厚生労働省科学研究費補助金　食品の安全確保推進研究事業
「ゲノム情報および機械学習を用いた腸管出血性大腸菌の高精度型別モデルの構築」
(20KA3002)
研究分担報告書

研究分担者　伊澤　和輝（東京工業大学　情報理工学院）

研究要旨

腸管出血性大腸菌（EHEC）の高精度なサーベイランスを実現するためには、従来法である MLVA 型よりも詳細かつ安価で迅速な類別法が必要である。本研究では、EHEC のあるペアにおいて MLVA 型の差異から機械学習を用いて SNP 数を予測・類別指標とすることによりこれを実現し、高精度なサーベイランスに役立てることを目指した。機械学習モデルの構築には、東京工業大学が保有するスーパーコンピューター TSUBAME3.0 を用いた。

本研究では、まず日本の代表的な EHEC の血清型である O157 について機械学習モデルの構築を目指した。1 年目には 890 株、2 年目には 764 株のデータを追加し、最終的には合計 1636 株のデータを用いて機械学習モデルの作成を試みた。機械学習アルゴリズムには線形回帰モデル、回帰木モデル、勾配ブースティング回帰木を使用したが、勾配ブースティング回帰木モデルが最も精度が良く、SNP 数を連続値で予測する場合には R 二乗値が 0.8 以上の機械学習モデルを作成することができた。また株のペアの MLVA 型のデータを各 Clade ごとに分割し、各ペアの SNP 数を予測することも試みた。学習・予測の方針として、2 株間の SNP 数を連続値で予測する場合と、近縁株判定の指標である SNP 数 10 以下のペアか否かを予測するカテゴリの予測の場合を比較した。結果として、カテゴリの予測の場合の方が、連続値の予測の場合よりも精度が高かった。

3 年目においては、前年度までに得られていた O157 の機械学習モデル構築に用いた知見を利用し、O26、O111 の MLVA データを用いた学習・予測を試みた。O26 については 585 株、O111 については 285 株のデータを用いて、それぞれ約 17 万ペア、約 4 万ペアの MLVA 型のデータを用いた。前年度までの結果から、機械学習アルゴリズムとして勾配ブースティング法を使用した。学習・予測の方針として、2 株間の SNP 数を連続値で予測する場合と、近縁株判定の指標である SNP 数 10 以下のペアか否かを予測するカテゴリの予測の場合を比較した。結果として、カテゴリの予測の場合の方が、連続値の予測の場合よりも精度が高かった。

今後は本研究で作成した予測モデルを実際に新規株の近縁株判別に使用し、予測モデルの有用性について検討する。

A. 研究目的

腸管出血性大腸菌（enterohemorrhagic Escherichia coli: EHEC）は、国内で年間3,000名以上の感染者が報告され、毎年数名の死者が報告されている公衆衛生上重要な食中毒菌である。そのため、発生源の特定や伝播経路を明らかにするために、高精度なサーベイランス法が必要とされている。

従来のサーベイランスで用いられている分子型別手法（反復配列多型解析法：MLVA 法）はゲノム中に存在する複数のリピート配列のパターンによって菌株を型別する手法であり、迅速かつ安価であるが、ゲノム中の特定部分だけを用いるため、型別能には限界がある。一方、高精度なサーベイランスを実現する手法として、全ゲノム情報を用いた単一塩基多型（SNP）解析が存在するが、高い型別能を有する一方で迅速性や費用面で従来法に劣っている。

本研究では、MLVA 型および菌株情報から、全ゲノムレベルの型別情報を推測するモデルを、人工知能の一種である機械学習を用いて構築することを目指す。

B. 研究方法

1. O157 の機械学習モデル構築

2013年から2021年に分離されたEHEC O157の約1636株についてのMLVA型データと任意の 2 株間の SNP 数のデータ（約 130 万ペア）を研究代表者の李謙一氏から提供いただいた。

全データを用いた学習・予測においては、任意の2株間のSNP数のデータのうち、25%を機械学習モデルの評価用として分割し、残りの 75%を機械学習モデルの構築用のデータとして用いた。

Clade ごとの予測においては、任意の 2 株間の SNP 数のデータのうち、Clade 2、3、7、8 の各 Clade 内のペアのみを抽出した。各 Clade において、25%を機械学習モデルの評価用として分割し、残りの 75%を機械学習モデルの構築用のデータとして用いた。

予測結果として、各株ペア間の SNP 数を直接計算する連続値の予測と、各株ペアが 10 SNP または 20 SNP を閾値とした場合に近縁株であるか否かを予測するカテゴリの予測を行った。

機械学習モデルの構築には東京工業大学が有するスーパーコンピューターである TSUBAME 3.0 の環境を利用した。

連続値予測の最適化関数には平均二乗誤差（squared error）、カテゴリ予測の最適化関数には逸脱度（deviance）を用いた。

2. O26・O111 の機械学習モデル構築

2013年から2021年に分離されたEHEC O26の585株、O111の285株についてのMLVA型データと任意の2株間のSNP数のデータ（それぞれ約17万ペア、約4万ペア）を研究代表者の李謙一氏から提供いただいた。

任意の2株間のSNP数のデータのうち、25%を機械学習モデルの評価用として分割し、残りの 75%を機械学習モデルの構築用のデータとして用いた。

予測結果として、各株ペア間の SNP 数を直接計算する連続値の予測と、各株ペアが 10 SNP または 20 SNP を閾値とした場合に近縁株であるか否かを予測するカ

