(一財)日本建設情報総合センター研究助成事業

第 2022-7 号

下水処理施設における機械学習の利活用に関する予測手法の開発

報告書

東京大学　紀　佳淵

2024 年 2 月

目次

研究者紹介

紀　佳淵（き　かえん）Jiayuan Ji
博士（工学、東北大学、2019 年 9 月）
現職：東京大学　未来ビジョン研究センター　特任講師

主な論文

1） Ji, J., Du, R., Ni, J., Chen, Y., Hu, Y., Qin, Y., Hojo, T., & Li, Y. Y. (2022). Submerged anaerobic membrane bioreactor applied for mainstream municipal wastewater treatment at a low temperature: Sludge yield, energy balance and membrane filtration behaviors. Journal of Cleaner Production, 355.

2） Ji, J., Chen, Y., Hu, Y., Ohtsu, A., Ni, J., Li, E., Sakuma, S., Hojo, T., Chen, R., & Li, Y.-Y. (2021). One-year operation of a 20-L submerged anaerobic membrane bioreactor for real domestic wastewater treatment at room temperature: pursuing the optimal HRT and sustainable flux. Science of the Total Environment, 775, 145799.

3） Ji, J., Ni, J., Ohtsu, A., Isozumi, N., Hu, Y., Du, R., Chen, Y., Qin, Y., Kubota, K., & Li, Y.-Y. (2021). Important effects of temperature on treating real municipal wastewater by a submerged anaerobic membrane bioreactor: Removal efficiency, biogas, and microbial community. Bioresource Technology, 336(March), 125306.

4） Ji, J., Sakuma, S., Ni, J., Chen, Y., Hu, Y., Ohtsu, A., Chen, R., Cheng, H., Qin, Y., Hojo, T., Kubota, K., & Li, Y. Y. (2020). Application of two anaerobic membrane bioreactors with different pore size membranes for municipal wastewater treatment. Science of the Total Environment, 745, 140903.

5） 五十棲直子, **紀佳淵**, 李玉友, 嫌気性 MBR を用いた実下水のメタン発酵処理に及ぼす温度の影響, *土木学会論文集*, 2020 年 76 巻 7 号, III_173-III_179.

主な受賞

- 2023.10.24, Young Water Professional Award, 国際水協会.
- 2023.03.03, JSPS HOPE フェロー, 日本学術振興会
- 2022.05.25, 建設工学研究奨励賞, 財団法人建設工学研究振興会
- 2016.09.26, Professional Master for Sustainable Environment, IELP, 環境科学研究科, 東北大学
- 2016.01.09, 優秀賞, 日本水環境学会第 3 回東北支部研究発表会

１．はじめに

１．１研究背景と課題形成

伝統的な下水処理方法である従来の活性汚泥法は、100 年以上にわたって使用され発展し続けてきた（Jenkins and Wanner, 2014）。活性汚泥法は、安定した処理性能と、都市廃水のような大量の廃水に対応する優れた能力を持ち、信頼できるプロセスであることが証明されている。しかし、微生物の代謝に必要な酸素を供給するために曝気が必要である。その結果、曝気プロセスには大量のエネルギーが消費され（Verrecht et al., 2008）、好気性反応によって大量の汚泥が生成し、定期的に除去して余剰活性汚泥とする必要がある（Liu and Tay, 2001）。嫌気性処理は、エネルギー回収を可能し、余剰汚泥量を削減するために、下水処理への適用が期待されている。実験室規模での開発やいくつかのパイロット的な応用はあるものの、本格的な応用にはまだ長い時間がかかると考えられる（Hu et al., 2022）。

機械学習は、アルゴリズムやニューラルネットワークの急速な発展以来、様々な分野で応用されている。最近では、下水処理プラントのモニターシステム、コントロールシステム、さらには生物学的処理プロセスにも活用されている（Ge, 2017; Li et al., 2022; Shi and Xu, 2018）。特に下水処理プロセスへの応用では、革新的なプロセスの開発を加速するのに役立つ。また、従来のモデリング手法では、汎化性の低さや複雑なパラメータ設定といった問題があるため、機械学習を導入することで、下水処理のモデリングに新たなアプローチ（図 1）を確立することも期待される（Bengio et al., 2013）。

上述したように、機械学習アプローチは、より優れたモデリング手法の実装を可能にし、その予測を通じて自治体の下水処理における嫌気性技術の実用化を前進させる。これまでの研究では、都市廃水の処理に使用される嫌気性膜分離活性汚泥の処理性能を予測するために、ディープニューラルネットワークを適用してきた（Li et al. 2022）。しかし、生化学反応結果の評価指標の違いによる平均相対誤差は 2.56～28.23%であったため、予測精度をいかに向上させるかは重要な課題となる。

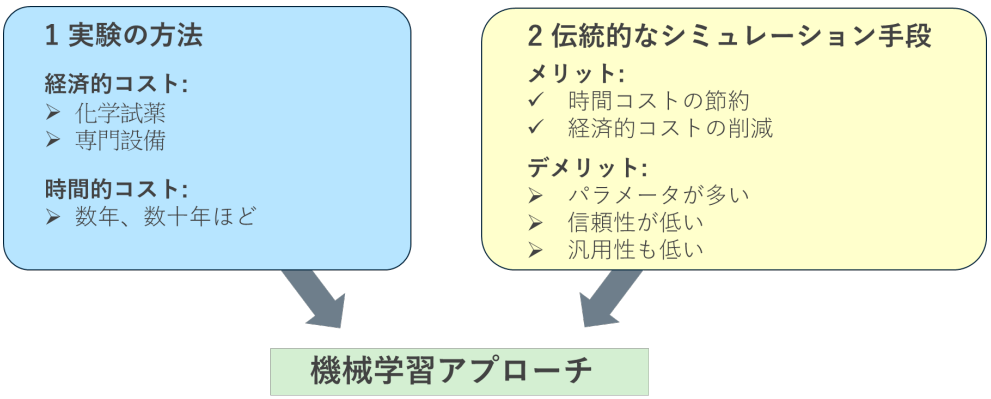

図 1　研究課題形成の概要

１．２研究目的

以上の問題点を解決するために、本研究では、データセットの構築と改善を検討する。具体的には、運転日数の追加と、トレーニングおよびテストのためのデータセットの再調整が含まれる。Extra Trees の機械学習アルゴリズムを、都市廃水の嫌気性処理のモデル化と予測に使用した。本研究により下水下処理プロセスにおける機械学習の適用には、データセットの構築により予測性能を向上させることが期待される。また、重要度分析を行い、各インプットパラメータについて、予測結果に与えられた影響を解明する。

