人工知能で切り開く植物科学の近未来

戸田陽介 [1,2]
[1]JST さきがけ
〒332-0012 埼玉県川口市本町４丁目１−８
[2]名古屋大学トランスフォーマティブ生命分子研究所
〒464-8602 愛知県名古屋市千種区不老町

Paving the Future of Plant Science with Artificial Intelligence

Yosuke Toda[1,2]
[1]Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan
[2]Institute of Transformative Bio-Molecules (WPI-ITbM),
Nagoya University, Chikusa-ku, Nagoya 464-8602, Japan
Key words: Deep Learning, Machine Learning, Plant Phenotyping, Plant Science
DOI: 10.24480/bsj-review.11c1.00190

近年の著しいハードウェアの性能上昇と低廉化，さらには機械学習におけるゲームチェンジングテクノロジーとも評される深層学習（ディープラーニング）の実用化によって，従来技術では想像もできなかった複雑なアルゴリズムをコンピュータに実装し，運用するハードルが下がった。最近では「AI を活用した」,「人工知能による」といったフレーズで，産学問わず様々なシチュエーションで我々の社会に顕在化している。当該技術を駆使することによって，我々の研究が大きく加速することには間違いない。しかしながら最近では，言葉だけが独り歩きし「人工知能を使えば何でもできる」,「人工知能が取って代わる」といった極端な考えが散見されるようになってきた。近年，植物科学・農学分野への情報科学の新技術の急速な流入が起きており，多様な研究が芽生え始めている，まさに黎明期である。人工知能とは何なのか, 現状技術でどのようなことが可能になるのか, 当技術を真に有効活用するため, 当該分野で情報を共有する必要があると考えていた。

そのような考えを前提とし，筆者は日本植物学会第 83 回大会において「人工知能で切り開く植物科学の近未来」と題する理事会主催シンポジウムを企画した。植物科学・農学分野において「機械学習」,「画像解析」,「特徴量学習」などが関連する研究テーマに携わり，著しい成果を挙げている方々にお声がけをし，参加頂いた。発表者には，自身の研究成果の発表に限定せず，関連分野を俯瞰した総説的な内容となるよう依頼した。さらには，総合討論としてパネルディスカッションを設けるなど，聴講者と対話的な形式とした。当日は予想を遥かに超える多くの聴衆が参加し，総合討論の時間が足りなくなるほど質疑が絶えることなく，大盛況な会として終了した。本会では複数のシンポジウムが時間的に重複しており，魅力的な演題もたくさんある中，敢えてこの挑戦的な内容となる本企画に足を運んでくださった参

加者の皆様に感謝申し上げたい。素晴らしい内容を講演してくださった演者の方々，本会の開催のきっかけとなった伊藤正樹先生，並びに本会のサポートをしていただいた植物学会の関係者の皆様にも併せて感謝申し上げる。

植物科学の「人工知能」との関わり方を考える

大倉史生 [1,2], 水谷未耶 [3], 野下浩司 [1,4], 戸田陽介 [1,5]
[1]JST さきがけ　〒332-0012 埼玉県川口市本町 4 丁目 1-8
[2]大阪大学情報科学研究科　〒565-0871 大阪府吹田市山田丘 1-5
[3]名古屋大学理学研究科　〒464-8601 愛知県名古屋市千種区不老町
[4]九州大学理学研究院生物科学部門　〒819-0395 福岡県 福岡市西区 元岡 774 番地
[5]名古屋大学トランスフォーマティブ生命分子研究所
〒464-8602 愛知県名古屋市千種区不老町

Past and Future of Plant Science with Artificial Intelligence

Fumio Okura[1,2], Miya Mizutani[3], Koji Noshita[1,4], Yosuke Toda[1,5]
[1]Japan Science and Technology Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan
[2]Graduate School of Information Science and Technology, Osaka University,
1-5 Yamadaoka, Suita, Osaka, 565-0871, Japan
[3]Division of Biological Science, Graduate School of Science, Nagoya University,
Chikusa-ku Nagoya 464-8602, Japan
[4]Department of Biology, Faculty of Science, Kyushu University,
744 Motooka, Nishi-ku, Fukuoka 819-0396, Japan
[5]Institute of Transformative Bio-Molecules (WPI-ITbM), Nagoya University,
Chikusa-ku, Nagoya 464-8602, Japan
Keywords: deep learning, machine learning, plant phenotyping
DOI: 10.24480/bsj-review.11c2.00191

１．「人工知能（Artificial intelligence）」

１－１．人工知能分野の歴史と主な技術

人工知能（artificial intelligence; AI) という言葉の初出は，ジョン・マッカーシー，マービン・ミンスキーらにより開催された 1956 年のダートマス会議 (McCarthy et al. 2006) に遡る。とはいえ，人工知能関連の技術開発や議論は，それ以前から様々な分野で行われてきた。例えば近年様々な分野において活用が進むニューラルネットワークの端緒である形式ニューロン (McCulloch and Pitts 1943) は，1940 年代から神経科学の分野で研究されてきたし，「知能を持つ機械」の判定を

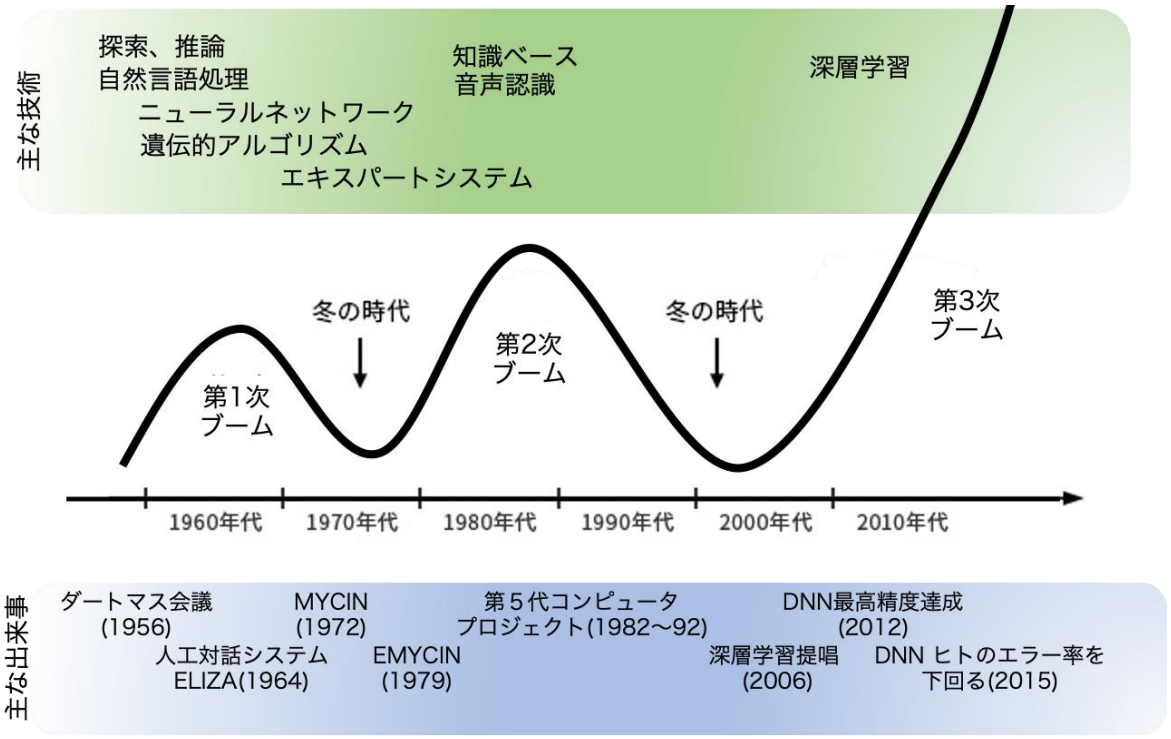

図 1 人工知能の歴史と技術 （平成 26 年版、28 年版 情報通信白書を改変）

行うためのチューリングテスト (Turing 1950) は 1950 年に提案されている[1]。以来，図 1 に示すように，人工知能分野で生み出された技術は，人々の過度な期待（あるいはビジネスへの有用性）と技術的困難の狭間で，社会における流行（AI ブーム）と終焉（AI 冬の時代）を繰り返してきた。植物科学における活用例を紹介する前に，人工知能分野に関連する主要な歴史や技術を，背景知識としてごく簡単に紹介する。

第一次 AI ブーム：ダートマス会議以降，1960 年代前後に盛んに（多額の資金を投入して）行われた研究群を「第一次 AI ブーム」と呼ぶことがある。数学の定理証明，画像認識，言語理解を含めた様々なタスクに対する研究が行われ，人間と同等の知能を備えた機械の実現も近い，などと楽観視されていた側面もある (Crevier 1993)。しかし，この時点では多くの問題において非常に小規模な例を示せたに過ぎず，実世界の問題は，当時の研究者が考えるほど簡単ではなかった[2]。

エキスパートシステム（第二次 AI ブーム）：1980 年代を中心に流行したエキスパートシステム (Buchanan and Feigenbaum 1981) は「第二次 AI ブーム」の立役者の一つであり，事前に用意した特定の専門分野に関する知識ベース（もし A ならば B である，などの知識の集合）を用いて推論

[1] チューリングテストは非常に有名である一方，反論もある。例えばサールの「中国語の部屋」。

[2] 問題を過小評価してしまうのは，どの分野でも研究あるあるだと思う。

を行い，専門家の代わりに問題を解決しようとするものである。知識ベースを準備することは非常に大変な作業であるが，タスクを絞ることにより，産業界においても様々な分野で実用的なシステムが開発された。ルールに基づく推論という点で，ルールベースのシステム（rule-based system）などと呼ばれることがある。

機械学習：一方，事前に用意された明示的なルールの集合からの推論とは異なり，データの集合を入力し，事前に設定した評価尺度を改善するように，モデルを近似させる枠組みを機械学習と呼ぶ (Mackay 2003) 。カテゴリ分類などに広く使われる「教師あり学習」と呼ばれる枠組みにおいては，事前に入力データ（例：画像，あるいは画像から計算された特徴）とともに，対応する正答データ（例：撮影された物体の種類や位置）を用意する。これらを「教師データ」と呼び，モデルは，入力データから正答データに近い出力を導き出すように学習される。機械学習の詳細な分類（教師あり学習，教師なし学習，強化学習）についての言及は避けるが，植物科学におけるデータ解析でも馴染みの深い回帰分析[3]（線形，Lasso，ロジスティック，など）のほか，クラス分類（この物体は「猫」である，など），類似したデータごとにグループ化するクラスタリング（k 平均法，など）などのタスクは，機械学習の一種といえる。最適化の対象となるモデルの構造や最適化手法は数多く提案されており，深層学習を含むニューラルネットワークはその一例である。その他，サポートベクターマシン（SVM） (Suykens and Vandewalle 1999)，ランダムフォレスト (Breiman 2001)，勾配ブースティング（GBM）による決定木学習 (Chen and Guestrin 2016; Ke et al. 2017) [4]などの手法は，深層学習ブームの現在も，学習データの少ない場面などで日常的に使われている。

深層学習（第三次 AI ブーム）：現在，我々は新たな，かつてない規模の「AI」ブームの只中にいる。現在の AI ブームは，機械学習，その中でもニューラルネットワークの一種である深層学習の実用化とともに進んでいる。1980〜2000 年代にわたる地道な研究の結果，いくつかの技術的ブレイクスルーと計算資源の飛躍的進歩を背景に，大規模なニューラルネットワークの最適化が可能となった (Hinton and Salakhutdinov 2006) ことで，多層のニューラルネットワークを用いた機械学習（＝深層学習）が実用化された[5]。2012 年，畳み込み演算を行うフィルタの学習を伴う畳み

[3] 最も単純な例の一つとして，最小二乗法による線形回帰は，データ集合から評価尺度（二乗誤差）を最小にする近似モデル（線形回帰モデル）を求める点で，機械学習の一種であると呼ぶこともできなくはない。

[4] 近年は，GBM の一つである XGBoost (Chen and Guestrin 2016) や LightGBM (Ke et al. 2017) が，比較的小規模の分類・回帰においてファーストチョイスとされることが多いのではないかと思う。

[5] ニューラルネットワークの仕組みについては，ここでは解説しない。デモを含む直感的な解説が https://playground.tensorflow.org/にある。また，本稿の最後で，学習のためのリソースをいくつか紹介する。

込みニューラルネットワーク（CNN） (Fukushima and Miyake 1982; LeCun et al. 1989)[6] の一種である AlexNet (Krizhevsky et al. 2017) が従来の機械学習による手法と比較して画像分類精度を圧倒的に改善し，現在の AI ブームの幕開けを告げた。SVM など従来の機械学習手法の適用時には，多くの場合，画像からの特徴抽出（エッジ検出，次元削減など）が事前に必要であり，その際にある種の専門知識が必要であった。しかし，深層学習は，他の機械学習手法と比較し次元数の大きな入力データ・非常にたくさんのパラメータを含むモデル（例えば，AlexNet の最適化対象のパラメータは 6000 万個以上）を比較的うまく最適化できる。そのため，CNN を用いた画像分類などのタスクにおいては，特徴抽出を省略してネットワークに画像を直接入力し，特徴抽出のための畳み込みフィルタを自動学習させるような使い方ができる。あわせて，深層学習の実装がライブラリ化され，容易に使用できるようになり，参入障壁が低くなっている（今では，高性能のマシン・環境構築を必要としないクラウド実行環境もある[7]）。機械学習や特徴抽出の専門性を必要としない深層学習の特性は，ブームの拡大に大きく寄与しており，深層学習を活用した囲碁の世界チャンピオンへの勝利や，コンテンツ生成，実社会への数多くの応用例など，その社会的インパクトは広く知られるとおりである。

１－２．結局，人工知能ってなに

「人工知能とは何か」，その明確な定義は非常に困難である。一般論として「知的である」と思われるようなタスクを機械が遂行する場合に「人工知能」と呼ばれることが多い。その点で，初期の人工知能分野においては，非常に簡単な問題が扱われ「人工知能」と呼ばれてきた。しかし，コンピュータの進歩により過去に扱われた問題が一般化することにより，それらの問題は「人工知能」であると呼ばれづらくなる （AI 効果）。「知能」あるいは「知的である」ことの定義自体が困難であり，しかも時代とともに変化するのである[8]。

近年流行を見せる深層学習で用いられる深層ニューラルネットワークはニューラルネットワークの 1 カテゴリであり，ニューラルネットワークは機械学習を実現する際に用いるモデルの一つである。さらに，機械学習は人工知能分野で扱われる技術の 1 カテゴリである。たとえば，エキスパートシステムなどのルールベースの手法は人工知能分野で生み出された技術であるが，機械学習とは呼ばれない。「人工知能」という語の定義の難しさや曖昧さは，ビジネス面で（ある意味で）有用であるかもしれないが，ことアカデミックな文脈では，具体的な手法群（こういう特