テゴリの予測を行った。

機械学習モデルの構築には東京工業大学が有するスーパーコンピューターである TSUBAME 3.0 の環境を利用した。

連続値予測の最適化関数には平均二乗誤差（squared error）、カテゴリ予測の最適化関数には逸脱度（deviance）を用いた。

C. 研究結果

1. O157 株ペアの各 MLVA 座位の差異の有無を学習データに用いた機械学習モデル

任意の 2 株間の SNP 数のデータで指定された 2 株において、各 MLVA 座位の差異の有無（17 座位）を特徴量として用い、回帰木および勾配ブースティング回帰木のアルゴリズムを利用して機械学習モデルを構築した。

回帰木のアルゴリズムを利用した結果を図 1 に示す。学習時のパラメーターとして回帰木の深さを深さ 2、深さ 5 を利用した。結果、深さ 2 の回帰木モデルでは、RMSE が 91.1、深さ 5 の回帰木モデルでは RMSE が 69.0 となった。これは直感的には各ペアの SNP 数の実測値に対し、深さ 2 の回帰木では 91 個、深さ 5 の回帰木モデルでは 69 個程度、SNP 数がずれた予測を行なっていることを示している。

また勾配ブースティング回帰木のアルゴリズムを利用した結果を図 2 に示す。学習時のパラメーターとして、勾配ブースティング回帰木の深さ 3 を利用した。RMSE が 61.0、R^2 値は 0.87 となり、回帰木のアルゴリズムを用いた場合よりも予測精度が向上した。

2. O157 株ペアの各 MLVA 座位データを学習データに用いた機械学習モデル

任意の 2 株間の SNP 数のデータで指定された 2 株において、2 株の各 MLVA 座位データ（34 座位）を特徴量として用い、線形回帰および勾配ブースティング回帰木のアルゴリズムを利用して機械学習モデルを構築した。

線形回帰のアルゴリズムを利用した場合、実測値と予測値の関係は図 3 のようになり、RMSE 値は 88.8 となった。

また勾配ブースティング回帰木のアルゴリズムを利用した結果を図 4 に示す。学習時のパラメーターとして、勾配ブースティング回帰木の深さ 3 を利用した。RMSE が 22.9、R^2 値は 0.98 となり、線形回帰のアルゴリズムを用いた場合よりも予測精度が向上した。

3. O157 株ペアの SNP 数を連続値で予測する機械学習モデル

任意の株ペアにおいて、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無、分離地の緯度・経度情報を特徴量として用い、勾配ブースティング回帰木のアルゴリズムを利用して機械学習モデルを構築した。

この結果を図 5 に示す。連続値の予測においては、Clade 2 では二乗平均平方根誤差（RMSE）は 3.8 となり、これは直感的には Clade 2 内の各ペアの SNP 数の実測値に対し±4 ヶ所程度増減した予測が行われていることを表す。同様に Clade 3 では RMSE が 4.8、Clade 7 では RMSE が 35.6、Clade 8 では RMSE が 4.9 となった。

また、近縁株の基準を 10 SNP、20 SNP

とした場合の混同行列を図 6、7 に示す。

再現率（Recall）は、実測値から近縁株と判定される株ペアのうち、どの程度を予測から近縁株と判定できるかを表した数値であり、本研究で最も重要視している数値である。

Clade 2、3、8 でも、近縁株の基準を 10 SNP から 20 SNP に広げると再現率が上昇していた。一方、Clade 7 では他の Clade に比べて著しく再現率が低かった。

4. O157 株ペアを近縁株か否かのカテゴリで予測する機械学習モデル

任意の 2 株において、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無、分離地の緯度・経度情報を特徴量として用い、勾配ブースティング決定木のアルゴリズムを利用して機械学習モデルを構築した。

近縁株の基準を 10 SNP、20 SNP とした場合の混同行列を図 8、9 に示す。

カテゴリの予測においてはどの Clade においても連続値の予測の場合よりも再現率が上昇しており、特に連続値の予測では難しかった Clade 7 における再現率が著しく上昇た。

5. O26 株ペアの SNP 数を連続値で予測する機械学習モデル

O26 の任意の株ペアにおいて、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無を、分離日間隔を特徴量として用い、勾配ブースティング回帰木のアルゴリズムを利用して機械学習モデルを構築した。この結果を図 10 に示す。連続値の予測においては、二乗平均平方根誤差（RMSE）は 44.8 となり、これは O26 血清型内の各ペアの SNP 数の実測値に対し平均的に±45 ヶ所程度増減した予測が行われていることを表す。

また、連続値予測において、近縁株の基準を 10 SNP 以内とした場合の混同行列を図 11 に示す。

再現率（Recall）（赤字）は、実測値から近縁株と判定される株ペアのうち、どの程度を予測から近縁株と判定できるかを表した数値であり、本研究で最も重要視している数値である。

O26 においての連続値予測においては再現率は 61.4%となり、O157 において Clade を分けて学習・予測した場合に比べて再現率が低かった。

6. O26 株ペアを近縁株か否かのカテゴリで予測する機械学習モデル

O26 血清型の任意の 2 株において、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無、分離日間隔を特徴量として用い、勾配ブースティング決定木のアルゴリズムを利用して機械学習モデルを構築した。

近縁株の基準を 10 SNP、20 SNP とした場合の混同行列を図 11 に示す。

カテゴリの予測においてはどちらの近縁株基準においても連続値の予測の場合よりも再現率が上昇していた。

7. O111 株ペアの SNP 数を連続値で予測する機械学習モデル

O111 の任意の株ペアにおいて、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無を、分離日間隔を特徴量とし

て用い、勾配ブースティング回帰木のアルゴリズムを利用して機械学習モデルを構築した。この結果を図 12 に示す。連続値の予測においては、二乗平均平方根誤差（RMSE）は 38.7 となり、これは O111 血清型内の各ペアの SNP 数の実測値に対し平均的に±39 ヶ所程度増減した予測が行われていることを表す。

