

一般口演

一般口演21

医療データ分析9（機械学習・テキストマイニング2）

2018年11月25日(日) 10:40 ～ 12:10 D会場 (4F 413+414)

[4-D-2-2] 深層学習に基づく検体検査の時系列データからのネフローゼ発症予測とその予測モデル解釈の試み

○山崎 陽平¹, 平木 秀輔^{1,2}, 杉山 治², 大寺 祥佑^{1,2}, 山本 豪志朗², 佐々木 博史², 岡本 和也^{1,2}, 南部 雅幸², 黒田 知宏^{1,2}
(1.京都大学大学院情報学研究科, 2.京都大学医学部附属病院)

ネフローゼ症候群は血液中蛋白量が減少することにより生じる疾患であり、重症化すると腎不全・血栓症・感染症等を合併する危険性が存在するため早期発見が望ましい。そのため、発症予測を機械学習で行う場合、「早期かつ予測精度が高い手法であること」と「予測結果に対する解釈が可能な手法であること」が非常に重要となる。本研究では検体検査による時系列データから機械学習を用いて解釈可能で早期かつ精度の高い発症予測モデルの構築を目指す。提案するネフローゼ発症予測モデルを構築するために、本研究では京都大学医学部附属病院を過去に受診した患者の匿名データセットから約2000名の検体検査12項目の時系列データを用いた。時系列情報を扱うために機械学習のモデルはシーケンシャルなデータのパターンを認識することが可能なRNN(Recurrent Neural Network)を採用し、予測精度の向上を図った。予測モデルの学習においては早期の予測が可能であるかを検証するために、ネフローゼ患者がネフローゼを発症した日付から1ヶ月以上前のデータのみを用いて予測を行った。またデータの前処理においては検体検査項目毎に分布や値域の差の影響を抑えるために正規化を行い、欠損値は0で補間した。提案するRNNを用いたネフローゼ発症予測モデルとSVMを用いて構築した発症予測モデルの精度を5分割交差検証により評価したところ、SVMを用いたモデルに比べてRNNを用いた提案モデルの方が予測精度が向上することが示唆された。また、構築したモデルをLIME(Local Interpretable Model-agnostic Explanations)を用いて解釈をし、導出された複数の簡易予測モデルに対しての考察を医学的見地から行った。

深層学習に基づく検体検査の時系列データからの ネフローゼ発症予測とその予測モデル解釈の試み

山崎 陽平^{*1}、平木秀輔^{*1,2}、杉山 治^{*2}、大寺 祥佑^{*1,2}、山本 豪志朗^{*2}、
佐々木 博史^{*2}、岡本 和也^{*1,2}、南部 雅幸^{*2}、黒田 知宏^{*1,2}

*1 京都大学大学院情報学研究科、*2 京都大学医学部附属病院

Interpretation trial of Nephrosis prediction model based on deep learning with clinical time-series data.

*Yohei Yamasaki^{*1}, Shusuke Hiragi^{*1,2}, Osamu Sugiyama^{*2}, Shosuke Ohtera^{*1,2}, Goshiro Yamamoto^{*2},
Hiroshi Sasaki^{*2}, Kazuya Okamoto^{*1,2}, Masayuki Nambu^{*2}, Tomohiro Kuroda^{*1,2}

*1 Graduate School of Informatics Kyoto University, *2 Kyoto University Hospital

Nephrosis is a disease characterized by abnormal protein loss from impaired kidney. The disease sometimes results in acute kidney injury, venous thrombosis, or severe infection due to lowered oncotic pressure and immunoglobulin loss. Therefore, early detection is preferable for patients. The early prediction of the nephrosis is essential and, in order to introduce the prediction in clinical practice, both "early prediction with high accuracy" and "interpretation of the prediction" are required. In this research, we aimed to develop an early prediction model using machine learning from clinical time series data, that can predict the rise of nephrosis with high accuracy as well as being able to interpret the prediction result for the doctor. In order to develop an accurate nephrosis prediction model, we collected time series of 12 anonymized specimen examination data with 2,000 patients who visited in Kyoto University Hospital. In order to deal with time series information, we adopted Recurrent Neural Network, which is capable of recognizing sequential data patterns, to improve prediction accuracy. For the interpretation of the prediction results, we introduced Local Interpretable Model-agnostic Explanations (LIME) which can interpret the trained prediction model with the combination of Linear Classifiers, and interpreted the trained RNN prediction model using LIME. In the experiment, we evaluated our proposed nephrosis prediction method and those with SVM and Random Forest, and verified that the developed mode has higher accuracy compared with those of SVM by 5-fold cross validation.