２．研究方法

２．１アンサンブル学習

アンサンブル学習は、複数の学習の弱学習器からの予測を組み合わせることで、モデルの性能を向上させる機械学習アプローチである。アンサンブル学習は、弱学習器を強学習器に集約することで、オーバーフィッティングのリスクを軽減し、モデルの頑健性を向上させ、複雑な問題に対してより良いパフォーマンスを達成する。弱学習器とは、特定の問題においてランダムな推測よりもわずかに優れたモデルを指す。複数の弱学習器を組み合わせることで形成される強学習器は、その問題で優れた性能を発揮し、場合によっては人間の専門家を凌駕することさえある。アンサンブル学習の有効性は、個々の学習者の多様性と密接に結びついている。多様性とは、様々な側面や方向性の違いを意味し、それによってアンサンブルのパフォーマンスが向上する。

一般的なアンサンブル学習法には、ブートストラップ集計法（Bagging）がある。これは、ランダム・サンプリング（置換あり）によって複数のトレーニングセットを生成し、それぞれのセットで弱学習器をトレーニングし、予測値を Voting や Averaging によって結合するものである。もう 1 つの方法はブースティング（Boosting）で、学習サンプルの重みを徐々に調整し、先行学習器によって誤分類されたサンプルを弱め、一連の弱学習器を生成する反復アプローチである。著名な Boosting algorithms には、AdaBoost、Gradient Boosting、XGBoost などがある。

アンサンブル学習の主な長所は、個々の学習器の限界を補い、全体的なパフォーマンスを向上させる能力を持つことにある。このアプローチは通常、実用的な問題解決シナリオにおいて、より確実で正確な予測をもたらす。

２．２決定木モデル

アンサンブル学習は、多くの場合決定木ベースモデルに依存している。決定木は機械学習の基本であり、分類と回帰の両方のタスクに応用されている。その基本原理は、特徴値に基づいてデータセットを再帰的に分割し、木のような構造を形成することである。決定木では、各内部ノードは属性または特徴を表し、各ブランチは決定ルールを表し、各リーフノードは

出力結果に対応する。構築プロセスでは、データセットを再帰的にバイナリ分割する。各ノードでは、分割のために特徴が選択され、データ集合を異なるサブセットに分割する。分割の目的は、各サブセット内のデータの純度を高め、同じクラスのサンプルを近づけることである。ルートノードからリーフノードへの各パスは決定規則を構成する。予測中、入力データは木の経路を横断してリーフノードに到達し、そのノードからの出力がモデルの予測となる。決定木は、その解釈のしやすさ、データの正規化要件からの免除、欠損値の取り扱いにおける柔軟性、分類と回帰タスクの両方への適用性、特定のシナリオにおける非線形関係を捉える能力で知られている。しかしながら、決定木は、特に木の深さがかなり深い場合、オーバーフィッティングを起こしやすい。オーバーフィッティングを軽減するために、枝刈りのような技術を採用することができる。さらに、決定木は異常値に対して敏感であり、過度に複雑な分岐をもたらす可能性がある。

決定木の変種や改良されたアルゴリズムには、ランダムフォレスト、勾配ブースティング木、XGBoost などがあり、複数の決定木を統合することよりモデルの性能を向上させる。エクストリーム・ランダム・ツリー（Extreme Random Trees）は、エクストラ・ツリー（Extra Trees）とも呼ばれ、バギング（Bagging：Bootstrap Aggregating）アンサンブル学習法に属する。バギングでは、ランダムサンプリングによって複数のトレーニングセットを生成し、それぞれのセットで独立した弱学習器をトレーニングし、最終的に Voting または Averaging によってそれらの予測値を結合する。

２．３Extra Tree モデル

エクストリーム・ランダム・ツリー（Extreme Random Trees、Extra Tree）は、決定木の基礎の上に、さらにランダム性を導入したものである。また、従来のランダムフォレストとは異なり、Extra Tree は、ノード分割の際に特徴をランダムに選択するだけでなく、特徴分割の閾値の選択もランダムにすることで差別化を図っている。この拡張により、Extra Tree のベース学習器の多様性が強化され、より大きなモデルのランダム性が導入される。要するに、Extra Tree は、Bootstrap Aggregating の概念とさらなるランダム性を組み合わせたアンサンブル学習アプローチを実現するものである。

２．４Extra Tree と深層学習

これまで、深層学習モデルが AnMBR 型嫌気性リアクターのシミュレーションに採用され、これらのモデルに伴うメリットが明らかにされてきた。しかし、小規模なデータセットでは、深層学習はオーバーフィッティングを起こす可能性がある。ツリーベースモデルとディープラーニングモデルは、機械学習における 2 つの異なるカテゴリーであり、モデル構造、学習メカニズム、適用シナリオに顕著な違いが見られる。

まず、2 つのモデルには構造上の大きな違いがある。Extreme Tree は決定木に基づくアンサンブルモデルであるのに対し、深層学習モデルはニューラルネットワークの構造化され

