃度の乳房などの影響により，石灰化としてカテゴリされたデータセットに石灰化を含まない画像が多少混在してしまうことを許容して分類精度を評価している．より詳細な評価のために，より正確な石灰化の教師ラベルによって検証が行われることが望ましい．

5.4 まとめ

本研究では，マンモグラフィ上乳腺石灰化有無を対象にして，EfficientNetおよびGrad-CAMを用いて分類精度と説明可能性について検証した．その結果，EfficientNetを用いたパッチ分類器は，AUC 0.8以上を達成し，Grad-CAMも石灰化の領域を強調する傾向が確認できた．今後の研究では，さらなる説明可能性の向上のためのメカニズムの改良が望まれる．

第6章　GCMを用いたマンモグラフィ上乳腺石灰化有無分類における説明可能性の検討

6.1　はじめに

深層学習モデルの推論根拠を人に説明するための可視化マップは，乳房石灰化の有無を推定する際に，推論結果の合理的な根拠を直感的に説明するために有効であると考えられる．しかし，可視化マップが必ずしも正確な関心領域を捉えているとは限らないため，推論結果の説明に可視化マップを適用しても十分な説明能力を果たせないことがある．我々の予備実験 [78]では，石灰化ありクラスに対応するGrad-CAMによる可視化マップは，石灰化以外の領域を捉えている場合があることが示された．より正確に石灰化の関心領域を捉えることができる可視化マップが実現できれば，臨床的な意思決定の質を向上させるための説明可能性の向上に貢献できる可能性がある．

荒井ら [79]は，Generative Contribution Mapping（GCM）という新しい説明可能な深層学習モデルを開発した．GCM は，一般物体を高精度に分類できる深層学習モデルであり，Class Weight Map（CWM）やClass Contribution Map（CCM）という可視化マップにより，推論結果の合理的根拠を直観的に説明する能力を有する．

我々は，GCM が石灰化の有無を高精度に分類し，かつ，正確な関心領域を提示する可視化マップを提供できると仮定した．本研究の目的は，マンモグラフィにおける乳房石灰化の有無の分類に対する GCM の有用性を検証することである．

6.2　画像分類の根拠を説明し易い深層ネットワーク：GCM

図　6-1に本研究で利用したGCM のアーキテクチャを示す．GCMは，オートエンコーダーをベースとした説明可能な深層学習モデルであり，一般画像を対象とした実験において高い分類精度を達成することができるだけでなく，推論結果に対する直感的な合理的根拠を説明する可視化マップとしてCWM とCCM を生成できる．CWMは入力データによって活性化された各クラスに対応する勾配であり，CCMはCWMと入力画像を積算した結果を出力する．

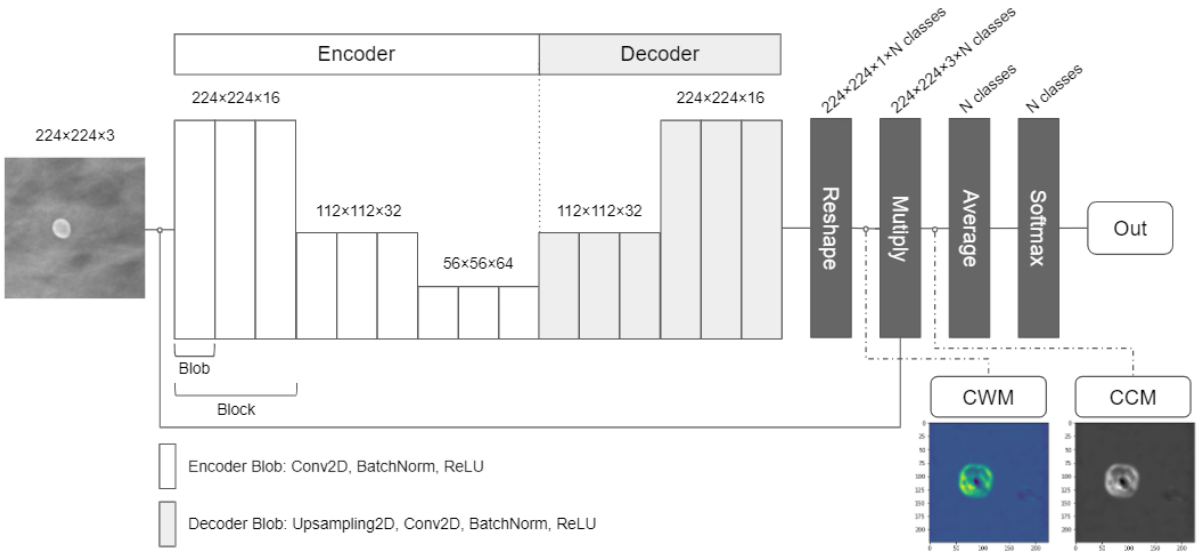


図　6-1 本研究で用いたGCMアーキテクチャ

6.3 GCMを用いたマンモグラフィ上乳腺石灰化有無推定

6.3.1 データセット

本検討では，108人の被験者のFFDMを含むINbreastデータセット [80]が利用された．INbreastデータセットは，CBIS-DDSMと異なり，データセットを作成した医師による石灰化の教師ラベル画像データが含まれており，石灰化に対してより正確な検討が行える．

データセットには，微小石灰化を有する90人の被験者（n = 308）が含まれている．このうち，1例の高濃度乳房症例は，乳房が石灰化や腫瘍などの所見と同様にマンモグラム上で高吸収領域として現れる非常に密な乳腺画像であるため，臨床経験10年以上の放射線科医師によって除外された．最終的に89人の被験者の微小石灰化所見を有するFFDM画像303枚が機械学習データセットとして選択された．

学習データセットは，マンモグラフィの投影方向の違いによる画像上の解剖学的構造が異なることを考慮し，CCとMLOに分割された．乳房の左右については，データセットのサイズを考慮して分割を行わなかった．

6.3.2 実験設定

GCM と EfficientNet-B0の深層学習モデルが比較のために用いられ，GCM-MLO，GCM-CC，EfficientNet-MLO，EfficientNet-CC の 4 種類の分類器が作成された．

本研究で用いた学習条件の設定を表 6-1に示す．

表 6-1 学習条件

Settings	GCM	EfficientNet
Total params	159,746	4,052,133
Number of layers	51	240
Input size	224×224×3	
Output activation function	SoftMax	
Optimizer	SGD	Adam
Loss function	Categorical cross-entropy	
Initial weight	Random initialization	
Batch size	32	
Learning rate	Max: 10^{-3}, Min: 10^{-5}	
Epochs	200	
Checkpoint	Maximum validation AUC	

入力サイズは224×224×3とした．最適化関数は，GCM では確率的勾配降下法（ stochastic gradient descent; SGD ），EfficientNet では Adamが使用された．初期重みはランダムな初期化を行った．学習時の訓練データのオーグメンテーション処理は，垂直回転と平行回転のみを行った．モデルの重みは，検証時の AUCが最も高くなった重みが採用された．

分類性能はAcc，AUC，感度，特異度によって総合的に評価された．可視化マップの精度は，CWM，CCM，GCM-GradCAM，EfficientNet-GradCAMを対象として，石灰化ラベル画像と正規化済みの可視化マップから求めた可視化マップAUC（ AUCvis ）にて評価された．図 6-2にAUCvisの計算方法を示す．AUCvisは，分類精度指標であるROC-AUCが教師ラベル配列と推定確率配列から求められるのに