Keywords: Nephrosis, Machine Learning, Decision Support Techniques, Data Interpretation, Statistical

1. 緒論

ネフローゼ症候群¹⁾は血液中蛋白量が減少することにより生じる疾患である。ネフローゼ症候群の発症は腎機能の低下を示しており、重症化すると腎不全・血栓症・感染症等を合併する危険性が存在するため早期発見が望ましい。ネフローゼの診断は、主に血液や尿の検体検査によって得られた尿蛋白や血清総蛋白の値を基準とするが、一方でネフローゼの発症の原因が不明であることも多く、個々の医師の経験に即した早期予測しか行えない。

本研究では、機械学習を用いたネフローゼの早期発症予測に取り組む。近年、機械学習の研究は深層学習手法の出現により飛躍的に発展しており、画像認識や音声認識、自然言語処理等の分野でも従来の機械学習モデルと比較して識別や予測の精度が大幅に向上している。本研究では、この深層学習の手法を用いて、過去の患者の検体検査値からネフローゼの発症を予測するモデルを開発する。

発症を予測するシステムを臨床現場で用いるためには、予測モデルの開発と同様に、予測モデルが返した結果の根拠が示されることが重要である。機械学習、特に深層学習を用いた予測モデルは重要な特徴量を機械自身が学習してくれる一方で、その意思決定プロセスは不透明で予測結果の根拠は不明である。そのため、医療業界など人の命に密接に関わる分野においては、予測精度の高さに関わらず予測結果が信頼されない場合が多い。医療業界で機械学習の予測結果を信頼して活用するためには、予測結果の根拠を医師が解釈可能な形式で提示することが重要になる。

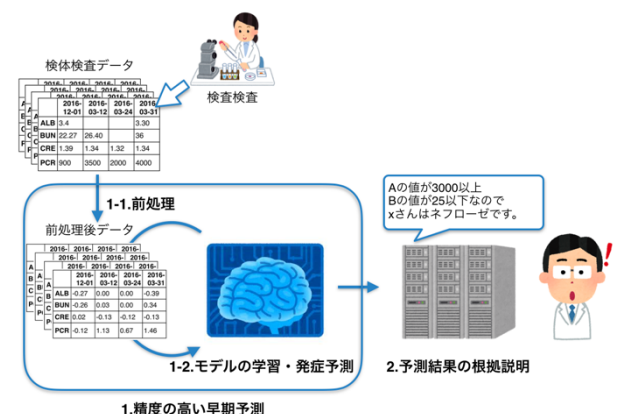
2. 開発目的

本研究では、機械学習を用いた医師が十分に信頼できる

ネフローゼの活性予測システムの開発を目的とし、その実現のため 1) 精度の高い早期発症予測と 2) 医師に解釈可能な形式での予測結果の根拠説明の二つを目的とする。1)については検体検査における時系列データ特有の課題を考慮した前処理やモデル選択により精度の向上を目指す。2)については複雑なモデルの予測結果の根拠を説明する手法を採用し、導出された結果を医師に理解できる形式で提供する。

3. システム概要

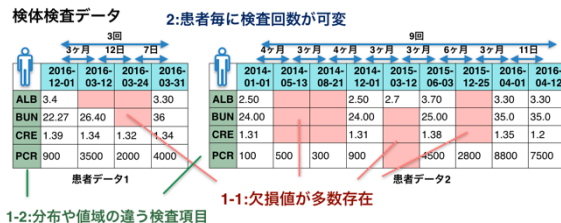
本システムによるネフローゼ活性予測の処理フローを図1に示す。ネフローゼの早期発症予測においては、早期発症予測の精度向上のため検体検査の時系列データに存在する複数の課題を解決する処理を行う。まず、モデルによる学習の前に適切な前処理を行う (図1、1-1)。



できる機械学習モデルにて学習、早期発症予測を行う (図1、1-2)。最後に予測結果に提案する根拠説明手法を適用することで、予測結果の根拠説明を導出する (図1、2)。

3.1 早期発症予測の精度向上

ネフローゼの早期発症予測のために、本研究では図2に示す処理を予測モデルの前処理として適用した。対象とする血液や尿の検体検査データは 1-1) 欠損値が多数存在する、1-2) 検査項目毎の値域や分布が異なる、2) 検査回数が患者毎に変である、といった課題が存在し、機械学習で予測を行う際に精度向上を妨げる原因となる。本研究では、1-1、1-2 の課題に対してデータの前処理を行い、2 の課題に対し