た組み合わせで構成される。次に、ツリー系モデルは、データセットをより純粋な部分集合に再帰的に分割し、特徴を選択し、不純物を減らす原理に基づいてノードを分割することによって学習する。対照的に、深層学習モデルは通常、ニューラルネットワーク内の重みを訓練し最適化するためにバックプロパゲーション・アルゴリズムを採用する。したがって、深層学習モデルの学習には、多くの場合、膨大な量のデータと計算リソースが必要となる。深層学習モデルが自動的に特徴を学習する能力を備えているのに対し、ツリー系モデルは一般的に明示的に特徴学習を行わないことに注意する必要がある。ディープ・ネットワークでは、モデルは階層的学習によって高レベルの特徴を抽出し、組み合わせることができる。そのため、ツリー系モデルの外挿能力は、深層学習モデルよりも弱い可能性がある。実用的なアプリケーションでは、ツリー系モデルは比較的小さなデータセットでよく機能することが多く、分類と回帰の両方のタスクに適している。一方、深層学習は、特に画像処理、音声認識、自然言語処理などの領域において、大規模な非構造化データセットを得意とする。深層学習モデルは、複雑なパターンの捕捉や特徴学習において強固な能力を発揮する。両モデルの特徴と比較は表 1 に示す。

表 1　本研究で使用された機械学習モデル ExtraTree と深層学習の比較

	ExtraTree	他の深層学習
解釈可能性	解釈しやすい	解釈しやすい
データ要件	小規模データセット	大規模データセット
計算要件	時間コストおよびリソース、低コスト	時間コストおよびリソース、高コスト

2．５データセットの構築とワークフロー

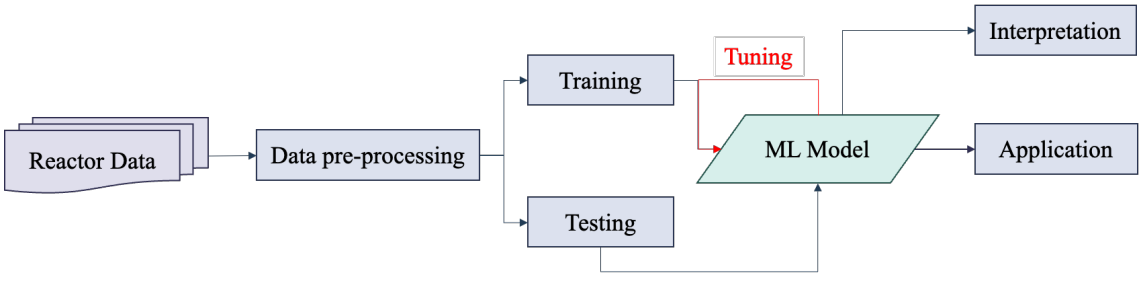

図 2　機械学習実行のフロー

2.4.1 データ取得

データ取得の段階では、実験やセンサーの測定値、処理プロセスのパラメータ、温度、他の情報など、反応器から様々なデータを得る。本研究では、主にリアクター運転時間、リアクター温度(inp1)、流入水温度(inp2)、環境温度(inp3)、流入水 pH(inp4)、流入水 COD 濃度(inp5)、流入水流量(inp6)、流出水 COD 濃度(outp1)、COD 除去率(outp2)などのデータを取り

込む。

2.4.2 データの前処理

欠損値をチェックし、データの前処理を行う。決定木ベースモデルの構造的特徴から、入力データの正規化は必要なく、生データをそのまま入力として使用できる。生データのプレビューは以下の通りである：

表 2　生データのプレビュー

	inp0	inp1	inp2	inp3	inp4	inp5	inp6	outp1	outp2
count	185	185	185	185	185	185	185	185	185
mean	182.94	25.49	17.83	17.41	7.17	376.2	0.2	44.78	87.57
std	101.77	1	5.49	6.4	0.16	88.85	0.06	9.82	3.6
min	2	20.8	8	6.8	6.89	202.88	0.07	19.49	75.82
25%	105	25	12.7	11.5	7.05	314.19	0.15	38.36	85.58
50%	172	25.3	18.6	17.9	7.18	367.71	0.19	43.6	87.97
75%	277	25.7	23	23.1	7.3	436.89	0.23	49	90.15
max	366	30	26.4	27.9	7.62	740.33	0.35	87.3	94.86

AnMBR 型のリアクターの評価指標の一部は正規分布に従わず、複雑な入力と出力の間には線形関係が存在しない。したがって、生データ間の関係を統計的に分析するためにスピアマンの順位相関分析を使用した。スピアマンの順位相関は、2 つの変数間の単調な関係を測定するために使用されるノンパラメトリック手法である。スピアマンの順位相関係数は以下の通りである:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

d_i は該当するデータポイントにおけるランクの差を表し、n はサンプルサイズである。スピアマンの相関は、線形関係ではなく単調関係を測定することに注意する必要がある。結果は次のようになる：

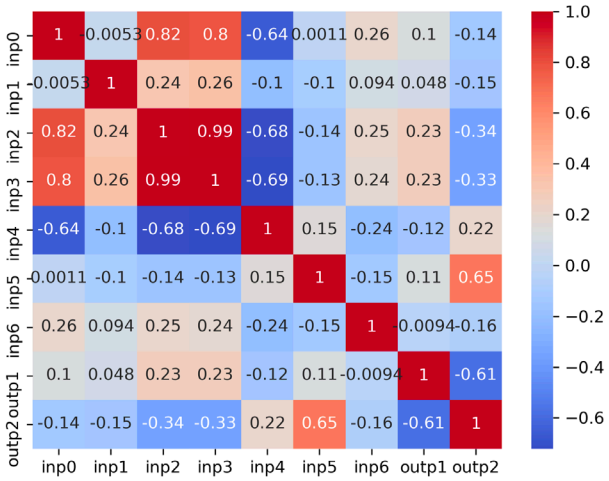

図 3　データのノーマライゼーション結果

2.4.3 データの分割

データセットをトレーニングセットとテストセットに分割することは、モデリング中にモデルの汎化能力を評価するために不可欠である。通常、データの大部分はトレーニングに使用され、テストにはより少ない部分が使用される。先行研究との比較を容易にするため、先行研究のトレーニングとテストのセットを機械学習モデルの構築とテストに利用した。データの分布を下図に示す：

