K-008

臨床インタビューデータを用いた自動うつ病診断
Automatic Depression Diagnosis from Clinical Interview Data
~ 機械学習モデルによる診断の根拠を示すための SHAP を用いた診断手法の提案 ~

前原 春佳 ‡　秦野 亮 ‡　西山 裕之 ‡
Haruka Maehara　Ryo Hatano　Hiroyuki Nishiyama

1. 序論

昨今，精神疾患症は重大な社会問題となっており，その代表例であるうつ病患者は世界で３億人以上にものぼる．現在のうつ病診断は，患者自身が記入した質問票の結果を用いた精神科医の問診によって行われており，患者の感覚や精神科医の能力に相当依存しているといえる．

こうした問題に対応し，定量的なうつ病診断実現のために，AI の活用に向けた研究も行われている．しかし，医療分野での AI の活用を考える際，AI が抱えるブラックボックス問題が障害となることがある．通常，医療機関で診断を受ける際，医師からどのような理由で診断が下されたのかという診断根拠が説明される．そのため，定量的な診断支援を目的としたシステムでも，患者が納得できるような診断根拠を明確に提示する必要があると考えられる．また，うつ病を治療する際には薬物治療も行われており，重要度や症状に合わせて薬が選択される．このような薬には副作用を含むものもあるため，症状や重症度を含めた正確な診断，適切な薬の処方が必要である．

そこで，本研究は，定量的なうつ病診断のための生体情報を用いた断機械学習モデルの開発を目的とする．うつ病と健常者の判別及び，うつ病の重症度や症状ごとの重症度も予測し，予測根拠の提示も行う．本研究の成果は，医療機関で利用する診断支援ツールとしての利用が期待できる．また，うつ病の兆候がありつつも，様々な理由で医療機関へ受診できていない人々に向けた自己診断ツールとしての利用も考えられる．その場合，人々が客観的に自身の状態を把握し，重症度が高いと想定される人々に対し，医療機関へ受診する後押しにもなり得る．本研究によって，うつ病兆候がみられる人々が適切な治療を受けるきっかけづくりとなることが期待できる．

2. 提案手法

本研究では，特徴量を抽出した後，うつ病か健常者かの２値分類を実施する．その後，うつ病の重症度および症状別の重症度を回帰モデルよって算出する．

2.1 データセット

本研究では，DAIC-WOZ データセット[1]を使用した．これは不安，うつ病，外傷後ストレス障害などの心理的苦痛を診断するために設計されたデータセットである．189 人の臨床インタビューデータとうつ病診断の際に用いられる質問票の１つである PHQ8 の回答が含まれる．データ内のうつ病患者は 133 人，健常者は 56 人であり，各参加者のビジュアルデータ，生音声データが存在する．ビジュアルデータは個人情報保護の観点より，生データは存在せず，インタビューを撮影した動画像から抽出されたさまざまな特徴量が時系列的に含まれている．また，音声データとビジュアルデータにはタイムスタンプが付与されている．本研究では，データに欠損がなく，各質問の回答が付与されている 136 人を対象としたデータセットを用いた．

2.2 前処理

まず初めに，音声データに対し，音声処理を実施した．付与されたタイムスタンプを用いて，インタビュアーと患者の音声を分割，患者の音声部分のみを抽出した．その後，無音区間を削除した．

次に，音声データとビジュアルデータからそれぞれ特徴量を抽出した．音声データは python の OpenSMILE3.0 ライブラリを用いて特徴量抽出を行った．抽出した特徴量は 8 種類のエネルギー特徴量，96 種類のスペクトル特徴量，11 種類の韻律特徴量を含む計 115 種類である．各特徴量について，インタビュー全体を通した平均値を取得し，音声特徴量として分析に用いた．ビジュアルデータからは表情を表す Action Unit(AU)，視線情報，頭部姿勢情報を使用した．視線情報と頭部姿勢情報は座標および回転角が付与されているため，それぞれインタビュー全体を対象とした平均値を取得し，ビジュアル特徴量とした．また，視線情報は座標を基に変位と速度を算出し，その平均値を使用した．頭部座標も同様に座標情報より変位を計算し，その平均値を加えた 54 種類をビジュアル特徴量とした．音声特徴量とビジュアル特徴量を合わせた 169 種類の特徴量を用いて分析を実施した．