て時系列情報を考慮したモデル選択をする事で、早期発症予測の精度向上に取り組んだ。

図2 検体検査データから発症予測を行う際に存在する課題

3.1.1 欠損値の補間

検体検査結果は測定期間が不規則であることから、欠損値を補間するには測定間隔を考慮した線形補間やガウス過程によるモデル化が用いられるが、これらの手法は単系列の情報しか考慮しない。一方で河添らは欠損値推定の際に補間ターゲットとなる系列以外の系列の情報を考慮するようにオートエンコーダを用いた補間手法を提案しているが、線形補間との大きな差異は観察されていないと結論づけている²⁾。本研究では測定間隔を考慮した線形補間を提案する。補間時に時系列情報を考慮することで予測精度向上に繋がることが期待される。また、平易な単系列情報による欠損値補間である平均値による補間も行い、二つの手法の性能を比較した。

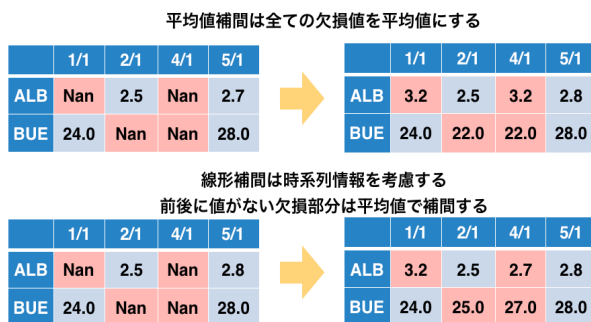


図3 欠損値補間手法の説明

3.1.2 検査項目の正規化

図3に示すように対象データは検査項目ごとに値域や分布が大きく異なる。そのため予測精度の向上のために、データの正規化を行う必要がある。一般的な正規化の手法として min-max normalization(min-max)と z-score normalization(z-score)が存在する。本研究の対象となるデータは外れ値を含むものが多い為、外れ値の値に大きな影響を受けない z-score を適用する。

3.1.3 可変な時系列情報を扱える早期発症予測モデルの選択

検体検査データから患者の病態や将来の値を予測する研究は数多くされている。倉沢らは欠損値を考慮したオートエンコーダを用いることで欠損値の補間と低次元特徴量の抽出を同時に行ない血糖管理の良好・不良を予測する手法を提案している³⁾。

本研究で対象とする検体検査データは多系列であり、検査回数は患者によって異なる。このように可変長の入力である時系列データであることを考慮して、予測モデルにニューラルネットワークの一種である Recurrent Neural Network(RNN)⁴⁾を採用する。RNN は通常のニューラルネットワークと異なり、隠れ層にループする重み付き結合を構成することで、連続したデータの関係性を考慮した学習が可能となる。単純な重み付き結合だけで構成した RNN のアーキテクチャ(simple RNN)では重み更新の際に勾配消失問題が生じ、長期的な時系列情報を考慮した学習ができない。この課題を解決するため、Long Short-term memory(LSTM)⁵⁾が提案されている。LSTM は RNN の一種であり、ネットワーク内部に入力ゲートや出力ゲートを構成することで、重み更新の際に必要な誤差信号だけを適切に伝搬させ、長期的なデータの学習を可能とする手法である。本研究では LSTM を早期発症予測モデルとして採用し、simple RNN と LSTM 二つのモデルで予測を行った結果を比較した。

3.2 医師に解釈可能な形式での予測結果の根拠説明

機械学習による予測結果の根拠説明は explain models と explain prediction の二つに大きく分類される。explain models は予測モデル自体がどのような性質を持っているかを説明することである。explain prediction は個々のデータに対する機械学習モデルの予測結果の根拠を説明することである。本研究では個々の患者の予測結果に対してそれぞれ根拠説明することが目的であるため、explain prediction を用いる。

近年、予測結果の根拠を説明する研究がいくつも取り組まれている。特に画像認識や自然言語処理における根拠説明に関する研究が多いが、時系列データに適用可能な手法もいくつか存在する。例えば Smilkov らが提案する Sensitive Analysis⁶⁾という手法は、入力された特徴量の中で変化させた際に出力に大きな影響を及ぼすものを重要な特徴量とみなし、それを勾配計算により求める。また翻訳モデルの根拠説明手法として Minh-Thang Luong らが提案した Attention⁷⁾は、出力を行う際に全ての隠れ層を利用し、その際に隠れ層毎に重み付けをすることでどの時系列の入力が重要であるかを判断するモデルである。