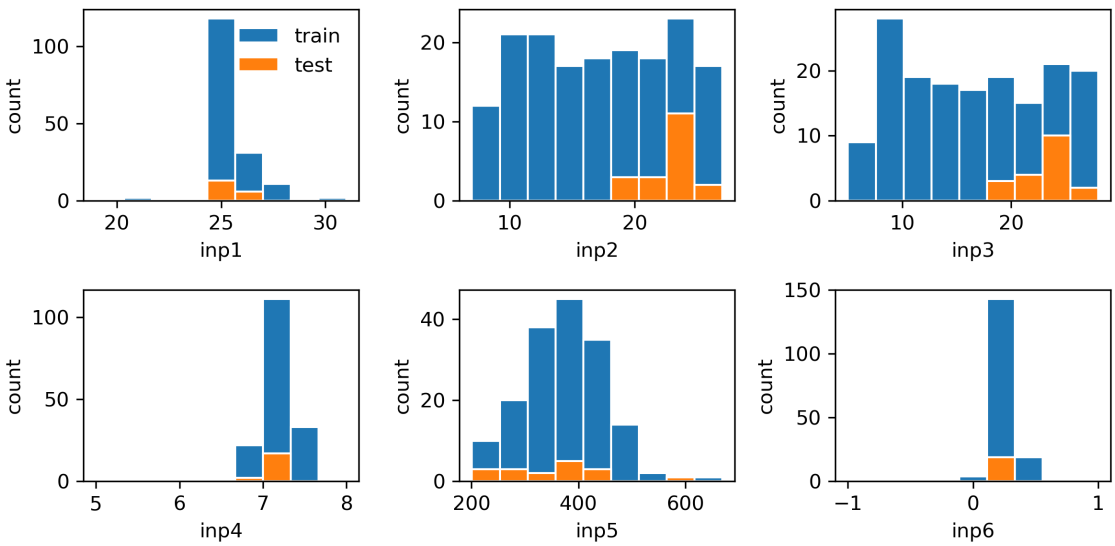

図 4　訓練用とテスト用のデータセットの構築

２．６機械学習モデリング

機械学習モデリングの段階では、モデルの選択、ハイパーパラメータのチューニング、そしてモデルの検証を行う。本研究では、モデルとして Extra Tree を使用し、パラメータ・チューニングを以下のように行った：

最大特徴数：最良の分割を見つける際に考慮する特徴の数を表す。

最大リーフノード数： フォレストの各ツリーにおけるリーフノードの最大数を決定する。リーフノードを制限することで、オーバーフィッティングを防ぎ、モデルの解釈性を高めることができる。リーフノードの上限に達すると、ツリーの成長は停止する。

エスティメーター数： フォレスト内のツリーの数である。各ツリーはトレーニングデータのランダムなサブセットを使用して構築する。推定子数が少ないと、推定子数の多いモデルと比較して予測性能が低下する可能性があるものの、データセットが小さい場合は計算効率が高くなる。

モデルの汎化性を高めるため、ハイパーパラメータのチューニング過程では、10 重クロスバリデーションと平均 2 乗誤差（MSE）が採用された。トレーニングとテストの間、データは均一に分割され、MSE は最小化された。MSE がこれ以上減少しなくなるまで学習を停止し、その後パラメータをモデリングに使用した。

２．７モデルの評価

テストセットを用いて学習済みモデルを評価するため、評価関数が必要である。本研究で利用される評価関数には、平均二乗誤差、平均二乗誤差、絶対誤差、平均パーセント誤差、決定係数が含まれ、以下のように計算した：

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\mathrm{RMSE} = \sqrt{MSE}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

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

MSE (Mean Squared Error) は，実際のオブザベーションy_i とモデル予測$\hat{y}_i$の間の差の 2 乗の平均である。これは，誤差の 2 乗により，より大きな誤差により高い重みを割り当てる。RMSE (Root Mean Squared Error) は，元の変数と同じ単位で表現される MSE の平方根である。より直感的に解釈できる指標を提供しながら、MSE の大きな誤差により大きな重みを与える。MAE (Mean Absolute Error) は、y_iと$\hat{y}_i$の差の絶対値の平均である。より大きな誤差に重みを与える MSE とは異なり、すべての誤差に等しい重みを与える。MAPE (Mean Absolute Percentage Error) は、y_iと$\hat{y}_i$の差の平均パーセンテージである。これは、パーセンテ

ージで平均誤差を測定する。また、決定係数（R^2）は，モデルが説明できる従属変数（目的変数）の分散の比率を表す。R^2 の値は、基本的に 0 から 1 の範囲であり、値が高いほど適合度が高いことを示す。よって、R^2 が 1 であれば、モデルがデータに完全に適合していることを意味とする。

モデルの性能を評価するためにこれらの一般的な指標を使用することに加えて、トレーニングセットの残差分布を評価するために QQ plot（Quantile-Quantile Plots）を採用した。数学と統計学では、残差は観測値と推定値の差である。QQ plot は、データが理論的な分布に適合しているかどうかをチェックするために使用される視覚的なツールである。QQ plot は、観測値の分位数と理論分布から期待される分位数を比較する。QQ plot は，データの正規性を検定するためによく使われる。横軸は期待分位数で，通常，標準正規分布から導かれる。これらの期待分位数は，仮定された理論分布に基づいている。縦軸は，観察された値の実際の分位を表す。データはソートされ，正規化され，対応する期待分位数と比較される。データが仮定された理論分布に適合している場合，QQ plot 上の点は，直線，典型的には 45 度の対角線に密接に一致するはずである。この直線からの偏差は、データの分布が想定された理論分布と一致していない可能性を示唆する。

