RIETI Discussion Paper Series 16-J-054

"声"だけで、うつ病はどこまで診断可能か？
～音声感情認識技術にアンサンブル型機械学習モデルを応用したうつ病スクリーニング機能に関する精度の検証

宗 未来
慶應義塾大学

竹林 由武
福島県立医科大学

関沢 洋一
経済産業研究所

下地 貴明
スマートメディカル株式会社

RIETI
Research Institute of Economy, Trade & Industry, IAA
独立行政法人経済産業研究所
http://www.rieti.go.jp/jp/

RIETI Discussion Paper Series 16-J-054
2016年9月

“声”だけで、うつ病はどこまで診断可能か？
～音声感情認識技術にアンサンブル型機械学習モデルを応用したうつ病スクリーニング機能に関する精度の検証[1]

宗　未来（慶應義塾大学医学部精神神経科学教室）
竹林由武（福島県立医科大学）
関沢洋一（経済産業研究所）
下地貴明（スマートメディカル株式会社）

要旨

近年、音声から感情を推測する技術が開発され、商業化されている（音声感情認識技術）。本研究では、この技術がうつ病の診断に活用できるかを検証した。オンライン調査で、約 2000 名の被験者に 2 ヶ月おきの 3 時点において音声を吹き込んでもらうと共に、うつ病のスクリーニングに使われている質問票に答えてもらい、収集したデータを解析した。最初に、得られた音声情報（パワースペクトル）から pitch、gain、power など 7 種類の音声パラメータを抽出し、個々の音声パラメータと抑うつ指標との間の関係における説明モデルを、3 種類の代表的なアンサンブル型の機械学習を競合させて構築した。具体的には、抑うつ評価尺度の PHQ-9 で 10 点以上を“うつ病”と定義した上で、時点 1 と時点 2 のデータを組み合わせて、SMOTE アルゴリズム(Synthetic Minority Over-sampling Technique)を用いて無作為抽出した 70%のデータで診断精度の高いモデルを構築し、それを使って、残りの 30%のデータについて、“うつ病”の診断精度を検証した。Random forest モデルを用いた機械学習の結果、診断精度の指標とされる ROC 曲線（受信者動作特性曲線：Receiver Operating Characteristic curve）における AUC（曲線下面積: area under the curve）において、性別や年齢といった属性データのみの場合の診断精度が中程度だったのに対して、音声解析のみ、あるいは音声解析と属性データを合わせた場合の方が、高精度でうつ病の診断が可能であることが確認された。しかし、これらの診断モデルを用いても、2 か月の時間間隔を経た時点 3 のデータを用いてのうつ病の診断や予測においては、十分な精度が得られなかった。以上のことから、音声感情認識技術には高い潜在性は示されたものの、更なる技術の改善が必要と考えられた。

キーワード：音声感情認識技術、うつ、診断、機械学習
JEL classification: I10, I31

[1] 本稿は、独立行政法人経済産業研究所におけるプロジェクト「人的資本という観点から見たメンタルヘルスについての研究 2」の成果の一部である。本稿の分析では、平成 27 年度「音声感情認識技術と心理指標・消費マインドとの関係を検証するための Web 調査」のデータを用いた。

１．背景

うつ病は、世界保健機関（WHO）により自殺危険因子でもある重大な障害として認知されてきた[1],[2]。うつ病は医学的治療だけでなく、予防治療、回復施策、保健施策といった面まで考慮せねばならず、その予備軍である未病うつまでを含めると、高額な社会的コストに繋がることも示されている[3]。シンプルで効率的、かつ信頼性の高いうつ病の診断方法が実現すると、それらのコストを大きく削減できる可能性に加えて、迅速な対応や人手をかけない自動的なトリアージといった画期的で有効なメンタルヘルスが実現すると考えられている[4]。そのような安価で容易なうつ病の早期発見の実現により得られる恩恵は図りしれない。また、簡易な抑うつのモニタリングが普及することは、生活における出来事がうつ病に及ぼす影響や因果関係に関する長期的な調査、あるいは治療法の精緻な評価を可能にする方法といった、うつ病研究にも新しい分野が拓かれることで既存の枠組みを超えた革新的な対応法が生まれる可能性も期待されている。

しかし、現在の標準的なうつ病の診断や重症度評価は、時間がかかる上に被験者や評価者の主観に影響を強く受けて客観性を欠き、また評価者の育成自体にも時間がかかるとの指摘は多い。専門家による構造的または半構造的な面接による評価や、自記式の質問事項に対する回答に基づいた様々な評価方法は、うつ状態の顕在的な定量評価には役立つとされるが、多くのうつ症状は、実際には直接測定することはできず、その評価は、ある程度の主観的傾向を帯びることになる。そして、そういった評価や診断の一致のためには多大な評価者への研修も要求される[5, 6]。そのため、診断の精度や治療効果を上げるための客観的な計測が可能な生物学的指標や観察可能な行動指標の組み合わせを見つけようとする研究がこれまでもなされている。

うつ病の生物学的指標の研究は、遺伝子変異[7]や、血漿タンパク質のような生物学的指標[8]、脳波や脳画像[9]といったいくつかの有望な成果を挙げている。しかしながら、今日までに、うつ病に対する決定的な生物学的指標は見つかっていない[7]。近々の研究では、うつ病に対する診断やモニタリングに役立つものとして行動指標という社会的信号処理の使用に期待がもたれてきている[10-12]。

人の一生のはじまりが『産声』からわかるように、声は生命的根源にも直結するインターフェースともいえる。例えば、乳児は『泣き声』から機嫌という感情情報を通じて空腹や体調異常といった生命活動に関わる重大情報を周囲に伝達するし、成長後も『声』から他人の気持ち（本音）を推し量りあうことで円滑な生活が実現するように、『声』は他人の感情を推し量る重要なツールになっている。研究領域においても、発話は、行動に基づいたうつ状態（病）の診断方法の中でも特に重要性が注目されているもののひとつである[13]。

発話がうつ状態（病）に影響を受けるという先行研究は多岐にわたり多く認められるが、それらは Cummins ら[13]のレビューによれば、精神状態と音声という観点からは 1) prosodic features (韻律的特徴：文脈によって変わる強調やリズムといった音声学的性質)、2) source features (音源的特徴：声帯の振動等)、3) formant features (フォルマント的特徴：

重軽やまろやかさといった音色)、そして 4) spectral features (スペクトラル的特徴：音声周波数を成分ごとに分析されたパターンの特徴)の４つが主に、典型的な精神状態に関係する音声研究のサンプルからの特徴抽出に利用されていると報告されている。

これが精神疾患であるうつ病水準の重度の抑うつ状態の判断に寄与するかどうかについての研究は、すべて英語によるものであったが先行研究では文献上認められ[14-16]、これらの研究では上記の組み合わせで機械学習の手法を用いてアルゴリズムを構築することでうつ病水準のうつ状態を有するか否かといった研究がなされている。特に、音声情報という複雑なパラメータの組み合わせから診断アルゴリズムを構築する際に、従来のようなあらかじめ想定された特定のモデルを仮定して最適のパラメータを求める回帰的アプローチではなく、事前にモデルの仮定はせずに入出力されたデータから最適の説明モデルを構築する機械学習アプローチが診断アルゴリズムには有用で、それらの精度が検証されるといった手順でなされる方法論が注目されてきている。