対し，教師ラベル配列を教師ラベル画像，推定確率配列を正規化した可視化マップ（0から1の間の信号値をもつ）として求められる．

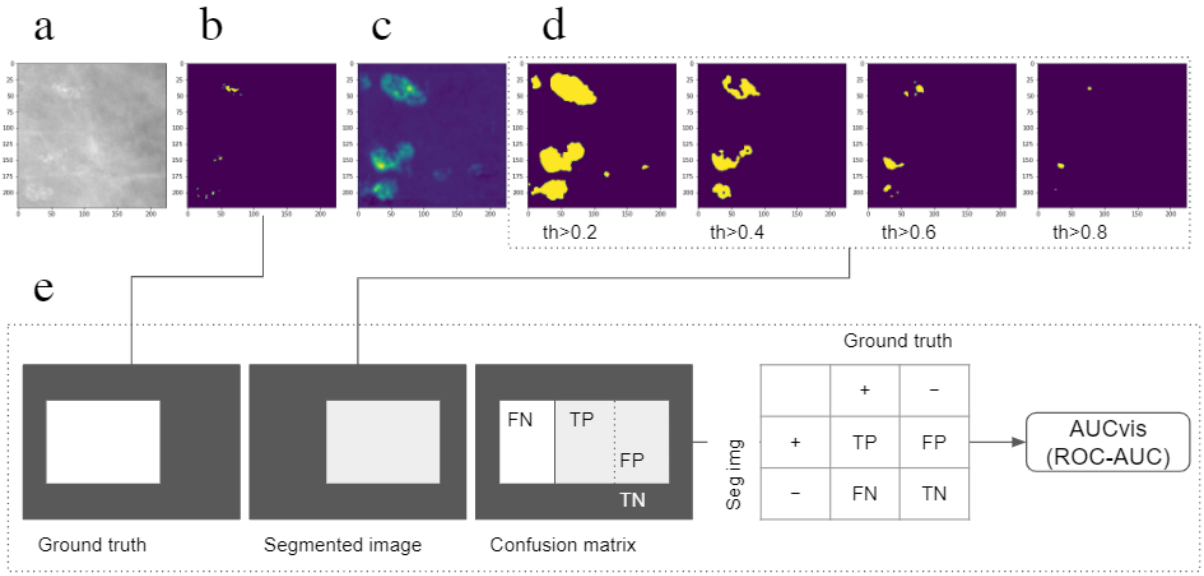


図　6-2 可視化マップAUCの計算

(a: original, b: calcification label, c: visualization map with min-max normalization, d: examples of segmented visualization map with thresholding, e: calculation of AUCvis).

6.3.3　　分類性能と可視化マップ精度

表　6-2は，各分類器の分類性能を示す．GCM の分類器は Acc と AUC がともに 0.9 を超える高い性能を示した．DeLong testの結果，EfficientNetによるモデルはGCMに比べ優位にAUCが高い結果となったが，GCMに基づくMLOとCCの分類器はともにAUC > 0.9を達成しており，一般に，AUC が 0.9 を超える分類器は優れた分類精度を有すると考えられるため，GCMも十分な分類精度を有していると考えられた．

表　6-2 各モデルの分類性能

Models	AUC (95% CI)	p-value	Accuracy	Sensitivity (Recall)	Specificity	Precision (PPV)	NPV
GCM (CC)	0.925 (0.917-0.933)	< 0.01	0.92	0.95	0.90	0.89	0.96
EfficientNet B0 (CC)	0.989 (0.987-0.991)		0.96	0.97	0.95	0.95	0.97
GCM (MLO)	0.949 (0.944-0.954)	< 0.01	0.91	0.98	0.86	0.85	0.98
EfficientNet B0 (MLO)	0.986 (0.984-0.989)		0.96	0.96	0.97	0.97	0.96

図　6-3はAUCvisの結果を示している．AUCvisの平均値は，GCM-CC-CWM 0.87，GCM-CC-GradCAM 0.86，EfficientNet-CC-GradCAM 0.44，GCM-MLO-CWM 0.87，GCM-MLO-GradCAM 0.87，EfficientNet-MLO-GradCAM 0.37 であった．GCM-CWMは，CCとMLOの両ケースにおいて，最も高い平均AUCvisを得ることができた．

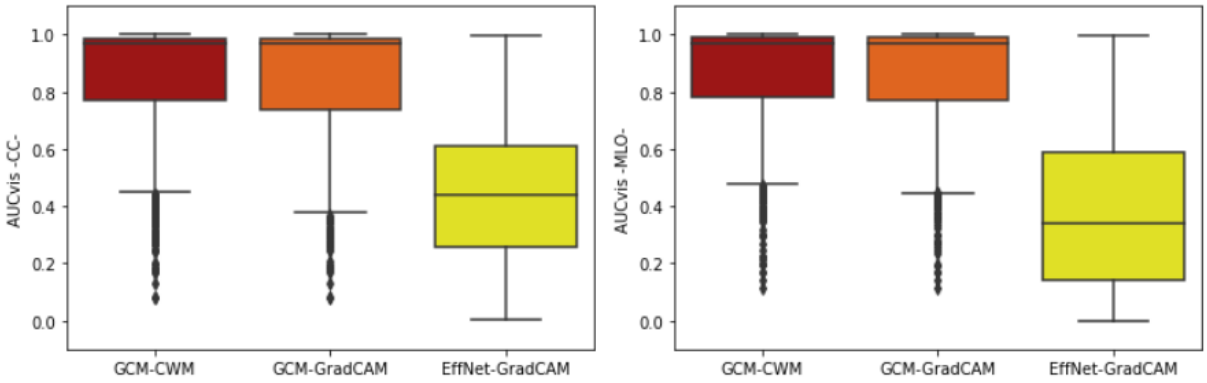


図　6-3 AUCvisの評価結果

図　6-4は可視化マップ4例を示す. 図中, (a) cluster, (b) microcalcification, (c) false negative cluster, (d) false positive noisy sign を示している. c（石灰化あり例）と d（石灰化なし例）は, GCM と EfficientNet のいずれの分類器も正確に推定することができなかった例である. GCM-CWMはEfficientNet-GradCAMよりも石灰化領域を明確に表現していることが分かる（Fig.7 a, b, c）. GCM-GradCAMも石灰化を明瞭に捉えているが, GCM-CWMの平均AUCvisはGCM-GradCAMのそれより若干高くなった.