残差（モデルが予測した値と実際のオブザベーションの差）については、分布はランダム、一様で、明確なパターンを示さないはずである。残差の QQ plot は、モデルの適合が正規分布の仮定に従うかどうかを調べるために使用できる。残差が正規分布している場合、QQ plot 上のポイントは、45 度の対角線にほぼ沿っている。ポイントが顕著な偏差または湾曲を示す場合、それは残差の正規分布の仮定との矛盾を示すかもしれない。

３．結果と考察

３．１ ExtraTrees の結果

本研究に、ExtraTrees を用いた機械学習の予測結果は図 5 に示す。この図では、先行研究にニューラルネットワークを使用した深層学習の予測結果も含めている。本研究に ExtraTrees を用いた予測結果は、先行研究のニューラルネットワークの予測結果と比較して、パフォーマンスが良いという結論が示唆された。また、比較のための定量分析を行うために、RMSE、MAE および R^2 の評価関数も求められ、表 3 に示している。

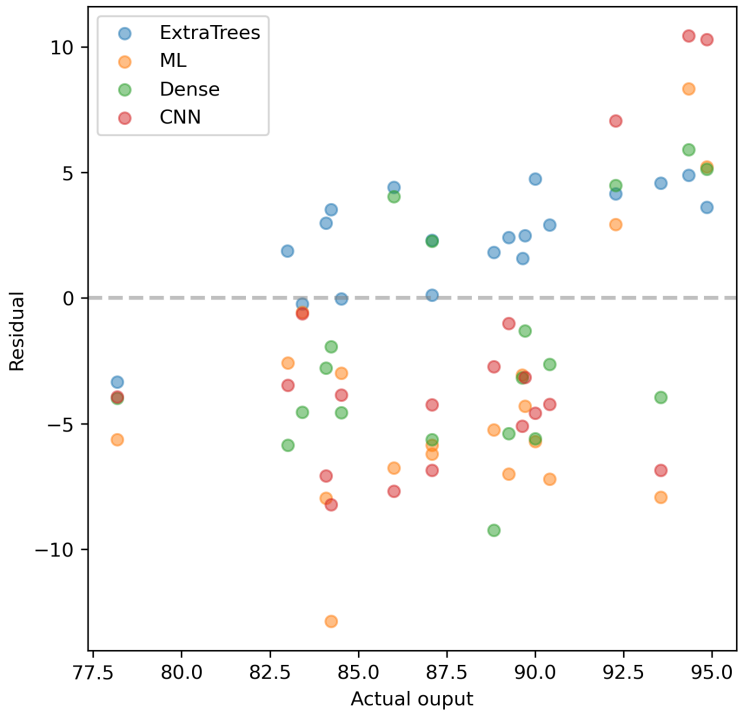

図 5　ExtraTrees とニューラルネットワークを用いた予測結果

表 3　各モデルを用いた予測結果による計算された評価関数

	Basic Machine Leaning Network	Basic Machine Leaning Network	DenseNet	ExtraTrees
MAE	5.71±2.73	5.34±2.77	4.34±1.83	2.74
RMSE	6.29±6.07	5.98±5.69	4.69±4.32	3.12
R^2	1.94 (0.89*)	2.28 (0.89*)	2.52 (0.93*)	0.46

*エラーデータを削除して更新された値。

上表により、ExtraTrees は過去に使用した三種類のニューラルネットワークと比べて、RMSE、MAE による予測パフォーマンスが良いということが明らかにした。しかし、R^2 のほうはニューラルネットワークのパフォーマンスが良いである。これに関して、深層学習を実行した結果は、R^2 の値が 1 に超えたことにより、オーバーフィットの可能性があるため、比較の意義がないと考えられる。

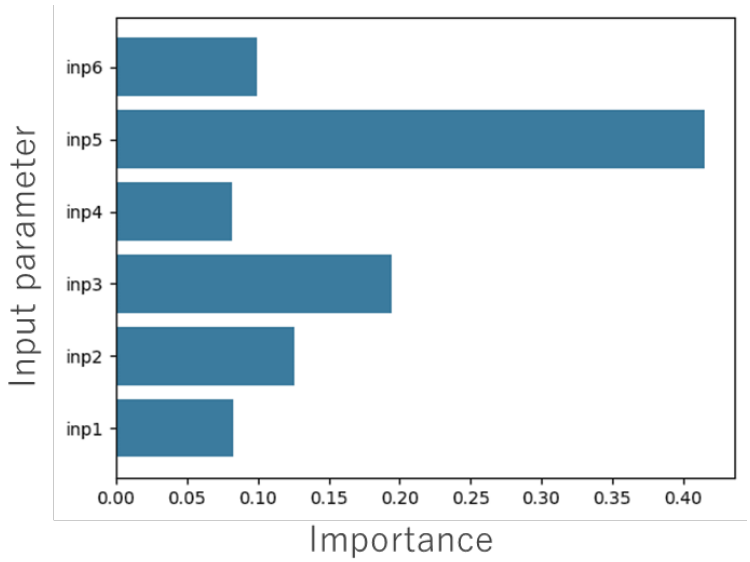

図 6　予測結果に基づいた重要度分析

重要度(importance)が分類や回帰に寄与した特徴量を把握できるため、図 6 に示す。重要度分析の結果より、流入水（すなわち inp5）の COD 濃度が予測結果に最も大きな影響を与えることがわかった。また、環境温度と流入水の温度（inp2 と inp3）も予測効果に大きな影響を与えることを示している。現在、実際の下水処理プラントの運転中、運転温度は一般的に記録されるが、環境温度はあまり測定されていない。本研究は、下水処理プロセスの有機物除去率を予測する場合、環境温度と流入水温度も重要なインパクトであることを示した。