[6] CNN の初出は Lecun らの LeNet (LeCun et al. 1989) ではなく，福嶋らによるネオコグニトロン (Fukushima and Miyake 1982) だ，と人工知能分野の研究者らはしばしば強調する。

[7] Google Colaboratory などのサービスが無料で利用可能である。本稿を読んで興味を持った読者は，インターネットさえあれば今すぐ活用することができる。https://colab.research.google.com/

[8] このあたりの議論は，人工知能学会の「教養知識としての AI」を参照されると良い https://www.ai-gakkai.or.jp/comic_no1/

徴量設計とこういうラベルで GBM を使って分類するように学習した，など）を挙げて議論されるべきであろう。よって，以降，これまで人工知能関連の分野に関わる技術や哲学の総称として人工知能（あるいは AI）という語を用いるが，個々の議論においてはできる限り具体的な手法を挙げるようにする。

有効活用するにせよ，あえて使わないにせよ，多くの分野で深層学習をはじめとした人工知能関連の技術と関わりが避けられなくなっている。その意味で，人工知能が社会や植物科学に何をもたらしたのか（また何をもたらそうとしているのか），その到達点を知っておくことは重要であろう。以降，本稿では特に画像解析や生命・数理科学・植物科学に関連する話題を通して「AI」の応用可能性の一端を明らかにするとともに，今後，植物科学分野が人工知能関連技術にどのように関わるべきかを議論する。

２．生命科学分野における数理モデル研究と人工知能

「人工知能とは機械を使って知的に推論しようとする試みである」 (Cartwright 1993) とも言われ，その推論の方法開発は数理モデル研究と不可分であった。また，知能という生物学的な情報処理や抽象化能力を人工的に模倣しようという人工知能研究は生命科学研究と関わりが深い。現在盛んに利用されている深層ニューラルネットワークもその名の通り神経科学的背景を持ち，特に最初期の研究は，膜電位のダイナミクスを記述する Hodgkin-Huxley 方程式 (Hodgkin and Huxley 1952) や FitzHugh-Nagumo 方程式 (Fitzhugh 1961; Nagumo et al. 1962) などに代表されるように神経細胞や神経回路の数理モデル的側面が強かった (McCulloch and Pitts 1943)。このように，人工知能は数理モデルと生命科学と深く結びつきながら発展してきており，生命科学分野でも様々な人工知能が活用されてきた。第 2 章では，第 1 章で述べた歴史を踏まえ，数理モデルと生命科学に関わりの深い人工知能について，「エキスパートシステム」，「遺伝的アルゴリズム」，「ニューラルネットワーク」の技術とともに，その応用例も含んで簡単に紹介する。

２－１．エキスパートシステム

前述の第二次 AI ブームの立役者となったエキスパートシステムは，1970 年代に開発され，1980 年代には商用化も行われた。専門家の知識やノウハウを人間が「もし○○ならば△△である」という規則でルールとして記述し，そのルールに従ってコンピュータに処理させることで問題の解決を目指すものである。特定の領域に限れば，エキスパートシステムは実用で成果を上げた初の人工知能技術といえる。科学の分野においても，装置の運転制御，分析手法の開発，分子モデリング，画像処理，ロボットによるサンプリング，診断システムなどで広く用いられ，一定の成果をあげている。医療分野では 1970 年代にスタンフォード大学で開発された細菌感染症診断を行うエキスパートシステムの“MYCIN”が有名な例として挙げられる (Shortliffe and Buchanan

1975)。MYCIN は，LISP 言語で記述されたかなり単純な推論（後向きの推論）エンジンと，約 600 のルールからなる知識ベースによって構築され，単純な質問を介して操作する。診断結果として，感染の可能性のある細菌のリストを確率とともに提示し，適した投薬を提案するシステムである。スタンフォード医科大学で行われた研究では，MYCIN は 65%の適切な結果を提示できた。5 人の医師による診断の適切性が 42.5%から 62.5%であったことからも，その成果は人間の専門技術に匹敵しうるとされた (Yu et al. 1979)。しかし，人工知能の誤診に対する責任を誰が負うのかという医療倫理に関する問題や，記述したルール以外の学習はできないという技術的な問題により，実用化には至らなかった。しかし，MYCIN の枠組みは，数年後に実装された E-MYCIN（非医療）がエキスパートシステムを用いる多くのアプリケーションに活用されることで第二次 AI ブームの火付け役になった。このように，知識を入力すれば入力するほど性能が向上するという長所を持ち，人工知能として素晴らしい成果を挙げたエキスパートシステムであるが，①知識を入力する専門家が不可欠，②専門家の持つ知識をすべて入力することが困難（例外や矛盾の存在），③抽象的な記述や主観的な記述を入力することが困難，という短所が存在する。これらの短所により，エキスパートシステムによる人工知能は人間の専門家を超えられない，と捉えられ，第二次 AI ブームは終息に向かっていった。

２－２．遺伝的アルゴリズム

生命科学研究者にとって，遺伝的アルゴリズム（Genetic algorithm, GA）はなじみ深い人工知能技術の一つではなかろうか。GA は進化的アルゴリズムの一つであり，生物の進化，特に自然選択，のプロセスを模倣し近似解を探索する (Holland 1975)。GA では，ある問題の近似解は数字や文字の配列で表現され，特定の配列（遺伝子型）を持つ個体を多数用意し，解の候補（集団 population）とする。近似解の「良さ」は適応度（fitness）として計算され，集団から適応度の高いものが優先的に選ばれ（選択 selection），配列の一部を個体間で交換する組み換え（crossover）や配列の一部をランダムに変化させる突然変異（mutation）を経て，次世代集団を作成する。この世代交代を繰り返し，より適応度の高い配列を得ることで近似解が探索できる。解を表現する配列，適応度関数，集団サイズ，選択の方法，交差の種類，突然変異率など様々な設定を検討する必要はあるものの，メタヒューリスティックな方法として多様な分野で活用されている。産業界では，ガスパイプライン制御，経路最適化，ロボット制御，プログラム自動生成，工場稼動計画などで広く用いられている。中でも N700 系新幹線の先頭形状設計おいて，約 5000 パターンのコンピューターシミュレーションの結果，最高速度を維持しつつ乗車人数を最大化する「最適解」を導き出した例として広く知られている。生命科学分野でも，バイオインフォマティクスでの遺伝情報解析やタンパク質の構造決定，物質合成経路の最適化など様々な応用例がある。GA は可能性のある解を並行処理することによって解に至るため，解に至る経路が毎回同じとは限らない。

しかし，何世代ものシミュレーションを繰り返すことで最適解に至るため，ノイズの多いデータやピークが複数ある複雑なデータでの最適解の抽出に対し，特に優れている。その「試行錯誤する」という性質ゆえに，本稿では取り上げないが，GA によって「人工生命」を創出する試みもなされてきた。GA 法は，従来法では解を得ることが難しいほど計算スケールが大きい場合に特に有用であるため，発表当初はコンピュータの処理速度の問題から，アルゴリズムの性能が疑問視されていたが，1990 年代の劇的なコンピュータの性能向上を受けて，再評価された。アルゴリズム上の「集団の大きさ」「適応度」などの変数の決定や方式の決定（確率的余り方式やエリート方式）はユーザーが行うので，生命科学研究においても実験結果などが重要な要素となるモデルといえる。

２－３．ニューラルネットワーク

第 1 章で述べたように，ニューラルネットワーク（Neural Networks, NN）が第 3 次 AI ブームを牽引している。前述のようにニューラルネットワークは生物学の数理モデルとして少なくとも 1930 年代から研究が行われ，ノーベル生理学賞を受賞した Hodgkin-Huxley 方程式は，ニューラルネットワークが電気回路によって表現できるとした画期的なモデルであった (Hodgkin and Huxley 1952)。人工知能分野の研究で用いられるニューラルネットワークは，神経細胞を「0」か「1」の値をとるものとしてモデル化し，それをネットワークにしたものである。生物学での「ニューラルネットワーク」と区別し，「人工ニューラルネットワーク」と称される場合があるが，本稿では多くの場合に倣い，単に「ニューラルネットワーク」と呼ぶ。

ニューラルネットワークは，機械が初めて「知識を学習できる」ようになった最初期のモデルの一つでもある。形式ニューロン (McCulloch and Pitts 1943) を応用し，学習可能としたニューラルネットワークのモデルを（単純）パーセプトロン (Rosenblatt 1958) と呼ぶ。単純な構造でありながら学習によってパターン認識を行うことから，「人工知能」の実現を期待させた。しかし，パーセプトロンには①線形分離可能なデータにしか用いることができない，②特徴を人間が（ある程度）教えなければならない，③精度を高めるためには膨大な数のデータを学習する必要がある，という問題があった (松田 2017)。この問題の指摘により一時は下火になったが，1986 年，Rumelhart らによってニューロン間の結合が多層化されたことにより非線形のデータも扱えるようになり①の問題は克服された (Rumelhart et al. 1986)。これにより飛躍的に「人工知能」の研究が進むと期待されたが，当時のコンピュータの計算速度では膨大なデータ量を多層のネットワークに学習させることが困難であった。1990 年代以降にコンピュータの計算能力が劇的に向上したことで問題③は克服され，2006 年に Hinton らによってさらに多層でありながら精度が落ちない深層ニューラルネットワークが提唱され (Hinton and Salakhutdinov 2006)，自動で特徴量抽出が可能になり②の問題も解決された。その結果，近年ニューラルネットワークは再び注目を集めている。

2012 年は，特にニューラルネットワークを用いた技術が世界に衝撃を与えた年である。創薬

分野においては，Merck 社が行った Merck Molecular Activity Challenge においてディープラーニングを用いた手法が最高精度を達成した (Ma et al. 2014)。このチームが化合物の活性予測についての専門家でなかったことも，当時注目を集めた。また，画像認識コンペティション ImageNet Large Scale Visual Recognition Competition (ILSVRC) において，AlexNet と呼ばれる CNN を用いた手法が，それまでの画像認識のデファクトスタンダードであった SIFT + Fisher Vector + SVM というアプローチ (Sanchez and Perronnin 2011) に大差をつけて優勝（前年のエラー率 26%に対し 16%と劇的に性能向上）した。さらに 2015 年には Google や Microsoft が人のエラー率である 5%を下回る手法を提案し (He et al. 2016)，「人工知能が人間を超えた」と騒がれた。画像認識において，2012 年のもう一つの有名な出来事といえば「Google の猫」であろう。Google の「Google X Labs」は，YouTube にアップロードされている動画から，ランダムに取り出した 200×200 画素の画像を 1000 万枚用意し，9 つの層からなるネットワークを用い，1000 台のコンピュータで 3 日間かけて学習を行った。その結果，特徴量を人間が教えることなく，猫の特徴に反応するニューロンが自動的に作られた。つまり，コンピュータは猫がどういうものであるか人間に教えられること無く，自力で特徴を抽出することに成功したといえる (Le et al. 2012)。「self-taught learning（自己教示学習）」による高精度の画像認識が可能になったことで，DNN を活用した画像認識は，現在，様々な分野で盛んに活用されている。

植物科学・農学においても，第二次 AI ブーム以降からニューラルネットワークが活用されてきた。藻類における種の自動認識 (Balfoort et al. 1992)，ニューラルネットワークとエキスパートシステムを組み合わせたモルト用大麦の最適な施肥量の提案 (Broner and Comstock 1997)，Hopfield モデルや前述の多層パーセプトロンによる入力画像からの植物形状の特徴量抽出やクラス分類 (Oide and Ninomiya 1998; Oide and Ninomiya 2000) など多岐にわたる。第三次ブームの中にある現在も，他分野の例にもれず植物科学・農学分野でも，ニューラルネットワークを用いた人工知能は注目を集めている。しかし，本章で述べたように，人工知能にはニューラルネットワーク以外にもそれぞれ特徴を持った多彩なアルゴリズムがある。次章以降では，特に画像定量技術に焦点をあて，植物科学・農学での具体的な応用例（ケーススタディ）を紹介しつつ，人工知能とどう付き合っていくべきかを議論する。

３．植物画像解析技術としての「人工知能」の活用例

３－１．画像解析・植物フェノタイピングと「人工知能」

情報には様々な形態があるが，画像は特に多くの情報を含む形態の一つである。画像に含まれる情報を解析し，必要なものを抽出するという作業は，現象の理解において有益である。そのため，生命科学においても，多くの分野で画像解析技術がこぞって活用されている。生物の画像か

ら取り出される情報の多くは，その生物に表れる性質である「フェノタイプ（表現型）」であり，これを正確に定量し，解析することが生命現象の理解には必須である。植物科学・農学の分野において「フェノタイプ」を正確に把握することは，特に育種や栽培管理などの観点から重要である。

画像に何らかの処理を施し，目的とする表現型を定量的または定性的に抽出する作業は一般的に画像解析と呼ばれるが，植物科学・農学分野においては，（画像を用いるにせよ用いないにせよ）植物の表現型を計測することを「植物フェノタイピング」と称している。画像を用いたフェノタイピング作業は，撮影条件，ノイズ，陰影など様々な要因によって定量が困難を極める場合があり，多くの機器や人的リソースを投じる必要があるため，かつては個人研究者では実現が難しいように思われてきた。しかし，前章までに紹介したように，機器やコンピュータ性能の向上や，画像解析ツール，深層学習を含む機械学習ライブラリの充実により，画像を用いて複雑なフェノタイプを自動・半自動で抽出するような取り組みが広く一般的に行われるようになった。近年は草丈や植物構造などの形状解析 (Minervini et al. 2016; Watanabe et al. 2017; Isokane et al. 2018)，果実や気孔などの器官検出 (Yamamoto et al. 2014; Toda et al. 2018)，ストレス・疾病検出 (Singh et al. 2016; Ghosal et al. 2018; Mohanty et al. 2016) などをはじめとした広い分野で画像解析が用いられており，「画像解析」や「フェノタイピング」といった言葉に，「人工知能」，「AI」といった単語がまるで枕詞のように使われるほど，当該分野に浸透している。しかし，人工知能関連技術の活用の幅が急速に広まるにつれ，各技術の仕組みを正しく理解しないことによる不適切なアルゴリズム設計や，実験条件などにおける問題がしばしば見受けられる。

本節では，人工知能（特に機械学習）の活用による画像解析を用いて「植物フェノタイピング」を行う際，研究者がどのようなアルゴリズムを設計すべきかを，植物の葉面積定量をケーススタディとして議論する。ここでは，特に代表的な画像解析手法を紹介・比較しつつ，いわゆる「人工知能」システムに必要なアルゴリズム設計過程（図 2）を解説しながら紹介したい。