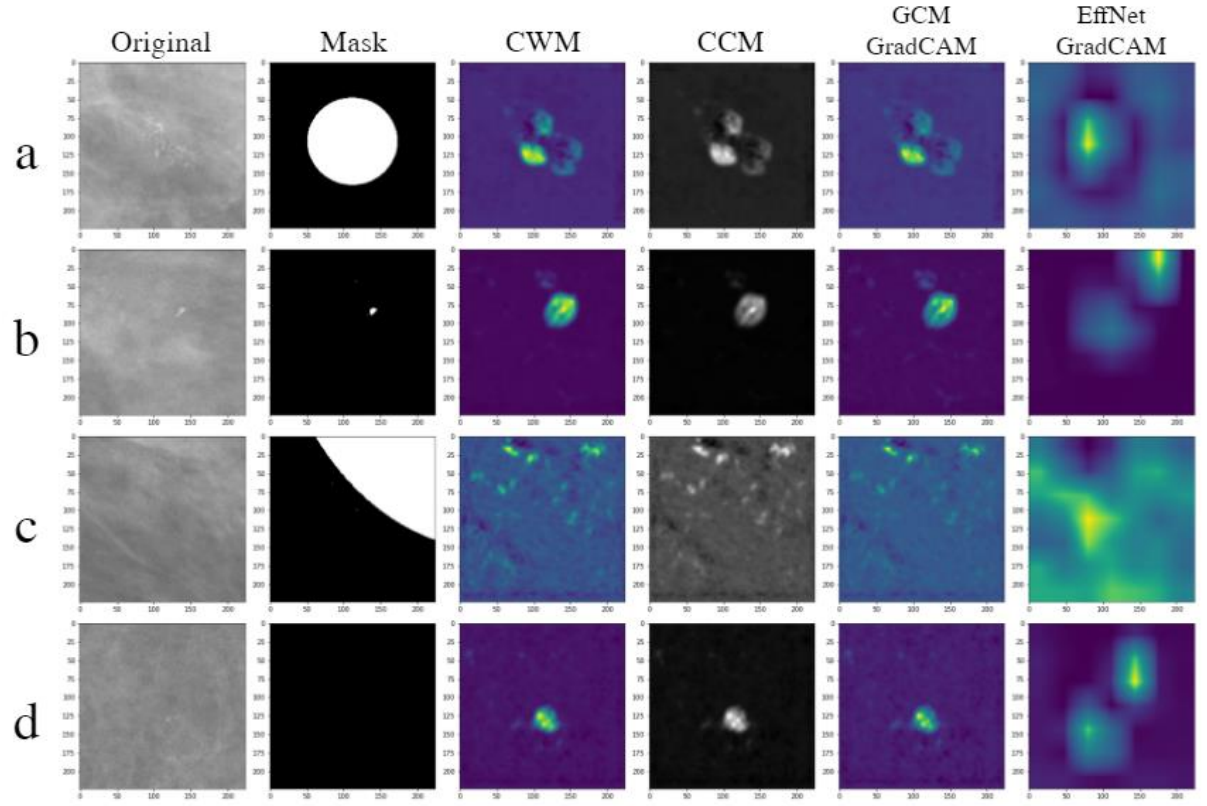


図　6-4 可視化マップ例

(a: cluster, b: microcalcification, c: false negative cluster, d: false positive noisy sign)

6.3.4　　考察

一般物体に対する分類を得意とする EfficientNet の分類性能は GCM よりも有意に高くなかったが, 本検討で用いたGCM は EfficientNet よりもネットワークパラメータと層数が少ない（表　6-1）. GCM の分

類性能は，最小化されたオートエンコーダーネットワーク部分を U-net [81] のような他のネットワークに置き換えることによって，さらに向上する可能性がある．

コントラストが低く，サイズが小さい微小石灰化は，GCM，EfficientNetともに偽陰性を生じさせる原因になった（図　6 -4 c）．より効果的な画像前処理を行うことで，この課題に対処できる可能性がある．

EfficientNetによるモデルは，分類性能は高く，可視化マップは不鮮明というミスマッチを生じさせている．これは他の一般的なCNNでも同様の傾向となることが考えられる．これに対し，本検証結果から分かるように，GCMでは関心領域を正確に捉える可視化マップを提供することから，このミスマッチを軽減しているといえる．また，GCMは，例え推論が偽陰性であったとしても，石灰化クラスの可視化マップ（GCM-CWM，GCM-CCM，GCM-GradCAM）において石灰化領域がシルエットになることが確認された．このことは，GCMが推定を誤ったとしても，GCMの可視化マップが視覚的な解釈をサポートできる可能性があることを示している．

石灰化を検出できる理想的なモデルの要件の1つは，経験豊富な放射線科医や乳腺外科医が容易に判断できない石灰化の有無を区別する，あるいは，意思決定を支援する能力を有していることである．この能力を検証するために，さらなる検証が必要である．

6.4　まとめ

本研究では，説明可能な深層学習モデルであるGCMが，マンモグラフィ上乳腺石灰化有無の分類に対して高い分類精度を示すとともに，GCMによる推論根拠を説明するための可視化マップは他の一般的なCNNによるGrad-CAMによる可視化マップに比べて，石灰化関心領域を正確に捉えることができる可能性があることを示した．

第7章　結論

画像分類を対象としたCADおよびRadiomicsベースのモデルは分類精度および説明可能性が求められる．本論文では，脳腫瘍染色体変異，乳腺石灰化を生物学的特徴の対象として，従来の表形式データを学習データとする機械学習アプローチと，深層学習アプローチそれぞれで検証した．

各章で得られた成果は以下の通りである．

7.1　Radomics特徴計算ライブラリ：RadiomicsJ

本論文執筆時点では，国内から発出されたオープンソースのRadiomics計算ツールはなく，この観点から，本研究にて開発されたRadiomicsJが国内初の成果となると考えられる．また，Radiomicsに利用される画像特徴の標準化や計算可能な特徴の増強は今後とも重要な課題となる．この課題について，RadiomicsJは国際的なデファクトスタンダードとなるIBSIリファレンスに準じて開発されていることや，IBSIに未収載の画像特徴であるFractal特徴が計算可能であるという長所がある．本研究では事例を通じて，他のRadiomics特徴計算ツールとの比較を行うことで，RadiomicsJから出力された豊富な特徴がモデル性能を高めたことを実証した．将来的にさらに充実したRadiomics特徴計算ライブラリとなることが期待される．

一方で，推定精度の高さの観点から，深層学習によるアプローチにも期待が集まっている．伝統的な画像特徴を用いた機械学習アプローチよりも，深層学習によるアプローチの方が期待した精度をもたらすケースは少なくない．これに対して，伝統的な画像特徴は，人が評価する主観的な見た目のパターンや人が目で見て気が付かない画像上の信号パターンに対する，人が理解可能な客観的指標となることから，ブラックボックスとなりやすい深層学習モデルと並行して検討されることで，精度と説明可能性の向上へ貢献できる．

7.2　脳腫瘍染色体変異を対象とした分類精度および説明可能性の向上のための進化計算による特徴選択

目的変数の推定に有効な特徴を取捨選択することによる推定精度向上，説明性や解釈性の向上，推定に影響の少ない特徴を削減（次元削減）することによる計算の高効率化などを目的として，従来の遺伝的アルゴリズムによる特徴選択にはない，選択個体カウント重要度を提案し，遺伝的アルゴリズムによる特徴選択後，さらに選択個体カウント重要度順の変数増加法にて最良の分類精度を達成するモデルを探索する手法を示した．

グレード2,3脳腫瘍の染色体変異有無を推定する実験にて検証した結果，本提案手法は，比較された特徴選択法に比べて，総合的によい性能をもつモデルの作成を可能にした．

7.3　EfficientNetを用いたマンモグラフィ上乳腺石灰化有無分類におけるGrad-CAMによる説明可能性の検討

分類精度の高さから第一選択肢になりつつある深層学習モデルを用いて，過去の事例から高い分類精度の実現可能性が見込まれるマンモグラフィ上乳腺石灰化有無を対象として，これまでに検証報告が少ないと考えられたEfficientNetとCBIS-DDSMデータセットの組み合わせの実験設定下において，マンモグラフィ上乳腺石灰化有無推定を対象とした分類精度および説明可能性の検証を行った．分類精度だけでなく，Grad-CAMを用いて推論根拠領域を可視化することで，説明可能性についても言及した点