近年、この技術の一部である 4)のスペクトラル技術を応用して、スマートフォンなどに日本語で吹き込まれた音声の物理的な特徴量と、その音声を聞いた評価者によってラベル化された情報から、機械学習によって得られたアルゴリズムに従って感情レベルを判別できる技術が開発されている（音声感情認識技術）[17，18]。我々の知る限り我が国においては、うつ病水準のうつ状態か否かを音声から高い精度で診断できる方法論はこれまで存在していない。そこで、本研究では音声感情認識技術を用いて精度の高いうつ病診断アルゴリズムを作成することを試みた。より具体的には、人による感情評価の情報は含めず、かつ先行研究のように多種の音声情報による組み合わせによる情報も利用せずに、スペクトラル特徴として純粋に周波数のみを利用して、かつ、複数の機械学習アプローチを競わせる形で、属性と周波数情報のみでうつ病水準の抑うつ状態の診断アルゴリズムを構築した。更に、そのアルゴリズムが診断アルゴリズムとして実際に妥当かどうかを検証した。

2. 方法

2-1. 調査対象者

本研究の調査対象者は、NTTコムオンライン・マーケティング・ソリューション株式会社（以下では「調査会社」と呼ぶ）のモニターとして登録している人々のうち、①２０～６９歳の日本全国在住の男女、②募集アンケートにおいて本研究に「参加する」と回答した人々、③募集アンケートの案内に応じて、音声の録音および音声ファイルのアップロードに成功した人々である。

2-2. 手続き

本研究は、筆者（関沢）が勤務する(独)経済産業研究所が調査会社と委託契約を締結することによって、遂行された。2015年9月10日に、調査会社が同社のモニターに対して募集アンケートを行い、本研究への参加に同意し、音声ファイルのアップロードに成功した人々を調査対象者として、以後の調査を行った。

第1回目の調査として、2015年9月24日から10月1日にかけて、調査会社が自社のモニターに電子メールを送ることによって調査参加者を募集し、参加に同意した人々に、「今日は朝から雨が降っています。」という文を音読してもらい、その音声をスマートフォンやパソコンなどに録音してもらった。録音された音声ファイルは調査会社のホームページ上にアップロードしてもらった。その後、属性と心理指標に関するアンケート調査に回答してもらった。

第2回目（2015年11月24日～30日）、第3回目（2016年1月25日～31日）として、第1回目に回答した人々に、第1回目と同じように、音声録音とアンケート調査に回答してもらった。録音する文は、第2回目は「日本には４７の都道府県があります。」で、第3回目は「日本で一番高い山は富士山です。」だった。

本研究は特定医療法人社団 慈藻会平松記念病院倫理委員会の承認を受けて行われた。

2-3. 評価指標

2-3-1. 属性に関する質問

属性に関する質問として、年齢・居住地・就業状況・学歴・婚姻状況・所得（8区分）・同居者数に回答してもらった。就業状況と所得について3回のアンケートの全てで質問しているが、居住地・婚姻状況・同居者数は数か月間で変化しにくいとの判断から第1回目のみで質問している。学歴は手違いにより第1回調査で質問しなかったために第2回目で質問している。また、回答日と回答時刻が調査会社により記録された。

2-3-2. 参照基準（Patient Health Questionnaire-9, PHQ-9)

PHQ-9 は、大うつ病性障害等の診断のために開発された質問票で[19]、日本語版は村松らが作成している[20]。多忙なプライマリケア医が短時間で精神疾患を診断・評価するためのシステムである PRIME-MD（Primary Care Evaluation of Mental Disorders）を Spitzer R.L らが開発し、さらに実施時間の短縮化のために PRIME-MD の自己記入式質問票版として Patient Health Questionnaire（PHQ)を開発した。PHQ はプライマリケア医が日常診療において遭遇する 8 種類の疾患の診断・評価ができるようになっている。PHQ の中から、大うつ病性障害モジュールの 9 個の質問項目を抽出したものが PHQ-9 である。PHQ-9 は多くの言語に翻訳され、妥当性および有用性が検討されており、村松らは Spitzer RL ら と PHQ-9 日本語版を再翻訳法によって作成し、妥当性研究を行っている[21]。英国における国立医療技術評価機構（NICE; National Institute of Health and Clinical Excellence）ガイドラインにおけるうつ病治療のガイドラインや米国精神医学会（American Psychiatric Association APA）がうつ病の評価尺度として PHQ-9 を推奨している。DSM-5（The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition)に対応する 9 個の質問から PHQ-9 は構成されており、過去 2 週間について、「全くない＝0 点」「数日＝1 点」「半分以上＝2 点」「ほとんど毎日＝3 点」となっている。合計点は 0～27 点で、0～4 点はうつ状態でない、5～9 点は軽度のうつ、10

～14 点は中等度のうつ、15～19 点は中等度～重度のうつ、20～27 点は重度のうつとなる。PHQ-9 で、10 点以上はうつ病（大うつ病性障害）水準のうつ状態と知られており、本研究でもその基準を採用した。

2-4. 音声データの抽出方法

発話音声を集音すると、時間領域における波形データが出力される。振幅を音波の特徴が表れるパラメータの一つとし、振幅が最大値となる時間を特徴量の時間成分として定義した。

また、周波数領域でのスペクトル解析には一般的に Fast Fourier Transform（FFT）が用いられる。一般的には関数 f (x) のフーリエ変換を F(w)、i は虚数単位とし、下記の式となる。

Fourier 変換：

$$\mathrm{F(k)} = \int_{-\infty}^{\infty} f(x) e^{2\pi \mathrm{i} k x}\, dx$$

音声の波形データを FFT にかけることで、パワースペクトルという、音の信号成分の強さを表す指標が得られる。これは、音声の特徴を表す指標の一つと考えられており、パワースペクトルが最大となる周波数のうち、最も低い周波数を基本周波数と定義した。

さらに、FFT 結果を時間単位で処理するために一般的に short-time Fourier transform、short-term Fourier transform（STFT）が用いられる。一般的には、w(t)は窓関数として、下記の式となる。

$$\mathrm{STFT}_{\mathrm{x,w}}(\mathrm{t.w}) = \int_{-\infty}^{\infty} x(r) w(r-t) e^{-\mathrm{iwt}}\, dt$$