また、連続値予測において、近縁株の基準を 10 SNP 以内とした場合の混同行列を図 13 に示す。

O111 においての連続値予測において再現率は 27.7%となり、O157 において Clade を分けて学習・予測した場合や O26 において学習・予測した場合に比べて著しく再現率が低かった。

8. O111 株ペアを近縁株か否かのカテゴリで予測する機械学習モデル

O111 血清型の任意の 2 株において、2 株の各 MLVA 座位データ（34 座位）、*stx1,2* 遺伝子の有無、分離日間隔を特徴量として用い、勾配ブースティング決定木のアルゴリズムを利用して機械学習モデルを構築した。

近縁株の基準を 10 SNP、20 SNP とした場合の混同行列を図 13 に示す。

カテゴリの予測においてはどちらの近縁株基準においても連続値の予測の場合よりも再現率が上昇していた。

D. 考察

1. O157 の機械学習モデルについて

各 MLVA 座位の差異の有無を特徴量として用いた場合、および各 MLVA 座位データを特徴量として用いた場合の両者において勾配ブースティング回帰木を用いた場合が最も精度が良かった。これは線形回帰や回帰木に比べ、勾配ブースティング回帰木のアルゴリズムが MLVA 型からの SNP 数の予測に適している可能性を示唆する。

また、各 MLVA 座位の差異の有無では特徴量として 17 座位のデータのみを用いていたが、各 MLVA 座位データを用いた場合では任意の 2 株の MLVA 座位全て（17 座位×2 = 34 座位）を用いることで予測精度が向上したと考えられる。各 MLVA 座位データを用いた場合の予測において、特徴量の重要度を比較すると、2 株間で対称的でない部分があり、MLVA 座位間に相互的な関係性が存在する可能性がある。

株全体からの学習・予測においては、最終的に RMSE が 22.9、R^2 値が 0.98 となる高精度な予測モデルの作成に成功したが、これは各株ペアの全体に対する予測精度である。本研究では、今後、近縁株の指標として 2 株間の SNP 数の差異が 10SNP 以下とするが、10SNP 以下の株のペアのみに着目した場合には予測精度は 3 割程度となり、十分な予測精度とは言えなかった。

そこで、本研究では O157 の株ペアを各 Clade に分けて学習・予測を行った。株ペアの SNP 数を連続値で予測する機械学習モデルの場合、Clade 7 での予測精度が他の Clade に比べて悪かった。これは、Clade 7 のデータセットには他の Clade ではそれほど多くない 200 SNP 以上の株ペアデータが多かったことが原因であると考えられる。連続値の予測においては、株ペアデータ全体に対して SNP 数の予測が最適化

されるため、200 SNP 以上のペアの学習・予測にあった最適化がなされることになる。この結果、RMSE が 36 程度と大きくなり、近縁株の閾値を大きく超えたため、近縁株の予測精度が悪かったと考えられる。

一方、カテゴリの予測では Clade 7 においても 60%以上の再現率が見られた。こちらの予測では、近縁株か否かの○×問題を解く学習・予測のため、データセットの中で 1%以下の近縁株についても、今回用いた特徴量から学習・予測が可能であったと考えられる。

2. O26・O111 の機械学習モデルについて

O26・O111 においては株ペアの SNP 数を連続値で予測する機械学習モデルの場合、O157 で Clade を分けて 予測した場合よりも予測精度が悪かった。これは、O157 においては Clade を分けることで、Clade 間ペアの SNP 数が大きいと思われるデータを学習・予測データから除くことができたことに起因している。現状では O26・O111 血清型においてはこのような Clade の別はなく、図 10、12 からもわかるように 2 株間で SNP 数が 1000 前後の遠縁株間データが含まれている状態である。また O26・O111 血清型においては、O157 今後に比べて既存のデータが少ないことも予測精度低下の原因と考えられる。今後、O26・O111 血清型のデータの追加により部分的に機械学習・予測の精度が上がることが期待される。

一方、カテゴリの予測においては、O26・O111 においても 75%以上の再現率が見られた。こちらの予測では、近縁株か否かの学習・予測のため、データセットの少なさ、また遠縁株が含まれたデータセットであっても、今回用いた特徴量から学習・予測が十分に可能であったと考えられる。

E. 結論

1. O157 の機械学習モデルについて

本研究では、2013 年から 2021 年に分離された国内 EHEC O157、1636 株についての MLVA 型データと任意の株ペアの SNP 数のデータから、MLVA 座位、*stx1,2* 遺伝子の有無、分離地の緯度・経度情報を特徴量として株ペアの SNP 数を予測する機械学習モデルの作成を試みた。

勾配ブースティング回帰木のアルゴリズムを用いた機械学習モデルは R^2 値で 0.98 を示し、高精度なモデルとなった。このことは MLVA 型と 2 株間の SNP 数の間に非線形の関係性があることを示唆している。

また Clade ごとにデータを分けた場合の連続値の学習・予測においては、特に Clade 7 において、SNP 数の大きい株ペアのデータに学習・予測全体が影響を受け、近縁株の予測がうまくいかない部分が見られた。

一方で、カテゴリでの学習・予測においては、Clade 7 においても精度良く近縁株を予測することができた。そのため、今後はカテゴリでの予測モデルを用いて、本機械学習モデルの実応用性を検討したい。

2. O26・O111 の機械学習モデルについて

本研究では、2013 年から 2021 年に分離された国内 EHEC O26・O111 血清型株についての MLVA 型データと任意の株ペア

のSNP数のデータから、MLVA座位、*stx1,2*遺伝子の有無、分離日間隔を特徴量として株ペアのSNP数及び近縁株か否かのカテゴリを予測する機械学習モデルの作成を行った。

