

一般発表 | Track 5: 高度なデータ活用・ドメイン応用（医療情報、教育、地理情報等）

🎵 2024年2月29日(木) 16:00 ~ 18:10 | 📺 T5-A オンライン（Zoom Events）

## 医療・ヘルスケア (2)

座長:横田 治夫(城西大学)

コメンテータ:白石 優旗(筑波技術大学)

16:00 ~ 16:05

イントロダクション

16:05 ~ 16:30

[T5-A-6-01] 福祉支援施設の支援記録からのインシデントの予兆検出手法の改良

\*山本 隆太<sup>1</sup>、上野 史<sup>2</sup>、太田 学<sup>2</sup> (1. 岡山大学大学院環境生命自然科学研究科、2. 岡山大学学術研究院環境生命自然科学学域)

16:30 ~ 16:55

[T5-A-6-02] レム睡眠のリアルタイム自動検出に関する研究

\*佐野 史弥<sup>1</sup>、国生 泰資<sup>1</sup>、堀江 和正<sup>2</sup>、古木 淳也<sup>4</sup>、鈴木 陽子<sup>3</sup>、阿部 高志<sup>3</sup>、北川 博之<sup>3</sup> (1. 筑波大学大学院 理工情報生命学術院 システム情報工学研究群 知識・データ工学研究室、2. 筑波大学計算科学研究センター、3. 筑波大学国際統合睡眠医科学研究機構、4. 筑波大学大学院人間総合科学学術院 人間総合科学研究群)

16:55 ~ 17:20

[T5-A-6-03] リアルタイム睡眠ステージ判定システムの開発と評価

\*国生 泰資<sup>1</sup>、堀江 和正<sup>2</sup>、鈴木 陽子<sup>3</sup>、阿部 高志<sup>3</sup>、北川 博之<sup>3</sup> (1. 筑波大学システム情報工学研究群知識・データ工学研究室、2. 筑波大学計算科学研究センター、3. 筑波大学国際統合睡眠医科学研究機構)

17:20 ~ 17:45

[T5-A-6-04] 特定健診データに含まれる自由記述に着目した予後予測

\*大塚 皇輝<sup>1</sup>、池之上 辰義<sup>2</sup>、福岡 真悟<sup>3</sup>、若宮 翔子<sup>1</sup>、荒牧 英治<sup>1</sup> (1. 奈良先端科学技術大学院大学 ソーシャルコンピューティング研究室、2. 滋賀大学 データサイエンス・AIイノベーション研究推進センター、3. 京都大学 医学研究科 人間健康科学系専攻)

17:45 ~ 18:10

[T5-A-6-05] AI 音楽療法士（さつきさん）の開発研究

\*小杉 尚子<sup>1</sup>、石井 健太郎<sup>1</sup>、児玉 直樹<sup>2</sup> (1. 専修大学、2. 新潟医療福祉大学)

# 福祉支援施設の支援記録からのインシデントの予兆検出手法の改良

山本 隆太<sup>†</sup> 上野 史<sup>††</sup> 太田 学<sup>††</sup>

<sup>†</sup> 岡山大学大学院環境生命自然科学研究科 〒700-8530 岡山市北区津島中 3-1-1

<sup>††</sup> 岡山大学学術研究院環境生命自然科学学域 〒700-8530 岡山市北区津島中 3-1-1

E-mail: <sup>†</sup>p0pm8dxf@s.okayama-u.ac.jp, <sup>††</sup>{uwano, ohta}@okayama-u.ac.jp

**あらまし** 上野らは、福祉支援施設の入所者が起こすトラブル（インシデント）の予兆検出のため、施設の支援記録本文を用いた BERT によるインシデントの予兆検出手法を提案した。また実験により、支援記録に含まれる日付などの付随情報を使用することで分類精度が向上する可能性も示唆した。しかし、入所者毎に個別に予兆検出器を学習したため、インシデントをあまり起こさない入所者に対して検出精度が悪かった。そこで本研究では、他の入所者の支援記録を学習に使用した分類器による実験を行い、入所者本人の支援記録のみで学習した分類器と精度を比較する。さらに本研究では、付随情報を利用することで分類精度が向上する可能性について検討した。実験の結果、分類対象者を除く全ての入所者の支援記録で学習する汎用分類器の精度が最もよく、施設 A で平均 F 値 0.790、施設 B で平均 F 値 0.753 となった。またその汎用分類器にインシデントの有無の付随情報を利用した場合、さらに精度が向上し施設 A で平均 F 値 0.825、施設 B で平均 F 値 0.772 となった。

**キーワード** インシデント予兆検出, BERT, 付随情報, 医療・ヘルスケア, 福祉支援施設

## 1 はじめに

近年、高齢者や障がい者が入所する福祉支援施設の需要が高まってきている。福祉支援施設では、入所者の人数に対する職員の数は限られており、支援には限度がある。そのため、入所者による器物破損や他者への傷害などといったインシデントが施設内で発生すると、業務の妨げとなり、入所者は十分な支援が受けられなくなる。しかし、インシデントを事前に察知することはベテランの職員でも難しいため、入所者が起こすインシデントの予兆を発見し未然に防止することは重要な課題である。

本研究では、福祉支援施設の支援記録からインシデントの予兆を検出する<sup>1</sup>。本研究では、実際に現場で働く施設の職員から見て、支援記録本文に利用者の衝動性が高い状態、あるいはネガティブな心理状態が現れているものをインシデントの予兆と定義している。本研究で利用する手法は、上野らが [1] で提案したものであり、上野らは福祉支援施設の利用者 10 人の支援記録を用いて、各入所者のデータのみで個別に学習した分類器の予兆検出性能を検証した。予兆の検出は、予兆の有無を示すラベルを用いた 2 クラス分類により実現しており、その際上野らは入力として、支援記録の記録本文のみでなく支援記録に含まれる日付などの付随情報を使用することで、分類精度が向上する可能性を示唆している。本研究では、上野らの手法を予兆の程度に応じた 4 値分類に拡張し、予兆を検出する。具体的には、支援記録の本文に加えて日付データなどの付随情報も利用した分類器を作成し、分類器に入力された支援記録を「予兆なし」「予兆（弱）」「予兆（強）」「予兆（両方）」の 4 クラスに分類することで予兆の検出を実現する。4 値分類への拡張により、施設

の職員が予兆の程度に応じてより適切な対応ができるようになる。検出には、自然言語処理モデルの 1 つである Bidirectional Encoder Representations from Transformers (BERT)[2] を利用した上野らの予兆検出器を拡張して利用する。上野らの研究では、分類器の学習に分類対象となる入所者の記録しか用いておらず、作成した分類器は、他の入所者の記録分類には向かない。そのため本研究では、他の入所者の支援記録を学習に使用した分類器による実験を行い、学習に入所者本人の記録しか用いない場合と精度を比較し、分類器の汎用性について考察する。さらに本研究では、分類精度向上のため、付随情報の有効な利用方法を検討する。

本稿の構成を述べる。2 節では関連研究を紹介する。3 節では、本実験で使用する支援記録と付与されたインシデントと予兆のラベルについて説明する。4 節では、本稿で用いるインシデントの予兆検出手法を説明する。5 節で実験について説明し、結果と考察を述べる。6 節では、まとめと今後の課題について述べる。

## 2 関連研究

### 2.1 BERT

Bidirectional Encoder Representations from Transformers (BERT)[2] は、Devlin らが提案した自然言語処理モデルであり、双方向 Transformer[3] というニューラルネットワークを利用している。BERT は大量の教師なしデータによる事前学習を行ったモデルを、比較的少量の教師ありデータでファインチューニング (FT) して個々のタスクに合わせて学習する。これにより事前学習済みモデルをそのまま適用するよりも性能の向上が期待できる。

BERT の特徴として、Transformer を利用し、文章を文頭

1: 本研究は、岡山大学の研究倫理審査専門委員会の承認を得て実施した。(研 2104-005)

と文末の双方向から学習することで、文脈を考慮できることが挙げられる。BERT ではこの双方向からの学習を、Masked Language Model (MLM) と Next Sentence Prediction (NSP) の2つのタスクを行うことで可能としている。MLM では、まず入力文章を単語ごとにトークンに分割し、その中からランダムに選ばれた 15 % のトークンを確率的に別のトークンで置き換える。具体的には、選択された 15 % のトークンのうち、その 80 % を [MASK] という特殊なトークンに、10 % をランダムな別のトークンに置き換え、残りの 10 % はそのままのトークンとする。その後 [MASK] トークンに置き換えられたトークンを予測し学習するのが MLM のタスクである。NSP は2つの入力文を受け取り、そのペアにおいて2つ目の文が1つ目の文の後続の文になっているかを予測するタスクである。具体的には、文の片方を 50 % の確率で他の文に置き換え、それらが隣り合っているか隣り合っていないか判別することで学習する。BERT は発表時に、General Language Understanding Evaluation (GLUE) ベンチマーク [4] の8つのタスク、Stanford Question Answering Dataset (SQuAD)[5] の v1.1 と v2.0, Situations With Adversarial Generations (SWAG)[6] の計 11 の自然言語処理タスクで最高精度を記録した。

## 2.2 インシデント検出

介護や福祉、医療分野におけるインシデントやトラブルへの対応は大きな医療事故や介護事故を防ぐためにも重要であり、事例の分析や予防手法の提案、支援システムの開発など様々な研究がされている。

例えば田中らは、大量に蓄積されている医療事故情報やインシデント情報をトラブルの未然防止に活用するための仕組みとして、機械学習を利用した階層型クラスタリングの方法を提案した [7]。具体的には、まず初めに日本医療機能評価機構が WEB 公開している医療事故データベースから必要情報をダウンロードして、ヒヤリハットデータベースを作成した。その後、それらに階層型クラスタリングを適用し、ヒヤリハットデータベース全体のクラスタ構造を各事例の類似性という視点で階層的に表示することで、インシデントをいくつかのグループに分類することを可能とした。このような自動分類により必要なクラスタを深堀できるなど、各医療機関でのインシデントへの事前対応が可能になることが期待できると報告されている。

峯崎らは、複数介護施設における事故報告書を使用して、テキストマイニングおよび機械学習によって事故の起こりうる状況を分析した [8]。テキストデータ部分について単語文書行列を作成して元のデータと結合し、事故の結果を目的変数としてランダムフォレスト法で重要な単語クラスタ及び重要な単語を抽出し、決定木学習によって可視化した。その結果、1. 出血は目に見える外傷が発見されて介護士が初めて気づく、2. 移乗などの際に転倒による打撲が発生する可能性がある、3. 食事時に誤嚥が起こりうる、4. 無断外出につながるいくつかの行動がある、などのいくつかの事故パターンを発見した。

中沢は、高齢者介護施設において、施設職員の支援や負担軽減を目的として、夜間徘徊者の発見、通報、声掛け、足止めを

総合的に実行する見守りロボットを開発した [9]。ロボットに搭載された赤外線カメラにより施設廊下を監視し、HOG 特徴量とフレーム間差分法による動体検知を併用することで徘徊者を検出した。また、その後施設職員へ無線連絡をすると同時に、ロボットが徘徊者に近づき、職員の音声で呼びかけることで職員が到着するまでの時間稼ぎが可能なシステムを構築した。

青木らは、介護 AI 記録分析ツール “けあらば”を開発した [10]。これは同社の開発した介護 AI 入力予測ツールによって作成された介護記録を対象とした記録分析ツールであり、施設利用者の日常生活における基本動作の評価値を推定することを可能とした。今後は介護記録から利用者の事故予兆や状況悪化の予測を行うことを目指していると報告されている。

上野らは、福祉支援施設で問題となるインシデントの予兆検出のために、BERT を利用した支援記録を予兆の有無で分類する手法を提案した [1]。具体的には、支援記録の記録本文とその付随情報を入力として、予兆の有無の2クラス分類を行うことで予兆の検出を実現している。実験では、利用者 10 人の支援記録を用いて提案する予兆検出器の精度を検証し、平均で 0.87 の F 値を確認した。また、8 人の利用者は付随情報を活用することで検出精度が向上することを確認した。

本研究は、支援記録を利用して施設利用者の介助の助けとする点では [10] と類似しているが、[10] は文例を選択することで介護記録が生成できる入力支援ツールがあるため、入力形態が統一化されている記録を利用している。一方本研究では、文章の記述形式といったものが定まっていない自由記述形式の記録を利用する。また、画像処理による支援システムを構築する [9] に対し、本研究では主に自然言語処理によって支援記録の文章を解析して用いる。さらに、[7] や [8] が介護施設や医療現場の、事故やインシデントの記録を分析するのにに対し、本研究では記録からインシデントの予兆の検出をする。

## 3 支援記録と付与ラベル

### 3.1 支援記録

本研究で用いる支援記録は、福祉支援施設の職員が各入所者に対して実施した支援について記録したものである。記録が書かれるタイミングは様々であり、支援の直後に書かれたものや業務終了時に一日分の支援をまとめて記録したものなどがある。また支援記録の本文は、文の長さや文字数、記述形式といったものが定まっていない自由記述形式の文章となっている。使用する支援記録の例を表 1 に示す。表 1 の通り、支援記録には先に説明した自由記述の本文以外にも複数の情報が含まれている。入所者 ID は施設入所者を判別するための 6 桁の数値である。処理日付はその記録の入力された年月日を意味する。入力者 ID は支援をして記録を書いた支援者を判別するための 6 桁の数値である。インシデントラベルはその記録に含まれるインシデントの情報を、予兆ラベルはその記録に含まれる予兆の情報をそれぞれ表している。これらのラベルは元の支援記録には含まれておらず、後述の基準に従って人手で付与されたものを本研究で使用している。

表 1 支援記録とラベルの例

| 入所者 ID | 入力者 ID | 処理日付       | 記録本文                                                   | 予兆ラベル | インシデントラベル |
|--------|--------|------------|--------------------------------------------------------|-------|-----------|
| 000111 | 000011 | 2020/04/12 | 寝ていることが多い。奇声もなく静かに過ごされている。                             | 予兆なし  | なし        |
| 000111 | 000012 | 2020/04/12 | 作業：朝礼を終え、イオン水を飲んだあとに職員への他害がある。クールダウンまでは早かったが、原因は不明である。 | 予兆（強） | 暴力#職員     |
| 000122 | 000011 | 2020/04/13 | 舎活動：居室で穏やかに過ごされている。（14：30）大きな声が聞こえ、見に行くと笑っている。         | 予兆（弱） | なし        |

表 2 インシデントとそのラベル

| 種類   | 対象       | ラベル           |
|------|----------|---------------|
| 器物破損 | 自分の所有物   | 器物破損#自分の所有物   |
|      | 他入所者の所有物 | 器物破損#他入所者の所有物 |
|      | 職員の所有物   | 器物破損#職員の所有物   |
|      | 施設の所有物   | 器物破損#施設備品     |
|      | その他      | 器物破損#その他      |
| 暴力   | 職員       | 暴力#職員         |
|      | 施設の入所者   | 暴力#施設の入所者     |
|      | その他      | 暴力#その他        |

### 3.2 インシデントの定義とインシデントラベル

本研究では、周囲に大きな損害を与える状況をインシデントと定義する。周囲に大きな損害を与える状況の例としては、暴力全般や水や唾をかけられるなどの職員や他の入所者が不快な思いをするものや、器物破損のような修復コストがかかるものなどが挙げられる。なお本研究ではインシデントは、暴力や器物破損が自己以外の対象に害を与える他害に限定する。そのため、自傷行為のような自身に害を与えるものはインシデントとしていない。この定義に従って、インシデントには表 2 のようなラベルが付与されている。一つの記録に対して複数のインシデントラベルが付与されることもある。このインシデントは福祉支援施設での支援業務を妨げるものであり、現場の施設職員の意見を参考に定められたものである。また本研究では、ここに定めたインシデントの種類を、予兆検出の分析には用いるが、インシデントの分類は行わない。

### 3.3 予兆の定義と予兆ラベル

本研究では、支援者である施設職員の判断に基づいて、支援記録本文に施設入所者の衝動性が高い、あるいはネガティブな心理状態が現れているものを予兆と定義する。予兆ラベルには「予兆（弱）」「予兆（強）」の 2 種類がある。「予兆（弱）」は、衝動性が高い、あるいはネガティブな心理状態が現れている状態、「予兆（強）」は衝動性あるいはネガティブな感情を抑えることができず、結果として周囲へ影響が出てしまっている状態と定義している。「予兆（弱）」の例としては、自傷行為、情緒が乱れている、大声を発する、などがあげられ、「予兆（強）」の例としては、他者への暴力、器物破損などのインシデントや、物に当たる行為などが挙げられる。一方で、やる気がない、暴言、騒ぐ、などはそれだけでは予兆とはせず、記述内容の文脈によって予兆か予兆でないかを判断している。例えば、気分が悪い、つらそうな表情、などは精神面の話であれば「予兆（弱）」であるが、体調面の話であればこれらは「予兆なし」となる。

また一つの記録に対して「予兆（弱）」と「予兆（強）」の予兆ラベルが同時に付与されることもあり、本研究においてはそういった記録を「予兆（両方）」として扱う。本研究では、この予兆ラベルを用いて、支援記録を「予兆なし」「予兆（弱）」「予兆（強）」「予兆（両方）」の 4 クラスへ分類を行う。

## 4 インシデントの予兆検出

### 4.1 インシデントの予兆の検出手法

図 1 に使用する予兆検出器の概略を示す。この検出器ではまず、入力された記録本文を BERT により文書表現ベクトルに変換する。また、支援記録から支援者 ID や日付情報などの本文以外の付随情報を取り出し、付随情報ベクトルに変換する。次に、これらの文書表現ベクトルと付随情報ベクトルを入力とする分類器によって支援記録の予兆を、「予兆なし」「予兆（弱）」「予兆（強）」「予兆（両方）」の 4 クラスに分類する。その際に使用する分類器は図 1 の右側に示す 4 層のニューラルネットワークである。この各層は全結合層であり、左から順に 256 次元、128 次元、32 次元、4 次元ベクトルへと線形変換される。活性化関数にはシグモイド関数、出力層には softmax 関数を用いている。損失関数には交差エントロピー誤差を採用し、Adam[11]によって最適化する。分類器のエポック数は 20、バッチサイズは 32 とする。

図 1 の右にある分類器の 1 層目では、文書表現ベクトルを 256～0 次元、付随情報ベクトルを 0～256 次元の全結合層へ入力して、ベクトルの次元を変換し、2 つのベクトルを合わせて 256 次元のベクトルとしている。また各層間では Dropout[12]を実行している。Dropout はニューラルネットワークの過学習防止のための手法の一つで、各層ごとに一定確率でランダムにノードの出力を 0 にすることで無効化して学習を進める。

分類器の出力結果は 0 から 1 の範囲で正規化された各ラベルに対する確率であり、その総和は 1 になる。本稿で行う 4 クラス分類では「予兆なし」「予兆（弱）」「予兆（強）」「予兆（両方）」の 4 つのクラスに対する確率が出力され、確率が最大のクラスに分類する。

図 1 の左の BERT の部分は、事前学習済みモデルを支援記録を用いてファインチューニングし分類に利用する。また本研究では、BERT 日本語 Pretrained モデル<sup>2</sup>を利用する。

また本研究で入力として使用する付随情報は表 3 に示す 5 種類である。日付情報として利用する月、日、曜日はそれぞれ 12 次元、31 次元、7 次元、入力者 ID は 100 次元、インシデント

2 : [https://nlp.ist.i.kyoto-u.ac.jp/?ku\\_bert\\_japanese](https://nlp.ist.i.kyoto-u.ac.jp/?ku_bert_japanese)