ネフローゼ発症予測における根拠説明では、「個々の患者毎にどの検査時のどの検査項目」が要因なのかを知る必要がある。そのため、どの特徴量が重要か判断できるが特徴量がどのような値になるべきか判断できない Sensitive Analysis や時系列の重要性のみで多系列なデータの重要性を判断できない Attention は本研究の要件に合致しない。以上の理由から本研究では予測結果の根拠説明に Local Interpretable Model-agnostic Explanations (LIME) を用いる⁸⁾。

LIME は explain prediction に分類される予測根拠説明手法であり、多系列な時系列データにも適用できるという特徴を持つ。LIME は予測したいデータ x の周辺からサンプリング

することで複雑なモデル f に近似させた解釈可能なモデル g を局所的に導出し予測根拠を説明する手法である。データ x の周辺で、複雑なモデル f に近似する説明用分類モデル g を学習させるために下記の目的関数を設定する。

$$\xi(x) = \arg \min L(f, g, \pi_x) + \Omega(g)$$

G は解釈可能なモデル f の集合で、 g は G のうちの一つのモデルである。 $L(f, g, \pi_x)$ はデータ x の周辺のモデル f と g の予測結果の誤差関数になり、 $\Omega(g)$ は説明用分類機の複雑さを表現する。つまり複雑なモデルとの誤差である $L(f, g, \pi_x)$ とモデルの複雑さを表す $\Omega(g)$ の和が最小となる識別機 g を求めて予測結果に対して解釈可能な説明をする。

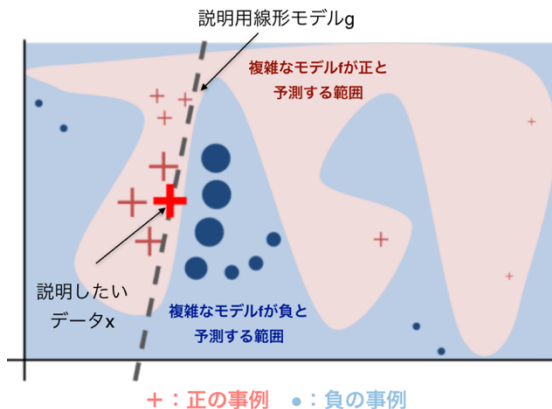


図4 LIMEによる根拠説明の動作原理
(出典: <https://arxiv.org/pdf/1602.04938.pdf>)

4. システム評価

本研究の目的は 1) 精度の高い早期発症予測と 2) 医師に解釈可能な形式での予測結果の根拠説明の二つである。1) に関しては、提案した複数の欠損値補間とモデル選択手法を組み合わせ、それぞれの精度を比較し評価する。2) に関しては、検体検査データを用いたネフローゼ予測モデルにLIMEを適用した根拠説明の結果を示す。

4.1 対象データ

京都大学医学部附属病院を過去に受診した患者の匿名データセットから検査回数が4回以上である約2000名の患者データを対象とする。また腎臓内科の医師の意見を参考にして、検体検査・尿検査のデータからネフローゼ発症と相関があると考えられる12項目の検査項目を抽出した。早期予測が目標であるため、ネフローゼ発症から1ヶ月以上前のデータ最長8回分を入力データとした。

またネフローゼ症候群の検査所見である尿蛋白 3.5g/日以上かつ、血清アルブミン(ALB)値 3.0g/dl 以下という条件を満たした際にネフローゼが発症したと判断した。

4.2 モデル設計

本実験に用いた simple RNN と LSTM は、隠れ層を一層とし、ユニット数は10とした。過学習を防ぐため、dropoutを設定し、その確率を0.3とした。また2クラスの分類問題であるため最終層のユニット数は2とし、softmax関数を適用した。誤差関数は交差エントロピー、最適化手法はAdam、ミニバッチサイズは16、学習エポックは50とした。またRNNモデルの比較対象となるベースラインモデルとして時系列情報を考慮しないモデルであるSVMとRandom Forestを用いた。実装はpython3系で行い、機械学習のライブラリはskikit-learn、ディープラーニングのライブラリはtensorflowをバックエンドとした。

kerasを用いた。

4.3 早期発症予測の精度検証

提案する手法により早期発症予測の精度が向上するかどうかを検証するために、平均値補間と測定期間を考慮した線形補間の2種類の欠損値補間、SVM、Random Forest、simple RNN、LSTMの4種類の識別モデル、それぞれを組み合わせた計8種類の識別手法を比較する事で提案手法が精度向上に貢献するかを検証する。