本研究では、得られた音源から、直接、以下の複数の指標を音声特徴量して抽出し、個々の音声指標と心理指標との間の関係を検証する方向とした。具体的には、対象発話音声に対し、時間幅=0.125 秒,頻度=0.03125 秒ごとに解析し、n 個の FFT 解析データを取得した。取得した n 個のパワースペクトルの平均を①power、基本周波数のうち最頻出を②pitch、pitch の示す強さを③gain、pitch の時間軸の変化を pitchMod、pitch の時間軸の変化総量を④pitchModDiff、gain の時間軸の変化を gainMod、gain の時間軸の変化総量を⑤gainModDiff、2 番目に低い周波数のうち再頻出を⑥pitch2、発話音声の長さを⑦ length として音声特徴量抽出を行い、各々の 7 つの音声指標と 6 つの属性指標を用いて (性別、年齢、配偶関係、就業状況、同居人数、回答時刻)を集積した上で、より至適な説明モデルを機械学習によって構築した。音声指標のうち、pitch, gain, power は対数変換が行われ、また、各指標は標準化された上で学習モデルに投入された。

2-5. 診断精度の指標

検証データに関して、複数の機械学習によって得られた分類結果と参照基準(PHQ-9)による診断結果からクロス集計表を作成し、感度および特異度、陽性的中率、陰性的中率、ROC 曲線（受信者動作特性曲線：Receiver Operating Characteristic curve）における AUC（曲線下面積: area under the curve）を算出することによって最も精度の高いものを最適モデルとして検証が行われた[22]。

感度や特異度は、検査や診断の妥当性を評価する代表的な指標である。例えば、ある疾患の検査を調べるとき、実際に疾患に罹っている人のうち異常ありと出る割合を感度、病気に罹っていない人のうち異常なしと出る割合を特異度と定義されている。一般的には、この感度と特異度が高い検査は信頼性が高いとされる。また、検査で異常ありと出た人のうち実際に疾患を罹っている人の割合を陽性的中率（Positive Predictive Value, PPV）、異常なしと出た人のうち実際に罹っていない人の割合を陰性的中率（Negative Predictive Value, NPV）と呼ぶ。例えば、感度が低いということは実際に病気の人を見逃す（偽陰性）危険性が高いことを意味し、特異度が低いということは、病気でもない人を病気と決めつけてしまう（偽陽性）危険性が高いことを意味する。一般に、感度と特異度はトレードオフの関係にあるとされ、感度が高いほど特異度が下がり、特異度が高いほど感度が下がる危険性が高いとされ、そのもっともバランスのとれた検査が理想とされる。

ROC 曲線は、y 軸に感度、x 軸に 1-特異度（=偽陽性率）を使用してグラフ化したものである（図１）。感度を最大化すると、対応して ROC 曲線の y 値が最大化し、特異度を最大化すると、対応して ROC 曲線の x 値が最小化する。アウトカムと独立変数の関係が完全に偶然だった場合、ROC 曲線は右上から左下への斜点線になり、一方で感度 1、特異度 1 の理想的な独立変数の場合、左上方に位置することになる。

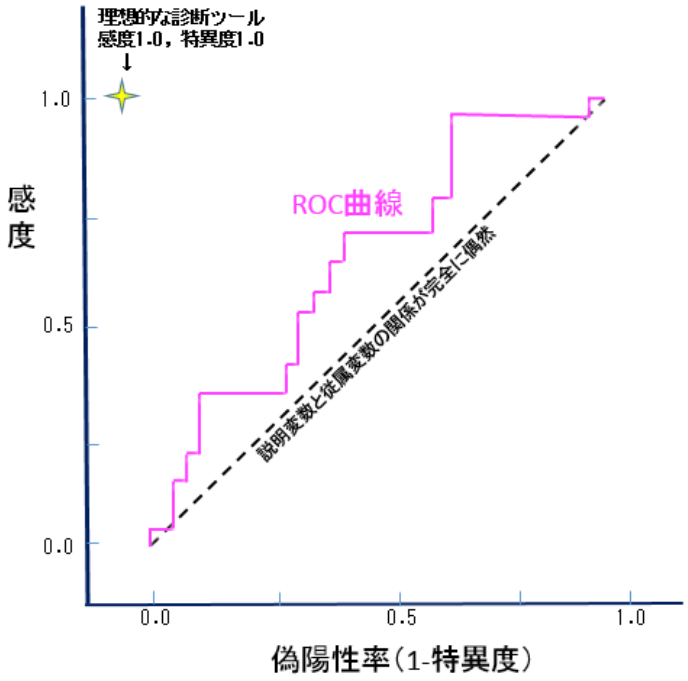


図 1　ROC 曲線（[23]より改編して引用）

この関係の度合いを評価するための指標が、ROC 曲線下面積(AUC: area under the curve)であり、0.5-1.0 の値をとる。診断能・予測能は AUC の値に基づいて一般に、0.9-1.0 で高い精度、0.7-0.9 で中程度の精度、0.5-0.7 で低い精度と解釈される[22]。

2-6. 機械学習によるアプローチとは

機械学習とは、データを入力して機械で解析を行い、そのデータから有用な規則や判断基準などを抽出してアルゴリズムを構築することで、出力未知の入力データに対しても予測を可能にする技術である。機械学習の問題は大きく分けて、訓練付き学習と、訓練無し学習に 2 分されるが、本研究では前者を採用した。これは、入力データが与えられたとき、それに対する出力を正しく予測することが目的で利用されるために、お手本となる訓練用データとも呼ばれる入出力ペアの事例が複数与えられ、これをもとに、新しい入力データが来たときに、それに対する正しい出力をするような関数を構築するものである。その結果、手持ちの訓練データ中には含まれていなかったような入力データに対しても、与えられた訓練データを一般化することで、出力未知のデータに対処可能とする汎化能力が高まるような、学習アルゴリズムを構築することで、予測を可能とする技術とも言える[24]。

例えば、回帰分析のような従来のような線形モデルでは、直線が最もよく説明するモデルであるといった線形性の想定の上で、目的変数を最もよく説明できる、つまり当てはまりのよい説明変数の係数（以下の式における β）が推定され、それを求めることで予測に応用されていた[25]。

$$y_i \sim \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots \beta_n x_{ni} + \varepsilon_i$$

機械学習では、下記の式で表されるような「構築された関数による出力と訓練データとの差」である誤差関数の影響を繰り返しの計算を重ねる中で最小化することで、より未知のデータのできる限り正しい診断を可能にし、予測精度を高める方法論である。一般に、線形モデルでは直観的に理解しやすいモデルが出来上がり、機械学習モデルでは必ずしもそうならないのが特徴のひとつでもある[25]。

$$E(\boldsymbol{\omega}) = \underbrace{\sum_{i=1}^{N} (t_i - f(\boldsymbol{x}_i))^2}_{\text{誤差関数}} + \underbrace{\lambda R(\boldsymbol{\omega})}_{\text{正則化項}}$$