が本研究の主な新規性であり，その検討の末に，GCMへ応用するアイディアに繋がったことも本研究の価値の1つである．

7.4 GCMを用いたマンモグラフィ上乳腺石灰化有無分類における説明可能性の検討

EfficientNetを用いたマンモグラフィ上乳腺石灰化有無分類におけるGrad-CAMによる説明可能性の検討の結果を踏まえ，分類精度と説明可能性の両立に有用であると考えられた説明可能な深層学習モデルであるGCMを用いたアプローチをEfficientNetによるアプローチと比較した．

GCMはAUC 0.9を超える分類精度を示し，GCMの可視化マップはEfficientNetのGrad-CAMに比べて石灰化の関心領域を正確に捉えることができた．本検証は，一般物体でのみ検証されてきたGCMを医療分野に応用した初めての研究であると同時に，GCMを石灰化有無に適用することによって存在診断の意思決定支援全般の説明可能性に対する課題に対して，GCMが応用可能である可能性を示唆している．GCMのさらなるネットワークの改良や応用が期待される．

謝辞

博士課程後期進学の機会をいただき，また，筆者の至らぬ点を，経験と心の厚みで受け入れて下さりつつ，研究におけるご助言を明晰に言葉にして筆者をご指導下さいました，長尾 智晴 教授に心よりお礼申し上げます．また，本論文をまとめるにあたり貴重なご指導とご助言をいただきました，森辰則先生，富井尚志先生，白崎実先生，白川真一先生に感謝申し上げます．

本研究に際してご支援いただきました長尾研究室の皆様にも感謝申し上げます．研究活動を通じて，さまざまなサポートをいただきました長尾研究室の秘書様方，大変お世話になりました．そしていつも見守り支えて貰っている両親と家族にはいくら感謝しても足りません．皆様のお力添えによって何とか本研究を終えることができました．本当にありがとうございました．

医用画像情報学は，医療へデータサイエンスを応用することで，さまざまな非効率やこれまでできなかったことを可能にすることができる力の一つであると信じています．本研究で培ったこの力を駆使して，世の中を少しでも良くしていくことに貢献できる可能性があると信じてこれからも精進致します．

参考文献

[1] 厚生労働省, “電子カルテシステム等の普及状況の推移,” *医療施設調査,* 2021.

[2] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun and J. Dean, ″A guide to deep learning in healthcare,″ *Nature Medicine,* vol. 25, pp. 24-29, 2019.

[3] G. Currie, K. Hawk, E. Rohren, A. Vial and R. Klein, ″Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging,″ *J Med Imaging Radiat Sci,* vol. 50, no. 4, pp. 477-487, 2019.

[4] M. N. Prakash, O.-M. Lucila , W. C. Wendy, “Natural language processing: an introduction,” *Journal of the American Medical Informatics Association,* 第 巻18, 第 5, pp. 544-551, 2011.

[5] K. Doi, ″Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential,″ *Comput Med Imaging Graph,* vol. 31, no. 4-5, pp. 198-211, 2007.

[6] H. Chan, L. Hadjiiski and R. Samala, ″Computer-aided diagnosis in the era of deep learning,″ *Med Phys,* vol. 47, no. 5, pp. e218-e227, 2020.

[7] P. Lambin, E. Rios-Velazquez , R. Leijenaar, “Radiomics: extracting more information from medical images using advanced feature analysis,” *Eur J Cancer,* 第 巻48, 第 4, pp. 441-6, 2012.

[8] H. ARIMURA and M. SOUFI, ″A Review on Radiomics for Personalized Medicine in Cancer Treatment,″ *MEDICAL IMAGING TECHNOLOGY,* vol. 36, no. 2, pp. 81-89, 2018.

[9] M.-B. Luis and A.-B. Angel, Imaging Biomarkers, Development and Clinical Integration, Valencia, Spain: Springer, 2017.

[10] H.-P. Chan, K. Doi, S. Galhotra, C. Vyborny, H. MacMahon and P. Jokich, ″Image feature analysis and computer-aided diagnosis in digital radiography,″ *Medical Physics,* vol. 14, p. 538-548, 1987.

[11] H.-P. Chan, K. Doi and C. Vyborny, ″Improvement in radiologists’ detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis,″ *Investigative Radiology,* vol. 25, pp. 1102-1110, 1990.

[12] P. Reel, S. Reel, E. Pearson, E. Trucco , E. Jefferson, “Using machine learning approaches for multi-omics data analysis: A review,” *Biotechnol Adv,* 第 巻49, p. 107739, 2021.

[13] M. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs and G. Cook, ″Introduction to Radiomics,″ *J Nucl Med,* vol. 61, no. 4, pp. 488-495, 2020.

[14] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. Eschrich, M. Schabath, K. Forster, H. Aerts, A. Dekker, D. Fenstermacher, D. Goldgof, L. Hall, P. Lambin, Y. Balagurunathan, R. Gatenby and R. Gillies, ″Radiomics: the process and the challenges,″ *Magn Reson Imaging,* vol. 30, no. 9, pp. 1234-48, 2012.

[15] R. Larracy, A. Phinyomark and E. Scheme, "Machine Learning Model Validation for Early Stage Studies with Small Sample Sizes," *Annu Int Conf IEEE Eng Med Biol Soc,* pp. 2314-2319, 2021.

[16] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology,* vol. 143, no. 1, pp. 29-36, 1982.

[17] M. K. K and E. P, "Algorithmic Splitting: A Method for Dataset Preparation," *IEEE Access,* vol. 9, pp. 125229-125237.

[18] E. B, "Bootstrap Methods: Another Look at the Jackknife," *Ann. Statist,* vol. 7, no. 1, pp. 1-26, 1979.

[19] v. B. Stef and G.-O. Karin, "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software,* vol. 45, pp. 1-67, 2011.

[20] A. Donders, G. Heijden, T. Stijnen and K. Moons, "Review: a gentle introduction to imputation of missing values," *J Clin Epidemiol,* vol. 59, no. 10, pp. 1087-91, 2006.

[21] W. Yoo, R. Mayberry, S. Bae, K. Singh, H. Q. Peter and J. J. Lillard, "A Study of Effects of MultiCollinearity in the Multivariable Analysis," *Int J Appl Sci Technol,* vol. 4, no. 5, pp. 9-19, 2014.

[22] K. Ron and H. J. George, "Wrappers for feature subset selection," *Artificial Intelligence,* vol. 97, pp. 273-324, 1997.

[23] v. d. M. Laurens and H. Geoffrey, "Visualizing Data using t-SNE," *J Machine Learning Research,* vol. 9, pp. 2579-2605, 2008.

[24] M. Leland, H. John and M. James, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv,* p. 1802.03426v3, 2020.

[25] E. SUZUKI, H. KOMATSU, T. YORIFUJI and E. YAMAMOTO, "Causal Inference in Medicine Part II," *Jpn. J. Hyg,* vol. 64, pp. 796-805, 2009.

[26] S. Quazi, "Artificial intelligence and machine learning in precision and genomic medicine," *Med Oncol,* vol. 39, no. 120, 2022.

[27] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician,* vol. 46, no. 3, p. 175-185, 1992.

[28] D. Cox, "The regression analysis of binary sequences (with discussion)," *J Roy Stat Soc B,* vol. 20, p. 215-242, 1958.

[29] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, p. 273-297, 1995.

[30] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

[31] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification And Regression Trees, New York: Routledge, 1984.