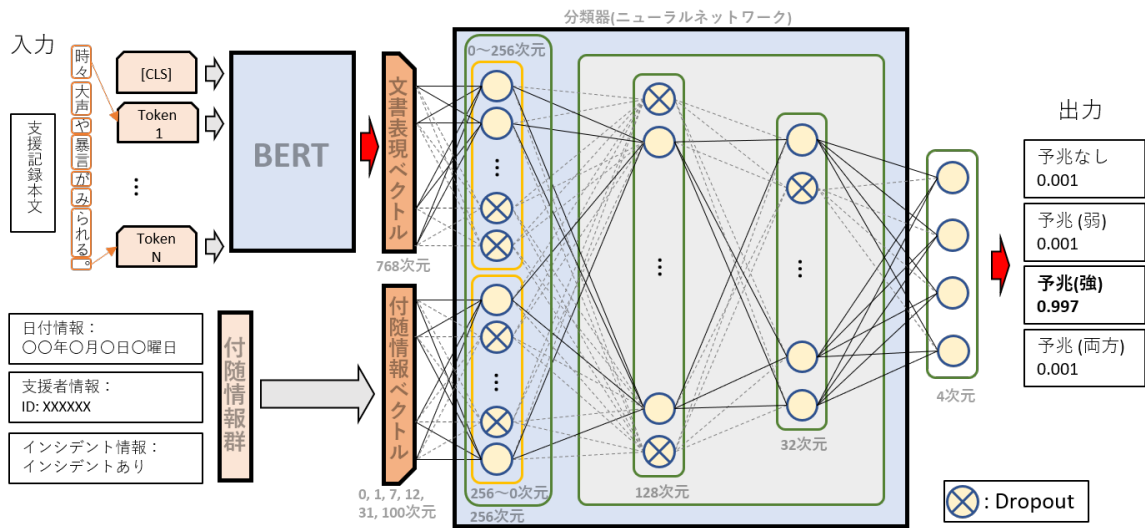


図 1 予兆検出器の概略図 [1]

表 3 付随情報と次元

| 付随情報      | ベクトルの次元 |
|-----------|---------|
| 月         | 12 次元   |
| 日         | 31 次元   |
| 曜日        | 7 次元    |
| 入力者 ID    | 100 次元  |
| インシデントの有無 | 1 次元    |

の発生の有無は 1 次元の One-hot ベクトルで表現する。

4.2 個人専用分類器と汎用分類器

上野らは、BERT のファインチューニングに分類対象となる入所者本人の記録のみを用いたため、本稿ではこの分類器を個人専用分類器と呼ぶ。本稿では分類器の汎用性について考察するため、この個人専用分類器の他に、他の入所者の支援記録をファインチューニングに用いた汎用分類器を使用する。汎用分類器は、分類対象者と同施設に入所している他の入所者全員の支援記録で学習をする汎用分類器（同施設）、分類対象者とは別の施設に入所している入所者全員の支援記録で学習をする汎用分類器（他施設）、分類対象者を除く全ての入所者の支援記録で学習をする汎用分類器（本人以外）の 3 つを使用する。

5 インシデントの予兆分類実験

5.1 実験に用いるデータ

インシデント予兆分類実験には、2 つの福祉支援施設の入所者 10 名ずつ、計 20 名の支援記録を用いる。この支援記録には、3 節で説明した「予兆なし」「予兆（弱）」「予兆（強）」の 3 つの予兆ラベルが付与されており、「予兆（弱）」と「予兆（強）」の両方のラベルが付与されている記録は「予兆（両方）」ラベルが付与されているものとして扱う。これらの支援記録は図 1 の BERT のファインチューニングや、分類器の学習に利用する。表 4 に支援記録の統計情報を示し、図 2 に入所者ごとの支援記録の各予兆ラベルの割合をまとめる。表 4 には、入所者ごとに、

各予兆ラベルの付与されている記録件数と全記録に占める割合を示している他、インシデントラベルが付与されている記録の件数とその全記録に占める割合についても記している。

5.2 実験内容

本研究では、以下の (1),(2) について実験する。

- (1) 汎用分類器による予兆分類
- (2) 付随情報を利用した予兆分類

まず実験 (1) では、他の入所者の支援記録を学習に用いた汎用分類器により予兆分類をし、その精度の評価、考察をする。この際、4 節で述べた個人専用分類器と汎用分類器（同施設）、汎用分類器（他施設）、汎用分類器（本人以外）の 3 つの汎用分類器を使用し分類実験をする。

個人専用分類器は、分類対象の入所者本人のデータを用いて五分割交差検定により分類精度を求め、その他 3 つの汎用分類器は五分割交差検定を行わず、各支援記録集合で学習した分類器に対して、テスト用データとなる分類対象者の記録全件分類精度を求める。

実験 (2) では、日付などの付随情報を利用して予兆分類をし、その精度を評価する。上野らの作成した付随情報を入力とする分類器は、入力の文書表現ベクトルと付随情報ベクトルをそれぞれ 128 次元のベクトルに変換し、利用している。実験 (2) では、このベクトルの次元を変更して、付随情報の有効性を比べる。具体的には、分類器 1 層目の 256 次元のうちの文書表現ベクトルと付随情報ベクトルの重み割合を、16:0~0:16 まで変化させ、その精度を比較する。重み割合が 16:0 の場合は文書表現ベクトルのみ、0:16 の場合は付随情報ベクトルのみを利用した分類となる。

また実験 (2) では、付随情報ベクトルの次元に文書表現ベクトルの次元を合わせた実験も行う。つまり、各付随情報ベクトルの次元はそのまま、文書表現ベクトルを付随情報ベクトルの次元まで削減し入力とする。例えば、付随情報が「入力者 ID」の場合は、文書表現ベクトルと付随情報ベクトルがそれぞれ

表 4 支援記録の統計情報

| 施設 | 人物 | 性別 | 記録件数 (割合)    |            |            |            |       | インシデント件数   |
|----|----|----|--------------|------------|------------|------------|-------|------------|
|    |    |    | 予兆なし         | 予兆 (弱)     | 予兆 (強)     | 予兆 (両方)    | 合計    |            |
| A  | a  | 女性 | 1,519(62.9%) | 646(26.7%) | 88(3.6%)   | 163(6.7%)  | 2,416 | 106(4.4%)  |
|    | b  | 女性 | 2,207(64.2%) | 999(29.0%) | 83(2.4%)   | 150(4.4%)  | 3,439 | 205(6.0%)  |
|    | c  | 女性 | 1,830(55.5%) | 607(18.4%) | 503(15.2%) | 360(10.9%) | 3,300 | 630(19.1%) |
|    | d  | 女性 | 1,510(78.6%) | 304(15.8%) | 14(0.7%)   | 93(4.8%)   | 1,921 | 87(4.5%)   |
|    | e  | 男性 | 2,800(73.7%) | 832(21.9%) | 83(2.2%)   | 82(2.2%)   | 3,797 | 156(4.1%)  |
|    | f  | 男性 | 2,702(81.3%) | 551(16.6%) | 40(1.2%)   | 30(0.9%)   | 3,323 | 52(1.6%)   |
|    | g  | 男性 | 2,564(75.7%) | 599(17.7%) | 67(2.0%)   | 157(4.6%)  | 3,387 | 99(2.9%)   |
|    | h  | 男性 | 3,034(88.2%) | 367(10.7%) | 31(0.9%)   | 8(0.2%)    | 3,440 | 15(0.4%)   |
|    | i  | 男性 | 3,216(84.3%) | 428(11.2%) | 94(2.5%)   | 77(2.0%)   | 3,815 | 73(1.9%)   |
|    | j  | 男性 | 3,320(85.6%) | 532(13.7%) | 14(0.4%)   | 13(0.3%)   | 3,879 | 6(0.2%)    |
| B  | k  | 男性 | 1,781(90.0%) | 63(3.1%)   | 121(6.0%)  | 37(1.8%)   | 2,002 | 149(7.4 %) |
|    | l  | 男性 | 1,927(96.3%) | 59(2.9%)   | 10(0.5%)   | 7(0.3%)    | 2,003 | 11(0.5%)   |
|    | m  | 男性 | 1,443(72.1%) | 157(7.8%)  | 379(18.9%) | 22(1.1%)   | 2,001 | 385(19.2%) |
|    | n  | 男性 | 1,621(81.0%) | 277(13.8%) | 78(3.9%)   | 26(1.3%)   | 2,002 | 56(2.8%)   |
|    | o  | 男性 | 1,667(83.3%) | 120(6.0%)  | 202(10.1%) | 14(0.7%)   | 2,003 | 174(8.7%)  |
|    | p  | 男性 | 1,774(88.7%) | 168(8.4%)  | 54(2.7%)   | 4(0.2%)    | 2,000 | 58(2.9%)   |
|    | q  | 男性 | 1,885(94.3%) | 95(4.8%)   | 17(0.9%)   | 3(0.2%)    | 2,000 | 18(0.9%)   |
|    | r  | 男性 | 1,546(77.3%) | 251(12.5%) | 144(7.2%)  | 62(3.1%)   | 2,003 | 206(10.3%) |
|    | s  | 男性 | 744(97.4%)   | 19(2.5%)   | 0(0.0%)    | 1(0.01%)   | 764   | 1(0.01%)   |
|    | t  | 男性 | 1,747(87.3%) | 91(4.5%)   | 135(6.7%)  | 28(1.4%)   | 2,001 | 133(6.6%)  |

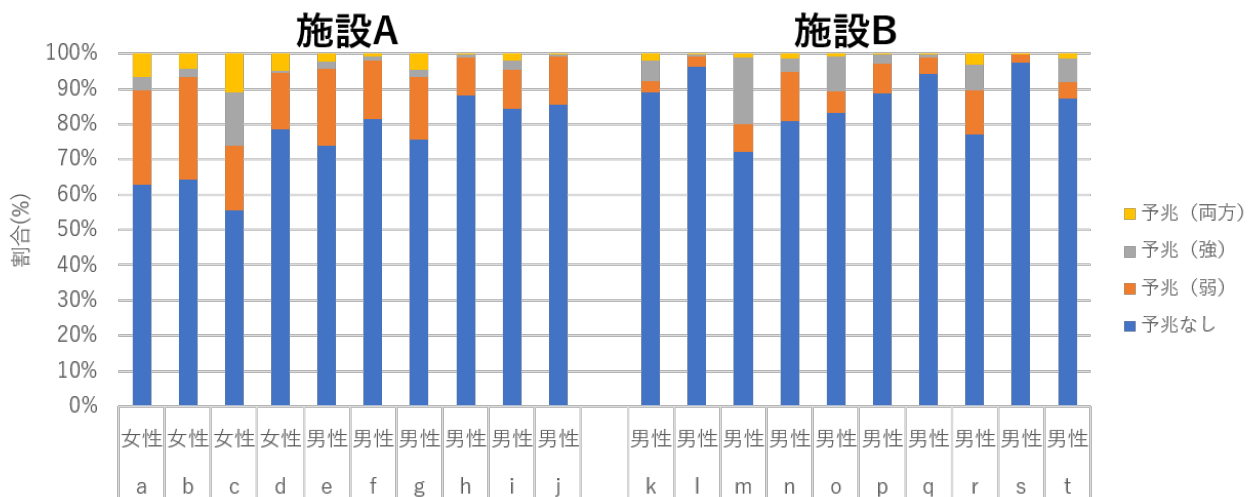


図 2 支援記録の予兆ラベルの割合

100 次元となり、図 1 の分類器 1 層目の次元は 256 次元ではなく 200 次元となる。この際、付随情報が「インシデントの有無」である場合のみ、文書表現ベクトルを付随情報ベクトルに合わせて 1 次元にせず、10 次元に削減して利用する。具体的には、図 1 の右側の分類器の一層目の次元を指定の次元とすることで実現している。これらの付随情報を用いた実験は、先の汎用分類器による予兆分類実験における汎用分類器 (本人以外) を用いて行う。

### 5.3 評価指標

分類器の評価には、以下に示す再現率 (Recall)、適合率 (Precision)、F 値 (F-measure) を利用する。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{F - measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3)$$

例えば、予兆ラベル「予兆 (強)」についての TP, TN, FP, FN は下記ようになる。

- TP: 「予兆 (強)」だと分類された記録のうち、実際に「予兆 (強)」である記録の件数
- TN: 「予兆 (強)」ではないと分類された記録のうち、実際に「予兆 (強)」でない記録の件数

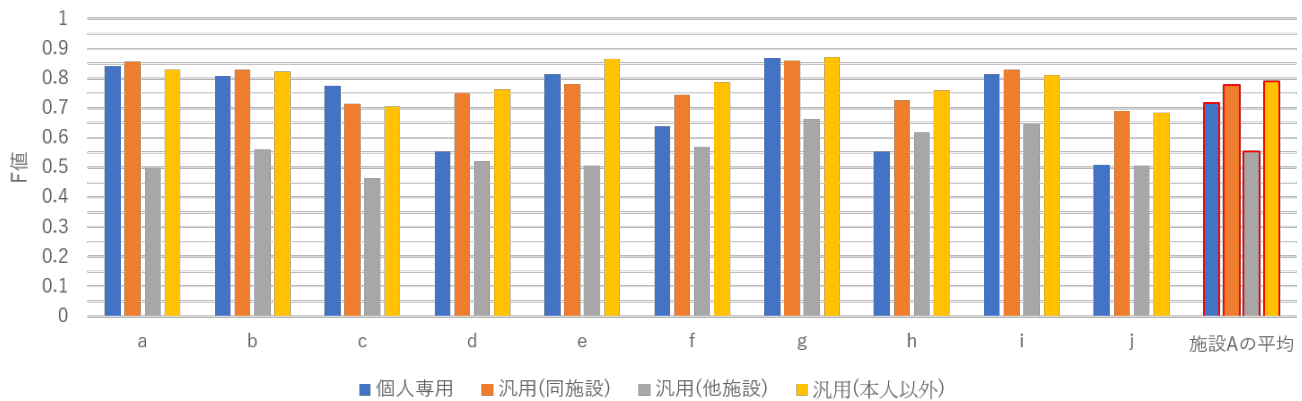


図 3 各入所者の分類器ごとの F 値 (施設 A)

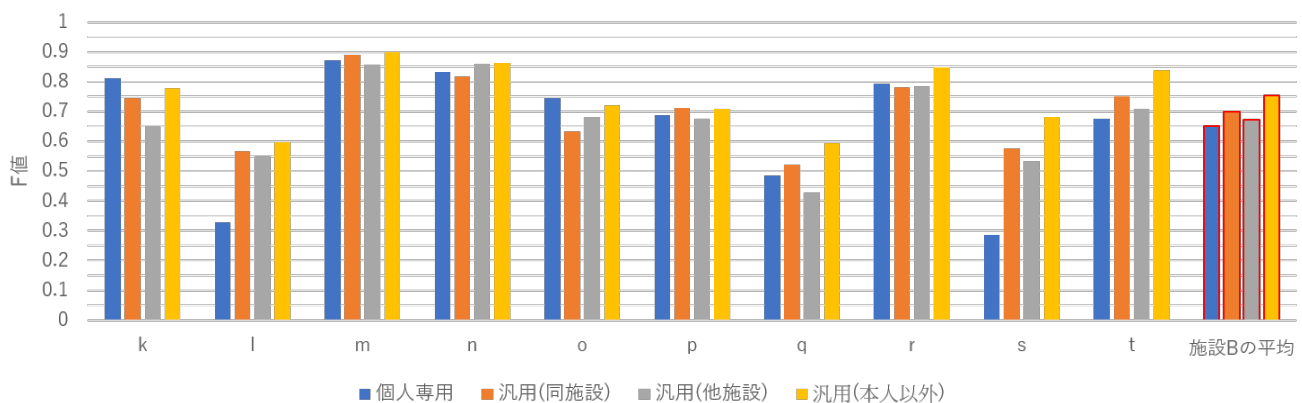


図 4 各入所者の分類器ごとの F 値 (施設 B)

- FP:「予兆 (強)」だと分類された記録のうち,「予兆 (強)」ではない記録の件数
- FN:「予兆 (強)」ではないと分類された記録のうち,「予兆 (強)」である記録の件数

よって再現率は全ての「予兆 (強)」である文章に対し,分類器が正しく「予兆 (強)」と分類したものの割合であり,適合率は分類器が「予兆 (強)」と分類した文章のうち,実際に「予兆 (強)」であるものの割合である。また F 値は再現率と適合率の調和平均となる。

以下の評価実験では,4つの予兆ラベルについて個別にこの値を計算した後,その平均を求める。

#### 5.4 汎用分類器による予兆分類実験

##### 5.4.1 実験結果

汎用分類器による予兆分類結果の F 値を図 3 (施設 A の各入所者)と図 4 (施設 B の各入所者)に示す。

図 3 と図 4 のいずれにおいても,入所者ごとの差異はあるが,各施設の平均で見ると汎用分類器 (本人以外) の精度が最も良く,次いで汎用分類器 (同施設) の精度が良いことが分かる。各入所者について見ると,施設 A の入所者 d, h や施設 B の入所者 l, q のように個人専用分類器の F 値が他の入所者に比べて小さくなっている人物は,個人専用分類器に比べて,汎用分類器 (同施設) や汎用分類器 (本人以外) の F 値が大きいことが

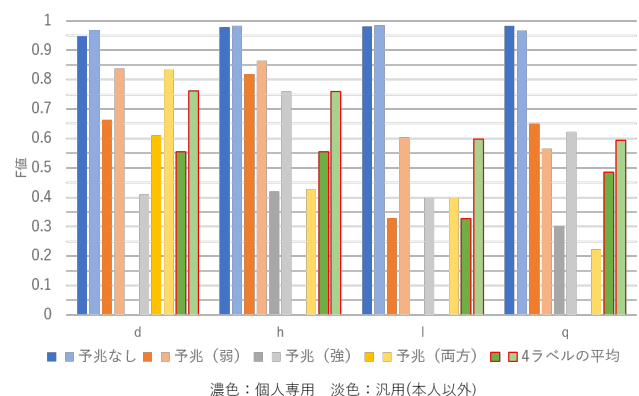


図 5 個人専用分類器と汎用分類器 (本人以外) の F 値の比較

分かる。また,図 3 の施設 A では汎用分類器 (他施設) の F 値が小さいのに対し,図 4 の施設 B の汎用分類器 (他施設) の F 値は,他の分類器とあまり差がないことが分かる。

##### 5.4.2 考察

両施設の平均の F 値で,汎用分類器 (本人以外) 及び,汎用分類器 (同施設) の値が大きいのは,先に述べた施設 A の入所者 d, h や施設 B の入所者 l, q のような人物の分類精度の影響がある。当該入所者の分類結果では,汎用分類器 (本人以外),汎用分類器 (同施設) の F 値が大きい。これらの人物は,表 4 や図 2 の支援記録の統計情報を見ると分かるように,「予兆 (強)」

表 5 「予兆（強）」の少ない入所者の支援記録の分類事例

|     | 記録本文                                           | 予兆ラベル  | 分類結果  |          |
|-----|------------------------------------------------|--------|-------|----------|
|     |                                                |        | 個人専用  | 汎用（本人以外） |
| (1) | 朝の着替えの際、自身のパジャマを割く。                            | 予兆（強）  | 予兆なし  | 予兆（強）    |
| (2) | 脱衣場で〇〇さんの頭を叩く。注意する。<br>突かれた事に怒った様子。            | 予兆（両方） | 予兆（弱） | 予兆（両方）   |
| (3) | 朝食後□□さんが食器を片付けないことに<br>腹を立て食堂内で追いかけて掴みかかろうとする。 | 予兆（両方） | 予兆（強） | 予兆（強）    |

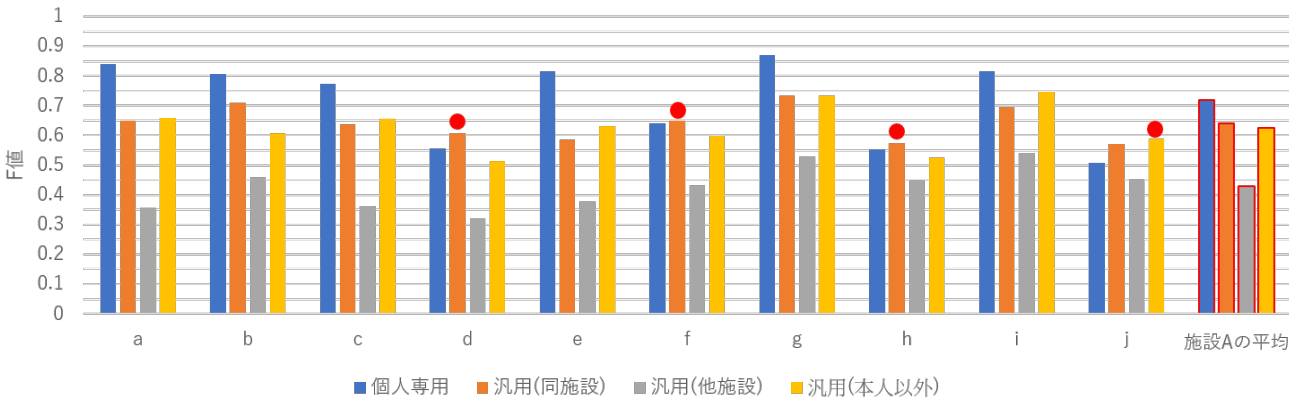


図 6 学習回数を揃えた実験における各入所者の分類器ごとの F 値 (施設 A)