３－２．ケーススタディ：画像解析による葉面積定量

定量性の必要なフェノタイピングにおいては，ハイスループットで再現性の高いシステムの構築が求められる。植物においても，機器や撮影技術の進歩に伴って大量の植物を自動で育成し，画像取得により植物生長の経時変化を追うためのシステムを構築する試みがなされている。理化学研究所で開発された全自動植物表現型解析システム RIPPS (Fujita et al. 2018) もその一つである。この RIPPS で被子植物シロイヌナズナを上面から撮影した画像を図 2A に示す。これを入力画像として葉面積を測定する場合，どのような方法が考えられるであろうか。既に述べたように，大量に取得された画像は多くの場合，陰影，ノイズ，微妙な撮影条件の差異などを含み，この中から如何にして必要な情報を定量的に取り出すか，が重要となってくる。以下ではこの点を主眼に置き，代表的な（かつ比較的シンプルな）画像解析の流れを解説する。

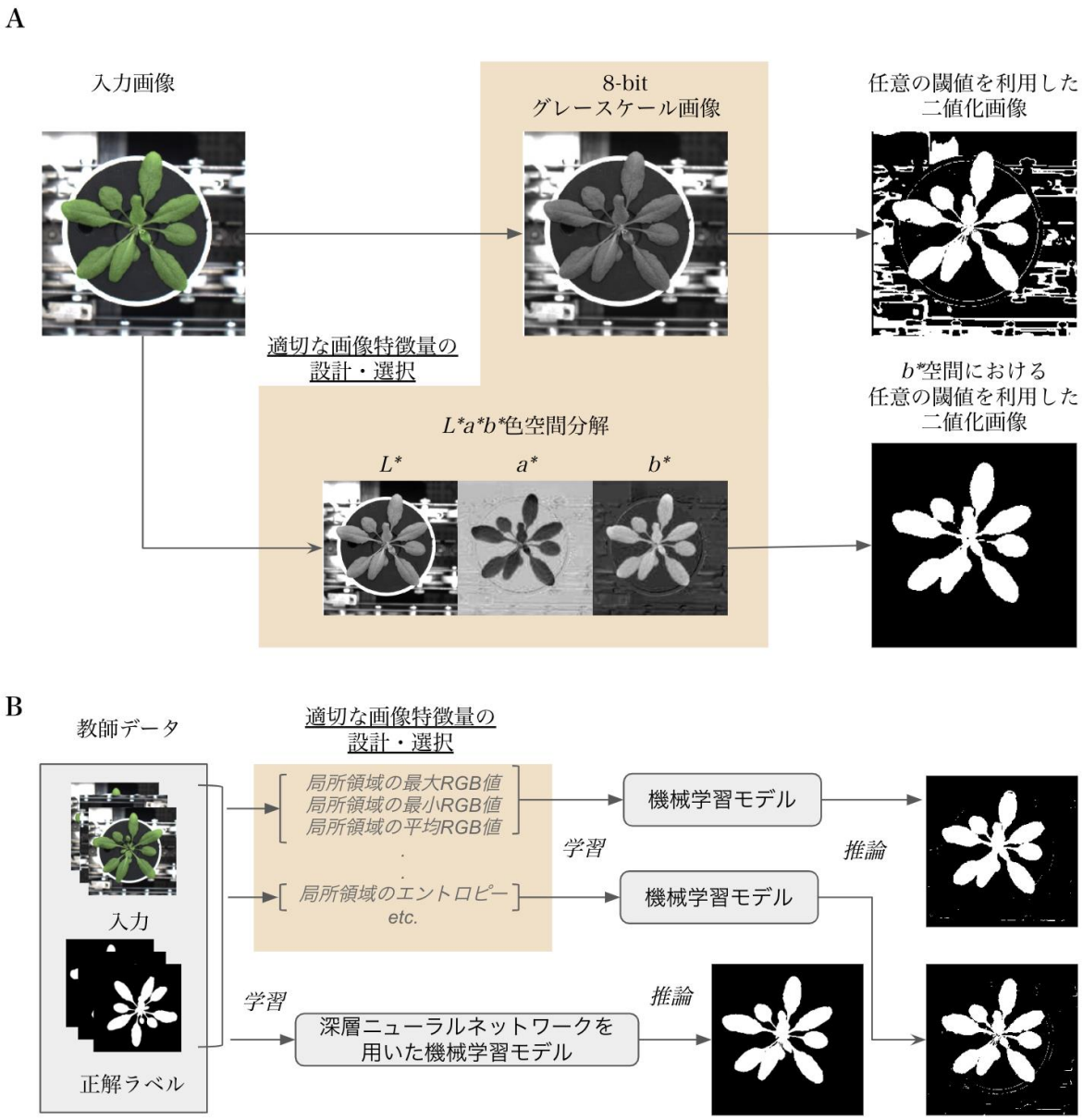

図 2　シロイヌナズナの葉面積定量を実現する画像解析技術

３－２－１．特徴量設計・選択と閾値の手動選択による手法

例えば図 2A のような入力画像をグレースケール変換（8bit に変換）すると，主に明るさの情報から構成されるデータに変換することができる。ここで，「一定の明るさの値を下回るピクセルは葉の領域，上回る領域は背景」と定義し，任意に定めた明るさの閾値に基づき，画像のフィルタリングを行うことで，葉領域のみを抽出することができる（図 2A 上）。これは，最もシンプルかつ伝統的な画像処理の手法であると考えられる。ここで，グレースケール変換を行う代わりに，入力画像を *L*a*b** 色空間に分解し，画像の黄色－青色成分を構成する *b** 空間の情報を利用する方法も考えられる（図 2A 下）。ここで利用した明るさ（明度）や色（色調）などのように，画像から情報抽出を行うために用いる情報を一般に特徴量と呼び，最終的に得られる情報の

精度は，利用した特徴量の種類および見出した閾値に大きく依存する。図 2A における二値化結果は，*b**空間を利用した方が優れているように見える。これは，今回用いた画像が背景領域に黄緑に近い色成分を含んでいないためである。この種の方法では，如何に閾値を見出しやすい条件で情報を取得できるか（この場合においては背景のノイズをいかに減らして閾値を設定しやすい画像を取得するか）も重要となる。適切な閾値を見出すことができれば，高速かつ明快に情報処理を行うことができるため，現在に至るまであらゆる分野で広く用いられるアプローチである。

３－２－２．機械学習を用いた閾値選択の自動化手法

前述の手法は，人間に処理の流れが理解しやすいことが利点の一つである。他方，ほとんどの過程を人間が行うため，特に，得られた画像が均質でない（色・明るさ・ノイズなどの点で）場合，入力により異なる作業が必要となり作業量が膨大となる。また，一義的な閾値よりも複雑な判断基準により判別・分離が必要な場合，手作業による基準の決定はしばしば困難となる。作業量削減および，複雑な判別基準を見出す観点から，機械学習を用いて一部の過程を代替する方法がある。

以下，伝統的な（深層学習以前の）機械学習手法を用いて画像から葉面の領域を抽出する一例を概説する。まず入力画像とそれに対応する望む出力結果（正解ラベルと呼ぶ；この例の場合では正しい葉面の領域である）の組み合わせ，いわゆる教師データセットを用意する（図 2B）。まず，利用すべき複数の特徴量（この例の場合は局所領域の RGB 値やエントロピー値など）を適切に選択し，機械学習器（例えば SVM や GBM など）を学習させる。機械学習器は，与えられたデータセットに対して目的（本節の場合は葉面の領域抽出）を達成するための適切な閾値，さらには特徴量の重み付けを行うことが可能となるので，教師データセット以外の画像に対しても，葉面の領域を抽出できるようになる（図 2B 上段）。画像解析に用いるのに適した特徴量やモデルに関しては，現在までに多様な画像特徴量と機械学習モデルが提唱されており，それぞれの特徴を理解して，それらを組み合わせることで，画像解析効率が格段に上昇し，利用可能な画像の種類も増加した。

３－２－３．深層学習を用いた特徴量設計・閾値選択の自動化手法

前章までででも述べた通り，深層学習も機械学習の一種であるが，前節の方法との大きな違いとして，画像解析における応用を考えると，人間が特徴量を設計しなくても良い場面が多くなることが挙げられる。前節の方法では，「葉面積定量を行うためには緑成分が重要かもしれない」というような観点の下，局所領域の RGB 値やエントロピー値を特徴量に選んでいたが，深層学習では，深層学習の過程でニューラルネットワークモデルが能動的に最適な特徴量を学習するような設計が可能である。少し乱暴な言い方をすれば，教師データセットと深層学習のモデルを用意すれば，目的とする一連の過程を一挙に達成する（end-to-end などと呼ばれる）ことが可能と

なる。その極めて高い利便性から，近年多くの研究者が自身の課題を解決するために深層学習を導入しつつあり，近年「人工知能を活用した」画像解析はこのような「深層学習を利用した」画像解析のことを指すことが多い。

３－３．植物画像解析における「人工知能」をもう一度考える

そもそも画像解析における人工知能とはなんなのか，ここでもう一度考えてみたい。前章までで述べてきたとおり，また，前章までででも場合によって表現が異なることからもわかるように，その言葉の定義は極めて曖昧である。本章でもう一度，分野を限って表現するならば，人工知能とは，「人間が従来行ってきた解析作業を大なり小なり代替してくれるアルゴリズム，パイプライン，またはソフトウェア」であると考える。つまり，技術の新旧を問わず，画像解析に携わってきた研究者はブームの前からずっと，人工知能を使い続けてきたのである。繰り返し述べてきたように，近年注目されている深層学習とは，機械学習の一部であり，「人工知能」のほんの一部である。

近年，多くのシチュエーションで「人工知能」を構成する中心要素として深層学習を利用することが増えてきた。例えば「圃場中の作物の穂の計測」や「作物の病害虫診断」など，およそ人が設計する特徴量では対応できないような複雑な課題には極めて有効である。しかしながらいまだ多くの画像解析タスクでは深層学習を必ずしも使う必要の無いことが多い。図 2 に示した葉の面積定量なども *L*a*b** 空間の利用が適切である場合もある。第 2 章でも述べたが，各々の「人工知能」には適した使い方がある。各々の特徴を理解し，適切な「人工知能」を最適な方法で利用することが肝要であろう。

また，深層学習特有の問題も多く残る。前章までで述べた開発の歴史を踏まえても，深層学習にはネットワークパラメーターの最適化や，適切なネットワーク構造を選ぶことなどが必要である。このために，実際にはユーザーが必要と予想される特徴量を事前にある程度把握して目的設定を行う必要がある。また，深層学習を行うために必要な教師データ（枚数）は従来の機械学習の手法と比べ数十倍から数百倍必要とされている。これについては，半教師あり/教師なし学習や，転移学習（transfer learning），データ拡張（data augmentation）など様々なアプローチで改善しつつあるもの，未だ多くの場合にデータの収集やアノテーションが問題となる。また，計算速度の速いコンピュータを用いても，学習にかかる時間が長く，他の手法のほうが効率的である場合も多い。深層学習を中心とした「人工知能」ブームに惑わされず，適切な技術を組み合わせて解析していくのが効率的な結果の取得には重要である。次章では，本章までの内容を踏まえて，我々研究者はどのように「人工知能」と付き合っていくべきか，考えたい。

４．「人工知能」に振り回されないために

人工知能関連の技術，とりわけ近年は深層学習が容易に使用できるようになり，植物科学をはじめとする様々な分野において応用されている。一方，近視眼的に流行に乗って「とりあえず深層学習を使う」ことや，「AI を使えばなんでもできるんでしょ？」などの発想が増えていることは否めない。また，近年の機械学習手法の仕組みを正しく理解しないことによる過度の恐怖が蔓延していることも事実である。本節では，「AI ブーム」に振り回されず，次世代の植物科学を創っていくためにはどのようなことが必要か，近年の（第 3 次）AI ブームの中核をなす深層学習によくある問題を踏まえて議論する。

４－１．深層学習の実際

議論の前提として，深層学習が実際に行うことを（ネットワーク的な図を使わずに[9]）まとめておく。機械学習の定義に則って説明するなら，深層学習は「たくさんの，かつ高次元（画像など）のデータ集合（学習データ）を入力」し，「事前に設定された評価尺度」を改善するように「（深層ニューラルネットワークで表現された）ものすごく高次元の関数」で近似するように学習するものであるといえる（この学習過程を，学習フェーズと呼ぶ）。学習フェーズで得られたモデルを用いて，任意の新たな入力（テストデータ）に対し，所望の出力に（上記評価尺度の基準で）近い結果を得ること（テストフェーズ，と呼ばれる）が，深層学習の目的である。基本的には，これまでの機械学習（極端に言えば線形回帰分析も含む）の延長線上の技術である[10]ため，実装は一般化したとはいえ，一歩間違えると科学的に正しくない結果や，本来の意図と異なる結果・結論を導きやすいことに注意が必要である。

４－２．深層学習（あるいは機械学習一般）活用時に直面する問題

ここでは，深層学習（あるいは機械学習一般）活用時に直面する問題について，主要な（？）2 点を挙げて議論する。

４－２－１．ブラックボックス性

深層学習が「ブラックボックス」である，という議論が多くなされている。機械学習は，入力データに対応して所望の出力（分類結果など）を得るようにモデルを最適化する。ここで，モ

[9] モデルがネットワーク状の構造をなすか否かは，ここでは本質ではない。

[10] そのため，現時点で（あるいは近い将来），AI が意思を持って人類を滅ぼしたり反乱を起こしたりすることはない。深層学習は人間の設計した評価尺度を改善するモデルを推定する最適化器にすぎない。より心配すべきは，人類自身が技術を悪用し，人類を滅ぼす可能性であろう。

デルが複雑であればあるほど，出力が得られたプロセス（なぜうまくいくのか・うまくいかないのかの判断根拠）を知ることは難しくなる。特に，深層学習は，適切に用いると他の機械学習手法より所望の出力に近い結果を得ることができることが多い一方，非常に多くの（数百万〜数億にもなる）パラメータを持つモデルを用いるため，人間が判断根拠を知ることが困難になる。すると，最適化されたモデルを用いて科学的なアプローチで結果を議論することが難しくなり，特に基礎科学や医療応用の分野において深層学習の活用を敬遠する一因となっている。これに対し，近年は深層学習（など）の判断根拠や付随する情報を可視化する説明可能 AI（Explainable AI; XAI）という枠組みに関する研究が盛んに行われている。特に，入力のどの部分に注目して判断が行われたかを可視化する GradCAM (Selvaraju et al. 2017) 等の手法や，ネットワーク中の各々のニューロンがどのようなパターンに強く反応するかを可視化する手法 (Olah et al. 2018) が広く使われる[11]。植物の疾病判別の可視化を対象とした可視化手法の比較研究 (Toda and Okura 2019) も行われており，興味を持った読者は参考にされたい。