連続値の学習・予測においては、近縁株の予測がうまくいかない部分が見られたが、カテゴリでの学習・予測においては、どちらの血清型においても精度良く近縁株を予測することができた。そのため、今後はカテゴリでの予測モデルを用いて、本機械学習モデルの実応用性を検討したい。

F. 健康危険情報

なし

G. 研究発表

1) 誌上発表

なし

2) 学会発表

MLVA結果と機械学習モデルを用いた腸管出血性大腸菌の遺伝的距離の予測
伊澤和輝、李謙一、泉谷秀昌、伊豫田淳、大西真、明田幸宏
（第42回日本食品微生物学会学術総会・2021年9月21日（火）～10月20日（水））

H. 知的財産権の出願・登録状況

1. 特許取得

なし

２. 実用新案登録

なし

図 1. 各 MLVA 座位の差異の有無を特徴量として回帰木を利用したモデル

横軸は SNP 数の実測値、縦軸は SNP 数の予測値を示す。

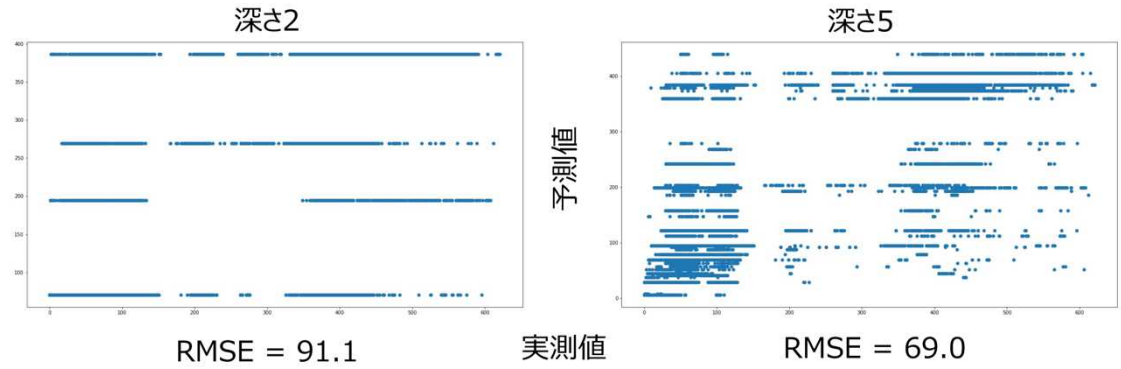


図 2. 各 MLVA 座位の差異の有無を特徴量として勾配ブースティング回帰木を利用したモデル

横軸は SNP 数の実測値、縦軸は SNP 数の予測値を示す。

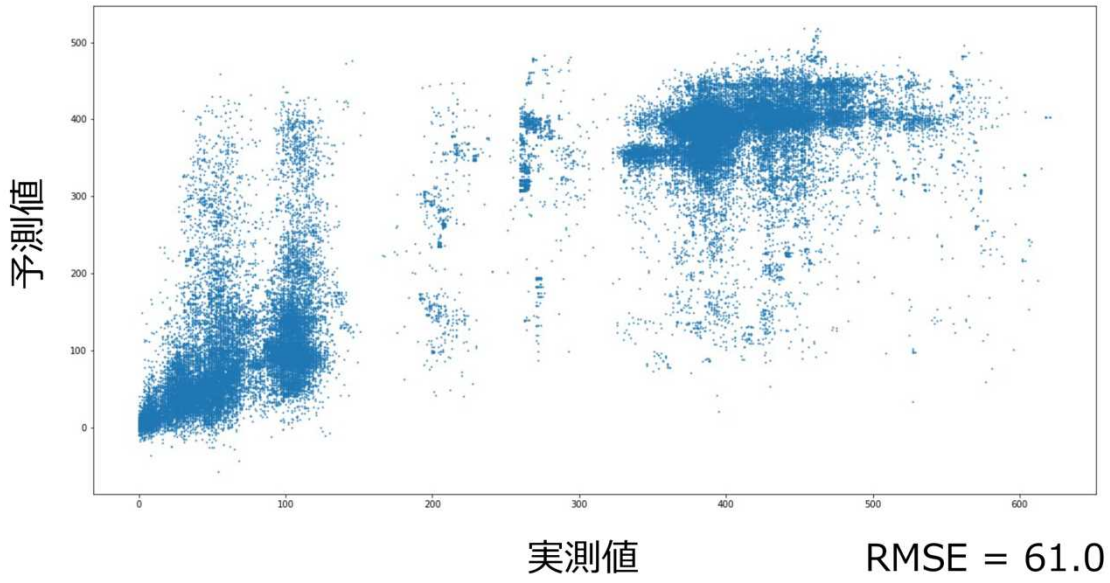


図 3. 各 MLVA 座位を特徴量として線形回帰を利用したモデル