[32] T. K. Ho, "Random Decision Forests," *Proceedings of the 3rd International Conference on Document Analysis and Recognition,* p. 278-282, 1995.

[33] S. Gallant, "Perceptron-based learning algorithms," *IEEE Transactions on Neural Networks,* vol. 1, no. 2, pp. 179-191, 1990.

[34] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences,* vol. 55, no. 1, pp. 119-139, 1995.

[35] C. Tony, H. G. Gene and J. L. Randall, "UPDATING FORMULAE AND A PAIRWISE ALGORITHM FOR COMPUTING SAMPLE VARIANCES," STANFORD UNIVERSITY, 1979.

[36] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature,* vol. 521, p. 436-444, 2015.

[37] S. Omer and R. Lior, "Ensemble learning: A survey," *WIREs DATA MINING AND KNOWLEDGE DISCOVERY,* vol. 8, no. 4, p. e1249, 2018.

[38] N. Japkowicz and M. Shah, "Part 6 Statistical Significant Testing," in *Evaluating Learning Algorithms: A Classification Perspective*, Cambridge University Press, 2011, pp. 206-291.

[39] E. DeLong, D. DeLong and D. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics,* vol. 44, no. 3, pp. 837-45, 1988.

[40] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," *npj Digit. Med,* vol. 5, no. 48, 2022.

[41] F. Rosenblatt, "The perceptron：A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review,* vol. 65, no. 6, pp. 386-408, 1958.

[42] D. Rumelhart, G. Hinton and R. Williams, "Learning Representations by Back-propagating Errors," *Nature,* vol. 323, no. Oct. 9, pp. 533-536, 1986.

[43] G. Hinton, S. Osindero and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput,* vol. 18, no. 7, pp. 1527-54, 2006.

[44] Y. LeCun, B. Boser, J. S. Denker and D. Henderson, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation,* vol. 1, no. 4, pp. 541-551, 1989.

[45] K. Fukushima, "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," *Biological Cybernetics,* vol. 36, no. 4, pp. 193-202, 1980.

[46] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision,* vol. 115, no. 3, pp. 211-252, 2015.

[47] M. Christos, F. H. Johan, S. Moein, S. Magnus and S. Kevin, "What Makes Transfer Learning Work for Medical Images: Feature Reuse & Other Factors," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 9215-9224, 2022.

[48] S. Kundu, "AI in medicine must be explainable," *Nat Med,* vol. 27, p. 1328, 2021.

[49] C. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med,* vol. 17, no. 195, 2019.

[50] B. A. Alejandro, D.-R. Natalia, D. S. Javier, B. Adrien, T. Siham, B. Alberto, G. Salvador, G.-L. Sergio, M. Daniel, B. Richard, C. Raja and H. Francisco, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *arXiv,* p. arXiv:1910.10045v2 , 2019.

[51] T. R. Marco, S. Sameer and G. Carlos, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," *arXiv,* 2016.

[52] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *arXiv,* 2017.

[53] D. Smilkov, N. Thorat, B. Kim, F. Viégas and M. Wattenberg, "SmoothGrad: removing noise by adding noise," *arXiv,* 2017.

[54] S. Mukund, A. Taly and Y. Qiqi, " Axiomatic attribution for deep networks," *In Proceedings of the International Conference on Machine Learning (ICML),* pp. 3319-3328, 2017.

[55] M. Grégoire, L. Sebastian, B. Alexander, S. Wojciech and M. Klaus-Robert, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognition,* vol. 65, pp. 211-222, 2017.

[56] D. Z. Matthew and F. Rob, "Visualizing and understanding convolutional networks," *In Proceedings of the European Conference on Computer Vision (ECCV),* pp. 818-833, 2014.

[57] J. T. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," *arXiv,* 2014.

[58] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. Müller and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS One,* vol. 10, no. 7, p. e0130140, 2015.

[59] R. S. Ramprasaath, C. Michael, D. Abhishek, V. Ramakrishna, P. Devi and B. Dhruv, "Grad-{CAM}: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision,* vol. 128, no. 2, pp. 336-359, 2019.

[60] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *arXiv,* 2015.

[61] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems,* vol. 47, no. 4, pp. 547-553, 2009.

[62] S. Arai, "Toward Explainable Deep Network Models," Ypkohama National University Institutional Repository, Yokohama, 2019.

[63] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2921-2929, 2016.

[64] Z. Fan, Z. Sun, S. Fang, Y. Li, X. Liu, Y. Liang, Y. Liu, C. Zhou, Q. Zhu, H. Zhang, T. Li, S. Li, T. Jiang, Y. Wang and L. Wang, "Preoperative Radiomics Analysis of 1p/19q Status in WHO Grade II Gliomas," *Front. Oncol,* vol. 11, p. 616740, 2021.

[65] H. W. Jarey, A. W. Kareem, V. v. D. Lisanne, F. Keyvan, F. T. Reid and D. F. Clifton, "Radiomic biomarkers of tumor immune biology and immunotherapy response," *Clinical and Translational Radiation Oncology,* vol. 28, pp. 97-115, 2021.

[66] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox and F. Prior, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *Journal of Digital Imaging,* vol. 26, pp. 1045-1057, 2013.

[67] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper and H. J. W. L. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research,* vol. 77, no. 21, pp. E104-E107, 2017.

[68] E. Pfaehler, A. Zwanenburg, J. de Jong and R. Boellaard, "RaCaT: An open source and easy to use radiomics calculator tool," *PLoS ONE,* vol. 14, no. 2, p. e0212223, 2019.

[69] M. Nolden, S. Zelzer and A. Seitel, "The Medical Imaging Interaction Toolkit: challenges and advances," *Int J CARS,* vol. 8, p. 607-620, 2013.

[70] R. Kikinis, S. D. Pieper and K. G. Vosburgh, "3D Slicer: A platform for subject-specific image analysis, visualization, and clinical support," in *Intraoperative Imaging and Image-Guided Therapy*, New York, Springer, 2014, p. 277-289.

[71] A. Zwanenburg, S. Leger, M. Vallières and S. Löck, "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping," *Radiology,* vol. 295, no. 2, pp. 328-338, 2020.

[72] K. Ninomiya and A. H, "Homological radiomics analysis for prognostic prediction in lung cancer patients," *Phys Med,* vol. 69, no. Jan, pp. 90-100, 2020.

[73] Z. Lin, R. Yang and K. Li, "Establishment of Age Group Classification for Risk Stratification in Glioma Patients," *BMC Neurol,* vol. 20, no. 310, 2020.

[74] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proceedings of the 36th International Conference on Machine Learning, PMLR,* vol. 97, pp. 6105-6114, 2019.

[75] M. S. H. Kumar, "Computer Aided Detection of Clustered Microcalcification: A Survey," *Curr Med Imaging Rev,* vol. 15, no. 2, pp. 132-149, 2019.

[76] L. Wilkinson, V. Thomas and N. Sharma, "Microcalcification on mammography: approaches to interpretation and biopsy," *Br J Radiol,* vol. 90, no. 1069, 2017.

[77] R. Lee, F. Gimenez, A. Hoogi, K. Miyake, M. Gorovoy and D. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data,* vol. 4, no. 170177, 2017.

[78] T. Kobayashi, T. Haraguchi and T. Nagao, "Classifying the Presence or Absence of Calcifications on Mammography using the EfficientNet: A Feasibility Study," *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech),* pp. 596-599, 2022.