４－２－２．バイアス・過学習・データリーク

深層学習（を含む機械学習一般）は本質的に，学習データに含まれる傾向（バイアス・偏り）を判断根拠として抽出し，学習するものである。もし，使用者が意図しない，学習データに特有の傾向が含まれる場合，これらがタスク（分類や回帰など）の判断根拠に含まれることがある。すると，学習済みモデルを他の環境で適用する場合にうまく働かなくなる（＝汎化性が失われる）。このような問題は，文脈により学習データに含まれるバイアス，あるいは過学習（過剰適合），データリークなどと絡めて議論される。最近は，大手 IT 企業の機械学習を用いたシステムに含まれる差別的なバイアス（学習データ自体にそのようなバイアスが含まれるため，男性を女性より優遇する採用提案システムができた，など）の話題を通じて，一般的に広く知られるようになった。実際，汎化性を毀損するような，使用者の意図しない傾向を判断根拠としてしまう例は数多くある。以下に架空の例を挙げて説明する。これらに類似するケースは様々な場面で発生するため，植物科学分野においても注意が必要であろう。

ケース 1：疾病検出　葉の疾病検出を行うシステムを構築するため，葉の疾病を含む画像を全世界の様々な圃場から収集したが，逆に，正常な葉の画像は持ち合わせていない。そこで，身近な圃場で正常な葉を撮影し，これらを学習データに含めて，疾病かどうかを判別するよう学習した。この場合，もしかすると機械学習モデルは，実際は背景に写った圃場の風景の違い（ある特定の「身近な圃場」かそうでないか）など，撮影設定に特有の特徴を学習している可能性がある。

[11] 以下に，直感的なデモを含む解説記事がある。https://distill.pub/2018/building-blocks/

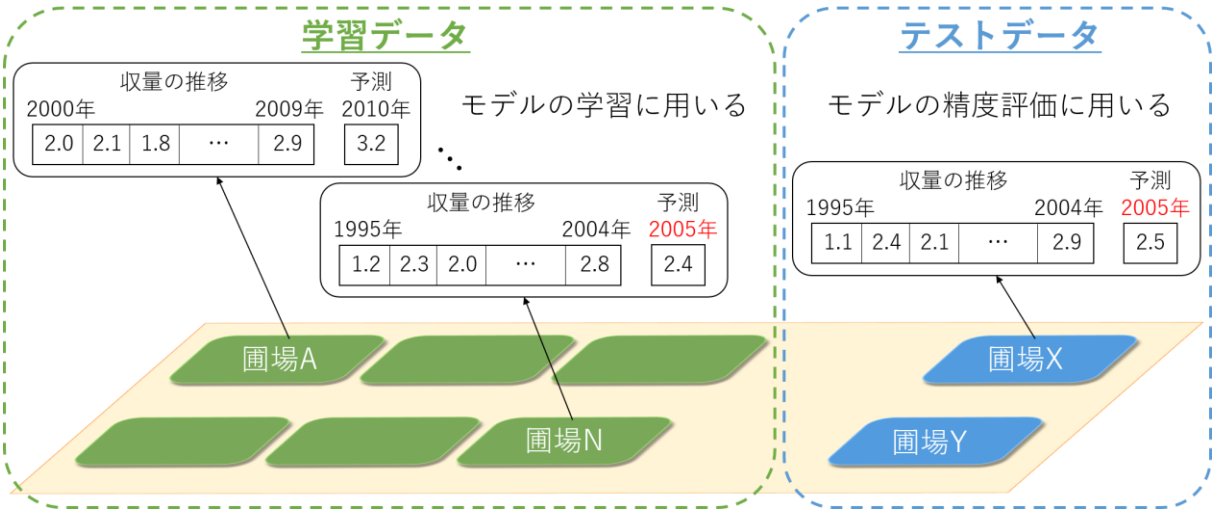

図 3 データリークにより，不当に高い精度が出やすい例（収量予測）：
同一年度の（類似した）データ系列が学習・テスト双方に含まれる。

ケース 2：収量予測　様々な圃場で長期間収集した環境情報・収量の時系列データを使い，収量予測モデルを作りたい（図 3）。例えば，推定対象年度より前 9 年分の環境情報と収量の推移，および推定対象年度の部分的な環境情報から，推定対象年度の収量を予測するようなモデルを学習することを考える。ここで，圃場 A〜N の収量・環境情報の推移を学習データ，圃場 X〜Y のものをテスト用のデータとし，推定対象年度はランダムに選択した。この場合，異なる圃場の同一年度のデータが学習・テスト双方に含まれ得る。つまり，（もしそれらの圃場が似たような環境にあるならば）テストデータとほとんど同じ系列が学習データに含まれるため，見かけ上高い精度を達成できるだろう。しかし，学習されたモデルは，未知の年度を対象とした予測には使えないものになっているかもしれない。

ケース 3：ハイパーパラメータの設定　機械学習には，学習に際して使用者が決めるべきパラメータ（ハイパーパラメータ）がたくさんある。例えば，深層学習においては学習の繰り返し数（エポック数）などが挙げられる。これらを決めるために，ハイパーパラメータを変えて学習を繰り返し，テストデータで評価尺度が最も改善されるものを採用することがあるが，特定のテストデータにのみ有効なモデルとなる可能性が高く，悪手である。ハイパーパラメータを決める必要があるなら，学習・テストいずれにも使われないデータ（validation データなどと呼ぶ）を使うべきである。機械学習と関わりの深い分野の論文でさえ，このようなミスはしばしば見られる。

ケース 4：手法の選定　多くの場合見逃されることが多いが，本来厳密に言うと，そもそも同じデータセットを使いまわして手法の検討を続けること自体，汎化性を毀損する可能性をはらんでいる。研究に使ったデータセットのみに有効な手法が選ばれる可能性が高くなるからである。本来，可能な限りバリエーションの異なる環境で取得されたデータセット，あるいは環境の異なる

複数のデータセットを使うべきであるが，データ収集の制約から，実際は難しいことが多い。

以上のように，機械学習において汎化性能があがらないことは永遠の課題であり，機械学習を専門とする分野の論文であっても，再現性の低さ，異なるデータセットにおける精度の低下（汎化性の低さ）が度々議論される。機械学習がうまく行えるのは，基本的に「モデルのパラメータが作る空間に分布するデータ点」の内挿である。学習データの分布に含まれない[12]ような入力に対するモデルの出力はあてにならない。つまり，機械学習を用いる場合は常に，「そのデータセット（あるいは検証方法）を使った場合，実際に活用したい場面で通用する（＝汎化する）モデルを学習できるのかどうか」を，データ取得の計画段階から慎重に考え続けることが重要である。様々な落とし穴があるため，「とりあえず」深層学習を適用し，驚くほど良い精度が出たとしても，その結果を鵜呑みにしてはならず，深く考察することを心がけたい。特に，データ取得や実装，実験設定が適切か，今一度確認することが重要であろう。あるいは，上述のような可視化手法を用いることで，意図しない部分に着目するようなモデルが学習されていないかどうかを（ある程度）確認することができる。

４－３．「人工知能」時代の植物科学

深層学習は，非常にパワフルかつ参入障壁の低い技術であり，機械学習の専門知識を持たずして，利用することができる。そして，これこそが現在の AI ブームの礎をなす。であるからこそ，植物科学は「AI」ブームに踊らされず，あくまで植物科学を発展させるべく，効果的に活用していくことが（あるいはあえて活用しないことも）必要であろう。

深層学習は，植物科学，なかでも植物フェノタイピングや遺伝子解析などにおいて強力なツールとなる。たとえば CRISPR を用いたゲノム編集が遺伝子解析の自由度を劇的に向上するツールであるように，深層学習をはじめとした技術群は，植物科学におけるデータ解析を効率化し，かつ人間にも見つけられないような特徴を抽出することを可能とするかもしれない。幸い，近年は初学者でも簡単に深層学習（あるいは他の機械学習も）を使い始めることができるため，必要に応じて，身近な問題から取り入れていくことができるだろう。

一方，上述のように機械学習（深層学習はとくに）には，ある意味でのブラックボックス性があることに注意したい。生理的・物理的・数学的にルールが自明であるタスクについて，機械学習的アプローチを取ることは，基本的にはおすすめできない[13]。特に解釈性が重視されるべき基礎科学分野において，ルールが既知である部分を，ブラックボックスに置き換えることには問題

[12] 一方，モデルのパラメータが作る高次元空間は人間に理解し難いため，何をもって「学習データの分布に含まれない」とするかは，非常に難しい議論である。

[13] 工学的な目的においては，ルールが自明な対象であっても，高速に推論できる深層学習の適用を選択することがある。

がある。近年，「AI」を使ったこと自体をアピールする研究や製品が数多く見られるが，「AI」を使うこと自体は，（技術的な）アドバンテージにはならない。深層学習（あるいは機械学習一般）はあくまでツールであり，実現したいことに適した手法を（それが「AI」と呼べるかどうかに関わらず）選択することこそが重要である。

「AI が仕事を奪う」という言説がある[14]。この真偽についてここでは議論しないが，科学研究の分野において，専門知識を持った研究者の重要性は，「AI 時代」においてより高まることが予想される。深層学習の解釈性の向上について，XAI などの研究がなされているものの，それらが行うのはあくまでモデルが注目した領域の可視化などにとどまる。モデルが学習した特徴の学術的（植物学的）な意味合いを説明し得るのは，専門知識を持った研究者である。今後，深層学習により，人間にはこれまで見つけられなかった詳細な特徴が（人間には理解し難い高次元のパラメータ空間内で）得られるかもしれない。しかし，これを植物科学に還元するためには，これらの特徴の可視化等を通じて，植物科学的な考察を付与できる研究者こそが欠かせないのである。

機械学習はデータに基づく最適化を行うものである。人工知能関連の技術の中で，人間が持つ知識（ヒューリスティクス）を扱う技術はエキスパートシステムなどのルールベースの手法に強みがある。近い将来，機械学習ベースの手法も，ヒューリスティクスを積極的に活用するような方向に進むべきであろう。例えば，「植物の発生ルール」を深層学習による植物の構造推定を行うモデルに組み込むことは，現時点では容易ではない。その点において，基礎科学分野における専門性は，AI 分野における新たな手法の構築にも寄与し得る。

「AI 時代」のいまこそ，植物科学研究者，および当分野に長年蓄積されてきた知識が重要である。「AI」に振り回され，植物科学を捨ててはならない。人工知能分野とともに歩んできた植物科学分野は，これからも人工知能分野とともに歩んでゆく。人工知能分野の劇的な進展を，植物科学の発展に活かすような使い方をすること（あるいはあえて活かさないことを選択すること）が最も重要である。

４－４．やりたくなった・勉強したくなった

本稿を読んで，機械学習・深層学習に興味を持った読者は，ぜひ触ってみることをお勧めする。多くのライブラリが Python で記述されており，最新の深層学習手法（モデル）の多くは実装が公開されているため，ツールとしての活用が比較的容易である。ここ数年，特に深層学習の研究は日進月歩どころか秒進分歩で進んでいるため，本稿では特定の手法を紹介することはしない（本稿が掲載される頃には陳腐化しているかもしれない）。代わりに，ここでは学習に活用できるリソースをいくつか紹介する。

[14] 著者の一人は，もし本当に AI が仕事を奪ってくれるなら奪ってほしい，悠々自適に暮らしたい，と考える。しかし現実には，研究者の仕事は増える一方である。

深層学習の初心者向けチュートリアルを含む記事は，インターネット上に多く存在する。一方，それらの多くが MNIST と呼ばれる文字認識データセットを用いたものであり，植物科学分野の研究者にとっては馴染みが薄く，応用との隔たりがある。本稿著者の戸田が制作した生物学者のための深層学習チュートリアル[15]では，深層学習の基礎的な活用方法を，植物を題材として進めることができる。比較的可読性の高い深層学習ライブラリである Keras と，オンラインでコードを実行できる環境である Google Colaboratory を用い，実行しながら学習をすすめることができる。深層学習に関するモデルは，日々より良いものが提案されており，専門家でも追いかけることが難しくなっている。Paper with Code[16]では，タスクごとに各種モデルの精度をランキング化している。これから挑戦するタスクに近いランキングを参照し，手法選択の参考とすることができる。

機械学習・深層学習の技術解説については，インターネット上に非常に多くのリソースがあり，特定のライブラリに特化した実装方法についても，インターネット・書籍ともに多数存在する。そのため，ツールとして機械学習・深層学習を使うための情報には事欠かない時代となった。一方，実装に特化しない，体系化された理論を学習したい場合には，以下のような本が多くの大学や研究室で講義・輪講の題材に挙げられているようである。ただし，いずれの本も，読みすすめるためには基礎的な解析・線形代数・統計の知識が必要である。

深層学習関連

- 導入編：ゼロから作る Deep Learning―Python で学ぶディープラーニングの理論と実装 (斎藤 2016)
- 学部レベル：深層学習（機械学習プロフェッショナルシリーズ） (岡谷 2015)
- 大学院レベル：深層学習（原題：Deep Learning） (Goodfellow et al. 2016)

機械学習一般

- 学部・大学院レベル：はじめてのパターン認識（通称：はじパタ） (平井 2012)
- 学部・大学院レベル：わかりやすいパターン認識（通称：わかパタ） (石井 et al. 2019)
  続・わかりやすいパターン認識 (石井 and 上田 2014)
- 研究者レベル：パターン認識と機械学習（原題：Pattern Recognition and Machine Learning, 通称 PRML） (Bishop 2006)

謝辞

本総説で紹介した研究の一部は，JST さきがけ「情報科学との協働による革新的な農産物栽培手法を実現するための技術基盤の創出」JPMJPR17O5（戸田），JPMJPR17O3（大倉），

[15] https://github.com/totti0223/deep_learning_for_biologists_with_keras

[16] https://paperswithcode.com/sota

JPMJPR16O5（野下）および，JSPS 若手研究 19K16163（水谷）の支援を得て遂行した。

参考文献

Bishop, C.M. 2006. Pattern Recognition And Machine Learning. New York: Springer.

Breiman, L. 2001. Random Forests. Springer Science and Business Media LLC.

Broner, I., & Comstock, C.R. 1997. Combining expert systems and neural networks for learning site-specific conditions. *Computers and Electronics in Agriculture* 19(1), pp. 37–53.

Buchanan, B.G., & Feigenbaum, E.A. 1981. Dendral and Meta-Dendral. In: *Readings in Artificial Intelligence*. Elsevier, pp. 313–322.

Cartwright, H.M. 1993. Applications Of Artificial Intelligence In Chemistry. Oxford: Oxford University Press.

Chen, T., & Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 785–794.

Crevier, D. ed. 1993. AI: The tumultuous history of the search for artificial intelligence, Basic Books.

Fitzhugh, R. 1961. Impulses and Physiological States in Theoretical Models of Nerve Membrane. *Biophysical Journal* 1(6), pp. 445–466.

Fujita, M., Tanabata, T., Urano, K., Kikuchi, S., & Shinozaki, K. 2018. RIPPS: A Plant Phenotyping System for Quantitative Evaluation of Growth under Controlled Environmental Stress Conditions. *Plant & Cell Physiology* 59(10), pp. 2030–2038.