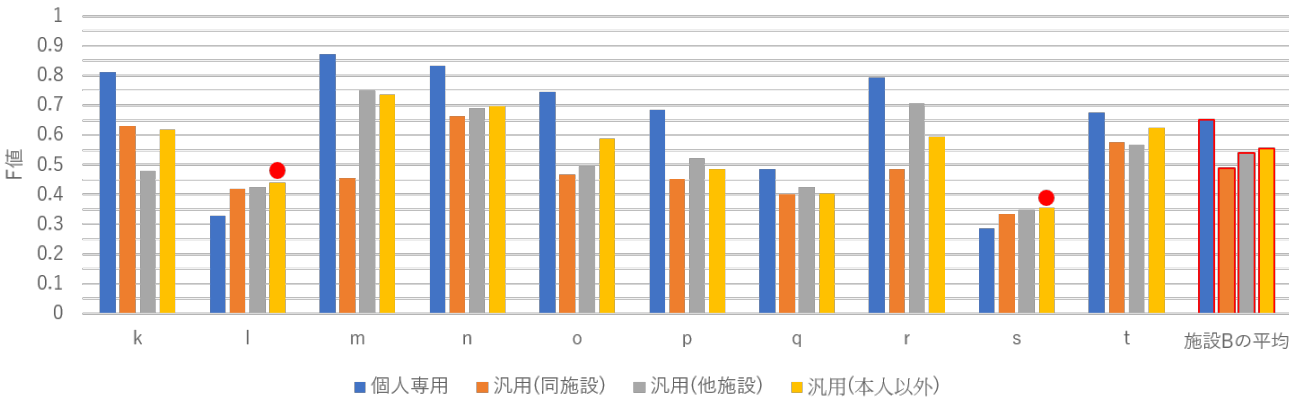


図 7 学習回数を揃えた実験における各入所者の分類器ごとの F 値 (施設 B)

を含む支援記録の割合が他の入所者と比べて小さい。そのため、自身の記録のみでの学習では「予兆（強）」や「予兆（両方）」ラベルの分類が困難である。図 5 に個人専用分類器と汎用分類器（本人以外）の予兆ラベル毎の F 値を示す。色の濃い棒が個人専用分類器、淡い棒が汎用分類器（本人以外）の F 値である。個人専用分類器の当該の入所者の F 値を見ると、「予兆（強）」や「予兆（両方）」ラベルの分類精度が悪いことが分かる。一方、汎用分類器（本人以外）の当該入所者の予兆ラベルごとの F 値は、個人専用分類器と比べて「予兆（強）」や「予兆（両方）」の値が相対的に大きくなっていることが分かる。

本実験において、予兆の少ない入所者について個人専用分類器で誤分類され、汎用分類器（本人以外）では正しく分類された記録本文の例を表 5 の (1)、(2) に、汎用分類器（本人以外）でも誤分類となってしまった例を (3) に示す。表 5 の (1) の例では、「自身のパジャマを割く」という「予兆（強）」である器物

破損の記述があるが、個人専用分類器では「予兆（弱）」に分類されている。(2) の例では、「頭を叩く」という「予兆（強）」の記述と「怒った様子」という「予兆（弱）」の記述があるため「予兆（両方）」であるが、個人専用分類器では、「予兆（弱）」に分類されている。汎用分類器（本人以外）では、このような「予兆（強）」を含む記録の誤分類が少なくなっている。

一方いずれも誤分類となってしまった (3) の例では、「追いかけて掴みかかろうとする」という「予兆（強）」の記述と「腹を立て」という「予兆（弱）」の記述があるため、「予兆（両方）」であるが、分類結果はどちらも「予兆（強）」となった。

また、各汎用分類器で学習に使用する記録の件数を、五分割交差検定を行う個人専用分類器の学習で使用する件数まで削減し、学習回数を揃えた分類実験を行った。分類結果の F 値を図 6（施設 A の各入所者）と図 7（施設 B の各入所者）に示す。図 6 と図 7 のいずれにおいても、各施設の平均を見ると個人専

表 6 施設 A と施設 B の記録本文の統計情報

| 施設   | 文字数   |     |       | 「予兆（強）」を含む記録の文字数 |     |        | 人物の名前 |                      | 「：」区切りの表記 |                       |
|------|-------|-----|-------|------------------|-----|--------|-------|----------------------|-----------|-----------------------|
|      | 最大値   | 最小値 | 平均    | 最大値              | 最小値 | 平均     | 件数    | 全記録に対する割合            | 件数        | 全記録に対する割合             |
| 施設 A | 1,047 | 0   | 56.00 | 1,047            | 12  | 130.73 | 2,671 | 8.16% (2671/32,717)  | 9,567     | 29.24% (9,567/32,717) |
| 施設 B | 809   | 3   | 44.80 | 448              | 6   | 53.91  | 4,689 | 24.97% (4689/18,779) | 188       | 1.00% (188/18,779)    |

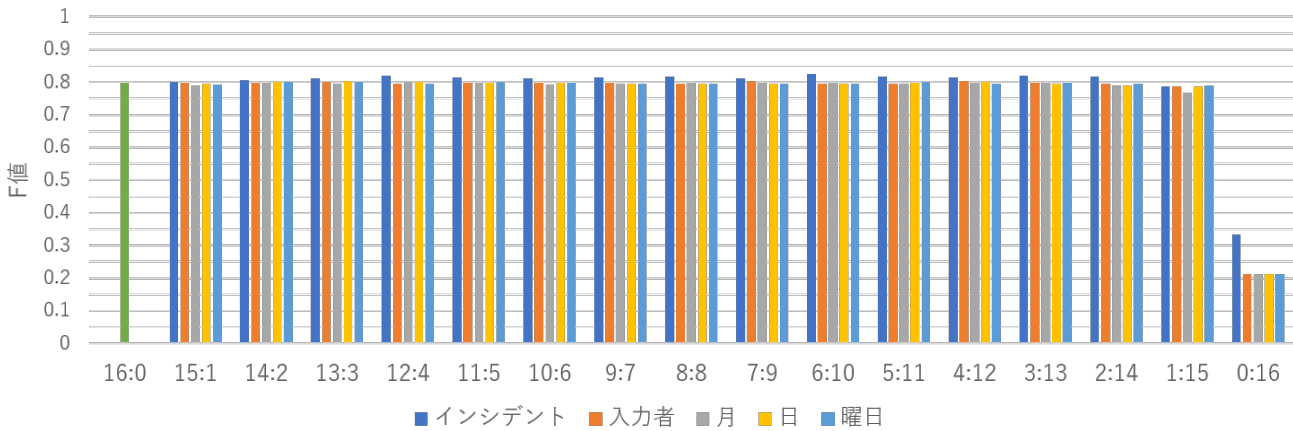


図 8 文書表現ベクトルと付随情報ベクトルの割合と施設 A の入所者の F 値の平均

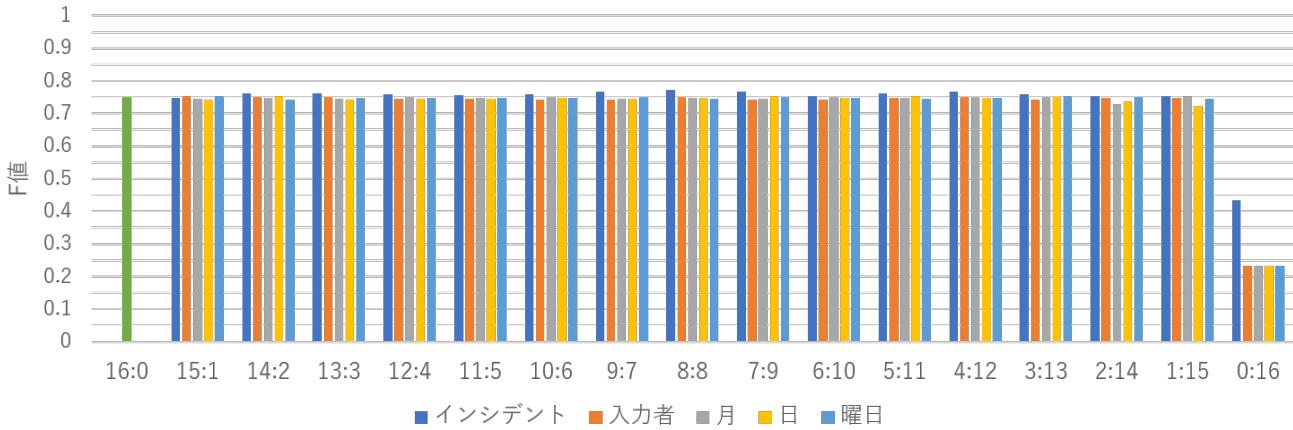


図 9 文書表現ベクトルと付随情報ベクトルの割合と施設 B の入所者の F 値の平均

用分類器の F 値が最大であることが分かり、各入所者についても個人専用分類器の F 値が最大となる人物が殆どである。よって、学習に利用する支援記録の件数が等しい場合は、学習に利用する記録は、本人の記録が良いことが分かる。しかし、図 6 と図 7 でグラフに赤丸を付けた入所者については、個人専用分類器と比べて汎用分類器（本人以外）や汎用分類器（同施設）の F 値が大きくなっている。これらの入所者は、先に述べた「予兆（強）」を含む支援記録が他の入所者と比べて少ない人物である。このことから、そのような「予兆（強）」を含む支援記録の少ない人物には、他者の記録を用いた学習も有効である。

また、図 3 と図 4 の比較でも、図 6 と図 7 の比較でも、汎用分類器（他施設）の分類精度が、両施設で大きく異なっていたため、施設ごとに記録本文の内容に差があると考え、分析を行った。具体的には、両施設の支援記録の文字数の最大値、最小値、平均値を調べ、さらに「〇〇さん」というような、人物の名前を変換した記述や、記録の初めに「舍活動：～」や「作業：～」

といった固有の表現を含む支援記録の件数を調べた。その結果を表 6 に示す。文字数から、施設 A の記録の方が文章が長い傾向にあることが分かる。中でも「予兆（強）」を含む記録の文字数については、施設 A は施設 B と比べ、最大値、平均値共に 2 倍以上の長さとなっている。また、人物名は施設 B の記録本文に多く出現し、「：」区切りの表記は施設 A の記録本文に多く出現しているように、施設間で記録の文章に差があることが分かった。

5.5 付随情報を利用した予兆分類実験

5.5.1 実験結果

図 1 の分類器 1 層目の 256 次元の文書表現ベクトルと付随情報ベクトルの重み割合を変化させた予兆分類実験の結果の F 値を図 8（施設 A の入所者の平均）と図 9（施設 B の入所者の平均）に示す。なおこれらの図のグラフの 16:0 は、付随情報を用いない文書表現ベクトルのみで分類した結果の F 値である。

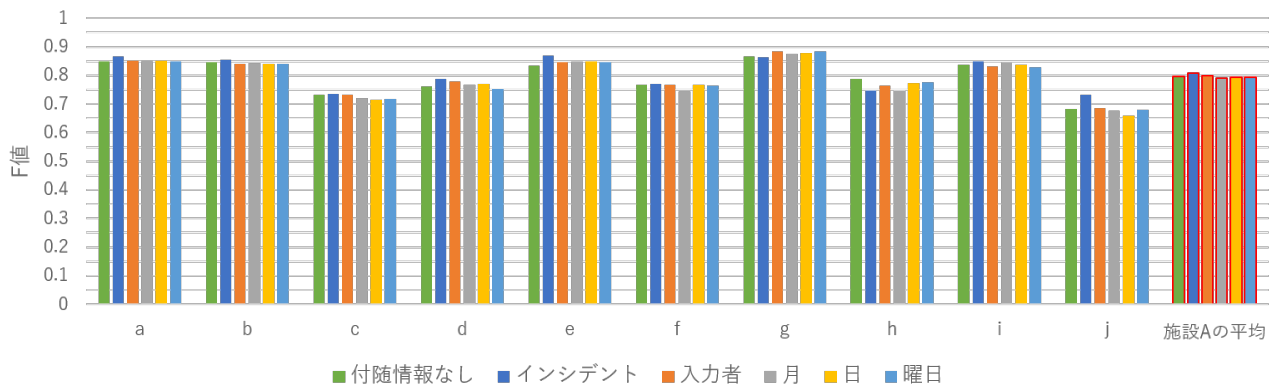


図 10 文書表現ベクトルを付随情報ベクトルに合わせ削減した実験の F 値 (施設 A)

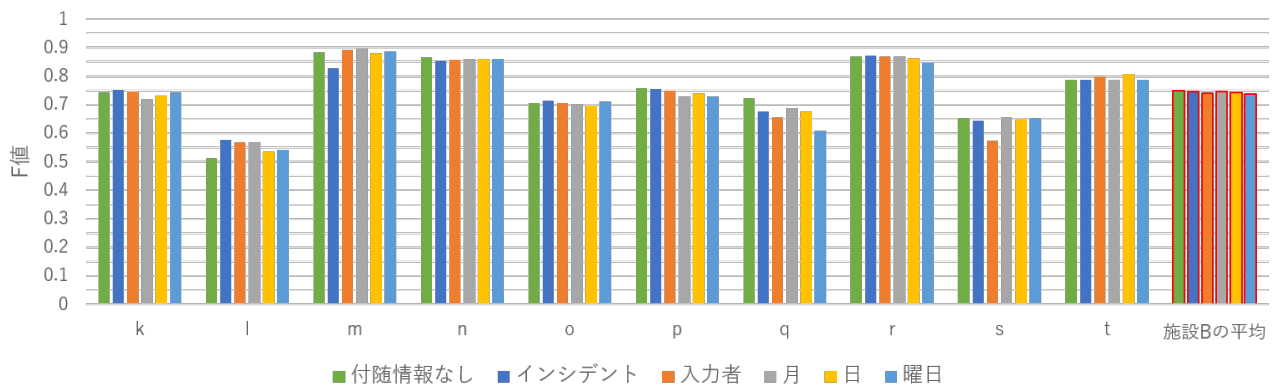


図 11 文書表現ベクトルを付随情報ベクトルに合わせ削減した実験の F 値 (施設 B)

0:16 はその反対に付随情報のみで分類した結果の F 値である。図 8 と図 9 のいずれにおいても、インシデントの有無以外の付随情報については、ベクトルの重み割合を変更しても、F 値がほとんど変化せず、0:16 においては F 値が大きく下がる。月、日、曜日の付随情報ベクトルのみを使用した場合は、分類結果として「予兆なし」のみしか出力しておらず、入力者 ID の付随情報ベクトルのみを使用した場合もほぼ「予兆なし」のみを出力した。一方、インシデントの有無の付随情報ベクトルを用いた場合は、F 値が大きくなっている場合がある。また、付随情報ベクトルのみを使用する場合においても、他の付随情報と比べると相対的に F 値が大きくなっている。

次に、文書表現ベクトルを付随情報ベクトルの次元まで削減して入力とした予兆分類の結果の F 値を図 10 (施設 A) と図 11 (施設 B) に示す。比較のため付随情報なし 16:0 の F 値も示している。図 10 と図 11 のいずれも、入所者ごとに結果の差異はあるものの、平均でみると付随情報なしの結果とあまり差がないことが分かる。

### 5.5.2 考 察

図 8 と図 9 より、インシデントの有無以外の付随情報はほとんど有効性が確認できなかった。インシデントの有無については、インシデントの付随情報ベクトルを使用した場合に F 値が 0.03 以上大きくなることのある入所者が施設 A で 10 名中 5 名、施設 B で 10 名中 4 名の計 9 名存在しており、分類精度の向上に有効な場合があった。

図 8、図 9 の予兆分類実験において、インシデントの付随情報ベクトルをベクトル割合 1:15 で使用した場合に分類精度が向上した入所者の分類事例を表 7 に示す。表 7 の (1) が付随情報ベクトルを使用した場合に正しく分類ができるようになった例、(2) が付随情報ベクトルを使用した場合に誤分類となった例である。表 7 の (1) の例では、「他利用者の手をはねたり押して」という「予兆 (強)」の記述があるにもかかわらず、文書表現ベクトルのみの場合では「予兆なし」に分類されているが、インシデントの付随情報ベクトルを用いた場合では正しく分類されている。(1) の支援記録は、「インシデントあり」となっており、正しく分類ができた。また例には示していないが、「予兆 (弱)」のみを含む記録を、文書表現ベクトルのみの場合は「予兆 (両方)」と誤分類、インシデントの付随情報ベクトルを用いた場合は「予兆 (弱)」に正しく分類ができた事例がある。この例の支援記録は、「インシデントなし」となっており、正しく分類ができた。

一方 (2) は誤分類の例で「××さんを引っ張ろうとする」という「予兆 (強)」の記述があり、文書表現ベクトルのみの場合は「予兆 (強)」と正しく分類出来ているが、インシデントの付随情報を使用した場合は「予兆なし」の誤分類となった。この例は、「インシデントなし」の支援記録で、その上 1:15 と付随情報ベクトルの比重が大きいため、「予兆なし」に誤分類した可能性がある。

文書表現ベクトルを付随情報ベクトルの次元まで削減し入力

表 7 インシデントの付随情報ベクトルを利用した分類事例

|     | 記録本文                                                                    | 予兆ラベル  | 分類結果       |               |
|-----|-------------------------------------------------------------------------|--------|------------|---------------|
|     |                                                                         |        | 文書表現ベクトルのみ | 付随情報あり (1:15) |
| (1) | 職員 (△△) に笑顔で付いて歩き、他の利用者が<br>職員の側に来たり手を引いたりすると、他利用者の手をはねたり押して移動させたりしていた。 | 予兆 (強) | 予兆なし       | 予兆 (強)        |
| (2) | 起床し血圧測定を行う。その後××さんを<br>引っ張ろうとする事が何度もあった。                                | 予兆 (強) | 予兆 (強)     | 予兆なし          |

とした図 10 と図 11 の予兆分類実験では、入所者毎に差はあったものの、両施設ともに平均の値は、付随情報があってもなくてもあまり差はなかった。そのため、本実験のようにベクトルの次元を揃えるだけでは、分類精度の向上はあまり期待できない。

6 ま と め

本稿では、上野らの提案した福祉支援施設の支援記録からのインシデントの予兆検出器を、本人以外の入所者の支援記録で学習して予兆検出実験を行った。予兆の検出は、支援記録の本文を利用した分類器を作成し、対象の支援記録を「予兆なし」「予兆 (弱)」「予兆 (強)」「予兆 (両方)」の 4 クラスに分類することで実現した。また、この分類に支援記録の本文以外の付随情報が有効かどうか実験した。

分類器の学習に本文以外の入所者の支援記録を利用する実験では、施設 A 入所者の全ラベルの平均 F 値は、個人専用分類器で 0.718、汎用分類器 (同施設) で 0.778、汎用分類器 (他施設) で 0.555、汎用分類器 (本人以外) で 0.790、施設 B 入所者のそれは、個人専用分類器で 0.652、汎用分類器 (同施設) で 0.699、汎用分類器 (他施設) で 0.673、汎用分類器 (本人以外) で 0.753 となり、汎用分類器 (本人以外) の分類精度が最も高かった。また、学習に本人以外の入所者の支援記録を利用した分類器は、予兆やインシデントの少ない入所者の予兆検出に有効であることが分かった。

分類器の学習に付随情報を利用する実験では、インシデントの有無の情報を利用した場合の予兆分類の F 値が最も大きく、施設 A の入所者では文書表現ベクトルと付随情報ベクトルの割合が「6:10」の場合に全ラベルの平均 F 値が 0.825、施設 B の入所者では、ベクトルの割合が「8:8」の場合に全ラベルの平均 F 値が 0.772 だった。インシデントの有無の付随情報は分類精度の向上に有効である可能性がある。

今後は、両施設間の支援記録文章の違いを調査し、分類器の汎用性の向上に有効な文章の特徴の発見を目指す。付随情報については、有用な可能性のあるインシデントの有無などの付随情報の有効な利用方法を検討する。

文 献

[1] 上野史, 松本典久, 太田学, “福祉支援施設の支援記録からの BERT によるインシデントの予兆検出,” 情報処理学会論文誌, Vol. 64, No. 11, pp. 1561–1566, 2023.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pretraining of deep bidirectional trans-

formers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” Advances in Neural Information Processing Systems 30, pp. 5998–6008, 2017.