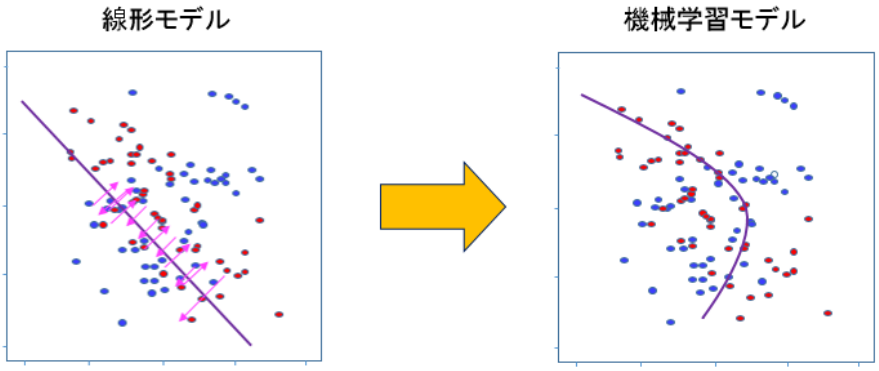


図 2　線形回帰モデルと機械学習モデルによる予測の違い
（[25]より改編して引用）

本研究では、訓練用データに対して 3 つの機械学習のアルゴリズムで、それぞれ診断モデルを構築し、その訓練用データで構築された診断モデルを検証データに適用し、その診断精度を 3 つの方法の精度を比較検証した。

実際には、バギング (bagging)法, ランダムフォレスト(random forest)法, ブースティング(boosting)法という 3 つの代表的なアンサンブル機械学習アルゴリズムを使用した。アンサンブル機械学習とは、個々に学習した複数の学習器（モデル）を融合させて汎化能力（未学習データに対する予測能力）を向上させ一つの学習器（モデル）を作成することで、識別能力が向上させていく機械学習のアプローチである。以下に各集団学習モデルの概要を述べる (詳細は、Hastie et al.[26])。Bagging は、与えられたデータセットから、ブートストラップ法によって複数のデータセットを作成し、そのデータセットを用いて作成した分類結果を統合することで精度を向上させる手法である。Random forest でも、ブートストラップ法によって複数のデータセットを作成する。そしてデータセットの各々で未剪定の最大の決定木を構成する。ただし、分岐のノードはランダムサンプリングされた変数の中

の最善のものを用いた上で、全ての結果を統合し、新しい分類器を構築する。ブースティング(boosting)は、与えた教師付きデータを用いて学習を行い、その学習結果を踏まえて逐次に重みの調整を繰り返すことで複数の学習結果を求め、その結果を統合し、精度を向上させる。個々の集団学習モデルのチューニングは Caret パッケージにより自動化し、個々のモデルで ROC の曲線下面積を指標として最適なモデルを選定した。そして、3 つのアルゴリズムのうち、訓練用データで最も高い診断精度が示されたアルゴリズムを用いて検証用データで診断精度分析を実施し、そこで得られたアルゴリズムによる診断と実際の診断結果のクロス集計から、診断精度を検証した。

Bagging、Random forest、Boosting による集団学習モデルの構築と検証には、R version 3.2.2[27]の Caret パッケージを用いた[28]。各集団学習モデルは Caret パッケージ内の組み込み関数である、ipred[29]、randomForest [30]、gba[31]を使用した。

2-7. データセット

本研究では、時点１～時点３のデータから 1 つの訓練用データと 2 つの検証用データを作成した。 機械学習モデルでは、予測精度の改善を試みる為にデータを複製する手法であるオーバーサンプリング法がデータセットに適用される。本研究のように、うつ群と非うつ群のようなクラスの構成比の偏りが大きい (全サンプルの 10%程度と非うつ群に比べてうつ群が極端に少ない)ような場合には、精度の高いモデルが構築できないという問題が知られている。これは、不均衡クラスの分類問題とも呼ばれ、単純な無作為オーバーサンプリングでは少数のクラス（本研究ではうつ群）が過学習を引き起こすことで予測モデルの汎用性が低下する現象が生じてしまう。そこで、既存のデータの単純な複製ではなく少数派サンプルの近傍データ周辺にノイズを加えたデータを増やしてサンプリングを行うことがその解決に有用であることが知られている。本研究では、そのオーバーサンプリング法の１つである SMOTE(Synthetic Minority Over-sampling Technique)［32］を用いた。最終的に、学習データは SMOTE によって新たに生成された少数派クラスのデータセット（うつ群）と，多数派のクラス（非うつ群）のデータセットを合わせたものを採用した。実際には、時点１、時点２のデータに N=200[オーバーサンプリングする割合(N)], k=5[k 最近傍(k-nearest)の値 (k)]とする SMOTE を適用しアウトカム分類の不均衡を補正したデータセットを生成した。このデータセットに対して、訓練用データが 70%、検証用データ 1 が 30%になるよう無作為抽出を行った。また、時点 3 のデータに対して同様に SMOTE を適用し、これを検証用データ 2 とした。

音声解析指標の有用性を検討するために、PHQ-9 による抑うつ状態をアウトカムとし、1) 音声指標と属性指標 (モデル 1)、2)属性指標のみ (モデル 2)、3)音声指標のみ (モデル 3) をそれぞれ説明変数とする 3 つのモデルで検討を行った。

加えて、時点 1 や時点 2 という“過去の音声”から時点 3 という 2 か月後の“未来のうつ状態”をどれだけ時間を越えて予測できるかという、本アルゴリズムによるうつ病診断の予測性能を評価するために、検証データ 1 における集団学習による診断結果と、時点 3 の

PHQ-9 による診断結果間での診断精度を検討した（＝予測データ）。

3.結果

3-1.属性

対象者の属性は表に示した通りである。参加人数の合計は 2273 人であり、男性（1478 人）が女性（795 人）の約 2 倍であり、平均年齢は 45.5（±10.9）歳だった。婚姻状況は既婚者が 60.7%と最大を占め、最終学歴は大学・大学院卒が 59.8%と最大であり、77.2%と大多数が就業者だった。年間の世帯収入は、300 万円以上～500 万円未満、500 万円以上～700 万円未満がほぼ同じ 2 割程度を占め最頻であった。PHQ-9 の平均点は、時点 1～3 を通じて、それぞれ軽度うつ状態を示すとされる 5 点を下回っており、健常範囲だった。

表 1　対象者の属性

性別	男性	1,478 (65.0%)
	女性	795 (35.0%)
年齢	平均(標準偏差)	45.5 (10.9)
婚姻状況		
	既婚	1,379 (60.7%)
	離婚	112 (4.9%)
	死別	15 (0.7%)
	未婚	767 (33.7%)
最終学歴		
	中学校その他	31 (1.6%)
	高校	410 (20.7%)
	短大・高専・専門学校	356 (18.0%)
	大学・大学院	1183 (59.8%)
就労状況		
	働いている	1,701 (77.2%)
	無職(求職中)	125 (5.7%)
	無職(求職中でない)	378 (17.2%)
PHQ-9		
平均(標準偏差)	時点1	4.2 (4.9)
	時点2	4.5 (5.2)
	時点3	4.3 (5.1)
10点以上	時点1	277 (12.2%)
	時点2	323 (16.0%)
	時点3	255 (13.8%)
年間の世帯収入		
	全くない(0円)	22 (1.0%)
	1円以上100万円未満	67 (3.0%)
	100万円以上～200万円未満	103 (4.5%)
	200万円以上～300万円未満	200 (8.8%)
	300万円以上～500万円未満	501 (22.0%)
	500万円以上～700万円未満	493 (21.7%)
	700万円以上～1，000万円未満	398 (17.5%)
	1，000万円以上（月額83.3万円以上）	267 (11.8%)
	答えたくない	222 (9.8%)
同居人数(自分自身を含む)		
	1名	467 (20.6%)
	2名	568 (25.0%)
	3名	555 (24.4%)
	4名	470 (20.7%)
	5名以上	213 (9.4%)