横軸は SNP 数の実測値、縦軸は SNP 数の予測値を示す。

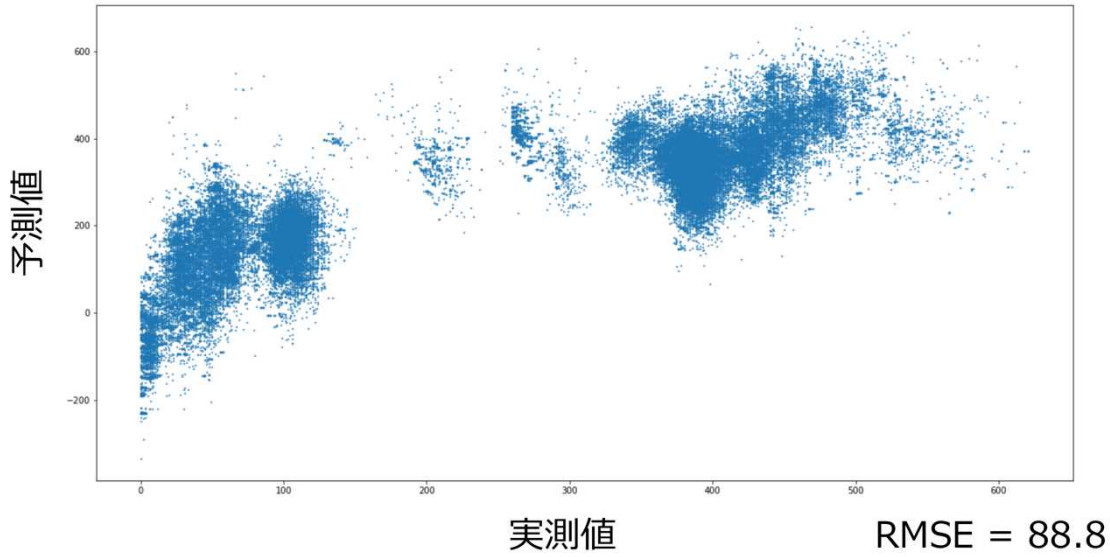


図 4. 各 MLVA 座位を特徴量として勾配ブースティング回帰木を利用したモデル

横軸は SNP 数の実測値、縦軸は SNP 数の予測値を示す。

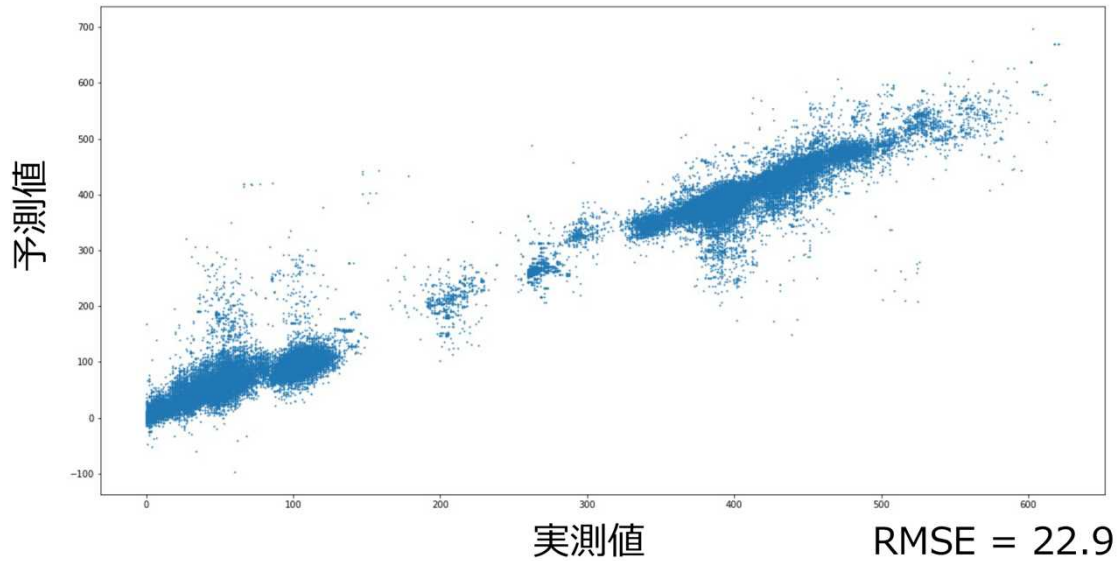


図 5. 連続値予測の機械学習モデルの予測結果

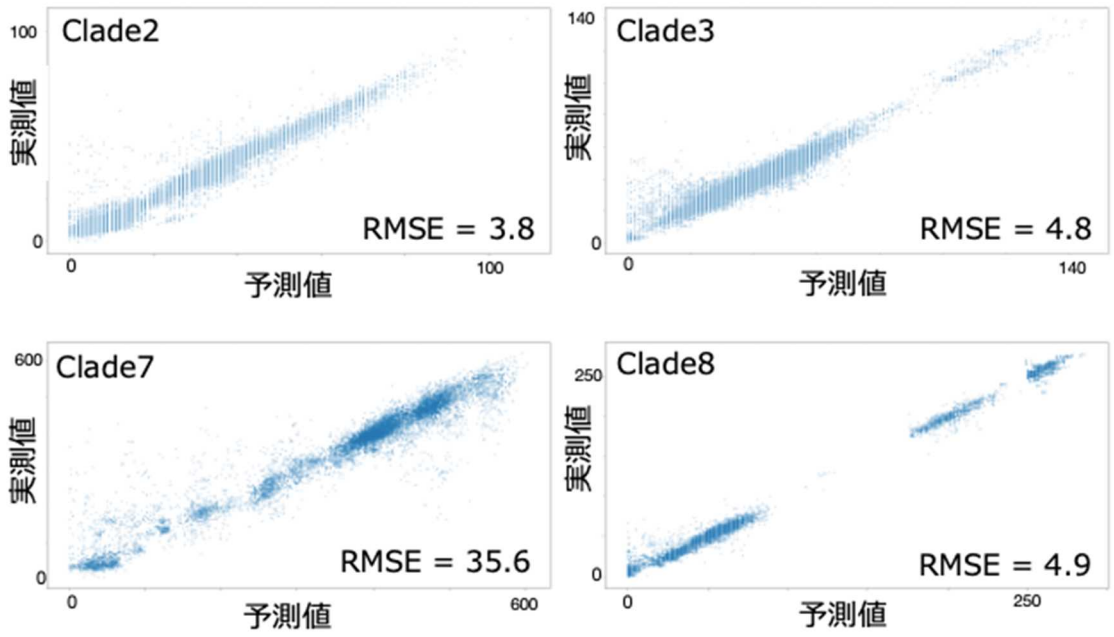


図 6. 連続値予測の機械学習モデルの予測結果（閾値 10 SNP の混同行列）

Clade2

実測値＼予測値	≦10	>10	
≦10	4,489	965	82.3%
>10	554	13,545	
	89.0%		

Clade3

実測値＼予測値	≦10	>10	
≦10	657	407	61.6%
>10	34	26,574	
	95.1%		

Clade7

実測値＼予測値	≦10	>10	
≦10	0	82	0%
>10	3	14,837	
	0%		

Clade8

実測値＼予測値	≦10	>10	
≦10	1,207	342	77.9%
>10	1	17,904	
	99.9%		

赤:Recall (再現率) , 青:Precision (適合率)