[4] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman, “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding,” arXiv preprint arXiv:1804.07461, 2018.

[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” arXiv preprint arXiv:1606.05250, 2016.

[6] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi, “SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference,” arXiv preprint arXiv:1808.05326, 2018.

[7] 田中健次, 坂東幸一, 津本周作, 佐藤直樹, “医療界でのインシデント情報を活用したリスク対応,” 横幹, Vol. 13, No. 2, pp. 84–90, 2019.

[8] 峯崎智裕, 松木萌, 井上創造, “介護サービス向上に向けた介護事故事例テキストの分析,” 情報処理学会論文誌, Vol. 58, No. 10, pp. 1701–1711, 2017.

[9] 中沢信明, “高齢者介護施設における夜間徘徊者の検出,” 社会安全とプライバシー, Vol. 3, No. 2, pp. 8–12, 2019.

[10] 青木三重子, 撫中達司, 豊川絢也, 竹内銀河, “科学的介護を支援する介護 AI 記録分析ツール「けあらぼ」,” 三菱電機技報, Vol. 95, No. 8, pp. 506–509, 2021.

[11] Diederik P Kingma and Jimmy Ba, “Adam: A Method for Stochastic Optimization,” arXiv preprint arXiv:1412.6980, 2014.

[12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Journal of Machine Learning Research, Vol. 15, No. 1, pp. 1929–1958, 2014.

# レム睡眠のリアルタイム自動検出に関する研究

佐野 史弥<sup>†</sup> 国生 泰資<sup>†</sup> 堀江 和正<sup>††</sup> 古木 淳也<sup>†††</sup> 鈴木 陽子<sup>††††</sup>

阿部 高志<sup>††††</sup> 北川 博之<sup>††††</sup>

<sup>†</sup> 筑波大学大学院 理工情報生命学術院 システム情報工学研究群 〒305-8573 茨城県つくば市天王台 1 丁目 1-1

<sup>††</sup> 筑波大学 計算科学研究センター 〒305-8573 茨城県つくば市天王台 1 丁目 1-1

<sup>†††</sup> 筑波大学大学院人間総合科学学術院人間総合科学研究群 〒305-8573 茨城県つくば市天王台 1 丁目 1-1

<sup>††††</sup> 筑波大学 国際統合睡眠医科学研究機構 〒305-8573 茨城県つくば市天王台 1 丁目 1-1

E-mail: <sup>†</sup>{f.sano,kokusho}@kde.cs.tsukuba.ac.jp,

<sup>††</sup>horie@kde.cs.tsukuba.ac.jp,<sup>†††</sup>{suzuki.yoko.gz,abe.takashi.gp}@u.tsukuba.ac.jp,

<sup>††††</sup>kitagawa@cs.tsukuba.ac.jp

**あらまし** 本研究では、レム睡眠のリアルタイム自動検出の機械学習モデル及びシステムアーキテクチャの提案を行う。既存の手法では計測済の睡眠データに対してレム睡眠を後から検出する手法を用いている。そのため、睡眠の計測中にレム睡眠を検出することが不可能であった。しかし、睡眠研究において、レム睡眠の重要性が認識されており、レム睡眠時に音や匂いなどを刺激を与えるなど、リアルタイムで被験者のレム睡眠状態を検出したいという要求が存在する。そのため、本研究では、睡眠ステージ判定に移動窓を適用させ、かつレム睡眠の検出に特化するようにファインチューニングしたモデルを提案し、評価を行う。

**キーワード** ヘルスケア, レム睡眠, 機械学習

## 1 はじめに

睡眠は人間の生活において極めて重要な役割を果たし、全生活時間の約 1/3 を占める。健康や日々の活動に直接的な影響を与えるこの生理現象は、睡眠障害により労働効率の低下、生活習慣病 (例えば糖尿病)、疫力の低下などを引き起こす可能性がある。したがって、睡眠パターンを正確に把握し、改善を促進することは非常に重要である。睡眠の評価には PSG(睡眠ポリグラフ検査) が用いられる。この検査では、被験者の脳波 (EEG)、眼電位 (EOG)、筋電位 (EMG)、血中酸素飽和度などの生体信号を一晩通して記録し、睡眠サイクルを評価する。実際に測定された生体信号を図 1 に示す。測定された生体信号から睡眠状態を推定する方法として睡眠ステージ判定がある。睡眠は大きく、覚醒状態、レム睡眠、ノンレム睡眠に分けられる。米国睡眠学会によると、睡眠は 5 つのステージ (W: 覚醒, REM: レム睡眠, ノンレム睡眠: N1, N2, N3) に分類される [1]。これらのステージ判定は生体信号における波の周波数や波形などの特徴から判断される。

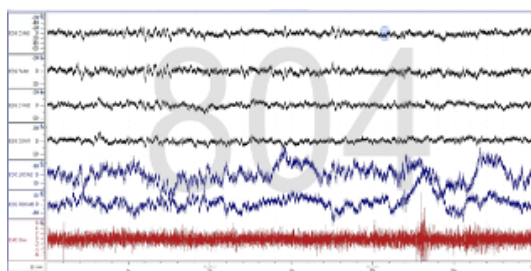


図 1: 実際に測定された生体信号

従来の睡眠ステージの判定は、生体信号を通常 30 秒単位で分割し、専門家が目視で各エポックにおける睡眠ステージを決定することが行われてきた。しかしながら一夜分のデータを分析する作業は専門的なスキルを要し、大きな労力と専門家に対する負担が伴う。この課題に対処するため、睡眠ステージの自動判定に関する研究が行われてきた。近年では、機械学習の進展により専門家と同等の精度を持つ判定モデルが提案されている。これを受け、近年では、睡眠中の被験者の睡眠状態をリアルタイムで分析したいというニーズも生じている。

本研究では具体的な応用先として、睡眠の基礎研究、特に、睡眠中の記憶や夢に関する介入研究に着目する。このような研究では睡眠中の被験者に対し、医師、技師がリアルタイムの判定を実行、レム睡眠中の被験者に対して刺激を与えるなどの実験が行われている。ここでは、このようなリアルタイムでのレム睡眠の自動検出 (レム睡眠の開始点、終了点の検出) を行う深層学習モデルの開発に取り組む。目標とする深層学習モデルは以下の条件を満たすようなモデルである。

- 確定レムの検出
  - － レム睡眠において特徴波である急速眼球運動を伴わないトニックレムと特徴波を伴う確定レムがある。以下の三つの全ての条件を満たした時確定レムであるという。
    - \* K 複合あるいは睡眠紡錘波を伴わない LAMF が混在する EEG 活動
    - \* 顎 EMG 低緊張
    - \* 急速眼球運動 (REM)
  - － 確定レムの際に刺激などを与えることが一般的である

ので、本研究では確定レムを検出する。これより先レム睡眠は確定レムを意味する。

- 誤検出が少ない
  - 非レム睡眠状態に刺激を与えることは良くないため、検出を頻繁に行うより誤検出を少なくさせることを重視。
- 誤差が少ない
  - レム睡眠開始時および終了時をできるだけ誤差なく判定することが求められる。

## 2 関連研究

リアルタイム性を重視した睡眠ステージ自動判定の関連研究として、Patanaiik らは EEG1ch を用いたリアルタイムステージ判定システムをウェアラブル上で構築している.[2] 正答率は 0.83 と高精度を記録しているが、多チャンネルの生体信号に対応できないという問題がある。

Koushik らは、信号計測と睡眠ステージの判定処理を別々のプロセスとして分離し、CFS 形式のファイルを介してデータをやり取りするアーキテクチャを提案している。このシステムでは約 5 ミリ秒ごとに 1 エポック分の睡眠ステージの判定が可能である。しかし、特定の時点でのステージ判定には、その直後の 5 エポック分のデータを反映させる必要があるため、実際にはリアルタイムでのステージ判定は 150s の遅延があるため行えない.[3]

## 3 提案手法

本研究の目的は、リアルタイムで高精度生体電位計測器 (Biosemi 社製 ActiveTwo) からデータを取得し、レム睡眠を検出するシステムアーキテクチャおよび、レム睡眠の検出モデル、モデルの適用方法の提案である。対象の高精度脳波計は Biosemi 社製 ActiveTwo である。

### 3.1 システムアーキテクチャ

リアルタイムで生体電位測定器からデータを取得し、レム睡眠を検出するシステムアーキテクチャは現段階では、図 2 のようなアーキテクチャを想定している。リアルタイム判定を行うサーバと生体電位計測器から被験者の生体信号を常時計測しサーバに送信するクライアントを用意する。生体信号をクライアント側から送信し、サーバ側で受け取りリアルタイムにレム睡眠かの判定を計測中に行い続ける。



図 2: 想定システムアーキテクチャ

### 3.2 移動窓を用いた基本判定モデル

基本判定モデルをリアルタイムに適用できるように、移動窓を基本判定モデルに適用させた。

#### 3.2.1 基本判定モデル

基本判定モデルは、本研究室によって提案された Sleep-CAM[4] である。Sleep-CAM は CNN ベースのモデルであり、図 3 に示すような特徴抽出層とステージ割り当て層で構成されている。このモデルは 30 秒の生体信号に対してステージ判定を行う。

入力には脳波  $EEG \times 6$ 、眼電位  $EOG \times 2$ 、顎筋電位  $EMG \times 1$  の 9 つの信号からなる。入力信号のサンプリング周波数は  $EEG$  と  $EMG$  は 200Hz、 $EOG$  は 100Hz である。

使用した重みは高精度脳波計 ActiveTwo で計測された 36 件のデータセットと簡易脳波計 PSG1100 で計測された 49 件のデータセットの計 75 件でトレーニングした重みである。ステージ判定精度はおおよそ 80% と高精度である。しかし、30s 単位で判定をおこなうため、リアルタイムに適していないという問題がある。それを解決する方法を次節で説明する。

#### 3.2.2 移動窓

リアルタイム性を重視するためのアプローチとして移動窓を用いた。移動窓を使用することで、図 4 のように入力窓を任意の時間でずらし、これを判定間隔と捉える。入力窓を 1s ずつずらすことで、基本判定モデルの問題点であった 30s ごとの判定を任意の短い時間で判定することを可能とさせる。入力窓 30 秒のステージ判定を行うそのステージ判定を「レム睡眠 (REM)」か「非レム睡眠 (W, N1, N2, N3)」に分け出力する。しかし、リアルタイムでレム睡眠を検出するには入力の最新の時点が「レム睡眠」か「非レム睡眠」かを判定することが求められるため判定の部分が適していない。判定の問題を解決する方法を次節で説明する。

### 3.3 ファインチューニングモデル

基本判定モデルとファインチューニングモデルの判定の違いを図 5 に示す。入力の最新の時点が「レム睡眠」か「非レム睡眠」かを判定するようにファインチューニングしたモデルである。基本判定モデルは入力窓に対する睡眠ステージ判定を行うモデルであった。この基本判定モデルの出力としては睡眠ステージの 5 値の出力となり、それを「レム睡眠」か「非レム睡眠 (Wake, N1, N2, N3)」としてレム睡眠か判定していた。そこで、図 5 のように入力窓の最新点が「レム睡眠」か「非レム睡眠」かを判定するモデルになるよう基本判定モデルのステージ割り当て層の Softmax のユニット数を 5 から 2 に変更しファインチューニングした。ファインチューニングモデルの構造は図 6 に示す。

### 3.4 重み層の導入

判定が最新点が「レム睡眠」か「非レム睡眠」かを判定するので最新点の重要性を強調することに重点を置いた。時間的な情報をより考慮させるためファインチューニングモデルのステージ割り当て層の GlobalAveragePooling 層の前に重みづけ

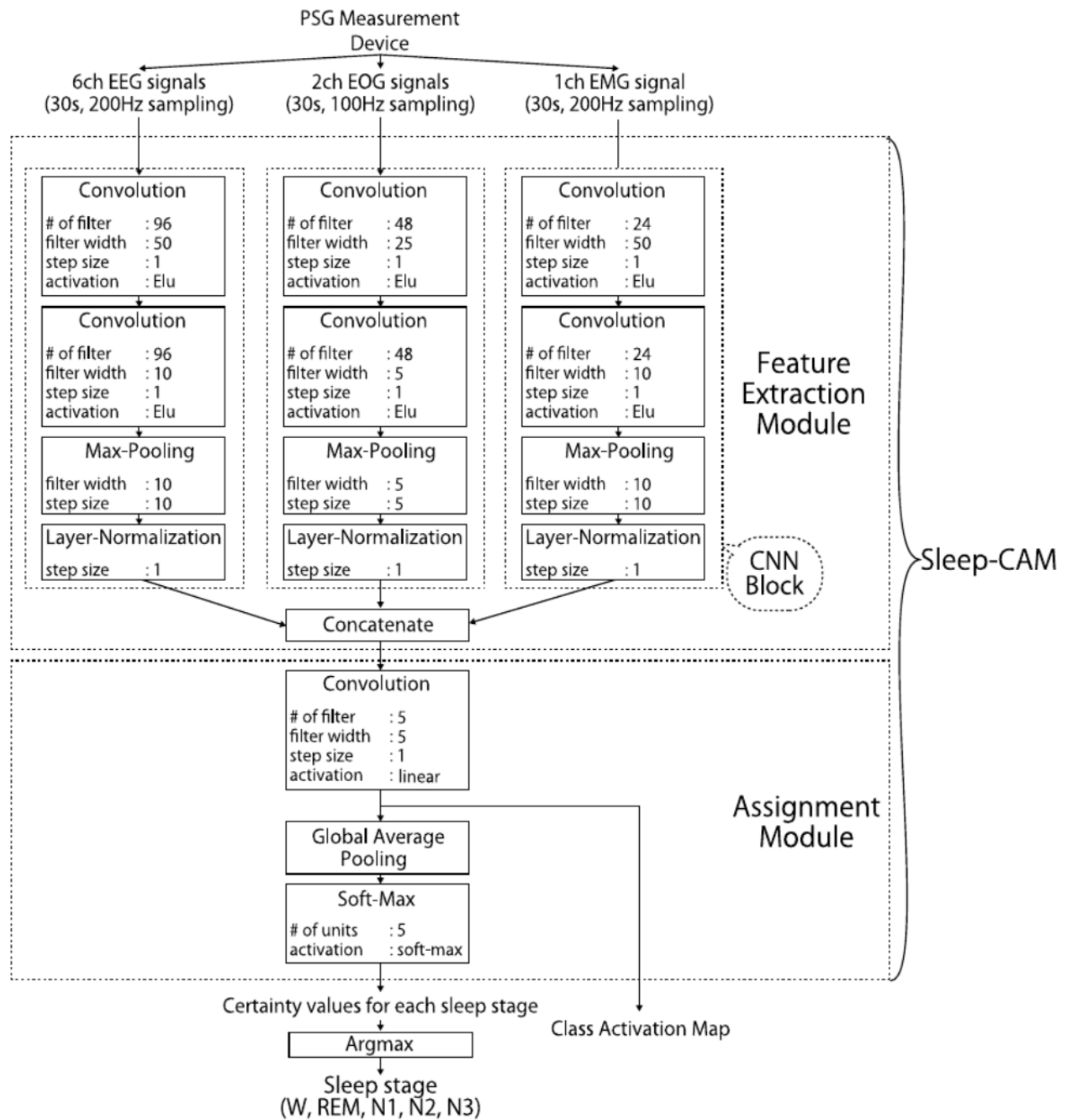


図 3: Sleep-CAM[4]

の層を導入した。時間的な情報を考慮するために線形重み層 (LinearWeight-Layer) とシグモイド重み層 (SigmoidWeight-Layer) を実装した。これらの層は、最新の状態ほど大きな重みを与えることで、最新の状態をより重視することが可能となる。

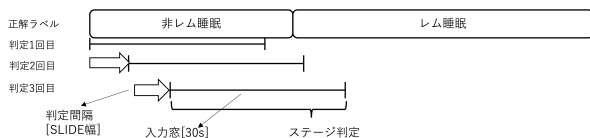


図 4: 移動窓

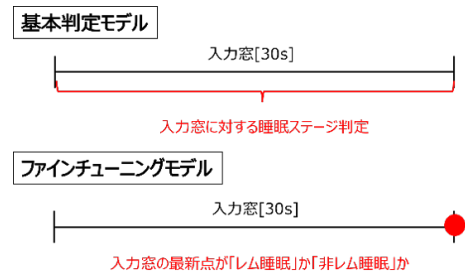


図 5: 判定の違い

(上) 基本判定モデルの判定

(下) ファインチューニングモデルの判定

3.4.1 線形重み層 (LinearWeight-Layer)

線形重み層は重み付けに最小値を 0，最大値 1 とする線形関数を使用し，入力各点の重要性を線形的に調整する層である．この線形重み層は，図 7b のように入力の最初の点を重み 0 を割り当て，最新点の重みを 1 とし，時間の経過と共に重要度が線形的に増加するようになる．これにより，モデルは時系列の進行に伴って最新の情報をより重視し，古い情報の影響を徐々に薄めることが可能となる．

3.4.2 シグモイド重み層 (SigmoidWeight-Layer)

シグモイド重み層は重み付けにシグモイド関数を使用し，入力各点の重要性を非線形的に調整する．シグモイド関数を用いることで，図 7a 入力最新点に対しては重要性を高め，より古いデータ点に対してはその重要性を徐々に低下させるような重み付けを行う．

このアプローチにより，モデルは最新のデータ点をより重視しつつ，入力データを通じた情報の流れを捉える能力を保持できる．

4 実 験

移動窓のスライド幅を 1s とし，基本判定モデルとファインチューニングモデルの比較を行った．使用したデータセットは提案手法のファインチューニングで使ったデータセットと同じである．実験環境について，OS は Ubuntu LTS 16.04.07，実装は Python3.8.0, TensorFlow 2.3.0 を使用した．また，CPU については Intel(R) Xeon(R) Gold 5122 CPU @3.60GHz，GPU については Nvidia Quadro P6000 を使用した．

4.1 データセット

実験に用いるデータセットは，計高精度生体電位計測器 ActiveTwo で測定され，専門家がレム睡眠の正解ラベルがついた 5 夜分のデータである．5 夜分の合計長は 41 時間 5 分であり，うちレム睡眠の時間は 9 時間 47 分 10 秒であった．各データの睡眠時間，レム睡眠の時間，非レム睡眠時間を表 1 に示す．

なおこのデータセットの前処理として，簡易脳波計 PSG1100 の電極間の電位差のチャンネルとなるよう ActiveTwo の単一チャンネルを用いて PSG1100 と同様の電極間の電位差のチャンネルになるように変換し，ダウンサンプリング，ノイズ除去，正規化が行われている．

表 1: データセットの詳細

| データ名  | 合計時間 [s] | レム睡眠 [s] | 非レム睡眠 [s] |
|-------|----------|----------|-----------|
| Data1 | 29671    | 5577     | 24123     |
| Data2 | 29431    | 7034     | 22426     |
| Data3 | 29731    | 6634     | 23126     |
| Data4 | 29611    | 9451     | 20189     |
| Data5 | 29311    | 7034     | 22306     |

4.2 ファインチューニング

異なる時間窓 (15s 及び 30s) を入力として用い，基本判定モ

デルをそのまま用いたモデル，線形重み層を追加したモデル，シグモイド層を追加したモデルの 3 つの異なるモデルに対してファインチューニングを行なった．全てのファインチューニングモデルにおいて図 6 に示されるようにファインチューニングの対象として基本判定モデルのステージ割り当て層の畳み込み層のみとし，特徴抽出層は基本判定モデルで重みを固定した．

モデルの汎化能力の評価を行うため，5 分割交差検証 (5-fold cross-validation) を実施した．各交差検証のサイクルで 1 件のデータがテストデータとして，別の 1 件が検証データとして，残りの 3 件が訓練データとして使用した．このアプローチにより，各データ点がモデル評価のために均等にテストデータとして利用され，同時にモデルの訓練と検証にも等しく寄与させた．モデルの学習において，学習率は  $5.0 \times 10^{-6}$  とし，損失関数はバイナリクロスエントロピーを，最適化アルゴリズムとして Adam を用いた．バッチサイズは 32 である．