3-2. 参照基準

訓練用データの全レコード数は 2642、アウトカムの度数は、うつ病群 1132 (43%)、非うつ病群が 1510 (57%)であった。検証用データ 1 の全レコード数は 1131 であり、アウトカムの度数は、うつ病群 485 (43%)、非うつ病群が 646 (57%)であった。検証用データ 2 の全レコード数は 1750 であり、アウトカムの度数は、うつ病群 750 (43%)、非うつ病群が 1000 (57%)であった。なお、3 つのモデル(音声指標+属性指標、属性のみ、音声のみ)では説明変数の組み合わせが異なるため、属性のみのモデルは音声指標における欠測が少ない分、全体の人数は多くなっている(訓練用データの属性指標のみモデル=2720, 検証用データの属性のみのモデル=1165)。

3-3. 訓練データ

訓練用データに対し 3 つの集団学習アルゴリズムを適用した結果、説明変数の組み合わせが異なる 3 つのモデルにおいて、Random forest アルゴリズムによる分類精度が他のアルゴリズムによりも、診断精度が高いことが示された。特に、音声指標に加えて属性データを加えた学習において、最も高い精度が示された (ROC: 0.92, 感度: 0.75, 特異度: 0.90)。この Random Forest モデルは、最終的に 10 の説明変数を用いて決定された。

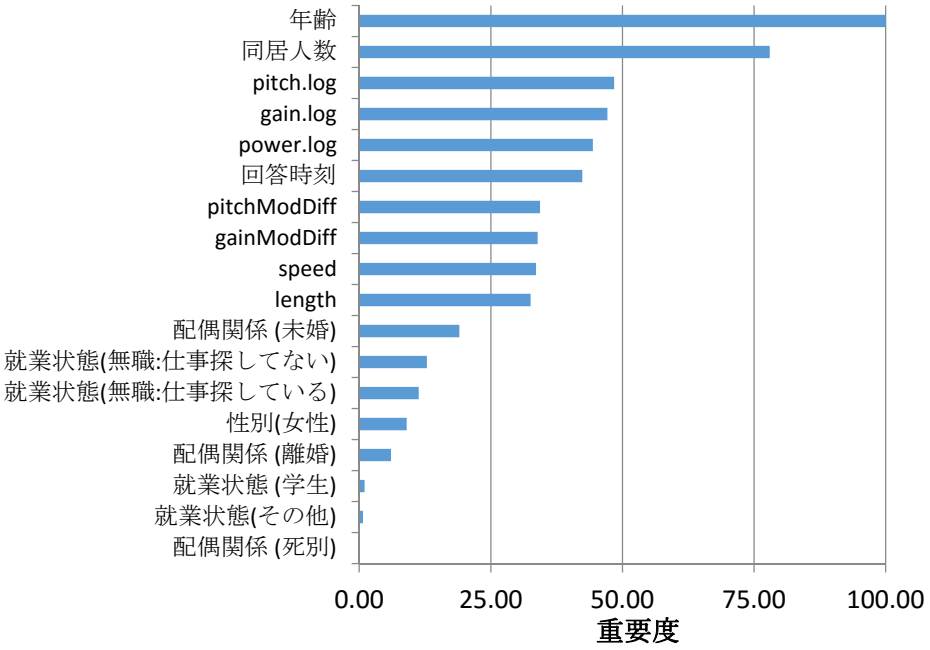
表 2 訓練データにおける集団学習による診断精度

	訓練データ (T1+T2: 70%)		
	ROC (AUC)	感度	特異度
モデル1: 音声指標+属性指標			
Bagging	0.90	0.74	0.89
Random Forest	0.92	0.75	0.90
Boosting	0.88	0.71	0.89
モデル2: 属性指標			
Bagging Tree	0.87	0.70	0.87
Random Forest	0.88	0.69	0.89
Boosting	0.86	0.67	0.88
モデル3: 音声指標のみ			
Bagging	0.88	0.72	0.86
Random Forest	0.90	0.73	0.89
Boosting	0.76	0.52	0.88

訓練データにおけるランダムフォレストモデルの説明変数の重要度を算出した。音声解析指標と属性指標を含んだ random forest による学習モデルにおける、変数の重要度を図 3

に示す。重要度は、当該モデルにおいてアウトカム分類への寄与の高さを反映する。年齢、同居人数に次いで、音声解析指標がアウトカム分類に寄与していることが示された。音声解析指標はすべて、年齢、同居人数、回答時刻以外の属性データよりも重要度が高いことが示された。このうち、上位 10 の説明変数が、最終的なクラス分類に使用されたことになる。

図 3　説明変数の重要度



3-4. 検証データ

検証データ 1 について、訓練データで高い精度を示した random forest model による分類を行った結果と参照基準による結果のクロス集計から診断精度を算出した (表 3 から表 5)。その結果、訓練データと同様、音声のみ、または属性のみのモデルよりも、音声データと属性データを組み合わせたモデルにおいて、診断精度が高いことが示された。ROC の値は 0.91 と高い精度を示していた。一方、検証データ 2 では、音声データと属性データを組み併せたモデルでは、属性データのみのモデルと比べて診断精度の向上が示されなかった (表 6 から表 8)。

表 3　検証データ 1 における音声指標と属性データ(モデル 1)による診断精度

	モデル1: 音声+属性			
		参照基準 (PHQ9)		
		抑うつ	非抑うつ	合計
指標検査 (random forest による分類)	抑うつ	354	63	417
	非抑うつ	131	583	714
	合計	485	646	1131
	精度指標		点推定値(95%CI)	
	感度		0.73 (0.69, 0.77)	
	特異度		0.90 (0.88, 0.92)	
	ROC (AUC)		0.91 (0.89, 0.92)	
	陽性的中率		0.85 (0.81, 0.88)	
	陰性的中率		0.82 (0.79, 0.85)	

表 4　検証データ 1 における属性データ(モデル 2)による診断精度

	モデル2: 属性のみ			
		参照基準(PHQ9)		
		抑うつ	非抑うつ	合計
指標検査 (random forest による分類)	抑うつ	348	88	436
	非抑うつ	151	578	729
	合計	499	666	1165
	精度指標		点推定値(95%CI)	
	感度		0.70 (0.66, 0.74)	
	特異度		0.87 (0.84, 0.89)	
	ROC (AUC)		0.87 (0.85, 0.89)	
	陽性的中率		0.80 (0.76, 0.83)	
	陰性的中率		0.79 (0.76, 0.82)	