精度は5分割交差検定(学習データ80%、テストデータ20%)により求めた。図6に各手法の予測精度の比較を示す。

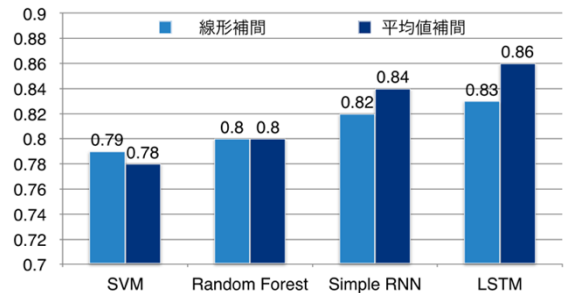


図5 各手法の予測精度の比較

結果、LSTMが予測モデルの中で最も精度が高くなった。またRNN、LSTMにおける補間手法の比較においては線形補間よりも平均値補間による欠損値補間の方が高い精度になった。

4.4 LIMEによる根拠説明

LIMEによる根拠説明の検証については、実行結果を医師に提示し、意見を求めることで予備的な検証を行った。あるネフローゼ患者の予測結果に対するLIMEを用いた根拠説明の結果を図7に示す。図7左の予測結果は、予測モデルが対象となるネフローゼ患者を将来的にネフローゼになるかどうかを判断した確率である。図7中右がLIMEを適用させた結果になる。LIMEの結果は条件式と貢献度のペアの集合で構成される。条件式は検査項目、検査時間、条件、検査値により構成される。

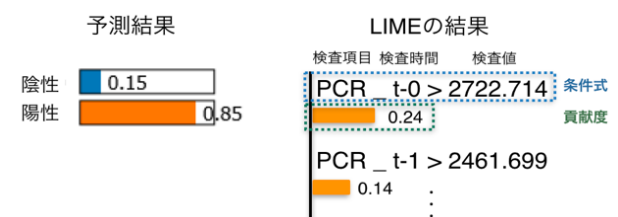


図7 LIMEによる予測結果の根拠説明

検査時間は $t-i$ で表記され、 i が0であれば直前の検査時のデータを示す。条件は等号か不等号になる。貢献度は対応する条件式が予測結果に貢献した値である。例えばLIMEの根拠説明の一つ目は「PCR_t-0 > 2772.714」が条件式でその下に記載される0.24が貢献度である。この例の場合、検査項目PCRにおける最も直前の検査値が2772.714より大きい値でないなら、陽性である確率が0.85から0.24低い0.61になる、ということを示す。この患者例では、ネフローゼの発症予測を行う際に直前のPCRやTPが重要な項目であることが示唆された。また、ネフローゼの発症予測を行う際に直前のPCRやALBが重要な項目であることが示唆された。なお、LIMEを適用させる際の欠損値補間の手法と予測モデルは、

早期発症予測において最も精度が高い平均値補間と LSTM を用いた。

5. 考察

実験結果より、本研究で採用した精度の高い早ネフローゼ早期発症予測、医師に解釈可能な形式での予測結果の根拠説明の手法が有効に動作したかを検証し、提案システム全体について考察する。

5.1 早期発症予測の精度向上手法の結果

検体検査の時系列データが持つ課題に対して前処理とモデル選択によって、早期ネフローゼ発症予測における精度向上に取り組んだ。精度向上のために提案した前処理とモデル選択の有効性について検証する。

まず、本研究では、検体検査の時系列データにおける補間手法として、平均値補間より線形補間が時系列情報を考慮できるため精度が向上するという仮説を立てた。しかし、実際には時系列情報を考慮できる予測モデルでは平均値補間の方が精度が高く、当初の仮説とは異なる結果を得た。この結果は、検査値の時系列変化が線形に変化していないことを示唆しているものと考えられる。より詳細に見れば、検査値によって、値の変化のパターンは異なると考えられ、それぞれの検査値の特性に応じた補間をする必要がある。

次に、正規化について、本研究では、z-score normalizationを採用したが、対象データの分布は非常に偏っていた。Z-score normalization は正規分布を持つデータに用いられることが多いため、偏りがある検査項目は対数を取り正規分布をとるデータに処理してから z-score normalization による正規化を行うことで、精度の向上が期待できる。