[79] S. Arai and T. Nagao, "Intuitive Visualization Method for Image Classification Using Convolutional Neural Networks," *IPSJ SIG Technical Report,* Vols. 2016-MPS-111 (10), 2016.

[80] I. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. Cardoso and J. Cardoso, "INbreast: toward a full-field digital mammographic database," *Acad Radiol,* vol. 19, no. 2, pp. 236-48, 2012.

[81] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI). Lecture Notes in Computer Science,* no. 9351, 2015.

研究業績リスト

論文誌

- Kobayashi T. RadiomicsJ: a library to compute radiomic features. Radiol Phys Technol. 2022 Jul 6. doi: 10.1007/s12194-022-00664-4. Epub ahead of print. PMID: 35792994.
- Kobayashi T, Haraguchi T, Nagao T. Classifying presence or absence of calcifications on mammography using generative contribution mapping. Radiol Phys Technol. 2022 Aug 21. doi: 10.1007/s12194-022-00673-3. Epub ahead of print. PMID: 35988097.
- Tomita H, Yamashiro T, Heianna J, Nakasone T, Kobayashi T, Mishiro S, Hirahara D, Takaya E, Mimura H, Murayama S, Kobayashi Y. Deep Learning for the Preoperative Diagnosis of Metastatic Cervical Lymph Nodes on Contrast-Enhanced Computed ToMography in Patients with Oral Squamous Cell Carcinoma. Cancers (Basel). 2021 Feb 3;13(4):600. doi: 10.3390/cancers13040600. PMID: 33546279; PMCID: PMC7913286.
- Tanuma T, Kobayashi T, Takaya E, Suzuki D, Inoue M, Yoshikawa T, Kobayashi Y. Object Detection Model Utilizing Deep Learning to Identify Retained Surgical Gauze in the Body on Postoperative Radiography: Phantom Study. Nihon Hoshasen Gijutsu Gakkai Zasshi. 2021;77(8):821-827. Japanese. doi: 10.6009/jjrt.2021_JSRT_77.8.821. PMID: 34421070.

国際会議発表

- T.Kobayashi, T.Nagao:Comparison of Feature Selection Methods for Bottom-up of 1p/19q Co-deletion Prediction Performance in Grade 2 and 3 Gliomas, 2022 5th International Conference on Digital Medicine and Image Processing (DMIP 2022), Kyoto, Japan, Nov. 10-Nov.13, 2022.
- T.Kobayashi, E.Takaya, et al: A Radiomics Approach for Differentiating between Benign and Malignant Breast Lesions on Breast Ultrasound Images, RSNA 2020 – 106th Annual Meeting, Scientific poster (Virtual), Chicago, US, Nov. 29- Dec. 05, 2020.
- T. Kobayashi, T. Haraguchi and T. Nagao, "Classifying the Presence or Absence of Calcifications on Mammography using the EfficientNet: A Feasibility Study," 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech), 2022, pp. 596-599, doi: 10.1109/LifeTech53646.2022.9754829.

国内学会発表

- 小林達明:“ラジオミクス特徴抽出ライブラリの開発”第78回日本放射線技術学会総会学術大会 パシフィコ横浜 神奈川, 2022.

付録 A 遺伝的アルゴリズムにおける選択個体カウントの実装例

Genetic algorithm for feature selection

```
# sklearn-genetic - Genetic feature selection module for scikit-learn
# Copyright (C) 2016-2022  Manuel Calzolari
# Copyright (C) 2022 Tatsuaki Kobayashi
#
# This program is free software: you can redistribute it and/or modify
# it under the terms of the GNU Lesser General Public License as published by
# the Free Software Foundation, version 3 of the License.
#
# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU Lesser General Public License for more details.
#
# You should have received a copy of the GNU Lesser General Public License
# along with this program.  If not, see <http://www.gnu.org/licenses/>.

"""Genetic algorithm for feature selection"""

import numbers
import multiprocess
import numpy as np
from sklearn.utils import check_X_y
from sklearn.utils.metaestimators import if_delegate_has_method
from sklearn.base import BaseEstimator
from sklearn.base import MetaEstimatorMixin
from sklearn.base import clone
from sklearn.base import is_classifier
from sklearn.model_selection import check_cv, cross_val_score
from sklearn.metrics import check_scoring
from sklearn.feature_selection import SelectorMixin
from sklearn.utils._joblib import cpu_count
from deap import algorithms
from deap import base
from deap import creator
from deap import tools
```

```
creator.create("Fitness", base.Fitness, weights=(1.0, -1.0, -1.0))
creator.create("Individual", list, fitness=creator.Fitness)


def _eaFunction(population, toolbox, cxpb, mutpb, ngen, ngen_no_change=None, stats=None,
                halloffame=None, verbose=0):
    logbook = tools.Logbook()
    logbook.header = ['gen', 'nevals'] + (stats.fields if stats else [])

    best_select = None

    # Evaluate the individuals with an invalid fitness
    invalid_ind = [ind for ind in population if not ind.fitness.valid]
    fitnesses = toolbox.map(toolbox.evaluate, invalid_ind)
    for ind, fit in zip(invalid_ind, fitnesses):
        ind.fitness.values = fit

    if halloffame is None:
        raise ValueError("The 'halloffame' parameter should not be None.")

    halloffame.update(population)
    hof_size = len(halloffame.items) if halloffame.items else 0

    record = stats.compile(population) if stats else {}
    logbook.record(gen=0, nevals=len(invalid_ind), **record)
    if verbose:
        print(logbook.stream)

    # Begin the generational process
    wait = 0
    for gen in range(1, ngen + 1):
        # Select the next generation individuals
        offspring = toolbox.select(population, len(population) - hof_size)

        # Vary the pool of individuals
        offspring = algorithms.varAnd(offspring, toolbox, cxpb, mutpb)

        # Evaluate the individuals with an invalid fitness
        invalid_ind = [ind for ind in offspring if not ind.fitness.valid]
        fitnesses = toolbox.map(toolbox.evaluate, invalid_ind)
        for ind, fit in zip(invalid_ind, fitnesses):
            ind.fitness.values = fit
```

```
        # Add the best back to population
        offspring.extend(halloffame.items)

        # Get the previous best individual before updating the hall of fame
        prev_best = halloffame[0]
        # print(prev_best)
        if best_select is None:
          best_select = prev_best
        else:
          best_select = np.vstack((best_select,prev_best))

        # Update the hall of fame with the generated individuals
        halloffame.update(offspring)

        # Replace the current population by the offspring
        population[:] = offspring

        # Append the current generation statistics to the logbook
        record = stats.compile(population) if stats else {}
        logbook.record(gen=gen, nevals=len(invalid_ind), **record)
        if verbose:
            print(logbook.stream)

        # If the new best individual is the same as the previous best individual,
        # increment a counter, otherwise reset the counter
        if halloffame[0] == prev_best:
            wait += 1
        else:
            wait = 0

        # If the counter reached the termination criteria, stop the optimization
        if ngen_no_change is not None and wait >= ngen_no_change:
            break

    return population, logbook, best_select


def _createIndividual(icls, n, max_features):
    n_features = np.random.randint(1, max_features + 1)
    genome = ([1] * n_features) + ([0] * (n - n_features))
    np.random.shuffle(genome)
```

```
    return icls(genome)


def _evalFunction(individual, estimator, X, y, groups, cv, scorer, fit_params, max_features,
                  caching, scores_cache={}):
    individual_sum = np.sum(individual, axis=0)
    if individual_sum == 0 or individual_sum > max_features:
        return -10000, individual_sum, 10000
    individual_tuple = tuple(individual)
    if caching and individual_tuple in scores_cache:
        return scores_cache[individual_tuple][0], individual_sum, scores_cache[individual_tuple][1]
    X_selected = X[:, np.array(individual, dtype=np.bool)]
    scores = cross_val_score(estimator=estimator, X=X_selected, y=y, groups=groups, scoring=score
r,
                             cv=cv, fit_params=fit_params)
    scores_mean = np.mean(scores)
    scores_std = np.std(scores)
    if caching:
        scores_cache[individual_tuple] = [scores_mean, scores_std]
    return scores_mean, individual_sum, scores_std


class GeneticSelectionCV(BaseEstimator, MetaEstimatorMixin, SelectorMixin):
    """Feature selection with genetic algorithm.
    Parameters
    ----------
    estimator : object
        A supervised learning estimator with a `fit` method.
    cv : int, cross-validation generator or an iterable, optional
        Determines the cross-validation splitting strategy.
        Possible inputs for cv are:
        - None, to use the default 3-fold cross-validation,
        - integer, to specify the number of folds.
        - An object to be used as a cross-validation generator.
        - An iterable yielding train/test splits.
        For integer/None inputs, if ``y`` is binary or multiclass,
        :class:`StratifiedKFold` used. If the estimator is a classifier
        or if ``y`` is neither binary nor multiclass, :class:`KFold` is used.
    scoring : string, callable or None, optional, default: None
        A string (see model evaluation documentation) or
        a scorer callable object / function with signature
        ``scorer(estimator, X, y)``.
```