３．２運転日数の配慮

COD 除去率の予測精度を向上させるため、データセットの構築を検討した。下水処理プロセスは連続的なプロセスであるため、まず運転日数をインプットとして追加することを検討した。図 7 と図 8 は、元のデータセットと運転日数を追加した後の QQ Plot の結果をそれぞれ示している。この結果から、両方のデータセットを使用した場合、フィッティングが良好であること、機械学習の性能が高いことが分かった。一方、2 つのデータセットを比較すると、運転日数を増やした場合の方が、元のデータセットよりも優れたことを示した。

元のデータセットと運転日数を追加した後のデータセットをより良くかつビジュアルに比較するために、評価係数を計算し、表 4 に示した。この表により、運転日数を増やした後、R^2 が著しく増加し、MSE、RMSE などの評価指数が一定的な程度に減少していることが分かった。

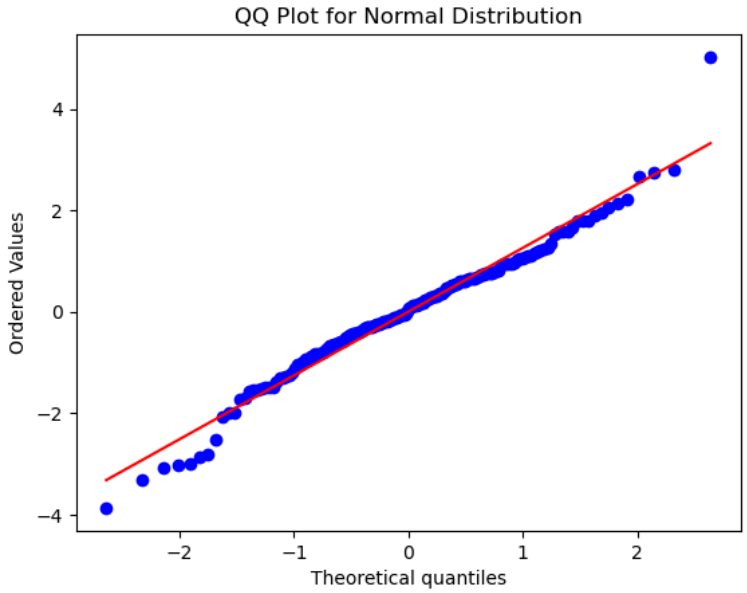

図 7　予測結果の QQ Plot 分析

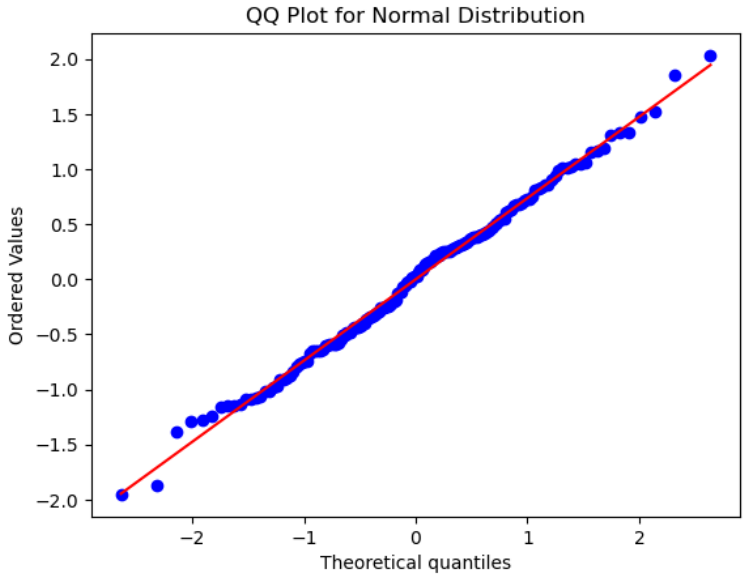

図 8　運転日数を追加した予測結果の QQ Plot 分析

表 4　運転日数を追加した予測の評価関数

	R2	MSE	RMSE	MAE	MAPE
デフォルト	0.46	9.73	3.12	2.74	0.03
運転日数追加後	0.63	6.62	2.57	2.12	0.02

さらに、重要性の分析（図 9）では、inp5・流入 COD 濃度が依然として最も重要なパラメータであることが示されている。運転日数も、環境温度と流入水温度に加えて、ほぼ同

程度の役割を果たしていることが判る。

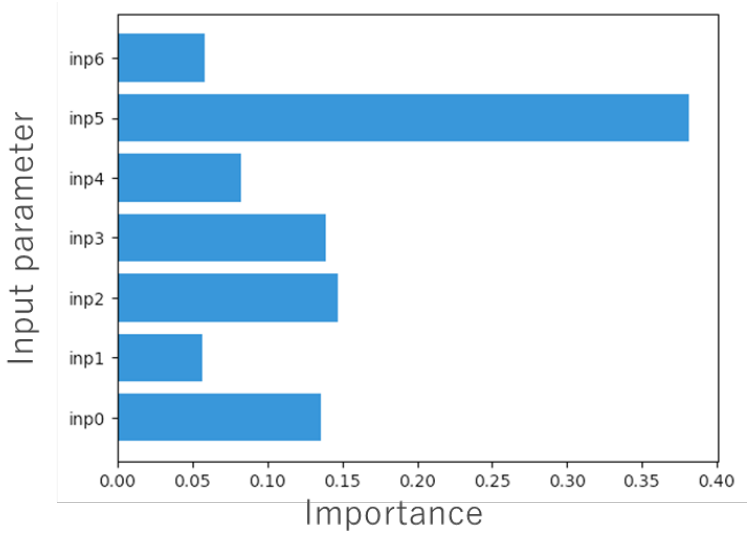

図 9　運転日数を追加した予測の重要度分析

３．３データセットのリサンプリング

運転日数を増やした後、トレーニングセットとテストセットの再配分（リサンプリング）を検討した。元のデータセットはすでにランダム割り当ての原則に従って構築されていたが、データ量が少なくなったためか、テストセットは比較的集中していた。調整後、テストセットはより均等な分布になった（図 10 参照）。