Fukushima, K., & Miyake, S. 1982. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: Amari, S. and Arbib, M. A. eds. *Competition and Cooperation in Neural Nets*. Lecture notes in biomathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 267–285.

Ghosal, S., Blystone, D., Singh, A.K., Ganapathysubramanian, B., Singh, A., & Sarkar, S. 2018. An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences of the United States of America* 115(18), pp. 4613–4618.

Goodfellow, I., Bengio, Y., & Courville, A. 2016. Deep Learning. The MIT Press.

He, K., Zhang, X., Ren, S., & Sun, J. 2016. Deep residual learning for image recognition. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 770–778.

Hinton, G.E., & Salakhutdinov, R.R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313(5786), pp. 504–507.

平井有三 2012. はじめてのパターン認識, 森北出版.

Hodgkin, A.L. & Huxley, A.F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology* 117(4), pp. 500–544.

Holland, J.H. 1975. Adaptation In Natural And Artificial Systems: An Introductory Analysis With Applications To Biology, Control, And Artificial Intelligence. U Michigan Press.

石井健一郎 & 上田修功 2014. 続・わかりやすいパターン認識―教師なし学習入門―, オーム社.

石井健一郎, 上田修功, 前田英作 & 村瀬洋 2019. わかりやすいパターン認識(第 2 版), オーム社.

Isokane, T., Okura, F., Ide, A., Matsushita, Y. & Yagi, Y. 2018. Probabilistic Plant Modeling via Multi-view Image-to-Image Translation. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2906–2915.

Ke, G., Meng, Q., Finley, T., et al. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In: *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*. pp. 3146–3154.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. 2017. AlexNet 2012 ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 60(6), pp. 84–90.

LeCun, Y., Boser, B., Denker, J.S., et al. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1(4), pp. 541–551.

Le, Q.V., Ranzato, M., Monga, R., et al. 2012. Building high-level features using large scale unsupervised learning. In: *Proceedings of International Conference on Machine Learning (ICML)*.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. and Svetnik, V. 2014. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model* 55(2), pp. 263-274.

Mackay, D.J.C. 2003. Information Theory, Inference And Learning Algorithms. 1st ed. Cambridge, UK: Cambridge University Press.

McCarthy, J., Minsky, M.L., Rochester, N. and Shannon, C.E. 2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine* 27(4).

McCulloch, W.S. & Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics* 5(4), pp. 115–133.

Minervini, M., Fischbach, A., Scharr, H. & Tsaftaris, S.A. 2016. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern Recognition Letters* 81, pp. 80–89.

Mohanty, S.P., Hughes, D.P. & Salathé, M. 2016. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science* 7, p. 1419.

Nagumo, J., Arimoto, S. & Yoshizawa, S. 1962. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE* 50(10), pp. 2061–2070.

Oide, M. & Ninomiya, S. 2000. Discrimination of soybean leaflet shape by neural networks with image input. *Computers and Electronics in Agriculture* 29(1–2), pp. 59–72.

Oide, M. & Ninomiya, S. 1998. Evaluation of Soybean Plant Shape by Multilayer Perceptron with Direct Image Input. *Ikushugaku Zasshi* 48(3), pp. 257–262.

岡谷貴之 2015. 深層学習 (機械学習プロフェッショナルシリーズ), 講談社.

Olah, C., Satyanarayan, A., Johnson, I., et al. 2018. The building blocks of interpretability. *Distill* 3(3).

Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6), pp. 386–408.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. 1986. Learning representations by back-propagating errors. *Nature* 323(6088), pp. 533–536.

斎藤康毅 2016. ゼロから作る Deep Learning ―Python で学ぶディープラーニングの理論と実装, オライリージャパン.

Sanchez, J. & Perronnin, F. 2011. High-dimensional signature compression for large-scale image classification. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1665–1672.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 618–626.

Shortliffe, E.H. & Buchanan, B.G. 1975. A model of inexact reasoning in medicine. *Mathematical Biosciences* 23(3–4), pp. 351–379.

Singh, A., Ganapathysubramanian, B., Singh, A.K. & Sarkar, S. 2016. Machine learning for high-throughput stress phenotyping in plants. *Trends in Plant Science* 21(2), pp. 110–124.

Suykens, J.A. & Vandewalle, J. 1999. Least squares support vector machine classifiers. *Neural Processing Letters* 9(3), pp. 293–300.

Toda, Y. & Okura, F. 2019. How convolutional neural networks diagnose plant disease. *Plant Phenomics*, Article ID 9237136.

Toda, Y., Toh, S., Bourdais, G., Robatzek, S., Maclean, D. & Kinoshita, T. 2018. Deepstomata: facial recognition technology for automated stomatal aperture measurement. *BioRxiv* 365098.

Turing, A.M. 1950. Computing machinery and intelligence. *Mind* 49, pp. 433–460.

Watanabe, K., Guo, W., Arai, K., et al. 2017. High-Throughput Phenotyping of Sorghum Plant Height Using an Unmanned Aerial Vehicle and Its Application to Genomic Prediction Modeling. *Frontiers in Plant Science* 8, p. 421.

Yamamoto, K., Guo, W., Yoshioka, Y. & Ninomiya, S. 2014. On plant detection of intact tomato fruits using image analysis and machine learning methods. *Sensors* 14(7), pp. 12191–12206.

Yu, V.L., Buchanan, B.G., Shortliffe, E.H., et al. 1979. Evaluating the performance of a computer-based consultant. *Computer Programs in Biomedicine* 9(1), pp. 95–102.

AI を活用した生物情報アプリを用いた生物モニタリングの社会実装

藤木庄五郎
株式会社バイオーム，〒600-8813 京都府京都市下京区
中堂寺南町１３４番地 ASTEM ビル８階

AI implementation in biodiversity monitoring, using smartphone application

Shogoro Fujiki
Biome Inc., ASTEM8F, 134 Chudoji Minamimachi, Shimogyo-ku, Kyoto, 600-8813, Japan
Keywords: Citizen science, Biodiversity monitoring, Deep Learning, Data annotation,
DOI: 10.24480/bsj-review.11c3.00192

１．はじめに

IPBES（2019）では，現在，人間活動により動植物 100 万種が絶滅危機リスクにあると報告されており，生物多様性の包括的な保全が喫緊の課題である。そのためには，生物多様性のモニタリングとデータに基づく確かな保全策の実行が求められる。しかし一方で，生物多様性を広域で評価する実用的なモニタリング手法は開発が遅れており（Fujiki et al. 2016），中でも，国レベルでの広域地上調査は 10 年以内に実用水準に到達することは困難（Goetz et al. 2015）という指摘もある。こうした背景から，実用的な生物多様性広域モニタリング手法の確立と，データベースの構築に向けた方法論の議論とその実践が重要である。

近年のカメラ技術と画像処理パッケージの進歩により，特に中大型哺乳類・鳥類研究の文脈で，カメラトラップを用いた調査が急速に拡大している（Swanson et al. 2015; Black et al. 2017; Jones et al. 2018）。カメラトラップを用いた調査は，比較的簡便かつ定期的な定量データの取得に適しており，広域での生物分布調査への応用が期待されている。しかし一方で，大量に得られる画像の解釈にかかる労力が無視できないレベルで大きいこと，小型動物や植物分布調査等に適しているとは言い難いことから，網羅的な生物多様性広域モニタリング手法として活用するには至っていない。

こうした背景の中で，Citizen Science（市民科学）の考え方を応用した取り組みが注目を集めている（宮崎 2016）。市民科学とは，多様な背景をもつ市民が研究者と連携しながら，科学研究の多様なプロセスに参画することによって，科学への貢献だけでなく，社会的な課題・要求にも応えていくための方法論を検討する新興の学術領域である（宮崎 2016）。南極のカメラトラップを用いた「ペンギン・ウォッチ」プロジェクト（Jones et al. 2018）では，少なくとも 7 万 3802 点のペンギンの画像分類と，それに関連するメタデータ（日付，時刻，気温情報を含む）のアノテーションを市民ボランティアが手動で実施した。こうした事例は，深層学習におけるアノテーションの困難さを克服し，画像を用いた生物調査をより簡便に実施

できる可能性を示唆している。

さらに，昨今のスマートフォン・タブレット端末の急速な普及を背景に，市民科学は，網羅的な分類群の生物分布データ収集にも応用が進み始めている。すなわち，スマートフォン・タブレット端末で市民が撮影した位置情報付きの生物写真を収集する取り組みである。生物写真を AI により自動で同定し，データベースとして蓄積することができれば，植物を含む網羅的な生物分布の広域モニタリングが飛躍的に進展する可能性があると考えられる。本稿では，画像解析 AI を実装したスマートフォンアプリによる市民参加型調査の事例を紹介し，生物多様性モニタリングの今後の展望について考察を行う。

２．スマートフォンを用いた市民参加型の動植物調査

カリフォルニア科学アカデミーが 2008 年から運営する iNaturalist（http://www.inaturalist.org/ 2020 年 1 月 1 日確認）プロジェクトでは，生物の発見情報を投稿・共有できるスマートフォンアプリを公開し，自然愛好家と研究者が協力して，これまでに全世界で約 3,000 万件の生物確認情報や写真を収集している。国内では，環境省生物多様性センターが 2015 年から生物の観察情報を集め・提供するサービス「いきものログ」を運営し，登録したユーザーが生物情報を投稿し，共有できるシステムを提供している（竹原ら 2013）。また，株式会社バイオームが運営するスマートフォンアプリ Biome（バイオーム）は，2019 年 4 月に日本国内のみを対象に公開し，半年で約 12 万人が利用，30 万件以上の投稿数が確認されている（山口 2019 も参照）。Biome では，ゲーミフィケーションを導入することで，これまでの市民科学の主な対象であったセミプロ・自然愛好家だけではなく，今まで自然科学に縁のなかったユーザー層の参加を積極的に促し，市民科学の裾野を広げることを狙いとしている。

（a） （b） （c）

図 1

（a）iNaturalist の Google play 公開画像

https://play.google.com/store/apps/details?id=org.inaturalist.android
（b）いきものログの Google play 公開画像
https://play.google.com/store/apps/details?id=jp.go.biodic.ikilog.ikimonolog&hl=ja
（c）Biome の Google play 公開画像
https://play.google.com/store/apps/details?id=jp.co.biome.biome&hl=ja

２－１．市民参加型の動植物調査の課題

こうした市民参加型調査は, 参加する市民にとって明確なメリットが存在していなければ, 成功は担保され難く（Silvertown 2009），研究者にも市民にもメリットのある双方向型の市民科学プロジェクトの構築が持続可能性の確保に必要不可欠であることが指摘されている（宮崎 2016）。市民からのデータ投稿に対し明確なメリットを提供する方法として，景品や賞金などによる外発的動機付けと，満足感，達成感などの内的動機付けが想定される。ユーザーが投稿する魚類の画像が図鑑に登録されるサービス WEB 魚図鑑においては，名前の分からない個体の同定結果を知りたいという欲求を満たすことができること，マスメディア等への写真提供による著作権収入を得られ得ること，出版される魚類図鑑や科学論文に自身の名前が掲載される可能性があることなど，いくつかのユーザー側のメリットが存在していることで，持続的な運営が可能になっていることが指摘されている（宮崎 2016）。ただし，外的動機付けにおいては継続的な報酬の維持ができない場合, アンダーマイニング効果, すなわち, 内発的動機づけによって行われた行為に対して，外発的動機づけを行うことによって動機づけが低減する現象（Deci 1971; Murayama et al. 2010）が発生する可能性があるため，留意が必要である。

また, 市民科学プロジェクトにもとづくデータを科学的成果として活用しようとする際に, 二つの大きな問題を考慮する必要がある。すなわち，サンプリング・バイアスとデータの質の二点である（宮崎 2016; Dickinson et al. 2010）。筆者が関与するスマートフォンアプリ Biome では，市民から提供されるデータに，場所的・時間的制約があること，目につきやすい種に投稿が偏っていることが確認されている。特に人口密度と投稿数の相関は強く，解析に用いる際にバイアスの除去が不可欠である。誤同定についても，ある程度の頻度で発生していることが確認されている。提案機能・通報機能を用いたユーザー間の相互監視のもと，自浄されていく傾向がみられるが，最終的に教師データ等として利用する際には専門家あるいはパラタクソノミストによる確認が必要である。ただ，同定の根拠となる写真資料が残されていることで，後から誤同定を緩和することが可能であることから，データのトレーサビリティに留意し，履歴・根拠をアーカイブとして保存していくことが重要である。

ユーザーメリットの提供及びデータの質の担保において画像解析 AI が果たす役割は極めて大きい。名前の分からない個体の同定結果を知りたいというニーズは Biome においても極めて強く，同アプリのユーザーメリットの一翼を担っていると言える。また，スケーラビリティと持続性を担保するため，データの同定精度をユーザー依存的にするのではなく，システムとして精度担保を行うべきであり，その観点からも画像解析 AI の発達は必要不可欠であると結論付けられる。

２－２．AI 画像解析を用いた生物の同定

画像解析を用いた生物種同定には複数の課題が残っている。特に，１．十分な量のアノテーション済み画像の確保，２．写真による同定の根本的な限界，が挙げられる。Biome に用いている同定アルゴリズムでは，上述の課題を解決・緩和するため，画像に依存させすぎないことを基本思想に設計されている。すなわち，畳み込みニューラルネットワーク（convolutional neural network: CNN）による画像の特徴量の深層学習と並行して，画像に付加されているメタ情報をモデル化し学習に用いる試みの導入である。ここで言うメタ情報とは，位置情報，撮影日時，およびそれらに紐づけられる環境条件（気温，降水量，植生パターン等）を指す。これにより，すでに 7 万 9 千種類に及ぶ大量のカテゴリ―に属する画像をそれぞれ準備するのではなく，場合によっては種よりも上位の分類概念（例えば属や科）に留めて学習を行うことが可能になった。画像による分類を上位概念に留め，メタ情報によって補完的に同定を完遂することができるため，大量の画像を事前に準備することなく，比較的高い精度で生物の同定をすることが可能になった。

生物の画像解析においては，そもそも写真による同定に限界があることが想定されるが，これもメタ情報の導入により改善がみられることが分かった。これは，深層学習の画像分類精度の限界を画像以外のパラメータで補うことができるという事例として，今後さらなる検証を予定している。

３．今後の展望

生物の画像及びメタ情報による AI モデルの進歩とともに，生物多様性モニタリングの在り方が今後大きく変わることが予想される。これまで半手動的に種同定を行ってきたカメラトラップによる中大型哺乳類・鳥類調査も大部分が自動化される可能性がある。また，Unmanned Aerial Vehicle（UAV）等の空撮技術の発達により，樹木調査の大幅な効率化に寄与できることが考えられる（Onishi et al. 2018）。さらに，水中ドローン及び水中動画撮影技術の進歩は, これまで労力が大きかった水中での生物調査の風景を大きく変えるかもしれない。市民参加型調査においても，これまでのように特定の人が特定のアプリで参加するのではなく，不特定多数の人が WEB や SNS 上にアップした画像を解析することで，より多くの人が生物モニタリングに参加できる仕組みに発展することも考えられる。

市民科学の分野においては，これまで多くの場合，科学者と市民の連携においてのみ解釈が行われてきたが，これからは企業，行政も巻き込んだ形でのありかたを考えることが重要だと考える。生物多様性の保全という，困難かつ，挑戦な課題に取り組むにあたって，組織や個人の枠を超えて，産官学民が連携して一丸となって取り組む姿勢や規模感が必要とされている。その実施をスムーズに行えるプラットフォームを構築することこそが今一番必要なアクションなのではないだろうか。

謝辞

本総説で紹介した取り組みは，公益財団法人京都産業２１平成３０年度次世代地域産業推進事業「生物の名前判定ＡＩを用いた生物ビッグデータの構築」の支援を得て遂行した。

引用文献

Black, C., Rey, A. R., & Hart, T. 2017. Peeking into the bleak midwinter: Investigating nonbreeding strategies of Gentoo Penguins using a camera network. *The Auk: Ornithological Advances*, *134*(3), 520-529.