fit_params : dict, optional
    Parameters to pass to the fit method.
max_features : int or None, optional
    The maximum number of features selected.
verbose : int, default=0
    Controls verbosity of output.
n_jobs : int, default 1
    Number of cores to run in parallel.
    Defaults to 1 core. If `n_jobs=-1`, then number of jobs is set
    to number of cores.
n_population : int, default=300
    Number of population for the genetic algorithm.
crossover_proba : float, default=0.5
    Probability of crossover for the genetic algorithm.
mutation_proba : float, default=0.2
    Probability of mutation for the genetic algorithm.
n_generations : int, default=40
    Number of generations for the genetic algorithm.
crossover_independent_proba : float, default=0.1
    Independent probability for each attribute to be exchanged, for the genetic algorithm.
mutation_independent_proba : float, default=0.05
    Independent probability for each attribute to be mutated, for the genetic algorithm.
tournament_size : int, default=3
    Tournament size for the genetic algorithm.
n_gen_no_change : int, default None
    If set to a number, it will terminate optimization when best individual is not
    changing in all of the previous ``n_gen_no_change`` number of generations.
caching : boolean, default=False
    If True, scores of the genetic algorithm are cached.
Attributes
----
n_features_ : int
    The number of selected features with cross-validation.
support_ : array of shape [n_features]
    The mask of selected features.
generation_scores_ : array of shape [n_generations]
    The maximum cross-validation score for each generation.
estimator_ : object
    The external estimator fit on the reduced dataset.
Examples

An example showing genetic feature selection.

```
>>> import numpy as np
>>> from sklearn import datasets, linear_model
>>> from genetic_selection import GeneticSelectionCV
>>> iris = datasets.load_iris()
>>> E = np.random.uniform(0, 0.1, size=(len(iris.data), 20))
>>> X = np.hstack((iris.data, E))
>>> y = iris.target
>>> estimator = linear_model.LogisticRegression(solver="liblinear", multi_class="ovr")
>>> selector = GeneticSelectionCV(estimator, cv=5)
>>> selector = selector.fit(X, y)
>>> selector.support_ # doctest: +NORMALIZE_WHITESPACE
array([ True  True  True  True False False False False False False False False
       False False False False False False False False False False False False], dtype=bool)
"""
def __init__(self, estimator, cv=None, scoring=None, fit_params=None, max_features=None,
             verbose=0, n_jobs=1, n_population=300, crossover_proba=0.5, mutation_proba=0.2,
             n_generations=40, crossover_independent_proba=0.1,
             mutation_independent_proba=0.05, tournament_size=3, n_gen_no_change=None,
             caching=False):
    self.estimator = estimator
    self.cv = cv
    self.scoring = scoring
    self.fit_params = fit_params
    self.max_features = max_features
    self.verbose = verbose
    self.n_jobs = n_jobs
    self.n_population = n_population
    self.crossover_proba = crossover_proba
    self.mutation_proba = mutation_proba
    self.n_generations = n_generations
    self.crossover_independent_proba = crossover_independent_proba
    self.mutation_independent_proba = mutation_independent_proba
    self.tournament_size = tournament_size
    self.n_gen_no_change = n_gen_no_change
    self.caching = caching
    self.scores_cache = {}
    self.importance = None

@property
def _estimator_type(self):
    return self.estimator._estimator_type
```

```
def fit(self, X, y, groups=None):
    """Fit the GeneticSelectionCV model and the underlying estimator on the selected features.
    Parameters
    ----------
    X : {array-like, sparse matrix}, shape = [n_samples, n_features]
        The training input samples.
    y : array-like, shape = [n_samples]
        The target values.
    groups : array-like, shape = [n_samples], optional
        Group labels for the samples used while splitting the dataset into
        train/test set. Only used in conjunction with a "Group" `cv`
        instance (e.g., `GroupKFold`).
    """
    return self._fit(X, y, groups)

def _fit(self, X, y, groups=None):
    X, y = check_X_y(X, y, "csr")
    # Initialization
    cv = check_cv(self.cv, y, classifier=is_classifier(self.estimator))
    scorer = check_scoring(self.estimator, scoring=self.scoring)
    n_features = X.shape[1]

    if self.max_features is not None:
        if not isinstance(self.max_features, numbers.Integral):
            raise TypeError("'max_features' should be an integer between 1 and {} features."
                            " Got {!r} instead."
                            .format(n_features, self.max_features))
        elif self.max_features < 1 or self.max_features > n_features:
            raise ValueError("'max_features' should be between 1 and {} features."
                             " Got {} instead."
                             .format(n_features, self.max_features))
        max_features = self.max_features
    else:
        max_features = n_features

    if not isinstance(self.n_gen_no_change, (numbers.Integral, np.integer, type(None))):
        raise ValueError("'n_gen_no_change' should either be None or an integer."
                         " {} was passed."
                         .format(self.n_gen_no_change))

    estimator = clone(self.estimator)
```

```
# Genetic Algorithm
toolbox = base.Toolbox()

toolbox.register("individual", _createIndividual, creator.Individual, n=n_features,
                 max_features=max_features)
toolbox.register("population", tools.initRepeat, list, toolbox.individual)
toolbox.register("evaluate", _evalFunction, estimator=estimator, X=X, y=y,
                 groups=groups, cv=cv, scorer=scorer, fit_params=self.fit_params,
                 max_features=max_features, caching=self.caching,
                 scores_cache=self.scores_cache)
toolbox.register("mate", tools.cxUniform, indpb=self.crossover_independent_proba)
toolbox.register("mutate", tools.mutFlipBit, indpb=self.mutation_independent_proba)
toolbox.register("select", tools.selTournament, tournsize=self.tournament_size)

if self.n_jobs == 0:
    raise ValueError("n_jobs == 0 has no meaning.")
elif self.n_jobs > 1:
    pool = multiprocess.Pool(processes=self.n_jobs)
    toolbox.register("map", pool.map)
elif self.n_jobs < 0:
    pool = multiprocess.Pool(processes=max(cpu_count() + 1 + self.n_jobs, 1))
    toolbox.register("map", pool.map)

pop = toolbox.population(n=self.n_population)
hof = tools.HallOfFame(1, similar=np.array_equal)
stats = tools.Statistics(lambda ind: ind.fitness.values)
stats.register("avg", np.mean, axis=0)
stats.register("std", np.std, axis=0)
stats.register("min", np.min, axis=0)
stats.register("max", np.max, axis=0)

if self.verbose > 0:
    print("Selecting features with genetic algorithm.")

with np.printoptions(precision=6, suppress=True, sign=" "):
    _, log, best_select = _eaFunction(pop, toolbox, cxpb=self.crossover_proba,
                          mutpb=self.mutation_proba, ngen=self.n_generations,
                          ngen_no_change=self.n_gen_no_change,
                          stats=stats, halloffame=hof, verbose=self.verbose)
if self.n_jobs != 1:
    pool.close()
    pool.join()
```

```
    # Set final attributes
    support_ = np.array(hof, dtype=np.bool)[0]
    self.estimator_ = clone(self.estimator)
    self.estimator_.fit(X[:, support_], y)

    self.generation_scores_ = np.array([score for score, _, _ in log.select("max")])
    self.n_features_ = support_.sum()
    self.support_ = support_

    self.importance = np.sum(best_select, axis=0)

    return self

@if_delegate_has_method(delegate='estimator')
def predict(self, X):
    """Reduce X to the selected features and then predict using the underlying estimator.
    Parameters
    ----------
    X : array of shape [n_samples, n_features]
        The input samples.
    Returns
    -------
    y : array of shape [n_samples]
        The predicted target values.
    """
    return self.estimator_.predict(self.transform(X))

@if_delegate_has_method(delegate='estimator')
def score(self, X, y):
    """Reduce X to the selected features and return the score of the underlying estimator.
    Parameters
    ----------
    X : array of shape [n_samples, n_features]
        The input samples.
    y : array of shape [n_samples]
        The target values.
    """
    return self.estimator_.score(self.transform(X), y)

def _get_support_mask(self):
    return self.support_
```

```
    @if_delegate_has_method(delegate='estimator')
    def decision_function(self, X):
        return self.estimator_.decision_function(self.transform(X))

    @if_delegate_has_method(delegate='estimator')
    def predict_proba(self, X):
        return self.estimator_.predict_proba(self.transform(X))

    @if_delegate_has_method(delegate='estimator')
    def predict_log_proba(self, X):
        return self.estimator_.predict_log_proba(self.transform(X))
```