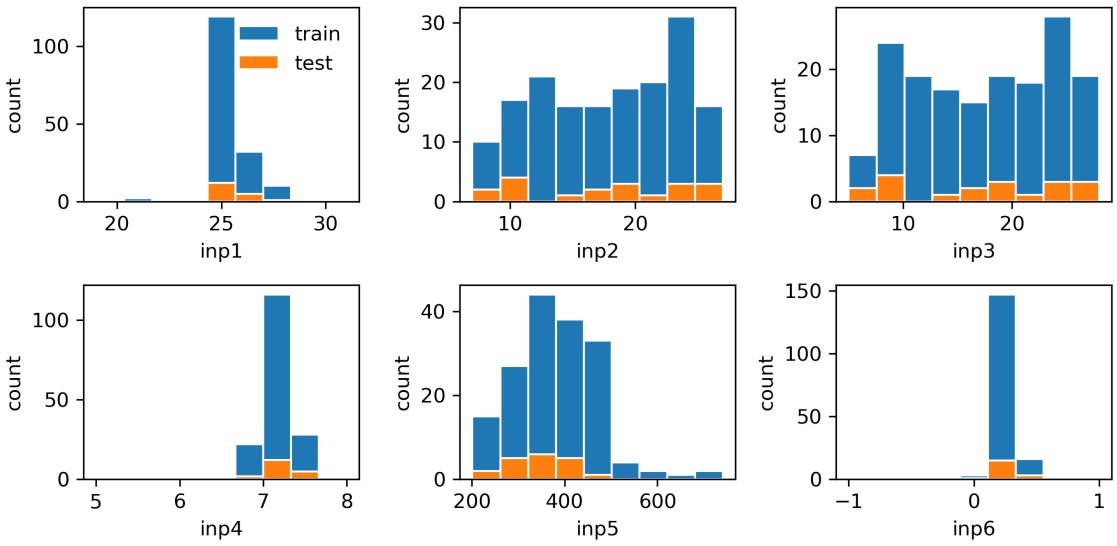

図 10　データセットのサンプルを調整したデータの構成

データセットを調整した後の QQ Plot を図 11 に示す。全体として、ポイントはうまくフィットしている。表 5 の結果により、R^2 が大幅に増加し、MSE や RMSE など他の評価指標もすべて増加した。よって、リサンプリングにおいて、高い予測精度を得られることが実現された。

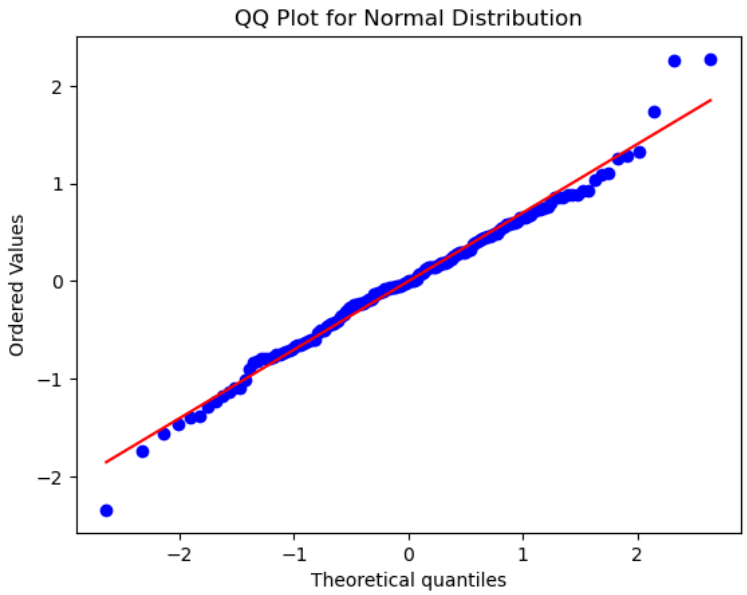

図 11　データサンプルを調整した予測結果の QQ Plot 分析

表 5　サンプルの調整を行った予測の評価関数

	R2	MSE	RMSE	MAE	MAPE
デフォルト	0.46	9.73	3.12	2.74	0.03
サンプル調整後	0.66	4.69	2.16	1.89	0.02

データセットを調整した後に得られた予測値の重要度分析は、元のデータセットの場合とは少し異なるが、傾向は基本的に同じである。流入 COD 濃度は、依然として機械学習の予測結果に影響を与える最も重要な要因である。次いで、環境温度と流入水温であることが得られた。その他パラメータの重要性はほぼ同じであることも分かった。

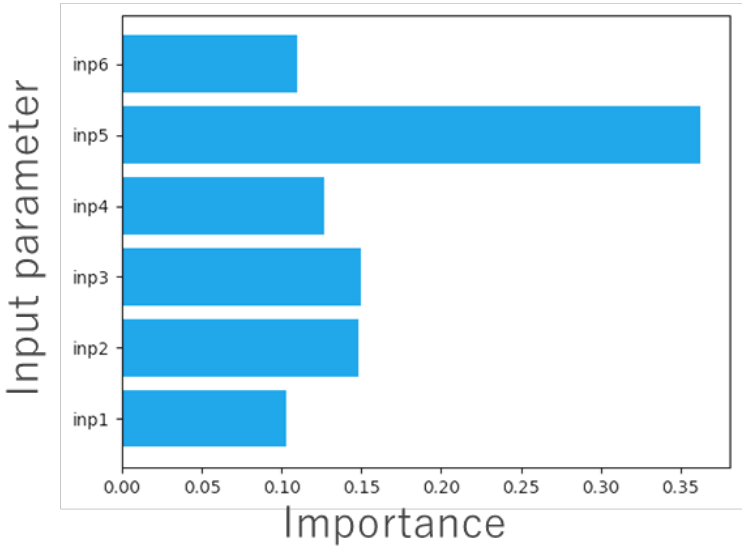

図 12　サンプルを調整した予測の重要度分析

３．４予測精度の向上

以上のことを踏まえ、両方のシナリオを併用することを検討した。表 6 に、いくつかのシナリオの評価関数をまとめたものである。そこで、R^2 は 0.7 と大幅に改善され、MSE、RMSE などの他の評価指標も減少していた。