最後に，抽出した特徴量のうち，予測に寄与する特徴量を見つけるため，特徴量選択を実施した．今回は，ランダムフォレスト，XGBoost，Permutation Importance の 3 つの重要度算出法を使用した．重要な特徴量上位 50 種類を選び，各機械学習モデルにおいて最も性能が高くなる手法を選択した．

本研究の正解ラベルには，うつ病診断に用いられる質問票(PHQ8)の回答を使用する．PHQ8 は 8 つの質問から構成されており，各質問でうつ病に関連する症状について問われている．具体的な症状は，興味の欠落，気分の落ち込み，睡眠障害，疲労感，食欲不振，劣等感，集中力の欠落，ふるまいの変化等である．回答は整数値 0~3 の範囲をとり，数値が大きいほど症状が重症であり，小さいほど健常状態であることを示す．8 つの質問の回答を合計した値が患者の PHQ スコアであり，このスコアが大きいほどうつ病の重症度が大きいことを示す．具体的には，スコアが 10 以上でうつ病，10 未満で健常者と診断される．

2.3 機械学習

まず初めに，うつ病患者のデータを正事例，健常者のデータを負事例として 2 値分類を行った．分類タスクでは，ランダムフォレスト，XGBoost，決定木，Support Vector

Machine (SVM)，KNN を用いて性能比較を行い，うつ病診断に最適と考えられる機械学習モデルを選定した．

次に，症状別の重症度を診断するための回帰モデルの学習を行った．回帰タスクでは，ランダムフォレスト，XGBoost，SVM，KNN，ガウス過程回帰を用いて性能比較を行い，最適な機械学習モデルを選定した．その後，最も性能が高かったモデルに対して，解釈性を高めるために SHAP 値の算出も行った．

評価指標として，分類モデルでは正解率(ACC)，適合率(PRE)，再現率(REC)，F 値(F1)を用いた．また，回帰モデルは平方根平均二乗誤差(RMSE)，平均絶対誤差(MAE)，決定係数(R2)を用いた．過学習を防ぐために，分類タスク，回帰タスクともに 10 分割交差検証を実施した．以降の章では，各検証で算出した評価指標の平均値を示す．

3. 実験・結果

まず初めに，特徴量のスケールをそろえるためにデータの標準化を行った．また，データの不均衡に対処するため，オーバーサンプリングを実施した．分類モデル学習時は SMOTE を使用し，回帰モデル学習時には SMOTE を回帰タスク向けに派生させた SMOGN を使用した．

3.1 うつ病患者と健常者の分類タスクの結果

分類タスクの最良となった結果を表 1 に示す．

表 1　分類モデル性能評価

機械学習モデル	ACC	PRE	REC	F1
ランダムフォレスト	0.567	0.517	0.542	0.493
決定木	0.650	0.667	0.683	0.647
SVM	**0.750**	0.713	0.725	**0.696**
KNN	0.717	0.725	0.733	0.686
XGboost	**0.750**	**0.733**	**0.750**	0.635

ランダムフォレストは Permutation importance を用いた特徴量選択，決定木と XGboost はランダムフォレストを用いた特徴量選択，SVM，KNN は XGboost を用いた特徴量選択を行った際のモデルが最も高性能であった．SVM，KNN，XGboost は 70%以上の正解率でうつ病患者と健常者を分類した．