図 7. 連続値予測の機械学習モデルの予測結果（閾値 20 SNP の混同行列）

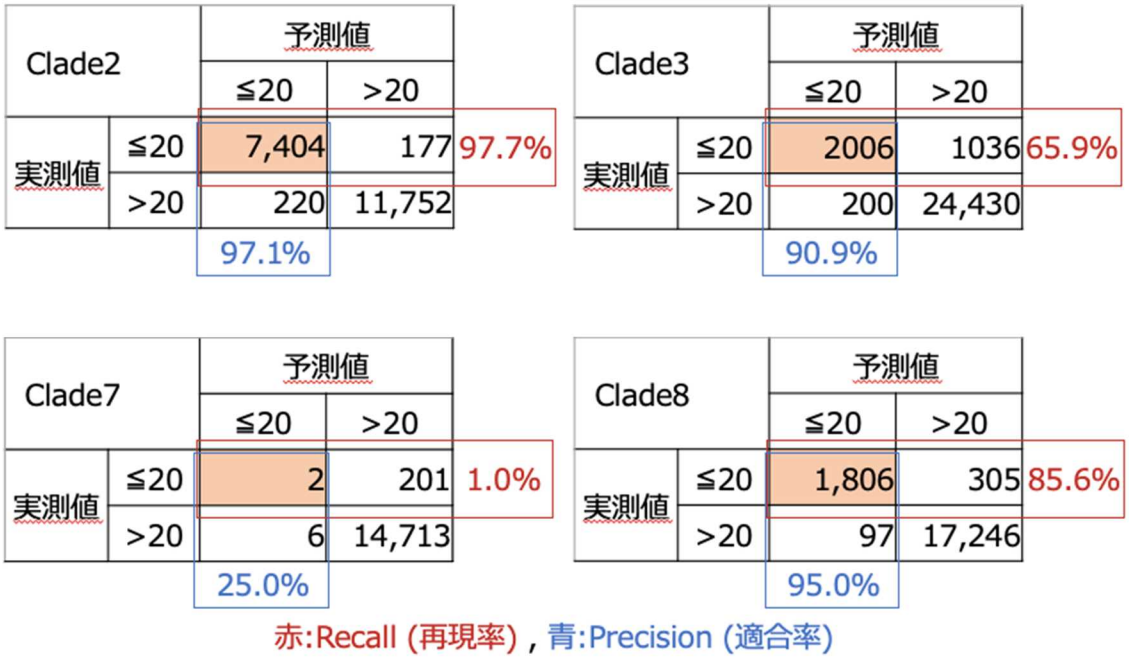


図 8. カテゴリ予測の機械学習モデルの予測結果（閾値 10 SNP の混同行列）