表 6　運転日数の追加およびサンプルの調整を行った予測結果の評価関数

	R2	MSE	RMSE	MAE	MAPE
デフォルト	0.46	9.73	3.12	2.74	0.03
運転日数追加後	0.63	6.62	2.57	2.12	0.02
サンプル調整後	0.66	4.69	2.16	1.89	0.02
運転日数追加＋サンプル調整	0.73	3.67	1.92	1.63	0.02

重要度分析の観点からは、流入水の COD 濃度が一番重要であるという結果が再び得られた。二番目に重要なものは運転日数であり、これは連続運転システムの予測精度にとって非常に重要であることが分かった。さらに、環境温度と流入水の温度も重要な役割を果たしていることが明らかにした。

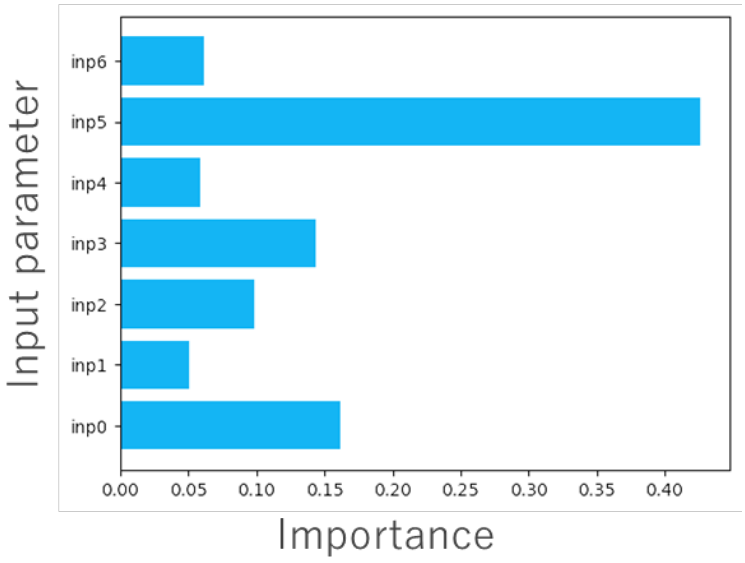

図 13　運転日数の追加およびサンプルの調整を行った予測の重要度分析

４．まとめと展望

本研究は ExtraTrees という機械学習アルゴリズムを用いて嫌気性下水処理プロセスに対する予測を実行した。得られた予測結果について、三種類の深層学習のニューラルネットワークモデルと比較してより良い予測性能を獲得してきた。

運転日数をパラメータ(inp0)としてデータセットの構築に配慮することにより、モデリングの予測性能を向上させることが得られた。運転日数を追加することが連続反応プロセスに対する予測は有意義であることが明らかにした。また、データセットの構築に対して、適切にリサンプリングを行うことで、モデルの効果を確認することができた。重要度分析の結果により、流入水の COD 濃度が重要であることほか、環境温度と流入水の温度も非常に重要なファクターということも判じた。将来的には、データの量を増やしながら、長・短期記憶（LSTM）の適用が計画している。また、機械学習アルゴリズムを用いた下水処理プロセスに対する高精度予測の実装も考えている。

謝辞

本研究に助成いただいた一般財団法人日本建設情報総合センターに、重ねて御礼申し上げます。データセットの構築に使用した実下水の嫌気性処理の実験データについて、東北大学大学院工学研究科環境保全工学研究室で所属した時に、長期運転実験や分析実験を一緒に行っていた方々に、感謝申し上げます。また、本研究の遂行に、機械学習の実行とExtraTreesの適用については英国マンチェスター大学のDr. Zhonghua Zheng、Mr. Junjie Yuにご協力いただきました。ここに記して謝意を表します。

参考文献

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Ge, Z., 2017. Review on data-driven modeling and monitoring for plant-wide industrial processes. Chemom. Intell. Lab. Syst. 171, 16–25.

Hu, Y., Cheng, H., Ji, J., Li, Y.-Y., 2020. A review of anaerobic membrane bioreactors for municipal wastewater treatment with a focus on multicomponent biogas and membrane fouling control. Environ. Sci.: Water Res. Technol. 6, 2641–2663. https://doi.org/10.1039/D0EW00528B

Jenkins, D., Wanner, J., 2014. Activated sludge-100 years and counting. IWA publishing.

Li, G., Ji, J., Ni, J., Wang, S., Guo, Y., Hu, Y., Liu, S., Huang, S.F., Li, Y.Y., 2022. Application of deep learning for predicting the treatment performance of real municipal wastewater based on one-year operation of two anaerobic membrane bioreactors. Sci. Total Environ. 813, 151920. https://doi.org/10.1016/j.scitotenv.2021.151920

Liu, Y., Tay, J.H., 2001. Strategy for minimization of excess sludge production from the activated sludge process. Biotechnol. Adv. 19, 97–107. https://doi.org/10.1016/S0734-9750(00)00066-5

Shi, S., Xu, G., 2018. Novel performance prediction model of a biofilm system treating domestic wastewater based on stacked denoising auto-encoders deep learning network. Chemical Engineering Journal 347, 280–290. https://doi.org/10.1016/j.cej.2018.04.087

Verrecht, B., Judd, S., Guglielmi, G., Brepols, C., Mulder, J.W., 2008. An aeration energy model for an immersed membrane bioreactor. Water Res. 42, 4761–4770. https://doi.org/10.1016/j.watres.2008.09.013