Deci, E. L. 1971. Effects of externally mediated rewards on intrinsic motivation. *Journal of personality and Social Psychology*, *18*(1), 105.

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. 2010. Citizen science as an ecological research tool: challenges and benefits. *Annual review of ecology, evolution, and systematics*, *41*, 149-172.

Fujiki, S., Aoyagi, R., Tanaka, A., Imai, N., Kusma, A., Kurniawan, Y., ... & Kitayama, K. 2016. Large-scale mapping of tree-community composition as a surrogate of forest degradation in Bornean tropical rain forests. *Land*, *5*(4), 45.

Goetz, S. J., Hansen, M., Houghton, R. A., Walker, W., Laporte, N., & Busch, J. 2015. Measurement and monitoring needs, capabilities and potential for addressing reduced emissions from deforestation and forest degradation under REDD+. *Environmental Research Letters*, *10*(12), 123001.

IPBES 2019. Nature's dangerous decline unprecedented; species extinction rates accelerating. https://www.ipbes.net/news/Media-Release-Global-Assessment. Accessed 20 May 2019

Jones, F. M., Allen, C., Arteta, C., Arthur, J., Black, C., Emmerson, L. M., ... & Miller, G. 2018. Time-lapse imagery and volunteer classifications from the Zooniverse Penguin Watch project. *Scientific data*, *5*, 180124.

宮崎佑介 2016. 市民科学と生物多様性情報データベースの関わり. *日本生態学会誌*, *66*(1), 237-246.

Murayama, K., Matsumoto, M., Izuma, K., & Matsumoto, K. 2010. Neural basis of the undermining effect of monetary reward on intrinsic motivation. *Proceedings of the National Academy of Sciences*, *107*(49), 20911-20916.

Onishi, M., & Ise, T. 2018. Automatic classification of trees using a UAV onboard camera and deep learning. *arXiv preprint arXiv:1804.10390*.

Silvertown, J. 2009. A new dawn for citizen science. *Trends in ecology & evolution*, *24*(9), 467-471.

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., & Packer, C. 2015. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific data*, *2*, 150026.

竹原真理, 佐藤直人, 大谷知生, & 鑪雅哉 2013. 環境省生物多様性センターにおけるウェブサイトを活用した生物多様性情報の収集・提供の取り組み. *日本生態学会誌*, *63*(1), 141-144.

山口泰博 2019. 環境問題をビジネスに変える 株式会社バイオーム代表取締役 藤木庄五郎. *産学官連携ジャーナル*, *15*(10), 29-30.

RFE を用いた植物収穫時品質に関与する特徴量分析手法

中西豪太[1]・峰野博史[2,3]
[1] 静岡大学大学院総合科学技術研究科，〒432-8011 静岡県浜松市中区城北３－５－１
[2] 静岡大学学術院情報学領域，〒432-8011 静岡県浜松市中区城北３－５－１
[3] 静岡大学グリーン科学技術研究所，〒422-8529 静岡県静岡市駿河区大谷８３６

Feature analysis method related to plant harvest quality using RFE

Gota Nakanishi[1] and Hiroshi Mineno[2,3]
[1]Graduate School of Integrated Science and Technology, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka, 432-8011, Japan
[2]College of Informatics, Academic Institute, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka, 432-8011, Japan
[3]Research Institute of Green Science and Technology, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka, Shizuoka, 422-8529, Japan
Keywords: feature engineering, recursive feature elimination, plant harvest quality
DOI: 10.24480/bsj-review.11c4.00193

１．はじめに

農業従事者の高齢化に伴い，高水準な日本の農業技術が新規就農者に継承されずに喪失してしまうことが懸念されている。 特に高糖度なトマトなどブランド価値の高い農作物を栽培する技術は熟練農家の長年の経験や勘によって培われており，会得するには長い年月を要する。そのため，高品質な農作物を栽培する技術は新規就農者に継承されることなく失われてしまう可能性が高い。この課題を解決するため，熟練農家が経験則と勘に基づいて判断していた多様な環境条件と植物の因果関係を明らかにする研究が行われている。例えば，光強度条件など栽培環境条件の違いが収穫時品質に与える影響の分析が行われている（浜本ら 2010, 東出 2018，望月ら 1999)。ただし，実際の現場では，植物の生育状態や生育過程，時間帯といった考慮も行われており，そのようなデータを長期間かつ高品質に経時計測することは困難なため，栽培技術の形式知化を十分に実現できているとは言い難い。また，栽培中の環境条件は多岐に渡るため，様々な環境条件と収穫時品質を網羅的に検証することも困難である。そのため，新規就農者が熟練農家の持つ栽培技術を継承することは一筋縄ではいかない。

そこで，収穫時の品質に影響の大きい特徴量をデータドリブンで機械的に選択できないか試みた。生育状況と時間帯を重畳した経時特徴量データを算出（中西ら 2018，水野ら 2018）し，決定木ベースの回帰手法を再帰的に用いて特徴量選択を適用することで，膨大な特徴量の中から収穫時の品質に影響の大きい特徴量を選択する。選択された特徴量を用いて収穫時の品質を推定できれば，生育状況と時間帯を考慮した分析の可能性を示すことができる。

２．関連研究

２－１．逐次変数選択法（Stepwise 法） (Bendel et al. 1977)

変数群から重要な変数を選択する代表的な手法として，逐次変数選択法がある。逐次変数選択手法とは, 変数増加法（Forward stepwise selection）と変数減少法（Backward stepwise selection）を組み合わせた変数選択手法である。変数増加法は最初に全ての変数に対して，モデルに加えた場合の p 値などのモデル自体の評価指標となる統計量を算出する。ここで，p 値を変数選択の評価指標と仮定すると，事前に決定された閾値を満たす指標のうち，最小の値を持つ変数をモデルに加えてモデル構築をする。この処理を繰り返してモデルに含まれる変数を増加させていく。閾値を満たす指標が算出されなくなった時点で変数の追加を終了する。このときの変数の組み合わせを最良の組み合わせとする。一方，変数減少法は，変数増加法とは逆に，最初に全ての変数をモデルに取り込んだモデルを作成する。その後，モデルに含まれる変数のうち，決められた閾値を満たさない指標を持つ変数をモデルから除去していき，閾値を満たさない指標が算出されなくなった時点で変数の除去を終了する。逐次変数選択法は，最初に変数増加法と同様に変数の追加を行う。変数の追加のたびに，モデルに既に含まれている変数の中で変数減少法と同様に除去が可能な変数があれば除去する。その後，追加及び除去する変数がなくなるまで変数の選択を行い，最終的に残った変数を最良の変数の組み合わせとする。しかし，逐次選択法は，より少ない変数でより効率的に予測することが目的であり，目的変数との因果関係を考慮した変数選択を行わないため，選択された変数と目的変数の因果関係の分析には向いていない。また，変数の取捨選択を繰り返すため，重要な変数の最適な組み合わせを得ることも困難である。

２－２．決定木ベースの機械学習手法を用いた変数選択

Random Forest (Breiman et al. 2001) や XGBoost (Chen et al. 2016) といった機械学習手法は，異なる決定木を多数作成し，その結果の平均値を求めることで，決定木の欠点である過剰適合を抑制するアンサンブル学習手法である。Random Forest は推定時に使用した特徴量の重要度を算出可能であり，重要度に基づいた特徴量選択が可能である。Random Forest の変数の重要度は，各決定木における変数の重要度の平均値によって算出される。まず，無作為にデータを選択して決定木を作成し，作成した決定木で使用されている 1 つの変数に関して，データの並び順をランダムに変更する。並び順の変更前後で，決定木の精度を比較し，大幅な精度の変化が観測された場合，重要な変数とする。最後に，多数の決定木にて同様にその結果の平均値を取り，Random Forest の変数の重要度とする。Random Forest の各変数の重要度は，変数全体から見た相対的な値であるため，変数の重要度を 0 にするスパースな推定を行わない。そのため，変数選択を行うには，閾値など人手を介した重要な変数を選別が必要となる。また，モデル構築時の重要度は，全特徴量に対する相対的な値であるため，一度の特徴量選択では重要な特徴量のみを選択することは困難である。これら課題を解決する手法として，モデル構築と特徴量削減を再帰的に行う RFE (Recursive Feature Elimination) (Guyon et al. 2002) という手法が提案されている。RFE は，モデル構築時の特徴量の重要度を基に特徴量を削減し続けることで，誤差指標が最良時の特徴量を選択することができる。

２－３．正則化項を用いた変数選択 （廣瀬 2016）

遺伝子解析など，変数の次元数がデータセット数に比べて遥かに大きいという課題に対処できる統計手法として，正則化項によるスパース推定を用いた変数選択手法がある。正則化項を用いたスパースな推定を行う代表的な手法として，Lasso (Least Absolute Shrinkage and Selection Operator）(Tibshirani et al. 1996) 回帰や Elastic Net 回帰 (Zou et al. 2005) などがある。Lasso 回帰モデルは，互いに相関の高い変数群が含まれている場合，推定時に相関の高い変数の中の 1 つだけが選択され，他の変数は回帰係数が 0 と推定され，選択されないという課題がある。つまり，相関の高い変数群の中で選択される変数は，推定を行うごとに変化してしまい，Lasso 回帰による変数選択は不安定とされている。この Lasso の課題を解決するため考案された回帰手法が，Elastic Net 回帰である。Elastic Net 回帰は，Ridge 回帰 (Hoerl et al. 1970) と Lasso 回帰を混合した回帰手法である。Ridge 回帰は, 相関の高い変数群を考慮できるため，Lasso 回帰の正則化項と Ridge 回帰の正則化項の強さのバランスをとることで，Lasso 回帰の課題を解決する。また，Lasso 回帰と Elastic Net 回帰においては，自動で変数の選別を行うため，閾値の決定など人手の介入なくとも変数選択を実現できる。しかし，Lasso 回帰や Elastic Net 回帰は線形的な回帰手法であるため，対象が植物の成長といった複雑な変数との非線形性な関係性を持つ場合には不向きな場合が多い。

２－４．関連研究のまとめ

以上のように，変数選択においては，p 値などの統計的な指標を算出した変数に対し，事前に設定した閾値を用いて選別することで変数選択を行う手法が多い。しかし，重要な変数を余すことなく, 選別することが可能な閾値を決定することは難しい。また, Lasso 回帰や Elastic Net 回帰では非線形な関係性の考慮が困難であるため，非線形な関係性を考慮可能な Random Forest や XGBoost といった機械学習手法によって誤差が最小となるよう再帰的にモデルを構築し，特徴量を削減する手法である RFE を適用することで，重要度の高い特徴量を機械的に選択できる可能性が高い。

３．RFE を用いた植物収穫時品質に関与する特徴量分析手法

３－１．概要

植物の生育状態と時間帯を考慮して収穫時の品質に影響の大きい経時的な特徴量を機械的に選択する手法を検討した．特に，目的変数に対する特徴量の重要度が算出可能である決定木ベースの回帰手法を再帰的に用いる特徴量選択手法を植物栽培環境のような対象に適用できるか検証した。代表的な決定木ベースのアンサンブル回帰手法として，Random Forest や XGBoost などを想定する。決定木ベースのアンサンブル回帰手法では，推定時に使用した特徴量の重要度を算出可能であるため，生育状態や栽培した時間帯を考慮した経時的な特徴量に対して，重要度を指標とした特徴量選択を行うことで，植物栽培環境における収穫時の品質に影響の大きな特徴量を機械的に選択する特徴量分析手法を検証する。

本手法の概要を図 1 (a) に示す。本手法は図 1 (a) に示すように，大きく A：特徴量算出と B：特徴量選択から構成される。A：特徴量算出では，経時的な特徴量データを栽培対象の農作物の生育ステージ毎に分割した後，時間帯ごとに分割する。次に図 1 (b) に示すように，分割された各経時特徴量データに対して基本統計量を算出することで，植物の生育状態と栽培した時間帯を考慮した特徴量の算出を行う。また，栽培期間を通した特徴量として，温度や日射量の積算値や，栽培開始からの経過日数など，栽培した季節の考慮が可能な特徴量を算出する。B：特徴量選択では，A：特徴量算出において算出された多数の特徴量に対し，機械学習を用いて再帰的に特徴量を削除する手法である RFE を用いて，収穫時の品質に影響の大きい特徴量の選択を行う。