4.3 モデルの精度比較

4.3.1 実験方法

リアルタイムな判定を想定し，入力窓を 1s ずつずらし，レム睡眠かどうかの判定を行う．モデルの評価指標は，分類問題で用いられる，Accuracy(正解率)，Precision(適合率)，Sensitivity(感度)，F1-score とした．Accuracy はモデルの全体的な正しさを表し，Sensitivity はレム睡眠を正しく識別するモデルの能力の指標であり，Precision はモデルが検出したレム睡眠が実際に正しかった割合を，F1-score は精度と感度の調和平均で，両者のバランスを取った単一スコアである．

4.3.2 実験結果

実験結果を表 3 に示す．表 3 におけるモデル状態はファインチューニング前後かを表し，モデルバリエーションにおいて，Normal は基本判定モデルをそのまま用いたモデル，LinearWeight は線形重み層を追加したモデル，SigmoidWeight はシグモイド重み層を追加したモデルである．正解率において全てのモデルで 90%を超えているため高精度であることがいえる．また，ファインチューニング後のモデルにおいて，入力 30s で線形重み層を追加したモデルが最も F1score が高くなっている．

ファインチューニング後の全てのモデルにおいて，ファインチューニング前のモデルと比較して，F1-score が低下してしまっていることが分かる．しかし，感度と適合率を見てみると，感度が低下し，適合率が向上していることがわかる．これは，ファインチューニング前より，レム睡眠を見逃す確率は高くなってしまったが，誤って非レム睡眠をレム睡眠と識別することが少なくなったことを示唆している．これより，ファインチューニングを行ったことでモデルは全体的な評価としてはファインチューニング前より劣ってしまうが，レム睡眠の判定がより慎重になり，正確にレム睡眠と判定できるようになり誤検出が少なくなったと考察できる．

次に一夜分の正解ラベルと判定結果を図 9 に示す．これはヒブノグラムという睡眠中の異なる睡眠段階を可視化したグラフであり，横軸は時間，縦軸は睡眠段階 (「レム睡眠」, 「非レム睡眠」) である．

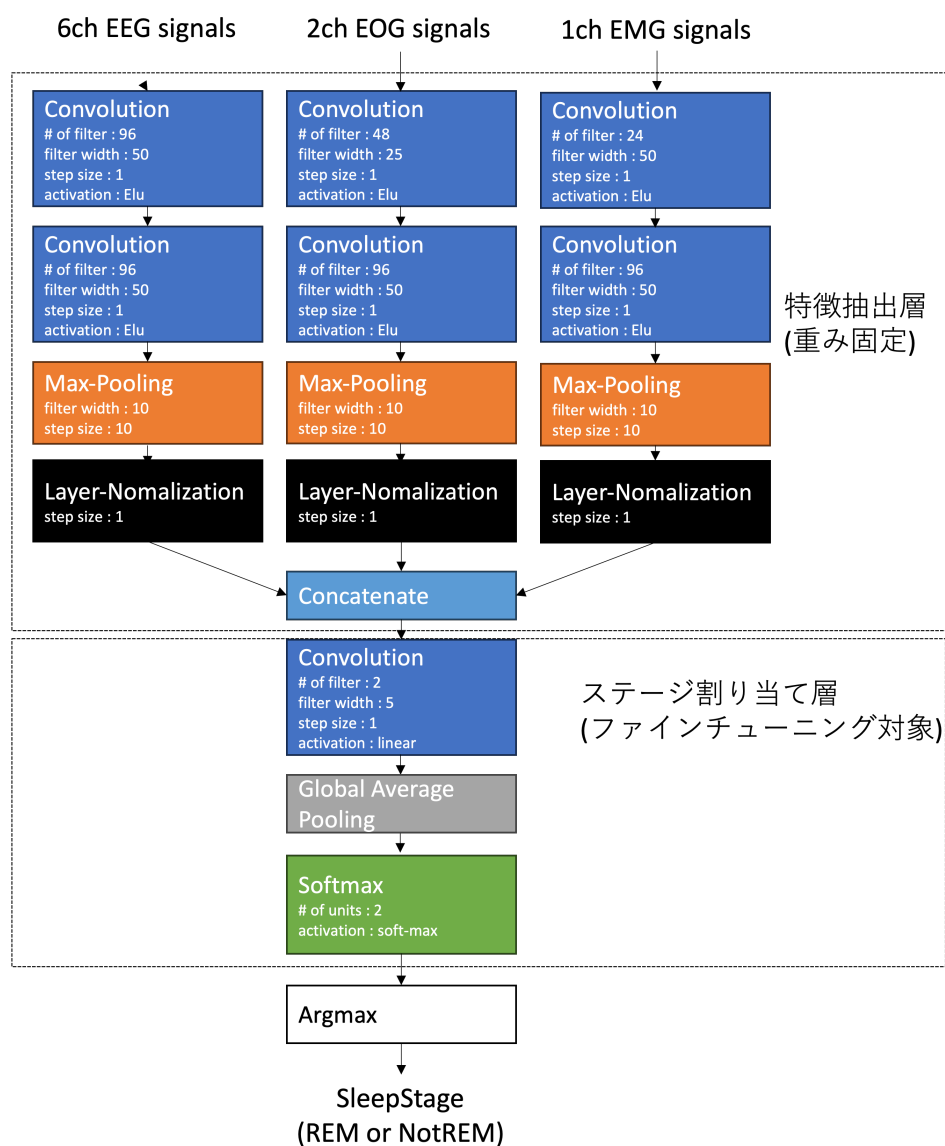


図 6: ファインチューニングモデル

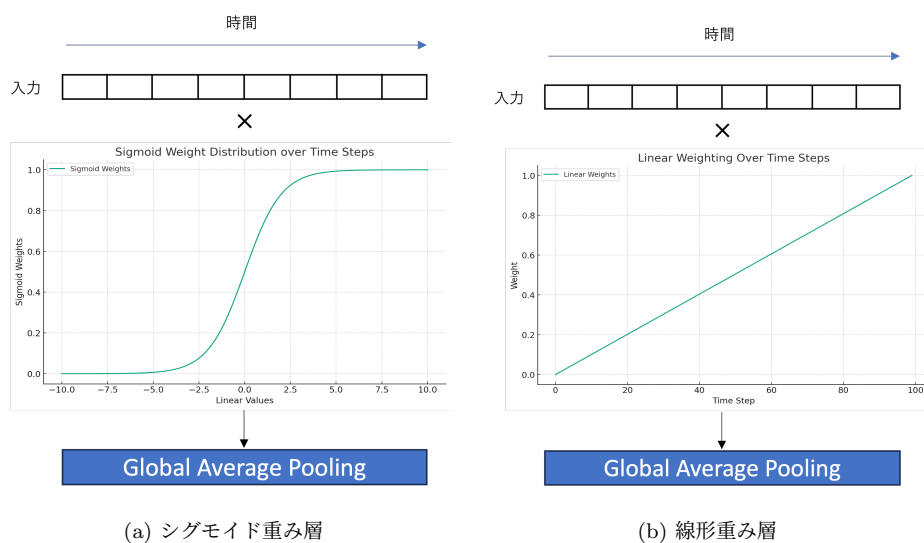


図 7: 重み層 (a) シグモイド重み層, (b) 線形重み層

図9に示される結果から、30秒の入力窓を持つモデルにファインチューニングを施した結果、より細かいレム睡眠段階を識別できるようになったことが明らかとなる。これは、ファインチューニングによりモデルのレム睡眠検出能力が向上し、睡眠サイクル中の短時間のレム睡眠も捉えられるようになったことを示唆している。一方で、15秒の入力窓を用いたモデルでは、レム睡眠と非レム睡眠の区別において一定の問題が確認された。特に、短期間のレム睡眠の識別が不確かとなり、レム睡眠と非レム睡眠の間での分類が短時間で不安定に振動する傾向が観察された。

#### 4.4 レム睡眠の開始点及び終了点の検出率

##### 4.4.1 実験方法

正解ラベルのレム睡眠の開始点と終了点を許容誤差内で判定できるか及びモデルによる判定の開始点と終了点から、判定を基準として正解ラベルが許容誤差内に存在するかを評価した。今回の実験において許容誤差は30sとした。正解ラベルのレム睡眠の開始点と終了点を許容誤差内で判定できるかについて図8を用いて説明する。図8はヒプノグラム(睡眠遷移図)の一部を切り抜いたものである。上のヒプノグラムが正解ラベル、下のヒプノグラムがモデルの判定である。正解ラベルから前後30s以内に開始点または終了点があれば、その正解ラベルの開始点または終了点を検出できたとみなす。同様に基準を逆にしモデルによる判定から前後30秒以内に正解ラベルが存在するかを検出した。

正解ラベルのレム睡眠の開始点と終了点を許容誤差内で判定できるかの評価指標として、感度 (Sensitivity)、検出回数 (Det.Count)、正解ラベルからの平均時間誤差 (Average)、標準偏差 (Std) を用いる。感度は、モデルが正解ラベルのレム睡眠の開始点及び終了点をどれだけ正確に検出できているのかを示す。判定のレム睡眠の開始点と終了点を基準として、許容範囲内に正解ラベルがあるかどうかの判定の評価指標として、適合率 (Precision)、判定回数 (Count)、前後30秒以内に正解ラベルが含まれている判定回数 ( $\text{Count} < 30\text{s}$ )、判定から正解ラベルの平均時間誤差 (Average)、標準偏差 (Std) を用いる。適合率は、レム睡眠の開始点及び終了点の検出においてモデルが正解ラベルと一致しているかを示す。

今回用いたデータセットの各データの開始点及び、終了点の回数を表2に示す。データセットにおける総開始点は279回あり、総終了点は277回であり、1データあたりの平均開始点および終了点の回数は55.8回、55.4回であった。

表 2: データセットにおける開始点及び終了点の回数

|         | 開始点 [回] | 終了点 [回] |
|---------|---------|---------|
| Data1   | 60      | 59      |
| Data2   | 37      | 37      |
| Data3   | 65      | 65      |
| Data4   | 64      | 64      |
| Data5   | 53      | 52      |
| Average | 55.8    | 55.4    |

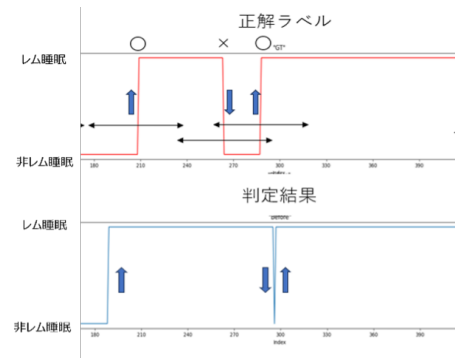


図 8: 判定方法

##### 4.4.2 実験結果

実験結果を表4, 6, 5, 7に示す。正解ラベルを基準とした結果から、開始点、終了点共に、ファインチューニングを行ったことで感度は向上していることがわかり、特に15sのSigmoid-Weightが0.9426, 0.9458と高い値を示した。また入力窓を狭くすることで平均誤差も短くなり標準偏差も小さくなっていることが分かる。一方で、判定基準の結果を見てみると、ファインチューニングを行ったことで、開始点及び終了点の判定回数が増加していることがわかる。これは短期間のレム睡眠の識別が不確かとなり、レム睡眠と非レム睡眠の間での分類が短時間で不安定に振動する傾向が要因として考えられる。それにより、開始点と終了点共に適合率が低下してしまっていることがわかった。これはモデルがレム睡眠に過剰に反応してしまっているということが考えられる。

## 5 まとめ

本研究では、リアルタイムにレム睡眠を検出するようなモデルを提案した。目標とする深層学習モデルに対して今回の結果は以下のようにまとめられる。

- 確定レムの検出

ファインチューニングを行ったことで確定レムの全体的な検出はあまり向上しなかったが、より細かな確定レムは捉えることができるようになった。

- 誤検出が少ない

ファインチューニングを行ったことで、全体的なレム睡眠全誤検出は少なくなった。ただし、開始点と終了点における誤検出は向上しなかった。

- 誤差が少ない

開始点と終了点の誤差はファインチューニングを行うことで改善された。

今後の課題として、以下が挙げられる。

- レム睡眠の開始点と終了点の検出精度のさらなる向上及びモデルの評価方法

- モデル判定におけるレム睡眠と非レム睡眠の間での短時間で不安定な振動の抑制
- レム睡眠を検出するシステムアーキテクチャの設計及び実装

## 6 謝 辞

本研究の一部は科研費: 基盤研究 C(JP20K03493), 挑戦的研究 (萌芽)(JP22K19802), 基盤研究 B(JP23H03399), 若手研究 (JP23K16988), AMED ムーンショット型研究開発事業 (JP21zf0127005) による。

## 文 献

- [1] American Academy of SLEEP MEDICINE, The AASM Manual for the Scoring of Sleep and Associated Events, “AASM Manual for the Scoring of Sleep and Associated Events”, (株) ライフサイエンス, 2018.
- [2] Amiy Patanaik, Ju Lynn Ong, Joshua J Gooley, So-nia Ancoli-Israel and Michael W L Chee, “An end-to-end framework for real-time automatic sleep stage classification, Sleep”, Vol. 41, No. 5, 2018.
- [3] Abhay Koushik, Judith Amores, Pattie Maes, “Real-Time Sleep Staging using Deep Learning on a Smartphone for a Wearable EEG”, IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN), 2019.
- [4] Kazumasa Horie, Leo Ota, Ryusuke Miyamoto, Takashi Abe, Yoko Suzuki, Fusae Kawana, Toshio Kokubo, Masashi Yanagisawa and Hiroyuki Kitagawa, “Automated sleep stage scoring employing a reasoning mechanism and evaluation of its explainability”, Scientific Reports 12, No. 12799, 2022

表 3: ファインチューニング前後のモデル性能比較

| モデル状態       | モデルバリエーション    | input window | Accuracy | Sensitivity | Precision | F1-score |
|-------------|---------------|--------------|----------|-------------|-----------|----------|
| ファインチューニング前 | N/A           | 30s          | 0.9233   | 0.967       | 0.773     | 0.859    |
| ファインチューニング後 | Normal        | 15s          | 0.9161   | 0.845       | 0.815     | 0.830    |
|             |               | 30s          | 0.9222   | 0.845       | 0.835     | 0.840    |
|             | LinearWeight  | 15s          | 0.9170   | 0.842       | 0.820     | 0.831    |
|             |               | 30s          | 0.9213   | 0.861       | 0.822     | 0.841    |
|             | SigmoidWeight | 15s          | 0.9108   | 0.817       | 0.814     | 0.816    |
|             |               | 30s          | 0.9170   | 0.850       | 0.815     | 0.832    |

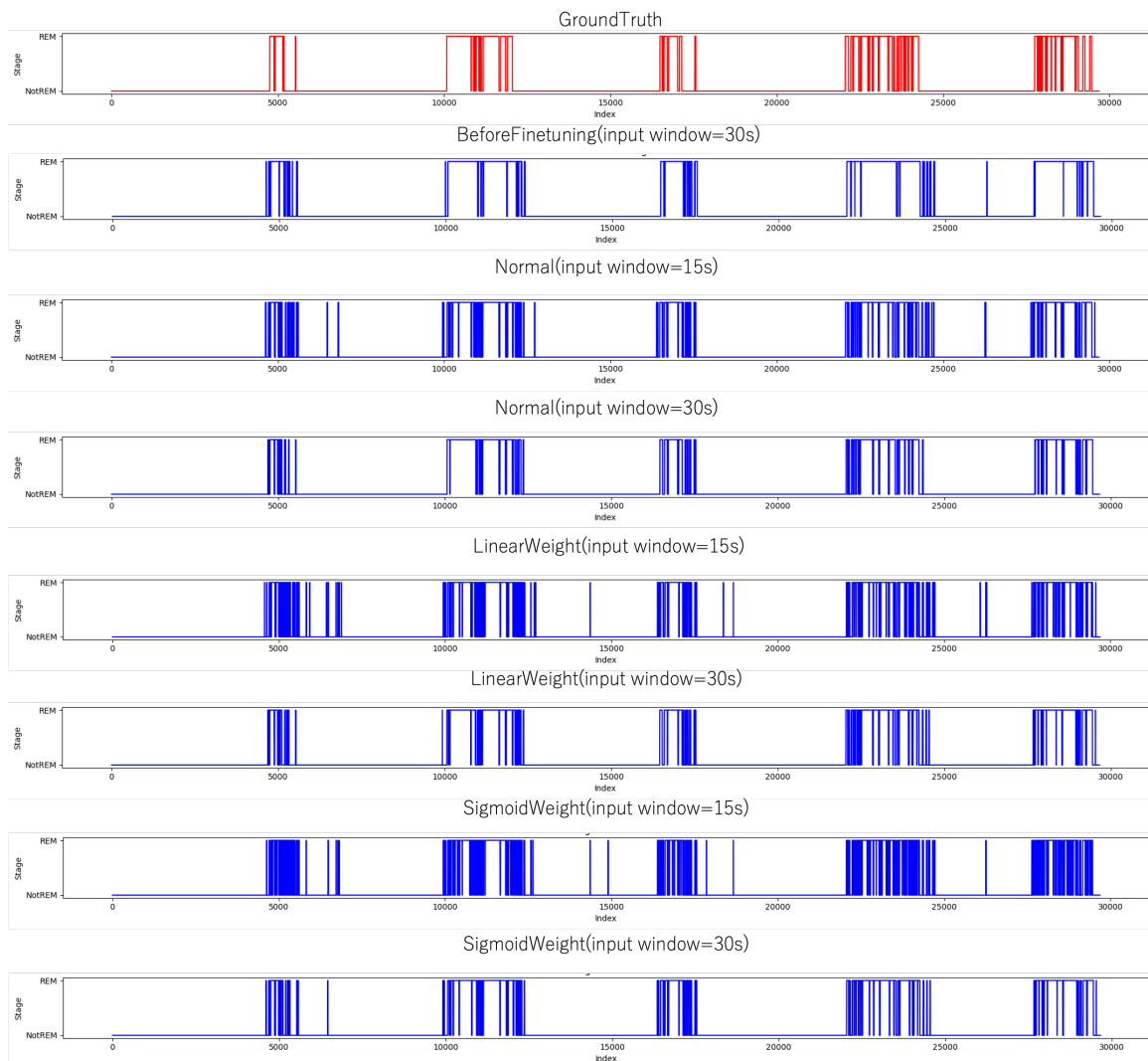


図 9: ヒプノグラム (睡眠遷移図)

表 4: 正解ラベルを基準としたレム睡眠の開始点における結果

| モデル状態       | モデルバリエーション   | input window | Sensitivity | Det.Count | Average | Std   |
|-------------|--------------|--------------|-------------|-----------|---------|-------|
| ファインチューニング前 | N/A          | 30s          | 0.4767      | 133       | +1.40   | 14.55 |
| ファインチューニング後 | Normal       | 15s          | 0.8781      | 245       | -0.61   | 10.31 |
|             |              | 30s          | 0.6308      | 176       | +4.92   | 13.88 |
|             |              | 15s          | 0.8924      | 249       | -0.90   | 9.88  |
|             | LinearWeight | 30s          | 0.6953      | 194       | +0.45   | 14.14 |
|             |              | 15s          | 0.9426      | 263       | -0.77   | 9.46  |
|             |              | 30s          | 0.7311      | 204       | -0.05   | 12.71 |

表 5: 判定基準としたレム睡眠の開始点における結果

| モデル状態       | モデルバリエーション   | input window | Precision | Count | Count<30s | Average | Std   |
|-------------|--------------|--------------|-----------|-------|-----------|---------|-------|
| ファインチューニング前 | N/A          | 30s          | 0.4120    | 381   | 157       | -0.09   | 14.91 |
| ファインチューニング後 | Normal       | 15s          | 0.3011    | 1448  | 436       | +0.75   | 15.15 |
|             |              | 30s          | 0.3618    | 644   | 233       | -2.45   | 15.37 |
|             |              | 15s          | 0.2995    | 1636  | 490       | +1.56   | 15.27 |
|             | LinearWeight | 30s          | 0.3409    | 880   | 300       | +2.71   | 15.22 |
|             |              | 15s          | 0.3015    | 1804  | 544       | +0.11   | 15.84 |
|             |              | 30s          | 0.3290    | 933   | 307       | +2.19   | 15.57 |