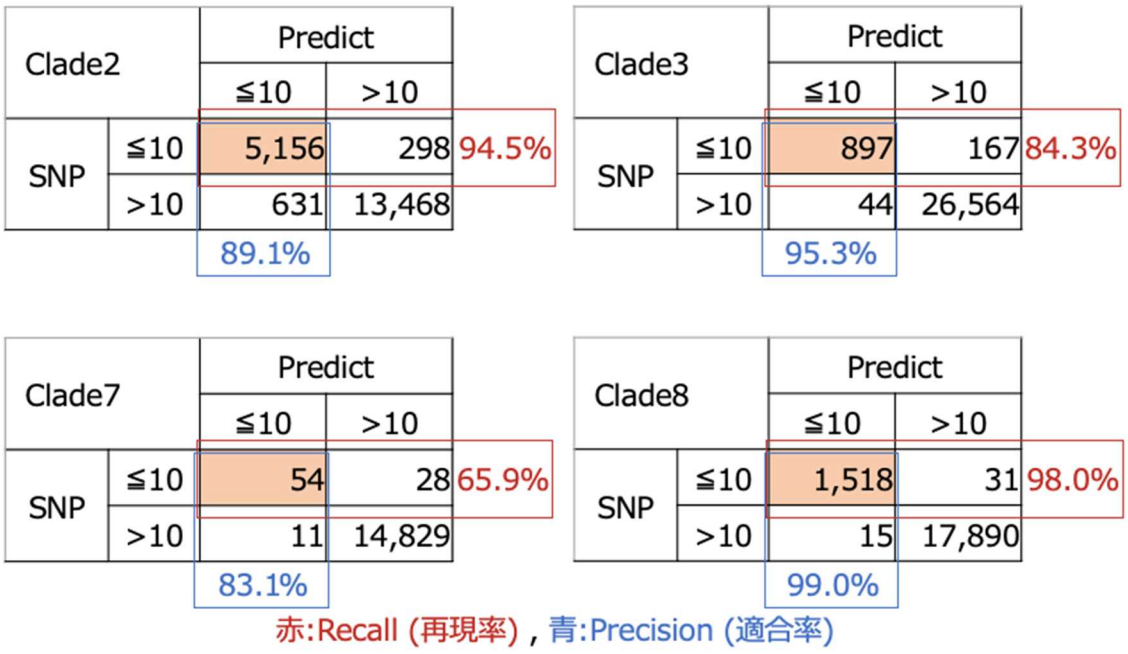


図 9. カテゴリ予測の機械学習モデルの予測結果（閾値 20 SNP の混同行列）

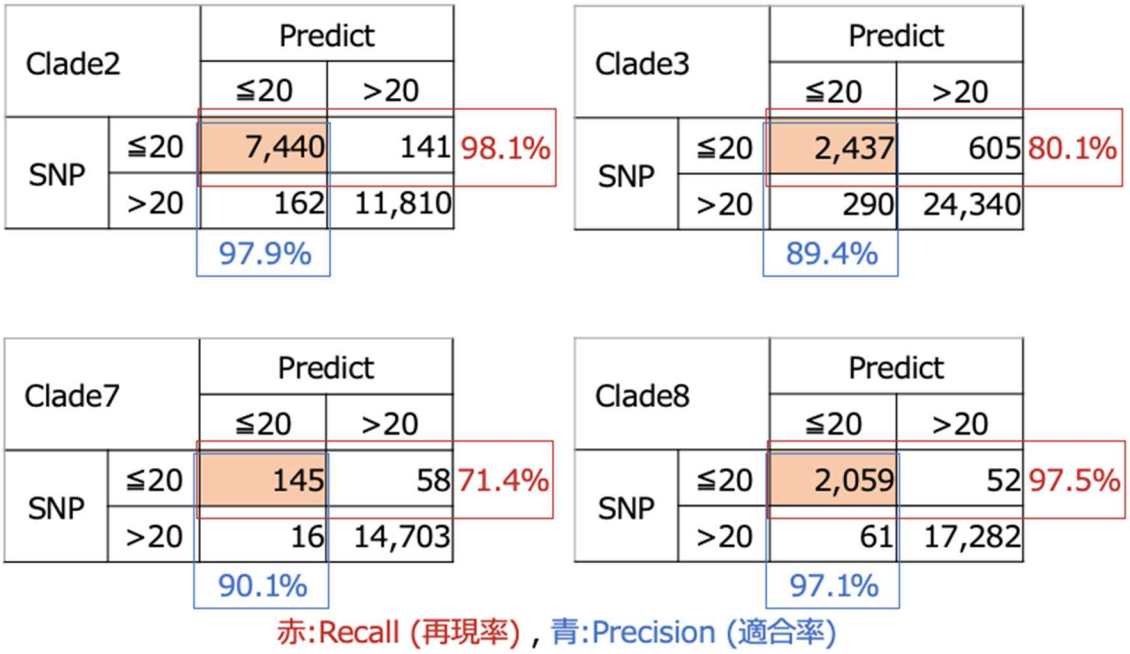
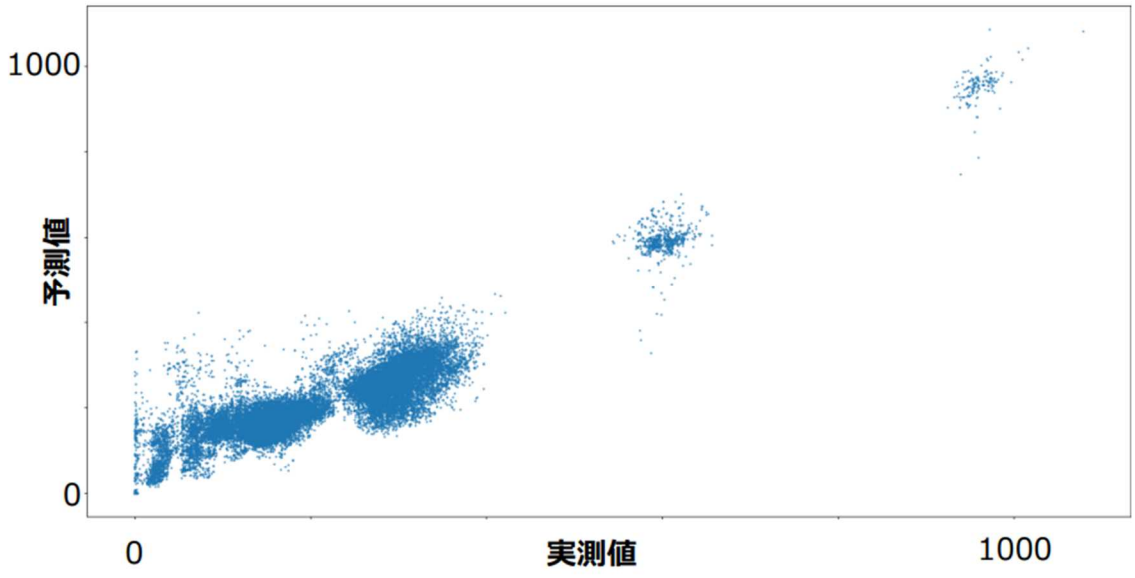