3.2 重症度予測に関する回帰タスクの結果

次に，うつ病の重症度を予測する回帰モデルの学習を実施した．RMSE と MAE は回帰モデルの予測誤差の大きさを示す指標であり，0 に近いほど精度が高いことを示す．また，R2 は回帰モデルが目的変数の変動を説明する能力を示す指標であり，1 に近いほど精度が高いことを示す．本研究のうつ病重症度を予測する回帰タスクにおいて，ランダムフォレストで特徴量選択を行った SVM モデルが最も高い性能を示した．その回帰モデルは RMSE が 4.51，MAE が 3.35，R2 が 0.310 であった．

3.3 症状別の重症度予測に関する回帰タスクの結果

症状別の重症度を予測する回帰タスクでは，疲労感の重症度を予測する回帰モデルが最も高性能を示した．その際，最も性能が高かったものはランダムフォレストで特徴量選択を行った SVM の回帰モデルであり，RMSE が 0.748，MAE が 0.635，R2 が 0.374 となった．そのモデルの SHAP 値を計算した結果を図 1 に示す．

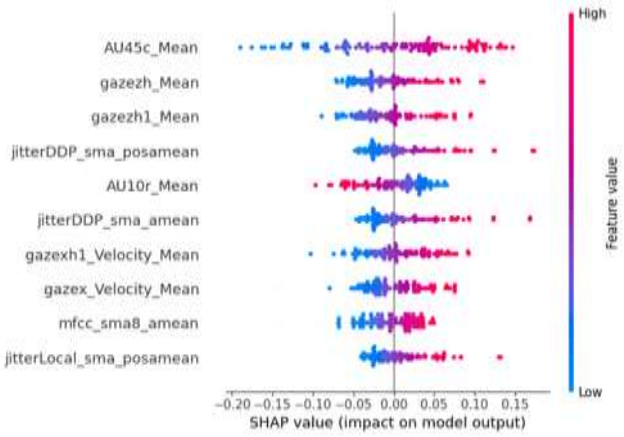

図 1　疲労度に関する質問の回帰タスクの SAHP 値

図 1 は横軸が目的変数の値の大きさ，縦軸が特徴量の重症度を示し，赤が正の相関，青が負の相関を示す．最も重要と判断された音声特徴量は韻律特徴量の 1 つで，声帯振動の不規則さを示す JitterDDP であった．最も重要と判断されたビジュアル特徴量は瞬きの頻度を示す AU45_C で，どちらも目的変数に対して正の相関があった．一方で興味や集中力の欠落に関する予測では，R2 値が 0 に近い値を取り，モデルがタスクにうまく適応しない結果となった．

4. 考察

分類タスクの結果より，音声特徴量やビジュアル特徴量はうつ病と健常者の分類に対して有効であった．

また，症状別の重症度予測の回帰モデルの SHAP 値より，疲労感の重症度が高い患者には，瞬きの増加や声のかすれといった症状がみられる可能性があると考察できる．このように，症状の重い患者に出現する兆候を調べる点において，SHAP の活用は有効であると考えられる．

一方で，疲労感以外の症状では，重症度予測に関する回帰モデルの性能が，有効と考えられる精度に達しなかった．原因として，回帰タスクと選択した特徴量が不適合であった可能性が考えられる．今後は精度向上のため，新たな特徴量の生成や時系列情報を取り入れる事により，回帰モデルのパフォーマンス向上を目指す．また，今回用いた機械学習モデルは表現力に限りがあるため，より複雑なモデルを使用するといったアプローチも検討していきたい．その際も，解釈性を損なわない工夫が重要であると考えている．

5. まとめ

本研究では，音声特徴量とビジュアル特徴量を用いた解釈性の高いうつ病患者の診断手法を提案した．これにより，うつ病患者と健常者を分類できた．しかし，重症度の予測タスクに関しては精度が低く，課題が残る結果となった．今後の展望として，新たな特徴量エンジニアリングの工夫によって，解釈性の高さを維持したまま，より高精度でのうつ病の重症度予測の実現を目指す．

参考文献

[1] Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum D., Rizzo S., & Morency, L. P., The distress analysis interview corpus of human and computer interviews. In Proc. of LREC'14, ELRA, pp.3123-3128, 2014.