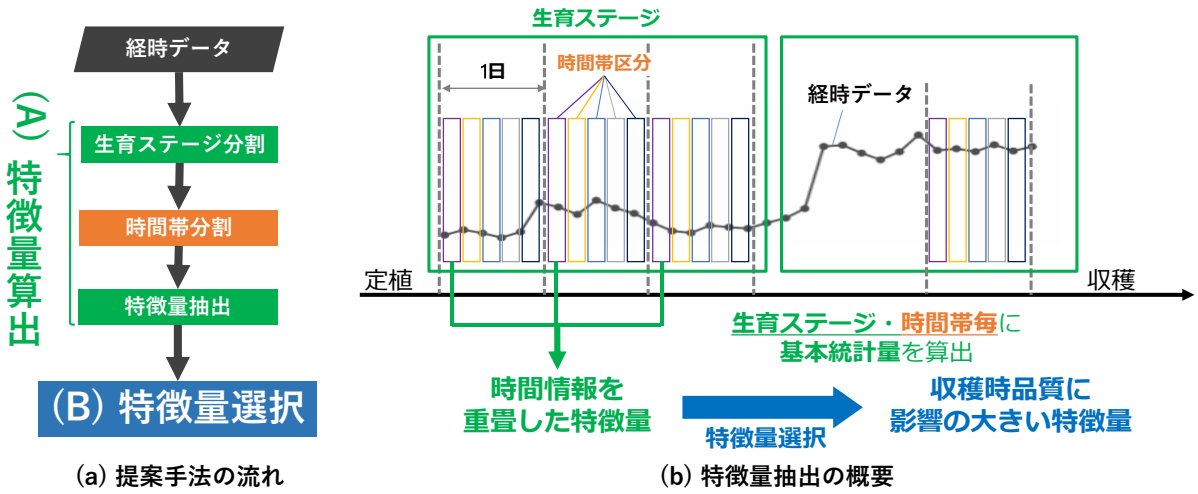

図 1 特徴量選択手法の概要

Fig. 1 Outline of the feature analysis

３－２．特徴量算出

特徴量算出では，経時特徴量データと生育状態，栽培した時間帯情報を重畳するため，経時特徴量データを生育状態と栽培時間帯ごとに分割する。分割の際には植物の状態や栽培した日のイベントを指標とする。生育状態は植物の開花から収穫までの栽培期間を対象とし，果実の生育ステージを指標に用いて分割する。栽培期間中の植物のイベントである開花や収穫を基準に分割を行うことで，果実の個体差を考慮した生育状態の分割が可能であると考える。また，栽培時間帯に関しては，日の出，日の入りを分割の指標に用いる。一般に植物は光合成によって果実の成長が促進される。そのため，日の出と日の入りを指標に用いることで，光合成の不可欠な要因である日射量の影響が考慮可能となると考える。次に生育状態と栽培時間帯に分割した各経時特徴量データに対して基本統計量の算出を行う。一般に熟練農家は特定の 1 日の植物の変化ではなく，長期的な変化を考慮した栽培技術を用いるため，基本統計量には大域的な変化を表現可能な積算，平均，最大，最小を用いることとした。 また，栽培を通じた特徴量として，植物の成長の指標に用いられる栽培期間の積算温度と，積算日射量や定植日からの日数を算出し利用することとした。

３－３．特徴量選択

Random Forest や XGBoost などのモデル構築時の特徴量の重み付けが可能な機械学習手法を用いてモデルを構築する。構築時に重要度の低い特徴量を削減する処理は，誤差指標を基準として再帰的に繰り返すことで特徴量選択を行う手法である RFE を特徴量算出で算出された特徴量に対して適用することで，機械的に目的変数に設定した収穫時の品質に対して重要な特徴量の選択が可能であると考える。また，特徴量選択の信頼性を高めるため，RFE で使用する外部学習器を複数使用することで，学習器別での特徴量選択結果の違いを考慮する。

４．基礎評価

実際に栽培した際のデータを用いて生育状態と栽培時間帯を重畳した特徴量の算出と，算出した特徴量から重要度の高い特徴量の選択，選択された特徴量を用いた収穫時品質の推定を実施した。本基礎評価では，2016 年 7 月から 2017 年 12 月にかけて複数のトマト（CF 桃太郎ヨーク）栽培農場にて低段密植栽培された 54 作分の栽培データを使用した。また，栽培期間に収集されたセンサデータとして，温度（℃），飽差（g/m^3），日射量（$J/m^2/s$），CO_2 濃度（ppm）を 1 分周期で測定した環境データを使用し，収穫時の品質の 1 つである収量は選果機を用いて 1 日毎に計測された値を使用した。特徴量算出での生育ステージの区分は，農学者の知見をもとに定植日から収穫開始日までを 5 等分し，生育ステージⅠから生育ステージⅤと設定した。栽培した時間帯の区分は，栽培した農場の日の出時刻と日の入り時刻を基準に，日の出時刻から日の入り時刻を 3 等分し，それぞれを時間帯 1 から時間帯 3 とした。また，日の入り時刻から翌日の日の出時刻までを 2 等分し，それぞれを時間帯 4，時間帯 5 と区分した（表 1）。また，特徴量を算出する際の基本統計量には，大域的な変化を表現可能である積算，平均，最大，最小を用いた。以後，図中の温度，飽差，日射量，CO_2 濃度は，それぞれ temp，vpd，solar，CO_2 と表記し，積算，平均，最大，最小はそれぞれ sum，ave，max，min と表記する。

特徴量選択で用いる RFE の外部学習器には，モデル構築時に特徴量の重み付けが可能な学習器である必要があるため，Random Forest，XGBoost，LightGBM (Wang et al. 2017) を用いて特徴量選択結果を比較した。選択された特徴量を用いた予測では，収穫開始日から 4 週間後までの 1 週間毎の総収量を予測対象とした。予測時に使用する学習器は Random Forest，SVR (Support Vector Regression) (Drucker et al. 1997)，XGBoost を用いた。また，予測結果の評価指標として，平均絶対誤差（MAE：Mean Absolute Error）（式 1）と平均二乗誤差平方根（RMSE：Root Mean Squared Error）（式 2），決定係数（coefficient of determination）（式 3）を用いた。実装は，Python3.7，scikit-learn

表 1 生育ステージと時間帯区分方法

Table 1 Growth stage and time zone division

生育ステージ	想定する生育状態	時間帯	範囲
Ⅰ	着花	1	日の出から日没（3 等分）
Ⅱ	開花前	2	
Ⅲ	開花	3	
Ⅳ	着果	4	日没から日の出（2 等分）
Ⅴ	果実肥大	5	

(Pedregosa et al. 2011) 0.21.0 を使用し，パラメータチューニングには Optuna を使用した。

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \quad (3)$$

４－１．相関分析

特徴量算出にて算出された特徴量の有効性を検証するため，算出された特徴量 400 次元（センサ値（温度，飽差，日射量，CO_2 濃度）× 生育ステージ（Ⅰ～Ⅴ）× 栽培した時間帯（1～5）× 基本統計量（積算，平均，最大，最小））と収量との相関行列を図 2 に示す。

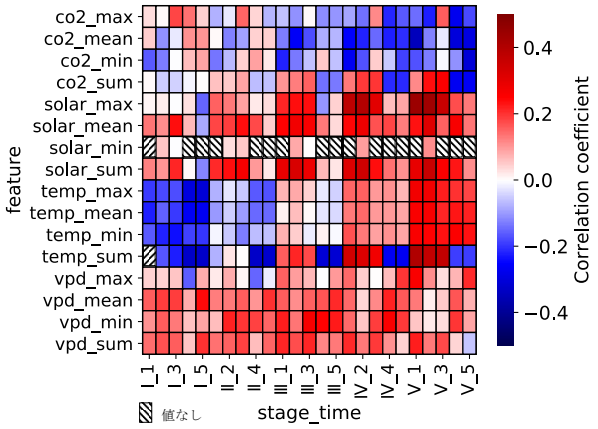

図 2 算出された特徴量と収量との相関

Fig. 2 Correlation between calculated feature quantity and yield

栽培期間の生育ステージⅢから生育ステージⅤの時間帯 1，時間帯 2 の日射量，温度との正の相関が特に高いことが分かる。生育ステージⅢは本実験では，開花時期を想定した生育ステージである。そのため，正の相関が高いことから，トマトは開花後の光合成の促進が収量増加に効果的である （東出 2010）という農学的な知見と一致することが分かる。また，時間帯に着目すると，時間帯 1,時間帯 2 の日射量，温度との相関が高いことが分かる。一般に光合成速度は午前中に最大値に達し，その後徐々に低下するという推移を示す。時間帯 1 は日の出時刻を指標として区分された時間帯である。そのため，日の出により光強度が増加したことで，光合成速度が光飽和点付近まで上昇したことで，光合成の促進に繋がったと考えられる。また，光合成には，光の他に温度，CO_2 濃度が重要である。これらの要因は光合成を行う上でそれぞれ限定要因となるため，時間帯 1，時間帯 2 において，温度との相関が高いという結果については，光と温度が光合成を行うために十分供給されたことで，光合成速度が大きくなり，光合成が促進され収量の増加に繋がったと考えられる。

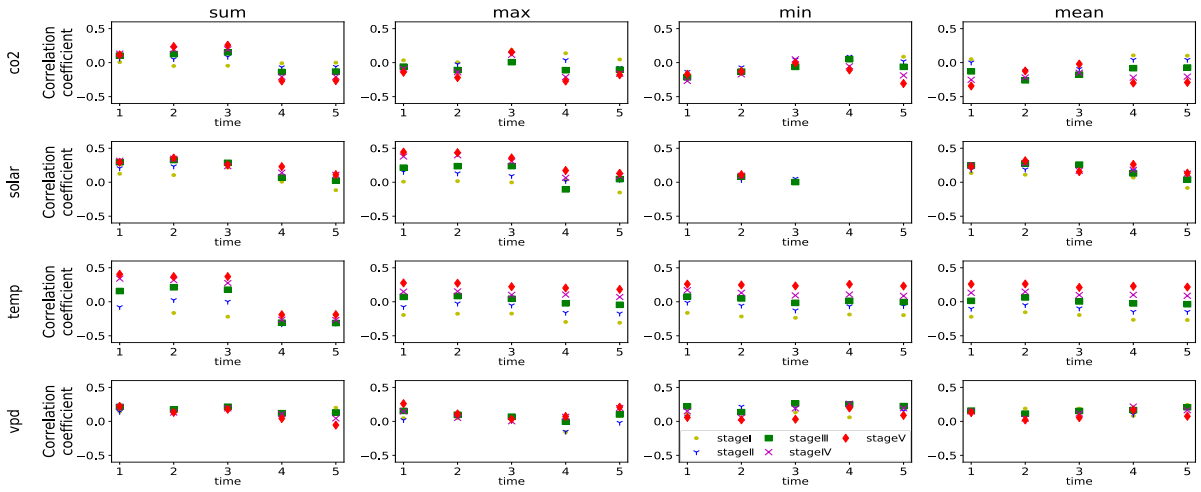

図 3 生育ステージ毎の相関

Fig. 3 Correlation by growth stage

算出した特徴量と収量との相関を，生育ステージ毎に図 3 に示す。図 3 における 2 行目に示す日射量に着目すると，栽培期間を通して時間帯 1 から時間帯 5 の積算値，最大値との相関が高い。特に開花前の積算日射量は収量との相関が高い（Pedregosa et al. 2011）ことが報告されており，開花時期を想定している生育ステージⅢの日射量の積算値との相関が大きい結果は妥当であると考える。また，前述の通り，果実の肥大が進む生育ステージⅣ以降の日射量との相関が高い点からも，算出された特徴量は時間情報の重畳が実現できていると考える。一方，図 3 の 4 行目に示す飽差に着目すると，栽培期間を通して飽差の最小値との相関が大きい事がわかる。つまり，飽差が小さいことで葉の気孔が開いた際の蒸発散速度が上昇し，収量に好影響だったと考えられる。よって，算出された特徴量は植物栽培における生育ステージや一日の時間帯といった時間情報を考慮した特徴量と言え，栽培期間中の環境要因が収穫時品質に与える影響の分析に有効である可能性が示された。

表 2 RFE のパラメータ

Table 2 Parameters of RFE

パラメータ	値
step	1
min features to select	10
cross validation	5
scoring	mae
regressor	XGBoost, LightGBM, Random Forest

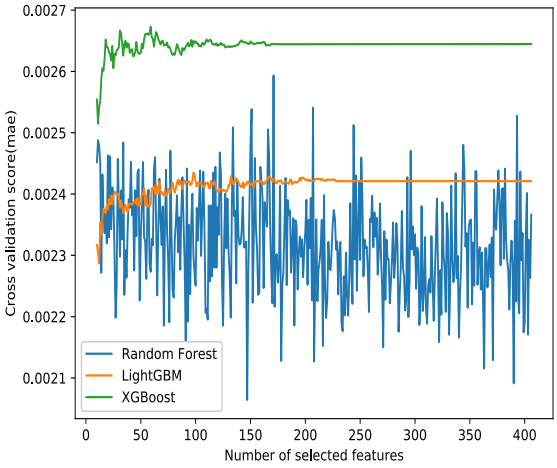

図 4 各学習器の特徴量選択推移

Fig. 4 Feature selection transition of each learner

４－２．特徴量選択の検証

算出された特徴量から，収穫時の品質に影響の大きい特徴量を選択するため，RFE を用いた特徴量選択を行った。RFE にて使用する外部学習機には XGBoost, LightGBM, Random Forest の 3 種類を用いた。RFE の評価には，収量を栽培した株数で除算し，1 株あたりの収量とすることで，生産者の栽培規模の違いを考慮した目的変数とした。RFE の各パラメータを表 2 に，各学習器の特徴量選択推移を図 4 に示す。また，各学習器を用いた特徴量選択結果を図 5 から図 7 に示す。

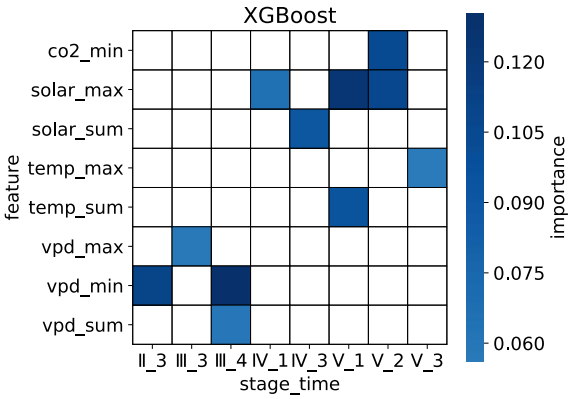

図 5 XGBoost による特徴量選択結果

Fig. 5 Feature selection result by XGBoost

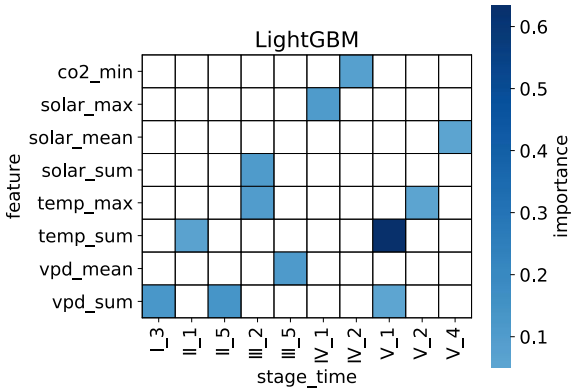

図 6 LightGBM による特徴量選択結果

Fig.6 Feature selection result by LightGBM

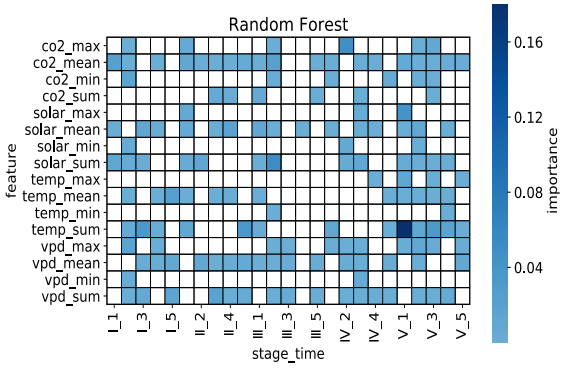

図 7 Random Forest による特徴量選択結果

Fig. 7 Feature selection result by Random Forest

図 4 より，Random Forest の特徴量選択推移が一定の範囲内で発散していることがわかる。Random Forest は無作為に選択された特徴量を用いて決定木を複数作成するアンサンブル学習であるため，特徴量の削減を繰り返すことで，決定木作成時の特徴量の組み合わせパターンが限定され，値が一定の範囲内で発散したと考えられる。図 5 に示す XGBoost による特徴量選択結果から，生育ステージⅤの時間帯 1 と時間帯 2 の日射量が，重要度の高い特徴量として選択されたことが分かる。時間帯 1，時間帯 2 は日の出時刻からの時間帯であるため，生育ステージⅤにおいて日の出による光強度の向上により，光合成速度が大きくなり，光合成が促進されたことが収量増加に繋がったと考えられる。また，同様に生育ステージⅤにおける日の出後の温度の積算値の重要度が高いという結果となった。植物の光合成促進を決定する要因である，温度，光強度，CO_2濃度のそれぞれが限定要因となるため，温度と光強度の両方が高まったことで収量が増加したと考えられる。植物は一般的に光合成速度が正午頃，つまり時間帯 2 に最大となり，その後は徐々に低下するという推移を示すこと （藤澤ら 2010）から，生育ステージⅤの日の出後から昼過ぎの光合成速度の増加が収量に対して重要であるといえる。生育ステージⅤは，