図 10. O26 についての連続値予測の機械学習モデルの予測結果



RMSE=44.8

図 11. O26 についての連続値･カテゴリ予測の機械学習モデルの予測結果（混同行列）

赤: 再現率（Recall）、青:適合率（Precision）

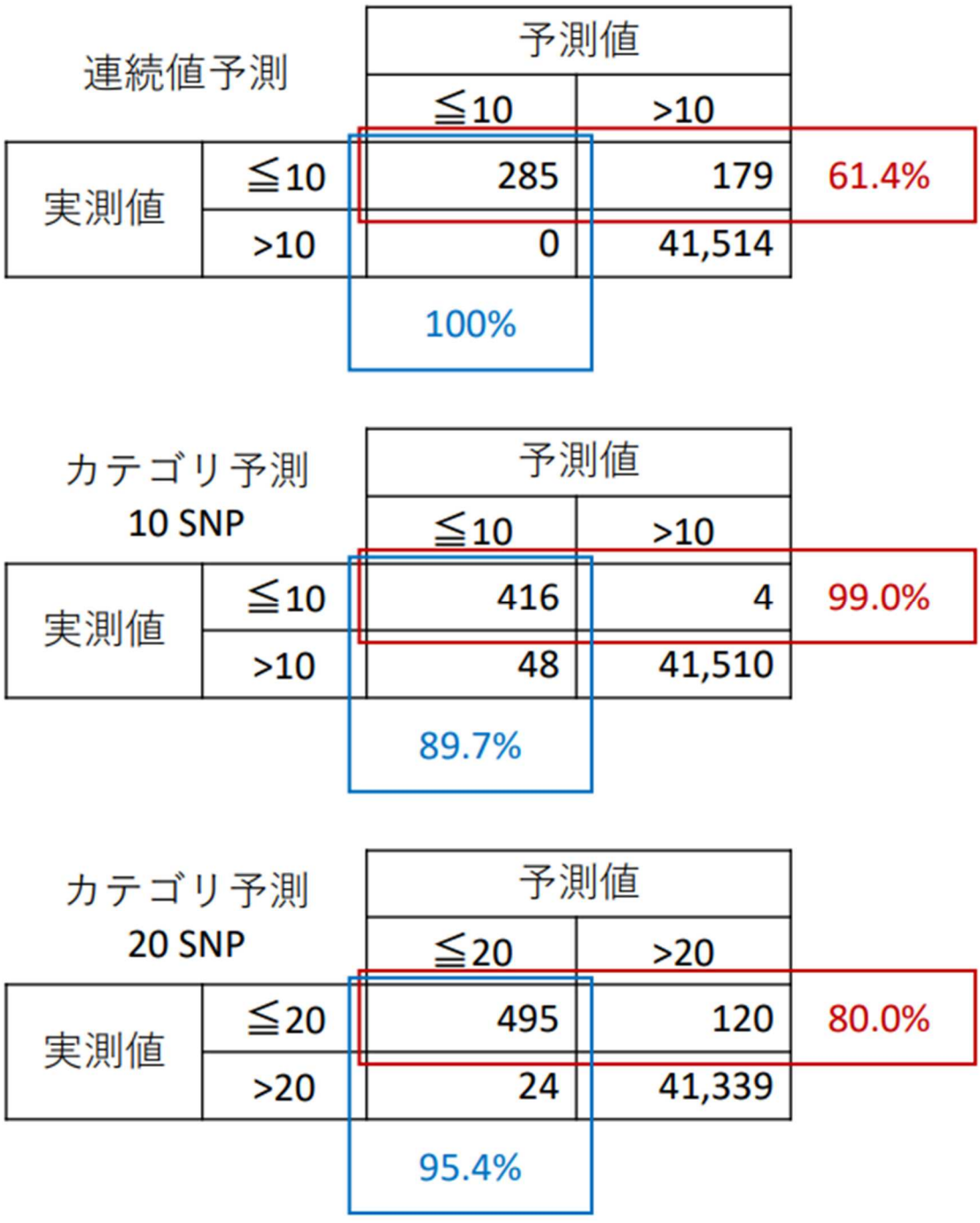
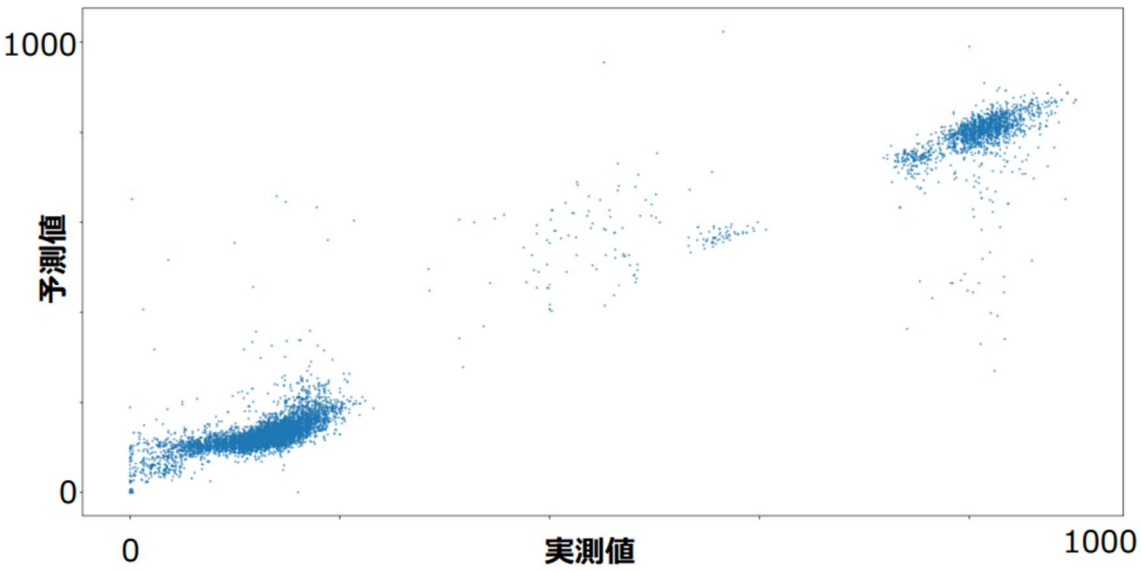


図 12. O111 についての連続値の機械学習モデルの予測結果



RMSE=38.7

図 13. O111 についての連続値･カテゴリ予測の機械学習モデルの予測結果（混同行列）

赤: 再現率（Recall）、青:適合率（Precision）