Usage: 重要度として選択個体カウントを取得する

```
def calculate_ga_selection():
    X_train, y_train, group = load_dataset()
    cv = StratifiedGroupKFold(n_splits=3, shuffle=True, random_state=76)
    estimator = reference_classifier()
    selector = GeneticSelectionCV(estimator,
                                  cv=cv,
                                  verbose=1,
                                  scoring='roc_auc', #scoring="accuracy",
                                  max_features= 30,
                                  n_population=50,
                                  crossover_proba=0.5,
                                  mutation_proba=0.2,
                                  n_generations=30,
                                  crossover_independent_proba=0.5,
                                  mutation_independent_proba=0.05,
                                  tournament_size=3,
                                  n_gen_no_change=10,
                                  caching=True,
                                  n_jobs=-1)# n_jobs = -1 default
    selector = selector.fit(X_train, y_train, group)
    features = X_train.columns.values[selector.support_]
    importances = selector.importance[selector.support_]
```

Usage: 選択個体カウント順に変数増加法にて最も精度の高くなるモデルを評価する

```
def evaluate_with_importances(feature_names, importances):
    # feature_names : array-like
    # importances : Importances with feature name order. array-like
    imp_df[‘f’] = feature_names
```

```
imp_df['i'] = importances
imp_sort = imp_df.sort_values(by='i', ascending=False, axis=0)
imp_sort = imp_sort.dropna(how='any')
sorted_cols = imp_sort['f'].values # to numpy
print("check sorted importance", sorted_cols[:5])

# Searching for a feature combination that yields highest performance.
num_of_feature = []
print('start evaluation from 1 to '+str(num_max_feature))
accs = []
aucs = []
senss = []
specs = []
TNs = []
FPs = []
FNs = []
TPs = []
aics = []
bics = []
log_losses = []
train_aucs = []
train_aics = []
train_bics = []
train_log_losses = []

# binary classification example
X_train, y_train, X_test, y_test = load_learning_dataset()

for i, nof in enumerate(num_of_feature):
    X_train_selected = X_train[sorted_cols[:nof]]
    X_test_selected = X_test[sorted_cols[:nof]]
    clf = eval_classifier() # To simplify, skip grid search.
    clf.fit(X_train_selected, y_train)

    # train auc
    train_proba_ = clf.predict_proba(X_train_selected)[:,1]
    fpr_train, tpr_train, _ = metrics.roc_curve(y_train, train_proba_, pos_label=1)
    train_auc_ = metrics.auc(fpr_train, tpr_train)
    train_aic, train_bic, train_log_loss = calc_aic_bic_logloss(y_train, train_proba_,k=nof)

    # test metrics
    proba_ = clf.predict_proba(X_test_selected)[:,1]
```

```
    fpr_, tpr_, th_ = metrics.roc_curve(y_test, proba_, pos_label=1)
    auc_ = metrics.auc(fpr_, tpr_)
    Acc = clf.score(X_test_selected, y_test)
    pred_ = clf.predict(X_test_selected)
    tn_, fp_, fn_, tp_ = confusion_matrix(y_test, pred_).ravel()
    aic, bic, log_loss_ = calc_aic_bic_logloss(y_test, proba_,k=nof)
    print("~" * 50)
    print("num of feature", nof)
    print('acc', Acc, 'auc', auc_, 'sens', tp_/(tp_+fn_),'speci',tn_/(tn_+fp_))
    print('tn_', tn_, 'fp_',fp_, 'fn_', fn_, 'tp_',tp_ )
    print('aic', aic, 'bic',bic, 'log_loss_', log_loss_)
    accs.append(Acc)
    aucs.append(auc_)
    senss.append(tp_/(tp_+fn_))
    specs.append(tn_/(tn_+fp_))
    TNs.append(tn_)
    FPs.append(fp_)
    FNs.append(fn_)
    TPs.append(tp_)
    aics.append(aic)
    bics.append(bic)
    log_losses.append(log_loss_)
    train_aucs.append(train_auc_)
    train_bics.append(train_bic)
    train_aics.append(train_aic)
    train_log_losses.append(train_log_loss)
evals = {
    "num of features":num_of_feature,
    "test_acc":accs,
    "test_auc":aucs,
    "test_senss":senss,
    "test_specs":specs,
    "test_TN":TNs,
    "test_FP":FPs,
    "test_FN":FNs,
    "test_TP":TPs,
    "test_aic":aics,
    "test_bic":bics,
    "test_log_loss":log_losses,
    "train_auc":train_aucs,
    "train_aic":train_aics,
    "train_bic":train_bics,
```

```
    "train_log_loss":train_log_losses
}
eval_df = pd.DataFrame(evals) # This is result
```