表 5　検証データ 1 における音声データ(モデル 3)による診断精度

	モデル3: 音声のみ			
		参照基準(PHQ9)		
		抑うつ	非抑うつ	合計
指標検査 (random forest による分類)	抑うつ	357	81	438
	非抑うつ	128	565	693
	合計	485	646	1131
	精度指標		点推定値(95%CI)	
	感度		0.74 (0.69, 0.77)	
	特異度		0.87 (0.85, 0.90)	
	ROC (AUC)		0.91 (0.89, 0.92)	
	陽性的中率		0.82 (0.78, 0.85)	
	陰性的中率		0.82 (0.78, 0.84)	

表 6　検証データ 2 における音声データと属性データ(モデル 1)による診断精度

	モデル1: 音声+属性			
		参照基準 (PHQ9)		
		抑うつ	非抑うつ	合計
指標検査 (random forest による分類)	抑うつ	524	174	698
	非抑うつ	226	826	1052
	合計	750	1000	1750
	精度指標		点推定値(95%CI)	
	感度		0.70 (0.66, 0.73)	
	特異度		0.83 (0.80, 0.85)	
	ROC (AUC)		0.84 (0.83, 0.86)	
	陽性的中率		0.75 (0.72, 0.78)	
	陰性的中率		0.79 (0.76, 0.81)	

表 7 検証データ 2 における属性データ(モデル 2)による診断精度

	モデル2: 属性のみ			
		参照基準 (PHQ9)		
		抑うつ	非抑うつ	合計
指標検査 (random forest による分類)	抑うつ	553	203	756
	非抑うつ	197	797	994
	合計	750	1000	1750
	精度指標		点推定値(95%CI)	
	感度		0.74 (0.70, 0.77)	
	特異度		0.80 (0.77, 0.82)	
	ROC (AUC)		0.86 (0.84, 0.88)	
	陽性的中率		0.73 (0.70, 0.76)	
	陰性的中率		0.80 (0.78, 0.83)	

表 8 検証データ 2 における音声データ(モデル 3)による診断精度

	Model3: 音声のみ			
		参照基準 (PHQ9)		
		抑うつ	非抑うつ	合計
指標検査 (random forest による分類)	抑うつ	306	240	546
	非抑うつ	444	760	1204
	合計	750	1000	1750
	精度指標		点推定値(95%CI)	
	感度		0.41 (0.37, 0.44)	
	特異度		0.76 (0.73, 0.79)	
	ROC (AUC)		0.86 (0.84, 0.88)	
	陽性的中率		0.56 (0.52, 0.60)	
	陰性的中率		0.63 (0.60, 0.66)	

3-5.予測データ

予測データでも、音声データと属性データを組み合わせたモデルは、属性データのみのモデルと比べて予測精度の向上が示されなかった(表 9 から表 11)。予測データにおいては、

特に陽性的中率が低下したことから、参照基準で非抑うつと診断される人を誤って抑うつと分類するエラーが高くなったといえる。

表 9 予測データにおける音声データと属性データ(Model1)による予測精度

	Model1: 音声+属性			
		参照基準 (PHQ9)		
		抑うつ	非抑うつ	合計
指標検査 (random forest による分類)	抑うつ	220	152	372
	非抑うつ	95	522	617
	合計	315	674	989
	精度指標		点推定値(95%CI)	
	感度		0.70 (0.64, 0.75)	
	特異度		0.77 (0.74, 0.81)	
	陽性的中率		0.59 (0.54, 0.64)	
	陰性的中率		0.85 (0.82, 0.87)	

表 10 予測データにおける属性データ(Model2)による予測精度

	Model2: 属性			
		参照基準 (PHQ9)		
		抑うつ	非抑うつ	合計
指標検査 (random forest による分類)	抑うつ	217	148	365
	非抑うつ	118	529	647
	合計	335	677	1012
	精度指標		点推定値(95%CI)	
	感度		0.65 (0.59, 0.70)	
	特異度		0.78 (0.75, 0.81)	
	陽性的中率		0.59 (0.54, 0.65)	
	陰性的中率		0.82 (0.79, 0.85)	

表 11 予測データにおける音声データ(Model3)による予測精度

	Model3: 音声			
		参照基準 (PHQ9)		
		抑うつ	非抑うつ	合計
指標検査 (random forest による分類)	抑うつ	201	165	366
	非抑うつ	106	492	598
	合計	307	657	964
	精度指標		点推定値(95%CI)	
	感度		0.65 (0.60, 0.71)	
	特異度		0.75 (0.71, 0.78)	
	陽性的中率		0.55 (0.50, 0.60)	
	陰性的中率		0.82 (0.79, 0.85)	

４．考察

当初、我々は、得られた音声の物理的な特徴量と、過去に音声を聞いた評価者によってラベル化された情報から機械学習によって構築されたアルゴリズムを用いて、音声からのうつ病診断に関する精度検証を試みた。しかし、その結果は、感度 0.32 [95%CI 0.28-0.35]、特異度 0.72 [95%CI 0.71-0.72]、陽性的中率（PPV）は 0.14 [95%CI 0.13-0.16]、陰性的中率（NPV）0.87 [95%CI 0.87-0.88]であった。これは、音声判定による分析では、実際にうつ病水準のうつ状態を有している人間の 31.5%しか探知できず、また、音声分析によってうつ病水準のうつ状態を有すると判定された人間において非うつ状態であるリスク（偽陽性率）は 28.5%ということを意味し、うつ病スクリーニングとしての判定ツールとしては感度が著しく低く、問題があると判断された。

そこで、より精度の高いうつ病診断アルゴリズムを作成する必要性が生じた。そのために、本研究では人による感情評価の情報は説明変数に含めず、かつ先行研究のような多種の音声情報による組み合わせによってではなく、スペクトラル特徴の中でも純粋に周波数から抽出した、時間当たりに音が伝えるエネルギーであるパワースペクトルから得られるパラメータのみを利用して、これに属性データを組み合わせて、Bagging、Random Forest、Boosting という複数の機械学習アプローチを競わせる形で、どこまで最適の精度を持つ診断アルゴリズムを構築できるかの検証を試みた。その結果、まずは時点１と時点２の音声サンプルを集めた中の 70%から構築された訓練データを対象にした解析からは、音声指

標＋属性データ、音声指標のみ、属性データのみ、のすべてのモデルにおいて、Random forest アルゴリズムによる分類精度が他のアルゴリズムによりも、診断精度が高いことが示された。特に、音声指標に加えて属性データを加えた学習において、最も高い精度が示された（ROC：0.92，感度：0.75，特異度：0.90)。そして、これを残りの 30%に対して妥当性を確認した検証データにおいても、やはり音声情報と属性情報の組み合わせによるモデルにおいて感度 0.73（95%CI 0.69 - 0.77)、特異度は 0.90（95%CI 0.88- 0.92)で ROC における AUC は 0.91（95%CI：0.89 - 0.92)と同じサンプル集団内における検証では高い精度での診断能を有することが示された。しかし、これらのアルゴリズムを用いても、約 2 か月の時間を経た時点 3 の音声データや、過去（時点 1 と時点 2）の声から未来（時点 3）のうつ病水準の抑うつ状態を予測するための予測データに対して精度を検証した場合に、音声データと属性データを組み併せたモデルが、属性データのみのモデルと比べて診断精度の向上が示されなかった。