果実が赤くなる完熟期を想定した生育ステージであるため，トマトは緑熟期に収穫時の 7 割程度の糖度を含有し，緑熟期以降に更に糖度が上昇する （石井ら 1994)。そのため，緑熟期として想定した生育ステージⅣ以降の光量を増加させることが，収穫時糖度を向上する要因になると期待できる。

一方，図 6 に示す LightGBM を用いた特徴量選択結果では，生育ステージⅤの時間帯 1 の温度の積算が重要な特徴量として選択された。この結果は，XGBoost による特徴量選択結果と一致するが，日射量の最大値は選択されなかった。また，LightGBM による特徴量選択結果では，生育ステージⅢの午前中の日射量の積算，温度の最大という光合成を促進させる要因が選択され，XGBoost による特徴量選択結果と異なり，生育ステージⅢの環境要因を重要視していることがわかる。

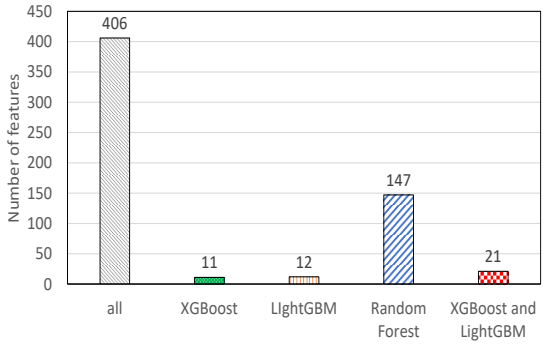

図 8 学習器別の選択された特徴量数

Fig. 8 Number of selected features by learner

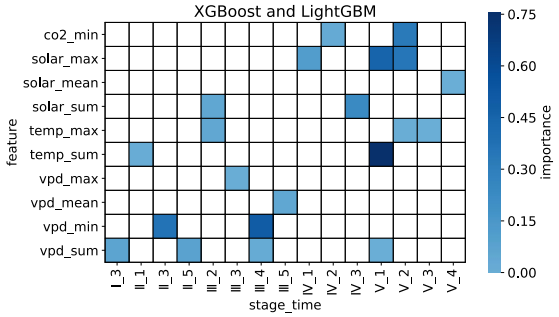

図 9 XGBoost と LightGBM による特徴量選択結果

Fig. 9 Feature selection result by XGBoost and LightGBM

図 7 に示す Random Forest を用いた特徴量選択結果では，147 次元の特徴量が選択され，他の学習器と異なる選択結果が散見された。このような学習器の違いによる特徴量選択結果の違いを考慮するためには，複数の学習器の結果を用いた特徴量選択が有効であると考える。

そのため，学習器毎の特徴量選択結果の違いを考慮できるよう，異なる学習器を用いた RFE の結果を組み合わせた特徴量選択を試みた。図 9 に XGBoost による選択結果の重要度と，LightGBM による選択結果の重要度をそれぞれ正規化し，平均値化した結果を示す。複数の学習器を用いた特徴量選択によって，単一の学習器による特徴量選択時の欠点である結果の信頼性の向上が期待できる。

４－３．選択された特徴量による収量予測

特徴量選択結果の検証として，選択された特徴量を用いた収量の予測性能を評価した。図 5, 図 6, 図 9 で示した各特徴量を用いた予測時の誤差と，全特徴を用いた推定誤差を比較することで，特徴量選択結果の妥当性を検証する。予測する収穫時品質として収量を設定し，栽培者やその栽培規模の違いを正規化するため，収穫開始日から 4 週間後までの 1 週間後ごとの収量 (kg) を株数で割った値を予測対象とした。

データセットに関しては，収集した 54 件の栽培データを，一般的に収穫が終了する収穫開始から 4 週間後まで 1 週間ごとに抽出し，それぞれ予測対象週に応じて生育状態と時間帯分割を行うことでデータの拡張を行った。拡張したデータに対して，予測対象週の偏りが発生しないよう，予測対象週ごとに学習データとテストデータを 8 対 2 の割合で分割し，データセットとした。収穫開始から 3 週間以内に収穫が終了した作に関しては，収穫終了週までのデータを用いた。

評価諸元と学習器のパラメータを表 3 と表 4 に，図 10, 図 11 に特徴量，学習器別の予測誤差を示す。全特徴量を用いた予測時誤差と比較して，RFE を用いて選択した特徴量のみを用いた場合でも同程度の誤差の予測が行えていることが分かる。また，RFE に使用した学習器に着目すると，単一の学習器による特徴量選択結果を用いた RFE (XGBoost) と RFE (LightGBM)では全特徴量を用いた予測と比較して予測精度として，MAE がそれぞれ約 12.4%，約 10.1%，RMSE がそれぞれ約 10.5%，約 8.0%悪化した。精度が悪化した原因としては，図 8 より選択された特徴量数がそれぞれ 11 次元, 12 次元と少

表 3 評価諸元

Table 3 Evaluation specifications

項目	内容，値
目的変数	収量(kg) / 株数(本)
number of cross validations	5
scoring index	RMSE
regressor	LightGBM, Random Forest, SVR, XGBoost

表 4 パラメータ探索範囲

Table 4 Parameter search range

LightGBM	
min data in leaf	100 - 500
number of leaves	10 - 300
max depth	1 - 10
number of estimators	10 - 300
learning rate	0.01 - 0.2
Random Forest	
max depth	100 - 500
number of estimators	100 - 500
SVR	
penalty parameter C	2e-10 - 2e11
kernel	rbf
gamma	2e-10 - 2e11
epsilon	2e-10 - 2e11
XGBoost	
min samples in leaf	100 - 500
number of leaves	10 - 300
max depth	1 - 10
number of estimators	10 - 300
min samples to split	100 - 500

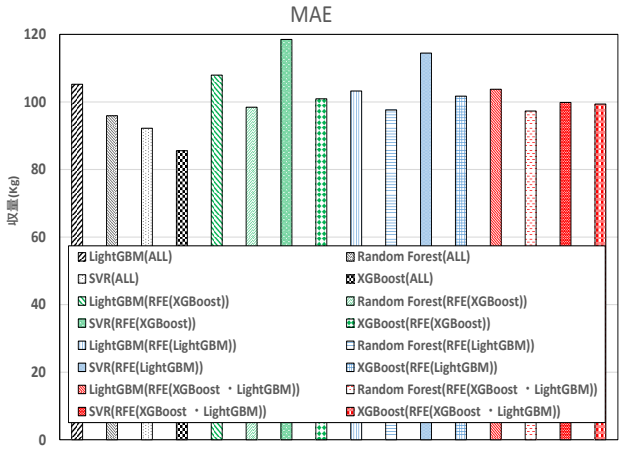

図 10 予測誤差（MAE）

Fig. 10 Prediction error (MAE)

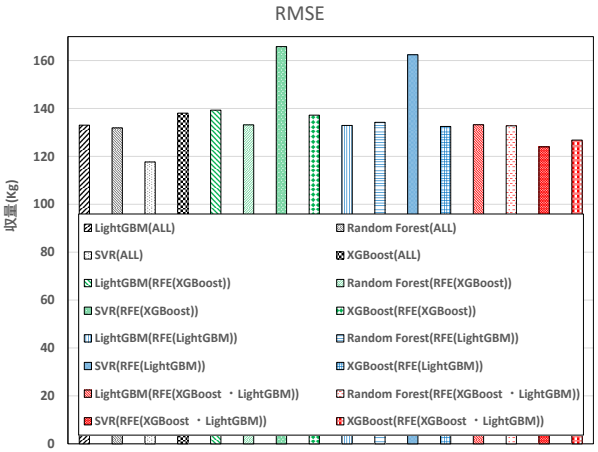

図 11 予測誤差（RMSE）

Fig. 11 Prediction error (RMSE)

なく，特徴量選択時に過学習してしまったと考えられる。しかし，複数の学習器を用いた RFE による特徴量選択では，もとの特徴量から 385 次元の特徴量が失われているにも関わらず，各学習器を平均して MAE が約 5.6%悪化，RMSE が約 0.7%向上と，単一の学習器による RFE と比較して性能を維持できていることが分かる。そのため，複数の学習器を用いた RFE によって選択された 21 次元による単純なモデルでの効率的な予測が行えたと考える。つまり，選択された特徴量は植物収穫時品質への影響の大きい特徴量であると言える。また，一般に熟練農家は植物の生育状態やその時間帯を自身の経験則から総合的に判断して栽培や環境制御を行うため，栽培期間中の特に重要度の高い要因を明らかにできれば，新規就農者の高品質な農作物栽培に役立つものと考える。

５．おわりに

本稿では，植物の栽培期間中の環境要因が収穫時品質与える影響や重要度を栽培データから明らかにするため，植物が成長とともに環境要因やその栽培時間帯が与える影響が異なることに着目し，植物の生育ステージや時間帯を考慮した特徴量を算出し，RFE による機械学習モデルベースの収穫時品質に影響の大きい特徴量選択手法を検討した。本手法を用いて実際に低段密植栽培トマトを栽培した際のデータを分析した結果，特徴量選択では午前中の日射量や温度といった光合成を促進させる要因が選択され，農学的な知見と一致する結果を得ることができた。また，選択された特徴量のみを用いて収穫時の品質である収量を予測した結果，算出された全ての特徴量を用いて予測した際と同程度の精度の結果を得られたことから，選択された特徴量は収量予測時に重要度の高い特徴量であり，この特徴量を考慮した計測や分析，栽培を意識することで新規就農者の栽培支援に役立てられると考える。

今後は，データセットの拡充だけでなく，灌水や降雨といった植物に対して影響の高い外部要因を含めた分析を行うことで，収穫時の品質予測の精度向上を目指す。特に栽培期間中の植物の生育状態を記録することで，より正確に生育状態を考慮した特徴量を検討できると考える。また，他の圃場にも本手法を適用することで，汎用性の評価も行うとともに目的変

数を糖度など他の収穫時の品質でも有効であるか検証を進めていく。

謝辞

本研究は JST さきがけ（JPMJPR15O5）の支援を受け実施されたものである。また，本研究にあたり，テラスマイル株式会社の金田千広様には農学的な観点からのアドバイスを頂いた。ここに感謝の意を示す。

引用文献

Bendel, R.B., & Afifi, A.A. 1977. Comparison of stopping rules in forward “stepwise” regression, Journal of the American Statistical Association, Vol.72, No.352, pp46-53.

Breiman, L. 2001. Random Forests, Machine learning, Vol.45, No.1, pp5-32.

Chen, T., & Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System, KDD ’16, pp.785-794.

Define-by-run hyperparameter optimization framework. https://optuna.org/ (アクセス日 2019-06-26).

Drucker, H., Burges, C.J.C., Kaufman, L., et al. 1997. Support vector regression machines, Advances in Neural Information Processing Systems, Vol.9, pp155–161.

藤澤弘幸ほか 2010. JM1,JM7,JM8 および M.9 台木がリンゴの葉の光合成速度に及ぼす影響, 園芸学研究, Vol9, No.2, pp.171-176.

Guyon, I., et al. 2002. Gene selection for Cancer classification using support vector machines, Machine learning, Vol.46(1-3), pp.389-422.

浜本浩, 星 岳彦, 山崎敬亮ほか 2010. 3 段取りトマト栽培における群落内補光の時間帯が収量に及ぼす効果と補光の経済性, 植物環境工学, Vol.22, No.2, pp.95-99.

東出忠桐 2010. 開花前の日射に基づいた夏秋トマトにおける週間収量変化の予測, 園芸学研究 別冊, Vol.9, No.1, p13.

東出忠桐 2018. 施設トマトの収量増加を目的とした受光と物質生産の関係の利用, 園芸学研究, Vol.17, No.2, pp.133-146.

Hoerl, A.E., & Kennard, R.W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, Vol.12, No.1, pp55-67.

廣瀬彗 2016. スパースモデリングとモデル選択, 電気情報通信学会誌, Vol.99, No.5, pp392-399 (2016).

石井孝典, 藤野雅丈, 矢ノ口幸夫 ほか 1994. トマト品種の果実成分と熟度の関係, 東北農業研究, Vol.47, No.1, pp.275-276(1994).

水野涼介, 柴田 瞬, 峰野博史 2018. 植物収穫時品質や収量に関連する経時特徴量分析手法の検討, 第 80 回全国大会講演論文集 2018.3, pp.121-122.

望月龍也, 石内伝治, 伊藤喜三男 1999. トマト果実における糖含量およびその栽培・環境条

件に対する安定性の品種間差異, 園芸学会雑誌, Vol.68, No.5, pp.1000-1006.
中西豪太, 水野涼介, 今原淳吾, 前島慎一郎 ほか 2018. 植物収穫時品質に関与する経時特徴量の検討, 情報処理学会マルチメディア, 分散, 協調とモバイル(DICOMO2018)シンポジウム, pp.75-81.
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, Vol.12, pp.2825-2830.
Tibshirani, R. 1996. Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society: Series B (Methodological), Vol. 58, No.1, pp. 267–88.
Wang, D., Zhang, Y., & Zhao, Y. 2017. LightGBM, Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics - ICCBB 2017, ACM Press, pp. 7–11.
Zou, H., & Trevor H. 2005. Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol.67, No.2, pp. 301–320.