表 6: 正解ラベルを基準としたレム睡眠の終了点における結果

| モデル状態       | モデルバリエーション   | input window | Det.Rate | Det.Count | Average | Std   |
|-------------|--------------|--------------|----------|-----------|---------|-------|
| ファインチューニング前 | N/A          | 30s          | 0.5126   | 142       | +9.40   | 12.61 |
| ファインチューニング後 | Normal       | 15s          | 0.9061   | 251       | +0.67   | 8.46  |
|             |              | 30s          | 0.6714   | 186       | +4.78   | 12.31 |
|             |              | 15s          | 0.9061   | 251       | +0.13   | 7.19  |
|             | LinearWeight | 30s          | 0.7075   | 196       | +2.58   | 10.15 |
|             |              | 15s          | 0.9458   | 262       | -0.43   | 6.94  |
|             |              | 30s          | 0.7509   | 208       | +1.39   | 9.68  |

表 7: 判定基準としたレム睡眠の終了点における結果

| モデル状態       | モデルバリエーション   | input window | Precision | Count | Count<30s | Average | Std   |
|-------------|--------------|--------------|-----------|-------|-----------|---------|-------|
| ファインチューニング前 | N/A          | 30s          | 0.4120    | 381   | 157       | -9.15   | 12.64 |
| ファインチューニング後 | Normal       | 15s          | 0.3035    | 1446  | 439       | -0.37   | 13.36 |
|             |              | 30s          | 0.3841    | 643   | 247       | -3.61   | 13.30 |
|             |              | 15s          | 0.2990    | 1635  | 489       | +1.54   | 13.68 |
|             | LinearWeight | 30s          | 0.3788    | 879   | 333       | -0.35   | 13.65 |
|             |              | 15s          | 0.298     | 1802  | 537       | +1.19   | 14.68 |
|             |              | 30s          | 0.3401    | 932   | 932       | -0.47   | 12.37 |

# リアルタイム性を考慮した睡眠ステージ判定システムの開発

国生 泰資<sup>†</sup> 堀江 和正<sup>††</sup> 鈴木 陽子<sup>†††</sup> 阿部 高志<sup>†††</sup> 北川 博之<sup>†††</sup>

<sup>†</sup> 筑波大学大学院 理工情報生命学術院 システム情報工学研究群 〒305-8573 茨城県つくば市天王台 1 丁目 1-1

<sup>††</sup> 筑波大学 計算科学研究センター 〒305-8573 茨城県つくば市天王台 1 丁目 1-1

<sup>†††</sup> 筑波大学 国際統合睡眠医科学研究機構 〒305-8573 茨城県つくば市天王台 1 丁目 1-1

E-mail: <sup>†</sup>{kokusho, horie}@kde.cs.tsukuba.ac.jp, <sup>††</sup>{suzuki.yoko.gz, abe.takashi.gp}@u.tsukuba.ac.jp,  
<sup>†††</sup>kitagawa@cs.tsukuba.ac.jp

**あらまし** 睡眠ステージ判定とは被験者の脳波や筋電位などの生体信号に現れる特徴をもとに、一定間隔で睡眠状態をラベル付していく作業であり、被験者の睡眠の質やサイクルなどを評価することを可能とする。これまで睡眠ステージ判定とは被験者ごとに測定された睡眠データ (生体信号) に対して、専門技師の目視による 30s 間隔のラベル付が行われてきたが、このアナログ的作業は時間及び人的コストを多く要し、睡眠研究を促進していく上で課題となっていた。これに対し近年、様々な睡眠ステージ判定の自動化手法が研究され、特にニューラルネットワークを用いた判定手法が注目を集めている。一方、被験者の睡眠の質の向上や睡眠障害への治療といった睡眠研究においては、睡眠ステージをリアルタイムに判定し被験者の睡眠状態に応じたフィードバックを行う等のニーズが存在する。そこで本研究では睡眠中の患者の睡眠状態を把握できるようなリアルタイムなステージ判定手法を提案する。具体的には移動窓を用いた連続的なステージ判定手法について実験・評価を行う。

**キーワード** ヘルスケア, 睡眠ステージ判定, 機械学習

判定モデルの 2 つについて判定精度を検証する。

## 1 はじめに

睡眠は人間の健康と活動に大きく関わっている。睡眠の質はヒトの日中活動のパフォーマンスに大きく影響し、不眠症やナルコレプシーなどの睡眠障害を伴うことも珍しくない。したがって、現在に至るまで睡眠に関する研究が盛んに行われてきた。睡眠の検査手法の一つに PSG (Polysomnography) 検査がある。PSG 検査を通して、被験者の終夜にわたる脳波や眼電図、筋電図などを記録することができ、そうして得られた生体信号を元に睡眠ステージ判定が行われる。従来のステージ判定手法は専門技師による波形の目視により行われ、人的及び時間のコストが多くかかることが課題となっている。そのような課題に対し、機械学習を用いた自動判定に関する研究が多くなされ、近年ではニューラルネットワークモデルを中心とした判定モデルを用いて高い精度で判定することが可能となってきた。一方で、このような睡眠ステージ判定をリアルタイムに行いたいといったニーズが存在する。リアルタイムに睡眠ステージ判定を行うことは、被験者の睡眠状態に対応した即座のフィードバックを可能にし、睡眠障害の治療や睡眠の質の向上などを目指す睡眠研究への活用が期待されている。しかし、これまでに提案された睡眠ステージ判定の自動化に関する研究は測定済の生体信号に対しステージ判定を行うことを想定しており、睡眠中の被験者の睡眠ステージをリアルタイムに判定する手法は未だ確立されていない。本研究では、機械学習に基づくステージ判定モデルを用いたリアルタイムなステージ判定システムを提案する。実験では、リアルタイムシステムの処理時間の検証や、従来の睡眠ステージ判定モデルと新たにリアルタイム用に学習した

## 2 前提知識

### 2.1 PSG 検査

PSG 検査は、睡眠脳波や筋電位、呼吸サイクルなど睡眠状態に関わる生体信号の計測を目的に行われる。PSG 検査は終夜を通して行われ、被験者には頭部や目元等の複数箇所に電極を取り付け、生体信号を記録していく。こうして得られた 1 夜分にわたる生体信号の時系列データを元に睡眠のサイクルや睡眠中の異常などを分析することができる。

### 2.2 睡眠ステージ判定

PSG 検査により得られた生体信号を元に睡眠ステージ判定が可能となる。睡眠ステージは大きく分けて REM 睡眠と Non-REM 睡眠に別れる。REM 睡眠とは急速眼球運動 (Rapid Eye Movement) が伴う睡眠状態であり、脳が活発に動き記憶の定着などが行われる。Non-REM 睡眠とは体、脳共に休んでいる状態であり成長ホルモンの分泌や脳の休息が行われる。[3]

さらに Non-REM 睡眠には睡眠の深さが浅い場合と深い場合が存在し、米国睡眠医学会 (AASM) が示すガイドライン [4] では Non-REM 睡眠を 3 段階に分割して定義している。ヒトは 1 夜を通してこれらのステージを往来しながら、睡眠を行う。このような睡眠ステージは PSG 検査より取得した生体信号に現れる、特徴波や振幅及び持続時間などの特徴を元に判別可能である。これらの判定基準についても、米国睡眠医学会 (AASM) が判定基準を定義しており、その基準を元に専門技師が生体信号に対し、ステージをラベリングしていく。

REM ステージでは、脳波信号 (EEG) の電位が低く 2-6Hz ののこぎり波が出現する。また、筋電位 (EMG) では顎筋電位が低く、眼球電図 (EOG) では急速眼球運動 (REMs) が発生する。Non-REM1 ステージでは、EEG では頭蓋頂鋭波が発生し、EOG では低速眼球運動が見られる。Non-REM2 ステージでは、EEG にて K-complex 波や紡錘波が発生する。Non-REM3 ステージでは、EEG にて 0.2-5Hz で高振幅なデルタ波が発生する。[4]

| 睡眠ステージ |  | 特徴                                |
|--------|--|-----------------------------------|
| Wake   |  | $\alpha$ 波 (8-13Hz), まばたき, 急速眼球運動 |
| REM    |  | 急速眼球運動, のこぎり波, 低顎筋電位              |
| N1     |  | 低速眼球運動, 頭蓋頂鋭波                     |
| N2     |  | K-complex波, 紡錘波                   |
| N3     |  | 低周波数波の活発化 (0. 5-20Hz)             |

図 1 睡眠ステージの判定基準となる生体信号の特徴例

一般にこのラベリングはエポックと呼ばれる通常 30s 幅の時間帯の生体信号に対し 1 ステージを付与する形式で行われる。ラベリング作業は睡眠検査技師による波形の目視を必要とし、被験者一人の 1 夜分の記録は約 8 時間、すなわち約 1000 エポックほどの長さにもわたり複数夜分の記録に対し睡眠ステージ判定を行うことはかなりの労力と時間を要することがわかる。

このような課題に対して、睡眠ステージ判定の自動化についての研究が多くなされている。

3 関連研究

3.1 ニューラルネットワークを用いた睡眠ステージ判定手法

現在に至るまでさまざまなヒトを対象とした睡眠ステージ判定モデルが提案されてきたが、その中でもニューラルネットワークを用いたモデルが高い精度を記録しており、注目が集まっている。Tsinaliset ら [8] により 2 層の畳み込み層を用いた CNN モデルが発表されている。しかし、1ch 分の EEG のみを入力とする簡素なモデルであるため正答率 0.71-0.76 にとどまっている。Supratak ら [9] により提案された DeepSleepNet は CNN と Bi-directional LSTM を組み合わせたモデルを用い、前後情報を考慮した推測を可能とした。結果として正答率は 0.81-0.86 を記録しているが、ステージの判定に 25 エポック分の情報が必要となる。Phan ら [10] により提案された SeqSleepNet についても正答率 0.87 を記録しており、高精度であるが、10-30 エポックの情報を必要とする。ヒト (専門技師) によるステージ判定の正答率としては、複数の専門技師間での判定結果の一致率である検者間信頼性が目安となる指標であり、その値は約 80% [5] ほどであるとされている。これらのことより、上記の研究はヒトと同等位以上の判定精度を達成できているといえるが、実際には被験者の年齢や睡眠障害の有無等により判定精度がバラついたり、深層学習であるため判定の根拠がブラックボックス化されているなどの課題は未だ残っている。

3.2 Sleep-CAM

Sleep-CAM モデル [13] とは、本研究グループ内にて開発された深層学習を用いたステージ判定モデルであり、判定根拠のブラックボックス化という課題に対し、その CAM 層 (Class Activation Map) を用いて判定根拠を提示することができるモデルである。そのアーキテクチャについて図 2 に示す。脳波信号 (EEG), 眼電図 (EOG), 顎筋電位 (EMG) に対しそれぞれ 2 層の畳み込み層を用いて、各信号に現れるステージを判定するための特徴量を抽出する (Feature Extraction Module)。得られた特徴量に対し、5 つのステージに対する確信度を出力すると共に CAM を用いて、30s エポック内の各時点での 5 ステージ確信度も出力する (Assignment Module)。CAM の出力を見ることで 30s エポック内のどの時点が判定に対し強い影響を及ぼしているか確認することができる。

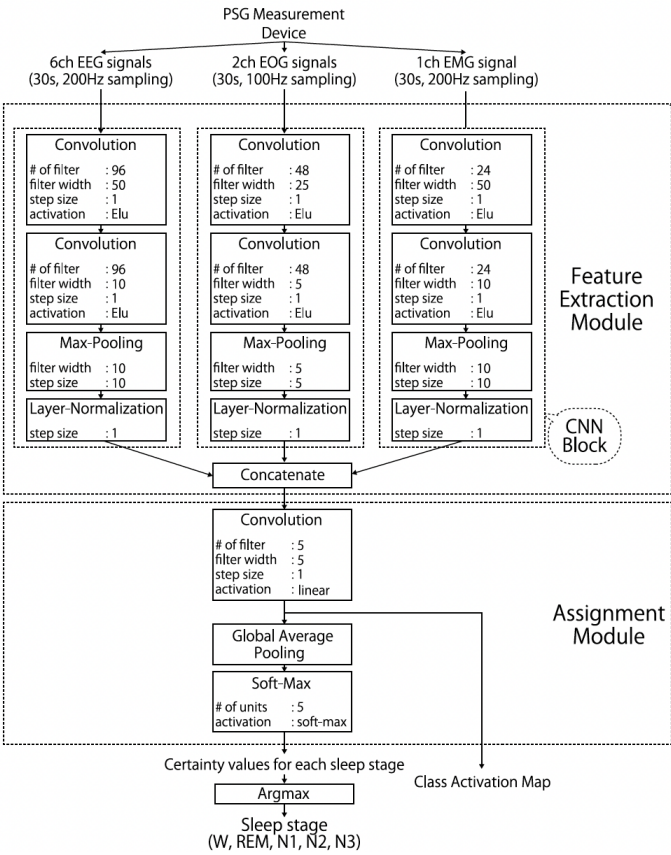


図 2 Sleep-CAM のモデルアーキテクチャ [13] より

3.3 リアルタイムなステージ判定手法

ステージ判定の自動化に関する研究は数多くなされてきたが、リアルタイム性を考慮したステージ判定に関する研究はまだ少ない。Koushik ら [11] は信号計測と睡眠ステージ判定処理それぞれの機能をプロセスレベルで分離し、両者間のデータ転送に CFS 形式のファイルを使用したアーキテクチャを提案しており、およそ 5ms 毎に 1 エポック分のステージ判定を下すことができる。しかし、ある時点でのステージ判定には直後 5 エポック分のステージ判定結果を反映させるモデルであるため、現実的にはリアルタイムにステージ判定を行うことはでき

ない. Patanaik [12] は 1ch の EEG を用いたリアルタイムなステージ判定システムをスマートフォン上で構築している. 1ch の EEG のみであるが正答率は 0.83 と高い精度を記録している. しかし, 多チャンネルの信号に対応できないという問題点があり, 頑健性に疑問が残る.

4 提案手法

4.1 リアルタイム判定システムの概要

リアルタイム性を考慮した自動睡眠ステージ判定システムを開発する上で, ”PSG 計測機器からデジタル信号を受け取り, ステージ判定をフィードバックするシステムの設計”と”機械学習モデルを用いてリアルタイムにステージ判定を行う判定する仕組み”の 2 つを考慮する必要がある.

4.2 PSG 計測機器と判定システムとのリアルタイム通信

一般的に PSG 検査は被験者の各部に電極を取り付け, 得られた電気アナログ信号をデジタル変換し, USB 等の有線を伝ってユーザ PC へと送信される. ユーザ PC では計測器専用のソフトウェアが付属されており, そのソフトウェアを通して, 生体信号のモニタリングやディスク保存などを行えるようになっている. こうした環境に対して, ステージ判定システムを組み込むことを考えたとき PSG 計測機器から直接信号を受け取ることは計測器の組み込みプログラムを修正する必要があり, 計測器の種類ごとにその操作を行うことは現実的ではない. また, SoC(separation of concerns)に基づき, リアルタイムな生体信号を送信する機能とステージ判定を行う機能を分離させ, 様々な計測機器に対応しやすいシステムを目指した. 以上を踏まえ, ユーザ PC 上のソフトウェアをサーバとし, ユーザ PC 上またはインターネット上の別ノード(クライアント)に対し随時計測した生体信号を送信するようなサーバクライアントシステムを採用した.

ユーザ PC 上のソフトウェアでは PSG 計測器から送信され, USB や WiFi 等を通して受け取った生体信号をクライアントプログラムへと TCP/IP を用いて送信する. クライアントプログラムではサーバから送られてくるパケットを遅延なく受け取り, リアルタイムにステージ判定をおこない, その結果をサーバプログラムまたは任意のノードへ送信する.

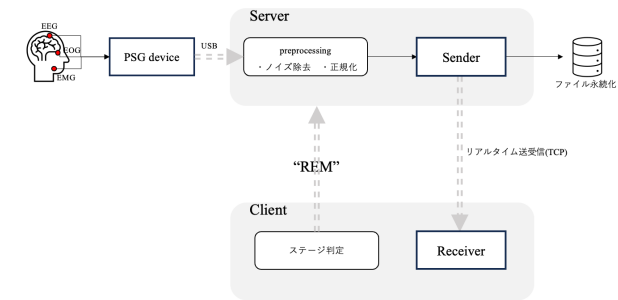


図 3 PSG 計測器からステージ判定結果提示までのアーキテクチャ図

4.3 リアルタイム判定を行う仕組み

本研究ではこれまで開発されてきた睡眠ステージ自動判定モデルを活用したリアルタイム判定手法を目指した. Sleep-CAM では入力サイズは 1 エポックであり, 出力は入力エポックに対する 5 ステージの確信度となる. このモデルをリアルタイムに拡張するため, エポック長の移動窓を用意する. この移動窓には随時最新の生体信号が追加され, 直近 30s までの生体信号を保存しておく. その移動窓に対し, あらかじめ設定した判定間隔ごとにステージ判定をおこなう. 以下に判定間隔を slide[s] と設定した場合のアルゴリズムを図 4.3 に示す. サーバから計測信号を受け取り, 判定に必要な 1 エポック (30s) まで受け取ると判定結果をサーバ等へ送信する. それ以後もパケットを受け取り続け, スライド幅 (slide[s]) ごとにステージ判定を行うことを繰り返す.

Algorithm 1 Mechanism of realtime sleep scoring in client program

```

S = slide[s], t = 0.0[s]
socket.connect()
data = []
while socket.isalive() == True do
    r = socket.receive(length = 1s)
    data = data.append(r)
    t = t + 1.0
    if t%S == 0 & data.length == 30.0s then
        stage = model.predict(data)
        socket.send(stage)
        data = data[t - 30 + S :]
    end if
end while

```

図 4 スライド窓を用いた判定手法

4.4 リアルタイム判定モデル

4.4.1 Sleep-CAM に基づくモデル

4.3 節で述べた通り, 今回設計した判定システムでは既存の自動判定モデルを組み込むことができる. Sleep-CAM モデル [13] は 1 エポックの入力に対し, 当該エポックのステージ判定を行うことができる. 本研究では, 従来の 30s エポックに対するステージ判定モデルを直接的にリアルタイム判定に適用してものとして, Sleep-CAM モデルを採用した自動判定システムを開発し, そのリアルタイム性を検証した.

4.4.2 Realtime-Sleep-CAM モデル

上に述べた Sleep-CAM モデルは, 従来の 30s のエポックを単位としてラベル付けされた正解データを学習データとして用いている. したがって, このデータでは, エポックの途中で睡眠状態の変化が起きているような場合に, その変化が正確には記述できていない. そこで, 1 秒単位でステージラベルが付与された正解ラベルを用意し, このラベルを用いて Sleep-CAM モデルを学習させた Realtime-Sleep-CAM モデルを作成した.