最後に、本研究では時系列情報を考慮できる RNN のアーキテクチャ(simple RNN と LSTM)を早期発症予測モデルとして提案した。結果としてベースラインとなる手法よりも高い精度を示し、時系列情報を扱うモデルを選択することが精度向上に繋がったことが示唆された。また、simple RNN より LSTM の予測精度が上回ったため、長期的な時系列情報を保持することが精度向上に繋がることも示唆された。

今後の課題としては、時系列情報を扱う LSTM 以外のモデルを適用することや、隠れ層を複数にする、ベイズ最適化によるハイパーパラメータのチューニングなどが考えられる。

5.2 LIME による予測結果の根拠説明

ネフローゼ早期発症予測における予測結果の根拠を説明する手法として、本研究では LIME を採用し、その結果の出力を行なった。LIME により出力された予測根拠の説明結果は、条件式とその貢献度のペアで表される。この形式について腎臓内科の医師に意見を求めたところ、何日に何の項目のデータが予測結果の根拠となったのかが一目で分かる形式ではないので、視覚的な工夫が欲しいとの意見を得た。今後、医師に理解できる形で LIME の予測根拠説明を提示する手法の検討が必要となる。

また、LIME による予測結果の根拠説明は、個々の患者に対する結果であるが、モデル全体の大局的な根拠説明は得られない。複数患者の根拠説明の事例を医師が考察した限りは、その根拠説明は重要と考えられる発症直前の PCR や ALB、TP に対する条件式が高い貢献度を示しており、一定の信頼ができるとの意見を得た。しかしながら、本実験での検証は、患者の一部事例の根拠説明を検証しただけに過ぎない。全患者の予測と根拠説明を行い、同じ根拠説明がどの程度行われたのか、貢献度の総和はどの程度が、といった指標

を計算することで大局的な根拠説明の信頼性を検証する必要がある。

5.3 提案システムがネフローゼ診療に与える影響の考察

本研究で開発した精度の高い早期発症予測、医師に解釈可能な形式での予測結果の根拠説明の二つの機能を有する提案システムが実際の診療現場にどのような優位性を持つかを検証する必要がある。現状は、医師に提案システムを使用した場合と使用しない場合に分けてネフローゼの診断予測を行って頂き、その結果を比較して有効性を検証していきたい。

6. 結論

本研究では、1) 精度の高い早期発症予測と 2) 医師に解釈可能な形式での予測結果の根拠説明の二つの機能を有するフローゼ活性予測システムの開発を目的とした。前者については検体検査による時系列データの欠損値の存在、検査項目毎の値域や分布が異なる、検査回数が患者毎に可変である、といった課題に対し、適切な前処理と時系列を考慮した予測モデルである RNN を用いることで精度の向上に取り組んだ。また後者は LIME を用いた予測結果の根拠説明を試み、その結果を示した。実験において、実装した simple RNN、LSTM の予測モデルと、Random Forest、SVM を用いた予測モデルを比較した結果、LSTM を用いた予測モデルが最も良い性能を示した。また、前処理における補間手法として、平均値補間と線形補間を比較したところ、平均値補間の方が良い性能を示した。LIME の根拠説明については、その結果を医師に提示したところ、違和感のある根拠でないとのコメントが得られた。LIME の結果の信頼性や、提案システムがネフローゼ診療に与える有効性は今後検証していきたい。

参考文献

- 1) 松尾清一, 斉藤喬雄, 田口尚ら. ネフローゼ症候群診療指針. 日腎会誌 2011;53(2):78-122.
- 2) 倉沢央, 林勝義, 藤昭典ら. Machine learning-based Prediction of poor glycemic control in patients with diabetes. 第 37 回医療情報学連合大会 37th JCM 2017;825-830
- 3) 河添悦昌, 倉沢央, 岩井聡ら. Comparison of Missing Value Estimation of Laboratory Tests between State Space Model and Deep Neural Network. 第 37 回医療情報学連合大会 37th JCM 2017;820-824.
- 4) T Mikolov, G Zweig. Context dependent recurrent neural network language model. 2012 IEEE Spoken Language Technology Workshop (SLT) 2012;234-239
- 5) Sepp Hochreiter, Jurgen Schmidhuber. LONG SHORT-TERM MEMORY. Neural Computation 1997;9(8):1735-1780.
- 6) D. Smilkov, N. Thorat, B. Kim, F. Viegas, M. Wattenbarg. SmoothGrad: removing noise by adding noise.
- 7) M. Luong, Hieu Pham Christopher, D. Manning. Effective Approaches to Attention-based Neural Machine Translation.
- 8) M Tulio Ribeiro, S. Singh, C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD 2016.