本研究は、我々の知る限り音声によってうつ病水準のうつ状態か否かの診断能を検証した日本語における初めての研究である。かつ、特に、属性データに加えて、純粋に周波数から得られた情報のみで、検証データ 1 において、高精度の診断アルゴリズムが構築できた。

発話とうつ状態の関係を探索する先行研究は、大きく 1）うつ状態の有無（＝健常かうつ状態かの識別）、2）うつ状態の重症度判定、3）うつ状態の得点予測（=未知の音声サンプルから、うつ状態を連続値で予測）の３領域に分類される[13]。本研究はこの定義に従うと 1)に分類され、これまで 3 件の先行研究がすべて英語音声を用いた研究が報告されている。Moore らによる韻律、声質、スペクトラル特徴および声帯の運動特徴から得られた解析では、男性で精度 0.91（感度 0.89、特異度 0.93）、女性で 0.96（感度 0.98、特異度 0.94）という高い精度が示されている[14]。ここでは特に声帯の運動特徴から得られた特徴が、うつ状態の識別に有用であったと強調されている。同様のアプローチを用いた Leo らは、精度が 0.50-0.70 というものだった[15]。この研究では、スペクトラル特徴よりも発話時に声道内に発生する瞬時的なエネルギーの変化を特徴量にするために開発された演算子である Teager Energy Operator（TEO)や、声帯の運動特徴の方が、スペクトラル的特徴よりも精度に寄与すると示していた。つまり、これら２つの研究では音声よりも筋緊張や喉頭の動きに関する情報の方がうつ状態との関連が大きいというものであった。さらに、やはり複数の特徴に機械学習を取り入れた Ooi らの研究では、対象は若者に限定されているが精度 0.73（感度 0.79、特異度 0.67)であった[16]。その研究でも声帯の運動特徴は非常に高い説明能力を有しているとの見解が記述されている。これらと比較しても、本研究では声帯の運動特徴を含む複数の特徴を利用せずに、これらを利用した過

去の研究と比較しても最も高い精度であったMooreの結果に匹敵する水準での診断精度が得られていた。

実際、訓練データにおけるランダムフォレストモデルの特徴量(各要因)の重要度を算出すると、音声解析指標はすべて、年齢、同居人数、回答時刻以外の属性データよりも重要度が高く上位にくることが示されていた。一方で、Mooreらは、彼等のツールが大規模なデータには汎用できないことに言及しているが、本研究も2か月後に聴取された時点3での音声データを利用した検証（検証データ2、予測データ）では診断における音声情報の寄与が示せておらず、これらの精度を高めることは今後の課題のひとつと考えられた。

今回、これだけの精度が得られた背景には、複数の機械学習を競合させて最も高い精度のものを選ぶという解析戦略を取り入れたことが挙げられる。これまでの先行研究では、主流の機械学習はSupport Vector Machines (SVM)[33, 34] とGaussian Mixture Models (GMM)[35]であった。どちらも代表的な機械学習のアプローチであるが、本研究ではBagging Tree、Random Forest、Boosting というどれも新しいアプローチを採用していた。特に、Random Forestは2001年に開発されたばかりで新しい上に精度が高いことで知られ、多領域の研究ではSVMよりも精度が高かったいという報告が認められる[36]。これまで発話音声とうつ病の研究では、我々の知る限りまだRandom Forestは利用されておらず、今回の結果にこれが寄与した可能性も考えられる。

本来、事前に特定のモデルを想定せずに入力データと出力データから、最適な説明モデルを構築するというのが機械学習の理念であることを考えると、事前に採用する特定の機械学習アルゴリズムを設定しないという今回のアプローチは、より機械学習の理念に近いアプローチを採用しているとも言え、結果的にそれがより高い説明能力を有する診断モデルの構築に寄与したかもしれないと推測できた。一方で、将来的には今回のモデルに加え、SVMやGMMで得られた精度との直接比較を行うことが、より精度の高いモデル構築の方法論の検証にもつながっていくと考えられた。

更に、他に今後の精度を高めるために念頭に置くべき点として、３つのポイントが挙げられる。ひとつめに、今回の収録した音声について、多くの音声において被験者の発声音以外の生活音等の非定常ノイズが多く録音されていた点である。このようなノイズが音声パラメータ抽出時に影響を与えた可能性が高く、診断精度が落ちた可能性が挙げられる。また、音声収録のインターフェイスが、パソコン、スマートフォンといった形で不揃いであったため、録音された音声の状態が利用者によってばらつきが出てしまったことも原因と考えられる。さらに、抑うつの心理指標として利用したPHQ-9については、そもそもがうつ病の評価のために開発された尺度で、質問者の過去2週間の心理を回想しながら回答するものであり、音声録音時での気分状態との時間的な隔たりがあったために、診断精度が下がった可能性も考えられる。そのため、今後の実験では、ノイズをクリアにする方法

論の検討、録音される音声についてある一定の基準となりうるよう、ハードウェア、ソフトウェアでの調整を行うこと、気分の時間変化に対応しているような適切な心理指標アンケートを利用することで、より一層の信頼精度を高める可能性が期待される。

５．今後に向けて

うつ病の現行の診断システムでは、その成功の鍵は患者の外観と話される情報から得られる臨床家のスキルと経験に依存しているのが現実である。そのため、シンプルで低コスト、自動的で客観的な診断補助が、医療場面だけでなくストレスチェック精度が導入された企業も含めた多様な場面での高いニーズとなっている。そういったツールが登場すれば、予防から治療に至るまで、うつ病対応に革新的な変化が訪れるとも考えられ、メンタルヘルスで苦しむ人々の QOL を向上させると期待される。複雑な症候群と考えられているうつ病という病態を考えた時、結果的には、たったひとつの決定的な生物学的、生理学的、行動的な指標が見つかるということはなく、精度の高い診断には、複数の指標による多軸的なアプローチが必要になるであろうと予測される。そして、発話はそれらのうつ病の重要な客観的指標の中で鍵となる重要な要因のひとつとして考えられる。そして、その発話もやはり、既述のように発話に関わる多様な要因の寄与が検証されている。まだまだ、これらの診断技術は多様な技術イノベーションを待たなければならず、過渡期と言ってよい時期である。しかし、最終的に目指すべきゴールは多軸的な診断技術の確立ではあろうともスポーツの世界で、「個のさらなるレベルアップが、チーム全体の力の底上げに繋がる」と喩えられるように、うつ病診断においても今回のように周波数だけからうつ病を診断できる精度を高めていこうとする個々の技術の精度を高めていく姿勢は、今後一層、期待される。