ただし、1 秒単位の正解ラベルの作成には多大な時間を要するためステージラベルは”Wake または N1”，”REM”，”N2 または N3”の 3 ステージとした。さらに検証のため、モデルへの入力サイズを 5s, 15s, 30s とした場合についてそれぞれ学習を行い、実験をおこなった。30s エポックごとにしか存在しない従来の正解エポックでは 30s より短い区間を入力とする判定モデルを作成することは難しかったが、1s 刻みの正解ラベルを用意したことで入力サイズを短くすることが可能となった。

5 実 験

5.1 実験の概要と目的

提案手法の評価のため、実際の PSG 検査を想定したリアルタイム判定を行う。この実験では、リアルタイム性の評価と判定精度の評価を目的としている。

実験 5.3.1 では、一定間隔でサーバより送信される生体信号に対し遅延なく判定が行えているかを検証する目的で、実際の PSG 検査を想定したリアルタイム測定実験を行う。

一方、判定モデルの精度検証については 30s エポック単位での判定精度と移動窓を用いたリアルタイムでのステージ判定精度の両方を考慮する必要がある。実験 5.3.2 では、30s エポックに対する Sleep-CAM モデルおよび Realtime-Sleep-CAM モデルの判定精度を検証する。この指標を用いて、作成したモデルの精度を関連研究などと比較することができる。本研究の主題であるリアルタイム判定 (30s より短い移動幅) での精度の検証については、実験 5.3.4 で行う。実験 5.3.3 では、Realtime-Sleep-CAM モデルについて、ステージ遷移の有無を区別したエポックについての判定精度、及びその影響について検証する。実験 5.3.4 では、実際に Sleep-CAM モデルと Realtime-Sleep-CAM モデルを用いてリアルタイム判定を行い、今回新たに用意した 1s 刻みの正解ラベルを用いてその判定精度を検証し、リアルタイム判定において Realtime-Sleep-CAM モデルに優位性があるかを検討する。

5.2 データセット

今回、モデルの学習及び精度検証では PSG1100(日本光電) を用いて測定されたオリジナルデータセットを用いる。計測は 20 ～30 歳の健常者を対象に行われ、サンプル数は 49 夜分 (53630 エポック) となる。取得したデータのうち、脳波信号 (EEG) を 6ch(F3-M2, F4-M1, C3-M2, C4-M1, O2-M1, O1-M2)、眼球運動信号 (EOG) を 2ch(E1-M2, E2-M2)、顎筋電信号 (EMG) を 1ch(EMG1-Z) をモデルへの入力として使用する。

| SleepCAM   |                     |
|------------|---------------------|
| # of Night | 49                  |
| # of epoch | 53630               |
| Stage      | W, N1, N2, N3 REM   |
| channel    | EEGx6, EOGx2, EMGx1 |

図 5 Sleep-CAM モデルの学習に用いるデータセットの概要

|            |                    | INPUT = 5s                                | INPUT = 15s | INPUT = 30s |
|------------|--------------------|-------------------------------------------|-------------|-------------|
| # of epoch | # of Night         | 5                                         |             |             |
|            | Overall            | 28749(original=511)                       |             |             |
|            | without transition | 27938                                     | 26206       | 24555       |
| Stage      | with transition    | 811                                       | 2543        | 4194        |
|            | channel            | WorN1, N2orN3, REM<br>EEGx6, EOGx2, EMGx1 |             |             |

図 7 Realtime-Sleep-CAM モデルの学習に用いるデータセットの概要

さらにこれら信号データセットに対し、検者間信頼性 92.4% の専門技師による AASM ルールに基づいた全エポックに対するラベル付が行われた。Sleep-CAM モデルの学習ではこのエポック単位のラベルを用いて学習を行う。

一方、49 夜分のうち 5 夜分について専門技師による 1 秒単位のラベルデータを新たに用意し、これらを用いて Realtime-Sleep-CAM モデルの学習をおこなう。1s 単位のラベルデータから学習およびテストデータを作成するために Realtime-Sleep-CAM モデルの入力サイズに合わせたエポックへと切り出しを行う必要がある。1 秒単位のラベルデータから入力サイズに合わせた正解ラベルを作成する手順を図 6 に示す。

エポックフリーなラベルから一定間隔で予測モデルの入力サイズ長のエポックをサンプリングするとき、ステージ遷移にまたがるエポックについては 2 つ以上のステージラベルが混在する状態となる。これについては、エポックに占める割合が大きいステージラベルを正解ラベルとしてサンプリングした。

さらに、5 夜分のラベルデータでは学習が不十分だと考えられるため、位置を 5 秒、10 秒、15 秒とずらした正解ラベルを用意し、データ拡張を行った。

また、Realtime-Sleep-CAM への入力エポックのサイズについては 5s, 15s, 30s の場合についてそれぞれ学習をおこなった。それに伴い、設定した入力サイズ、5s, 15s, 30s ごとに学習、テストデータを上記の手順で作成した。

5.3 実験結果

5.3.1 リアルタイム性の評価

リアルタイム性の評価では Sleep-CAM モデルを使用し、サーバからリアルタイムで送信されてくる信号データに対し、ステージ判定を遅延なく行えているか検証する。しかし、実際に被験者を用意し実験することは難しいため、あらかじめ計測済みの信号データを 1 夜分用意し、リアルタイム計測と同様な速度でクライアントプログラムへ信号データを送信するようにして実験をおこなった。また、実験条件として判定間隔であるスライド幅を 1s, 5s, 30s の場合でそれぞれ実験し計測を行った。

クライアントプログラムではスライド幅分の信号データをパケットで受け取り、ステージ予測を行い予測結果を送信することを繰り返す。そこで、パケット受け取りにかかる時間を TCP Time、ステージ予測と予測結果の送信にかかる時間を Scoring Time、TCP Time と Scoring Time を合わせたステージの判定間隔に相当する時間を Iteration Time としてそれぞれ計測した。実験に用いた 1 夜分のデータ長は約 9 時間 6 分にわたり、30s 幅では 1093 回、15s 幅では 2186 回 1s 幅では 6558 回の TCP Time 及び Scoring Time の計測が行われ、それらの平

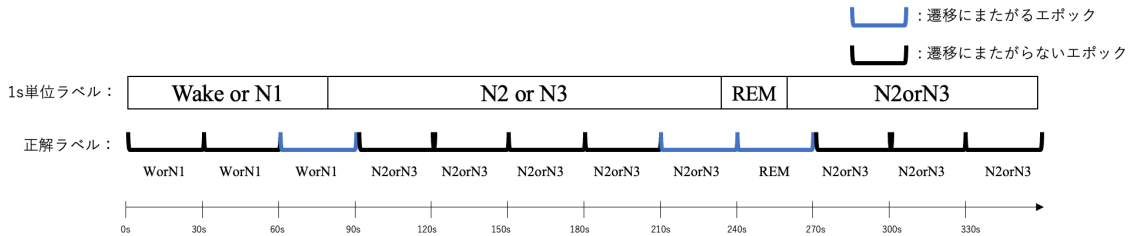


図 6 入力サイズ=30s の場合の、1 秒単位のラベルデータを元にした正解データの作成

均と標準偏差を算出した。実験に用いた OS は Ubuntu LTS 20.0.4 を使用し、実装は Python の Tensorflow-gpu-2.1.0 を使用した。また、CPU は Intel Core i3-12100, GPU は Nvidia GeForce GTX 1660 SUPER を使用した。

実験結果を図 8 に示す。結果を見てみるとまずどのスライド幅においても、Iteration Time は判定間隔がそれぞれ 30s, 5s, 1s のとき、30.128, 4.998, 0.978 となっているため、遅延なく判定が行えていることがわかる。さらにその内訳を見てみると、ステージ判定にかかる時間 Scoring Time は比較的小さく、TCP パケットを受け取る時間 TCP Time が支配的であることがわかる。TCP Time はサーバプログラムから送られてくる、設定したスライド幅分のパケットを受け取るまでの時間に相当することから、スライド幅分に近い時間を要する。Scoring Time については、どのスライド幅においても約 0.6s の判定時間を必要としていることがわかる。この時間は判定間隔の下限であり、0.6s より短い判定間隔でのリアルタイム判定は行えないことがわかる。

これらのことから、最大で判定間隔 1s のリアルタイムなステージ判定が遅延なく行えていることがわかる。

|             | Iteration |       | TCP    |       | Scoring |       |
|-------------|-----------|-------|--------|-------|---------|-------|
|             | Mean      | Std   | Mean   | Std   | Mean    | Std   |
| SLIDE = 30s | 30.128    | 0.122 | 29.512 | 0.085 | 0.618   | 0.072 |
| SLIDE = 5s  | 4.998     | 0.058 | 4.408  | 0.043 | 0.631   | 0.042 |
| SLIDE = 1s  | 0.978     | 0.046 | 0.389  | 0.033 | 0.619   | 0.041 |

図 8 リアルタイム性の評価結果

5.3.2 モデル単体での判定精度

Sleep-CAM モデルと Realtime-Sleep-CAM モデルの学習をおこなった。Sleep-CAM モデルでは 49 夜分のレコード (53630 エポック) に対し、7 分割交差検証をおこなった。正解ステージ数は 5 つで、'Wake', 'REM', 'N1', 'N2', 'N3' である。損失関数はクロスエントロピー関数とし、最適化手法は Adam(lr=5×10<sup>-6</sup>) を用いた。Realtime-Sleep-CAM モデルではデータ拡張をおこなった 5 夜分 (28749 エポック) のレコードに対し、5 分割交差検証をおこなった。正解ステージ数は 3 つで、'Wake または N1', 'REM', 'N2 または N3' である。損失関数及び最適化手法は Sleep-CAM モデルと同様である。

入力サイズ長のエポックに対する判定精度は、より細かな粒度でのリアルタイムな判定精度を直接評価することはできないが、関連研究と比較することでモデルの性能を判断することができる。リアルタイムな判定精度については、5.3.3 節で述べる。

まず Sleep-CAM モデルについて正答率 (Accuracy) と Kappa 係数及びを示す。

| Overall |       |
|---------|-------|
| ACC     | Kappa |
| 0.83    | 0.76  |

図 9 30s エポックに対する Sleep-CAM モデルの正答率, Kappa 係数

混同行列と再現率, 適合率は図 10 の通りとなった。

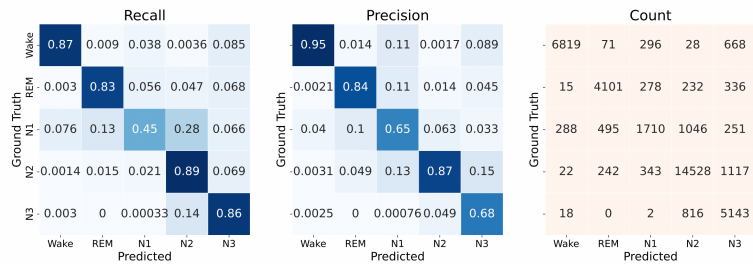


図 10 Sleep-CAM モデルの混同行列 (右) および再現率 (左), 適合率 (中)

関連研究と精度を比較したとき、1 エポック (30s) を用いた軽量のモデルとしては十分な判定精度をもっているといえる。また専門技師間の一致率が約 82% [5] であることから、専門技師と同等の正答率を持つといえる。各ステージについて、再現率及び適合率を見てみると N1 ステージの精度が低いのが、これはサンプルの中の N1 ステージの占める割合が少ないことに起因する。

次に Realtime-Sleep-CAM モデルについて正答率 (Accuracy) と Kappa 係数及びを示す。

|     | Overall |       |
|-----|---------|-------|
|     | ACC     | Kappa |
| 5s  | 0.84    | 0.66  |
| 15s | 0.85    | 0.69  |
| 30s | 0.86    | 0.72  |

図 11 各入力サイズ長のエポックに対する Realtime-Sleep-CAM モデル (入力サイズ=5s, 15s, 30s) の正答率, Kappa 係数

混同行列と再現率, 適合率は図 12, 13, 14 の通りとなった。

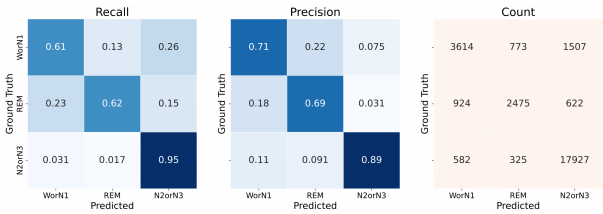


図 12 Realtime-Sleep-CAM モデル (入力サイズ=5s) の混同行列 (右) および再現率 (左), 適合率 (右)

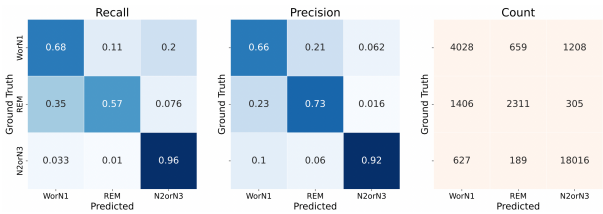


図 13 Realtime-Sleep-CAM モデル (入力サイズ=15s) の混同行列 (右) および再現率 (左), 適合率 (右)

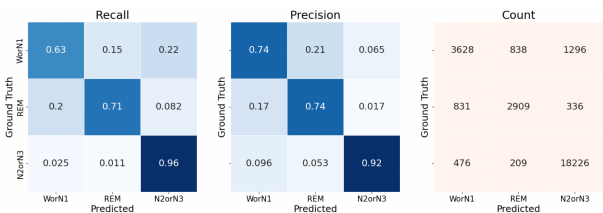


図 14 Realtime-Sleep-CAM モデル (入力サイズ=30s) の混同行列 (右) および再現率 (左), 適合率 (右)

正答率を見てみると 3 ステージ判定であるため, Sleep-CAM より数%高い判定精度となった. 図 12,13,14 の混同行列よりステージ学習に用いた正解ラベルのステージの分布が偏っており, 特に”N2orN3”ステージのラベル数が多いため, 結果として”N2orN3”ステージについて高い判定精度をもつモデルへと学習が行われたと考えられる. また, 図 11 の 30s ごとのエポック単位での判定精度を見てみると, 入力サイズが大きいほどモデルの精度は高くなった.

5.3.3 ステージ遷移の有無に対する Realtime-Sleep-CAM モデルの判定精度

5.2 節で述べた正解ラベルの作成方法では, エポック内で 1 つのステージが継続するものとエポック内でステージ遷移が生じるエポックが存在する. そこで, 複数ステージにまたがらないエポックに対する判定精度を”Without transition”, 複数ステージにまたがるエポックに対する判定精度を”With transition”として, それぞれ判定精度を求めた. 結果を図 15 に示す.

|     | Overall |       | Without transition |       | With transition |       |
|-----|---------|-------|--------------------|-------|-----------------|-------|
|     | ACC     | Kappa | ACC                | Kappa | ACC             | Kappa |
| 5s  | 0.84    | 0.66  | 0.84               | 0.67  | 0.56            | 0.26  |
| 15s | 0.85    | 0.69  | 0.87               | 0.73  | 0.59            | 0.32  |
| 30s | 0.86    | 0.72  | 0.9                | 0.78  | 0.64            | 0.42  |

図 15 ステージ遷移の有無による Realtime-Sleep-CAM モデルの判定精度

結果を見てみるとステージ遷移にまたがるエポックに対する判定精度が低いことがわかる. これについてはエポック内に 2 つ以上のステージが混在し, それに伴い Realtime-Sleep-CAM モデルの 3 ステージ確信度がバラついたため, ステージ継続エポックと比較して精度が下がることが予想される. 一方で, ステージが継続するエポックに対してはさらに高い判定精度をもつことがわかる.

5.3.4 リアルタイムな判定精度

今回提案したサーバクライアントシステムを用いて, Sleep-CAM モデル及び Realtime-Sleep-CAM のリアルタイムな判定精度を検証する. ただし, 正解ラベルの都合上”Wake or N1”, ”REM”, ”N2 or N3”についての判定精度を検証する. 判定間隔は 1s とし, relatime-Sleep-CAM モデルについては入力サイズが 5s, 15s, 30s の場合それぞれでリアルタイム判定をおこなった.

|                   | Overall |       |
|-------------------|---------|-------|
|                   | ACC     | Kappa |
| SleepCAM          | 0.85    | 0.75  |
| realtime-SleepCAM | 5s      | 0.86  |
|                   | 15s     | 0.86  |
|                   | 30s     | 0.84  |

図 16 Sleep-CAM モデル, Realtime-Sleep-CAM モデル (入力サイズ=5s, 15s, 30s) のリアルタイム判定に対する正答率, Kappa 係数

結果を見てみると, 入力サイズが 5s の場合の Realtime-Sleep-CAM モデルが最も高い正答率をもっていることがわかった. しかし, その差は 1%ほどであり, リアルタイム判定において Realtime-Sleep-CAM モデルが優位であると言い難い.

一方でステージが切り替わるタイミングに注目してみるとある一定の傾向が見られた. 図 17, 18 にその傾向が現れる睡眠経過図を示す. 図 17 では 02 : 00 : 10 に”N2 または N3”から”REM”への遷移が発生しているが, Realtime-Sleep-CAM モデルの入力サイズが 5s の場合誤差 2s, 15 の場合誤差 9s,30s の場合 24s となり, Sleep-CAM モデルでは誤差 20s となり Realtime-Sleep-CAM モデルの方が誤差が少ないといえる. さらに Realtime-Sleep-CAM モデルの入力サイズが小さいほどその遷移に対して誤差なく判定できていることがわかる. 図 18 では 01 : 52 : 36 におよそ 10s の短い”Wake または N1”ステージへの遷移が発生しているが, 入力サイズが 5s の場合のみ, その遷移を判定することができている.