利害相反：本演題に関して、筆頭研究者に開示すべき利益相反はない。

謝辞：本稿執筆にあたって、三重大学医学部の初村拓毅氏に大変有益なご指導をいただきました。この場を借りて、深く感謝の意を表します。

引用文献

1. Kessler, R.C., et al., *The epidemiology of major depressive disorder: Results from the national comorbidity survey replication (ncs-r).* JAMA, 2003. **289**(23): p. 3095-3105.
2. Üstün, T.B., et al., *Global burden of depressive disorders in the year 2000.* The British Journal of Psychiatry, 2004. **184**(5): p. 386-392.
3. Olesen, J., et al., *The economic cost of brain disorders in Europe.* European Journal of Neurology, 2012. **19**(1): p. 155-162.
4. World Health Organisation. *Prevention of Mental Disorders*. 2004 [cited 2016 6/18]; Available from: http://www.who.int/mental_health/publications/prevention_mh_2004/en/.
5. Blais, M. and L. Baer, *Understanding Rating Scales and Assessment Instruments*, in *Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health*, L. Baer and A.M. Blais, Editors. 2010, Humana Press: Totowa, NJ. p. 1-6.
6. Mundt, J.C., et al., *Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology.* Journal of Neurolinguistics, 2007. **20**(1): p. 50-64.
7. Schmidt, H.D., R.C. Shelton, and R.S. Duman, *Functional Biomarkers of Depression: Diagnosis, Treatment, and Pathophysiology.* Neuropsychopharmacology, 2011. **36**(12): p. 2375-2394.
8. Domenici, E., et al., *Plasma protein biomarkers for depression and schizophrenia by multi analyte profiling of case-control collections.* PLoS One, 2010. **5**(2): p. e9166.
9. Steiger, A. and M. Kimura, *Wake and sleep EEG provide biomarkers in depression.* J Psychiatr Res, 2010. **44**(4): p. 242-52.
10. Girard, J.M., et al., *Nonverbal Social Withdrawal in Depression: Evidence from manual and automatic analysis.* Image Vis Comput, 2014. **32**(10): p. 641-647.
11. Joshi, J., et al., *Multimodal assistive technologies for depression diagnosis and monitoring.* Journal on Multimodal User Interfaces, 2013. **7**(3): p. 217-228.
12. Scherer, S., et al., *Automatic audiovisual behavior descriptors for psychological disorder analysis.* Image and Vision Computing, 2014. **32**(10): p. 648-658.
13. Cummins, N., et al., *A review of depression and suicide risk assessment using speech analysis.* Speech Commun, 1015. **71**(10): p. 49.
14. E. Moore, I., et al., *Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech.* IEEE Transactions on Biomedical Engineering, 2008. **55**(1): p. 96-107.
15. Low, L.S.A., et al., *Detection of Clinical Depression in Adolescents’ Speech During Family Interactions.* IEEE Transactions on Biomedical Engineering, 2011. **58**(3): p. 574-586.
16. Ooi, K.E.B., M. Lech, and N.B. Allen, *Multichannel Weighted Speech Classification System for Prediction of Major Depression in Adolescents.* IEEE Transactions on Biomedical Engineering, 2013. **60**(2): p. 497-506.

17. 酒造正樹, *情動・感情判別のための自然発話音声データベースの構築.* 情報処理学会論文誌 2011. **52**(3): p. 1185-1194.
18. 門谷信愛希, *音声に含まれる感情の判別に関する検討.* Technical Report of IEICE, SP, 2000. **100**(522): p. 43-48.
19. Kroenke, K., R.L. Spitzer, and J.B. Williams, *The PHQ-9: validity of a brief depression severity measure.* J Gen Intern Med, 2001. **16**(9): p. 606-13.
20. 村松公美子, *Patient Health Questionnaire (PHQ-9, PHQ-15) 日本語版および Generalized Anxiety Disorder -7 日本語版 —up to date —.* 臨床心理学研究, 2014. **7**: p. 35-39.
21. Muramatsu, K., et al., *The patient health questionnaire, Japanese version: validity according to the mini-international neuropsychiatric interview-plus.* Psychol Rep, 2007. **101**(3 Pt 1): p. 952-60.
22. Akobeng, A.K., *Understanding diagnostic tests 3: Receiver operating characteristic curves.* Acta Paediatr, 2007. **96**(5): p. 644-7.
23. 大阪大学大学院医学系研究科老年・腎臓内科学腎臓内科. *Clinical Journal Club 5. ROC 曲線* 2009 [cited 2016 7/22]; Available from: http://www.med.osaka-u.ac.jp/pub/kid/clinicaljournalclub6.html.
24. 鹿島久嗣. *機械学習 (Machine Learning) の紹介.* 2016 [cited 2016 7/22]; Available from: http://www.geocities.co.jp/Technopolis/5893/machinelearning.html.
25. OZAKI, T.J. *「統計学と機械学習の違い」はどう論じたら良いのか「六本木で働くデータサイエンティストのブログ」.* 2015 [cited 2016 7/22]; Available from: http://tjo.hatenablog.com/entry/2015/09/17/190000.
26. Hastie, T., R. Tibshirani, and J.H. Friedman, *The elements of statistical learning: data mining, inference, and prediction, 2nd.* 2009, New York, NY: Springer.
27. Team, R.C. *R: A language and environment for statistical computing.* 2015; Available from: https://www.R-project.org/.
28. Tang, Y. and C. Candan. *caret: Classification and Regression Training. R package version 6.0-64.* 2016; Available from: https://CRAN.R-project.org/package=caret.
29. Peters, A. and T. Hothorn. *ipred: Improved Predictors. R package version 0.9-5.* . 2015; Available from: https://CRAN.Rproject.org/package=ipred.
30. Liaw, A. and M. Wiener, *Classification and regression by randomForest.* R news, 2002. **2**(3): p. 18-22.
31. others, G.R.w.c.f. *gbm: Generalized Boosted Regression Models. R package version 2.1.1.* 2015; Available from: https://CRAN.R-project.org/package=gbm.
32. Chawla, N.V., et al., *SMOTE: Synthetic Minority Over-sampling Technique.* Journal of Artificial Intelligence Research, 2002 **16**: p. 321-357.
33. V. Vapnik and A. Lerne, *Pattern recognition using generalized portrait method* Automation and Remote Control, 1963. **24**.

34. Tsochantaridis, I., et al., *Large Margin Methods for Structured and Interdependent Output Variables.* The Journal of Machine Learning Research archive 2005. **6**: p. 1453-1484

35. Reynolds, D., *Gaussian Mixture Models*, in *Encyclopedia of Biometrics*, S.Z. Li and A.K. Jain, Editors. 2015, Springer US: Boston, MA. p. 827-832.

36. Díaz-Uriarte, R. and S. Alvarez de Andrés, *Gene selection and classification of microarray data using random forest.* BMC Bioinformatics, 2006. **7**(1): p. 1-13.