【このようなステージ遷移に関する判定精度を検証するため

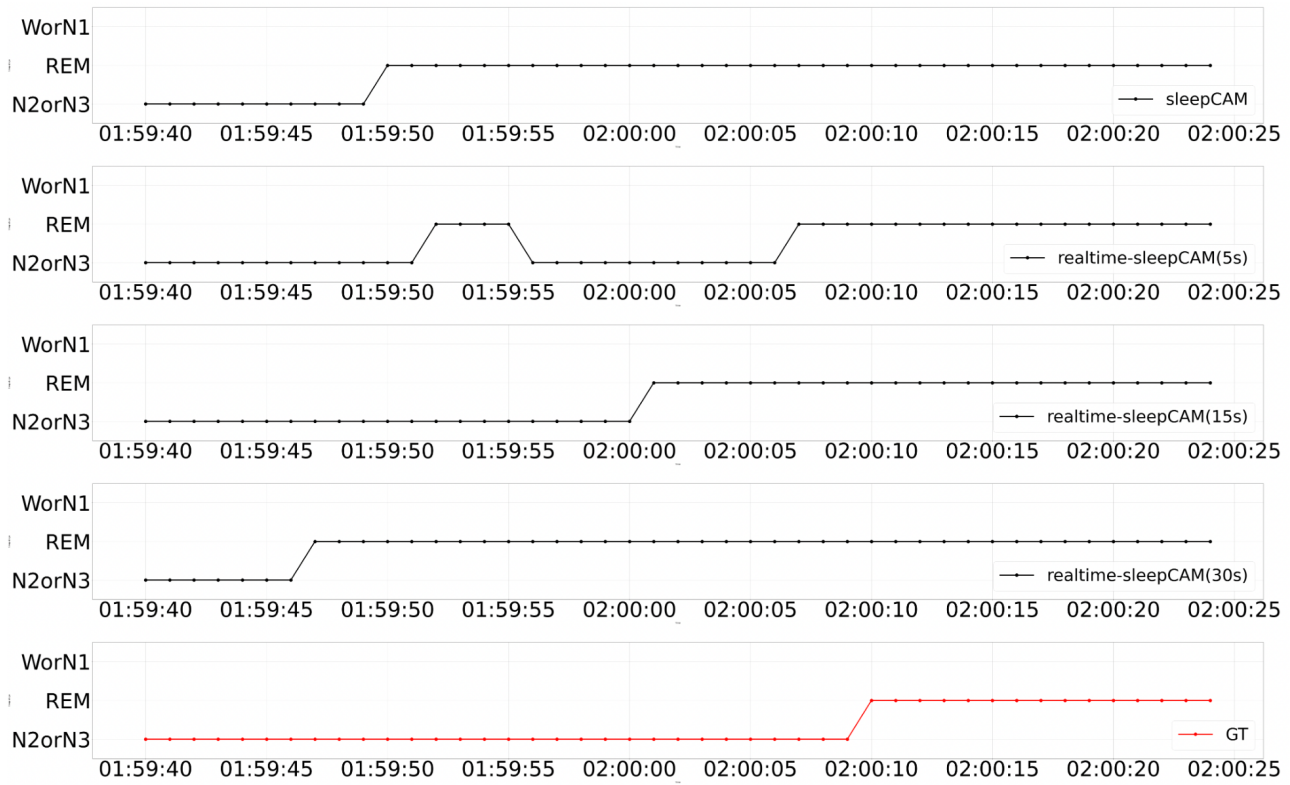


図 17 ステージ遷移 (02:00:10) に着目した正解ラベル (GT) と各入力サイズごとの判定結果

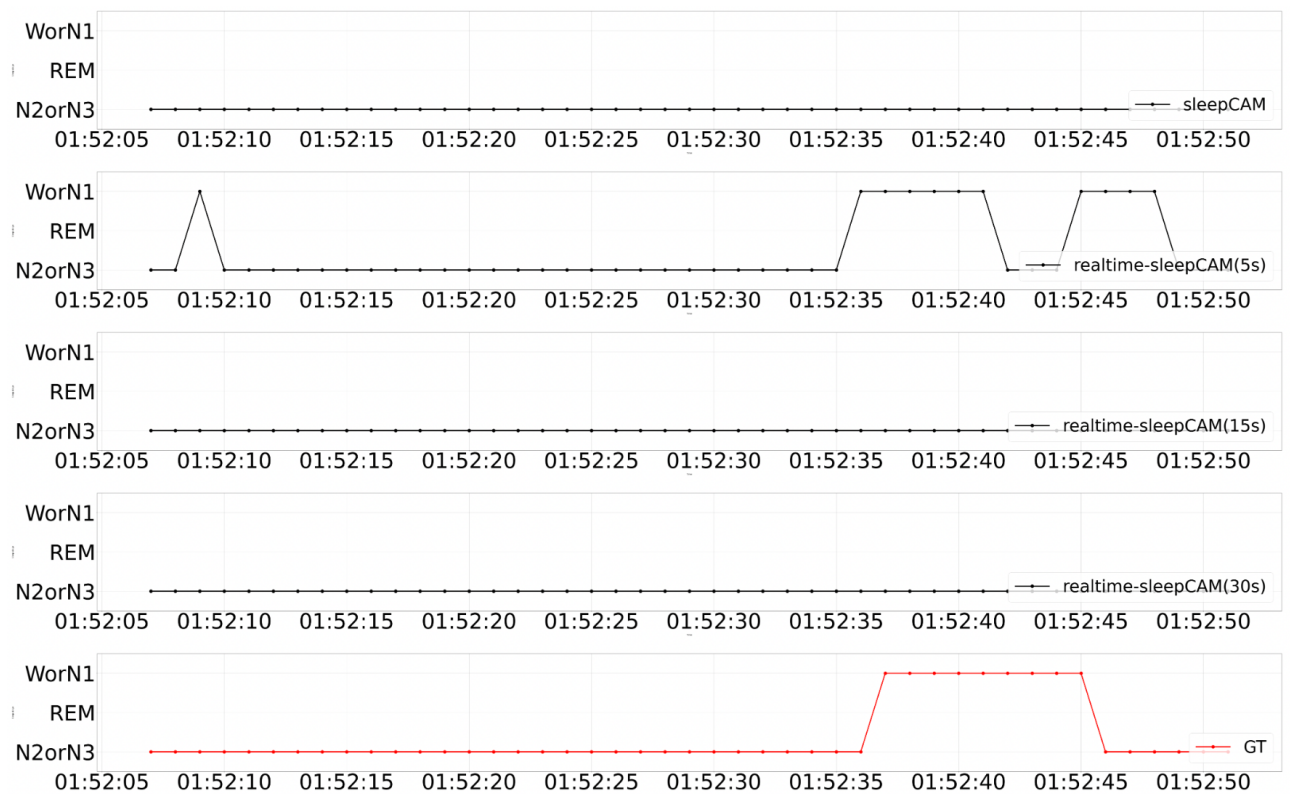


図 18 ステージ遷移 (01:52:36) に着目した正解ラベル (GT) と各入力サイズごとの判定結果

に正解ラベル (GT) 内に存在するステージ遷移の前後区間で各モデルが同様なステージ遷移判定を行えている確率を  $Recall^T$ , 各モデルの判定結果のうちステージ遷移判定の前後区間で正解ラベル (GT) にも同様なステージ遷移が見られる確率を

$Precision^T$ ,  $Precision^T$  と  $Recall^T$  の調和平均を  $F1^T$  として算出した。また前後区間の幅としては, 10s, 30s, 60s の場合それぞれについて各指標を求めた。結果を図 19 にしめす。

結果を見てみると, 区間幅 10s で見たとき  $Recall^T$ ,

|        |     | Sleep-CAM           |                        |                 | Realtime-Sleep-CAM(5s) |                        |                 | Realtime-Sleep-CAM(15s) |                        |                 | Realtime-Sleep-CAM(30s) |                        |                 |
|--------|-----|---------------------|------------------------|-----------------|------------------------|------------------------|-----------------|-------------------------|------------------------|-----------------|-------------------------|------------------------|-----------------|
|        |     | Recall <sup>T</sup> | Precision <sup>T</sup> | F1 <sup>T</sup> | Recall <sup>T</sup>    | Precision <sup>T</sup> | F1 <sup>T</sup> | Recall <sup>T</sup>     | Precision <sup>T</sup> | F1 <sup>T</sup> | Recall <sup>T</sup>     | Precision <sup>T</sup> | F1 <sup>T</sup> |
| window | 10s | 0.069               | 0.064                  | 0.066           | 0.399                  | 0.076                  | 0.128           | 0.168                   | 0.071                  | 0.100           | 0.079                   | 0.053                  | 0.063           |
|        | 30s | 0.243               | 0.222                  | 0.232           | 0.658                  | 0.159                  | 0.256           | 0.470                   | 0.214                  | 0.294           | 0.296                   | 0.201                  | 0.239           |
|        | 60s | 0.491               | 0.430                  | 0.458           | 0.783                  | 0.243                  | 0.371           | 0.644                   | 0.316                  | 0.424           | 0.515                   | 0.363                  | 0.426           |

図 19 各区間でのステージ遷移に関わる判定精度. 各区間での最良スコアを色付 (Recall: 緑, Precision: 青, F1: 赤)

$Precision^T$ ,  $F1^T$  共に入力サイズが 5s の Realtime-Sleep-CAM モデルが最良となった. しかし区間幅を広げていくにつれ,  $Precision^T$ ,  $F1^T$  について, 入力サイズが 15s の Realtime-Sleep-CAM(15s) や Sleep-CAM モデルなどが優れていることがわかる.  $Precision^T$  について Realtime-Sleep-CAM(5s) モデルが比較的低い値である理由として, 1s ごとのステージ判定結果が振れやすくなってしまうことが挙げられる. 図 20 に示す通り, 入力サイズが小さいモデルほどステージ遷移の判定回数が多いことがわかる.

| Sleep-CAM | Realtime-Sleep-CAM(5s) | Realtime-Sleep-CAM(15s) | Realtime-Sleep-CAM(30s) |
|-----------|------------------------|-------------------------|-------------------------|
| 1257      | 6019                   | 2689                    | 1752                    |

図 20 各モデルのリアルタイム判定結果ステージ遷移判定の回数

これらのことから, 入力サイズの大きい Sleep-CAM モデルや Realtime-SleepCAM(30s) モデルについては, 前後 30s の 60s という広い区間でみると判定精度が高い一方, 前後 5s の 10s という狭い区間でみると判定精度が低いことからリアルタイムな判定において 30s 以内の遅延が発生する可能性はあるが, 最もステージ遷移を捉えられているといえる. 一方, 入力サイズの小さい Realtime-SleepCAM(5s), Realtime-SleepCAM(15s) モデルについては, 短い区間 (10s, 30s) においては最も高い判定精度を持っていることから, ステージ判定結果が頻繁に振れてしまうため, Sleep-CAM や Realtime-SleepCAM(30s) よりステージ遷移の判定精度は低いが, その遅延は 5s, または 15s 以内に収まる可能性が高いことを示唆している. したがって, ステージ遷移の誤検出が少ないかつ, ステージ遷移判定の遅延が短いモデルを今後開発していくことが今後の課題として挙げられる.

## 6 ま と め

本研究ではヒトを対象としたリアルタイムなステージ判定システムを提案した. PSG 計測器側とステージ判定側をサーバ, クライアントとして分離し, 両者間で TCP/IP による生体信号の送受信を行うようなシステムを設計した. 一方で本システムの判定の仕組みとしては, 機械学習に基づく自動判定モデルを組み込み, 移動窓による判定手法を用いてリアルタイム判定を可能とした. このリアルタイム判定について処理時間を実験により確認したところ, 最小 1s でのリアルタイム判定が可能だと確認できた.

さらに新たに用意した 1s 刻みでのステージラベルデータを用いて, 既存のステージ判定モデル (Sleep-CAM モデル) を学習

させ, Realtime-Sleep-CAM を作成した. 1s 刻みのラベルデータを用いることで従来の 30s エポック単位でのラベルデータでは検証できなかった, リアルタイムなステージ判定に対する精度を算出することができた. 実験を通して, Realtime-Sleep-CAM モデルは 30s エポックに対しても Sleep-CAM モデルと同等の判定精度を確認でき, リアルタイムな判定精度では入力サイズを短くすることで, 最も高い判定精度を達成できることを確認できた.

今後の課題として, リアルタイム判定精度の改善が挙げられる. 本研究では, 5 夜分の 1s 刻みのラベルデータを作成し, リアルタイム判定モデルを学習させたが, そのラベルの事前分布が偏っていたため, ある特定のステージを優先して判定してしまう学習となってしまった. これについては, ラベルの分布を均一にするようなデータ拡張が必要だと考えられる. また, 汎用的な学習を目指すためにはより多くのラベルデータを作成することも重要である.

## 7 謝 辞

本研究の一部は, JSPS 科研費 JP23H03399, JP22K19802, AMED Grant Number JP21zf0127005, F-MIRAI 共同研究, (株) S'UIMIN 共同研究による.

## 文 献

- [1] Ferrie JE, et al, Sleep epidemiology - a rapidly growing field. Int J Epidemiol, 40(6) 1431-1437, 2011.
- [2] Andrew M. Gleason, A curvature formula, Amer. J. Math. 101 (1979) 86-93.
- [3] Roth Thomas, Characteristics and determinants of normal sleep, The Journal of clinical psychiatry vol. 65 Suppl 16 pp.8-11, 2004.
- [4] Iber, C., Ancoli-Israel, S., Chesson, A. L., and Quan, S., "The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, 2007.
- [5] Danker-Hopfe, H. et al, Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM, J. Sleep Res. 18(1) 74-84, 2009.
- [6] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, Going deeper with convolutions, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1-9, 2015.
- [7] Matthew D Zeiler and Rob Fergus, Visualizing and understanding convolutional networks, In European conference on computer vision, Springer, pp.1097-1105, 2014.
- [8] Orestis Tsinalis, Paul M. Matthews, Yike Guo, Stefanos Zafeiriou, Automatic Sleep Stage Scoring With Single-Channel EEG Using Convolutional Neural Networks, 2016

- [9] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo, Deep-sleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg, IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 25, No.11, pp. 1998-2008, 2017.
- [10] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chen, and Maarten De Vos. Seqsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging, IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 27, No. 3, pp.400-410, 2019.
- [11] Abhay Koushik, Judith Amores, Pattie Maes, Real-Time Sleep Staging using Deep Learning on a Smartphone for a Wearable EEG, IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN), 2019.
- [12] Amiya Patanaik, Ju Lynn Ong, Joshua J Gooley, Sonia Ancoli-Israel and Michael W L Chee, An end-to-end framework for real-time automatic sleep stage classification, Sleep, Vol. 41, No. 5, 2018.
- [13] Kazumasa Horie, Leo Ota, Ryusuke Miyamoto, Takashi Abe, Yoko Suzuki, Fusae Kawana, Toshio Kokubo, Masashi Yanagisawa and Hiroyuki Kitagawa, Automated sleep stage scoring employing a reasoning mechanism and evaluation of its explainability, Scientific Reports 12, No. 12799, 2022.

## 特定健診データに含まれる自由記述に着目した予後予測

大塚 皇輝<sup>†</sup> 池之上辰義<sup>††</sup> 福間 真悟<sup>†††</sup> 若宮 翔子<sup>†</sup> 荒牧 英治<sup>†</sup>

<sup>†</sup> 奈良先端科学技術大学院大学 先端科学技術研究科情報領域 〒630-0192 奈良県生駒市高山町 8916 番地の 5

<sup>††</sup> 滋賀大学 データサイエンス AI 推進センター 〒522-8522 滋賀県彦根市馬場 1 丁目 1-1

<sup>†††</sup> 京都大学 医学研究科 人間健康科学科系専攻 〒606-8507 京都府京都市左京区聖護院川原町 53

E-mail: <sup>†</sup>{otsuka.koki.om2,wakamiya,aramaki}@is.naist.jp, <sup>††</sup>saikawa@n-univ.ac.jp,

<sup>†††</sup>fukuma.shingo.3m@kyoto-u.ac.jp

**あらまし** 日本は少子高齢化の影響によって医療財源の逼迫が問題となっており、予防医療や実診療での適切な医療資源の配分が求められている。このような背景から、近年では機械学習を用いて患者の予後を予測することで、医療分野における意思決定を補助するような試みも登場してきた。一方で、機械学習はその予測過程がブラックボックスであり説明可能性という観点から医療分野での導入の障壁となっている。そこで本研究では自然言語の持つ説明可能性の高さに着目し、特定健診データから言語モデルを用いて予後の予測を試みた。特に、減量の予後予測に対する言語モデルの予測精度と説明可能性について検証を行った。

**キーワード** 自然言語処理, 医療情報学, 予後予測, 解釈可能性, 特定健診

### 1 はじめに

日本は少子高齢化や人口減、労働人口現象に伴う財政難など様々な問題を抱えている。これらに起因して日本の税収に占める社会保険料の増加が問題となっており、医療の人的・物的資源の効率的な利用が課題となっている。特に、日本では悪性新生物（がん）や心疾患といった一般的には生活習慣に起因する疾患が死因の上位を占めている [1]。これに対して、日本では 2008 年より生活習慣病の予防を目的として 40 歳から 75 歳までを対象に特定保険審査、及び生活習慣病リスクの高い患者に対しては特定保健指導を実施している。しかし、指導を行っても生活習慣が改善されない場合も多く、効果を疑問視する声もある。例えば運動や禁酒を薦められ、しばらく努力しても諦めてしまうといった事例も多い。どのような患者にどのように指導を行うかは、保健師・栄養士などの指導員の経験によることが多く、指導の質の向上と均質化が求められる。

このような中、近年急速に発達している機械学習技術に注目し、この技術を医療分野に応用するような研究も進んできた。特に、患者予後を予測することができれば、予防医療や実診療での医療資源の最適な利用に繋げることが可能となる。しかし、機械学習によって導き出された予測結果は、解釈性に乏しいという問題がある。基本的に機械学習モデルの予測過程はブラックボックスとなっており、人間が解釈を行うことが難しい。医療分野では治療をする上で解釈性の乏しい機械学習モデルの結果を信頼することは難しく、このことが医療分野における機械学習を用いた予測モデルの導入の障壁となっている [2]。

そこで今回は自然言語を有効に利用することで、医療現場における説明可能性という障壁を取り除けないかと考えた。医療データには様々な種類のデータが含まれており、検査値のよう

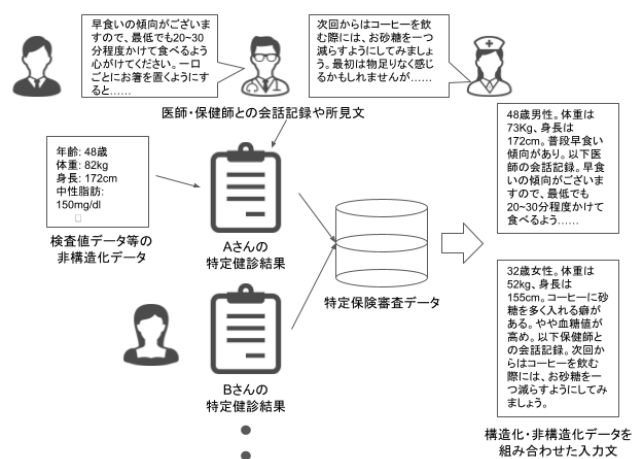


図1 本研究の概要: 特定健診データに含まれる様々な種類のデータを言語モデルが解釈できるように自然言語に変換し、予後を予測する。

な数値データに加え、医師の診断所見といった自由記述が多く含まれている。しかし、従来ではこれらの自然言語を機械学習モデルで扱うことが難しく、医学研究の材料としてあまり用いられてこなかった。数値データ単体を人間が解釈するためには十分な経験が必要とされるが、自然言語は比較的人間には解釈しやすいため、この特性を有効活用することで医療分野における機械学習の導入の一助となる。

このような背景から、本研究では自然言語の持つ解釈性に注目した予後の予測、及びその根拠となる情報を提示することにより、解釈可能性の検証を行った。予後予測の研究では検査値等の数値データや、カテゴリーカルな検査データを入力として用いることが一般的となっている。しかし、このようなデータを入力に用いた場合、どの項目が予測結果に寄与したのかを判定

することしか出来ず、その数値に対する解釈はそれを見た医師の経験に左右される場合がある。一方で、自然言語の場合数値データよりもコンテキストが含まれているため、解釈を行う上では有効に作用する。加えて、電子カルテに代表される診断記録には医師の自由記述による所見も多く含まれているものの、この部分の利活用はあまり進んでいない。

本研究では、特定保険審査データと言語モデルを用いて、予後が良好か不良かを当てる分類問題を設定し、予測性能の検証をした。本研究で定義する予後とは、次回健診時までの腹囲と体重の差分から判定をする。具体的には、予後良好は腹囲が2cm以上減少し、同時に体重が2kg以上減少している場合とし、予後不良はそれ以外の状態としている。加えて、分類結果の判断根拠となる自然言語部分を可視化することで、診断の判断根拠として使用できるかどうか検証を行った。

## 2 関連研究

### 2.1 健診データを用いた研究

健康診断は1年に1回定期的に実施されることから時系列的な変化を追いやすい。そのため、このデータを軸にして様々な他媒体のデータと組み合わせた研究が盛んに行われている。

恒川ら[3]は生活習慣病の発症を予測するにあたり、健診データとレセプトを利用した。レセプトとは、医療機関が医療保険者に対して行った検査や診断、処方した薬剤等の情報が記された診療報酬明細書のことである。彼らの研究では、生活習慣病の発症を予測するにあたり、1年に1回しかない健診データだけを用いた場合、生活習慣病になった時期の詳細が不明であるためレセプトを同時に使用した。作成されたデータには数値データや医師の自由記述など様々な形式のデータが含まれることになる。これらを解析した結果、生活習慣病の中でも糖尿病が識別しやすい病気であることが発見され、HbA1cや糖代謝判定といった項目が予測に大きく寄与していることがわかった。一方で、この研究では心電図の結果等に関する医師の自由記述項目がデータに含まれていたが、自然言語の処理を組み込むことができず今後の課題として紹介されている。

大場ら[4]も同様に健康診断データを利用して、生活習慣病の中でも特に糖尿病を予測するという研究を行った。彼らの研究では、予測精度だけではなく、医療分野での使用を想定し解釈可能性といった点にも注目している。健診データに加えて質問票への回答から予測に寄与するデータを明らかにしている。これを調査するにあたり、構築したモデルの入力として使われている属性をランダムに置き換えて変化を計測するPermutation importance[5]と、摂動を与えてその前後の予測結果の違いを分析するSensitivity analysisといった手法を利用している。前者は変化が大きかった場合、その属性が予測結果に与える影響が大きかったと判定することができるが、その解釈を行うことはできない。一方で、後者は摂動を与えた際の前後の変化を観察することで、どの程度の変化がどのような予測結果に結びいているか解釈することが可能となる。

### 2.2 マルチモーダルなデータ活用

健診データのように一度に複数項目の検査を実施すると、検査値データや心電図等の時系列データ、医師の所見などの自然言語等様々な種類のデータが作成される。それぞれ検査目的が異なるため、どの項目を用いて予測を行うべきかは医師などのドメインエキスパートによって選ばれる。そこで、例えば、モデルが数値データだけではなく、医師が診断時に感じ取った患者の状態等を記した自由記述データを同時に処理することができれば、より高精度な予測が可能になる。こういったことから、Agarwalら[6]はマルチモーダルなデータを一つのモデルで処理することで、COVID-19の患者の予後を予測する研究を行った。具体的には、一定期間後にまだ入院しているかどうか、人工呼吸器を装着することになるかどうかの2点を予測項目とした。これらの項目を予測する際に、性別や身長といったデータから飲酒喫煙の状態、検査値等の連続データを一つのベクトルとして埋め込むことでモデルの入力として利用している。

## 3 データセット

### 3.1 特定保険審査・特定保健指導データ

日本では生活習慣の変化により、高血圧症や糖尿病等のいわゆる生活習慣病の有病者が増加している。日本人の死因上位における悪性新生物や心疾患の原因には生活習慣病があるとされ、これらの対策が急務となっている。このようなことから日本は平成20年4月より生活習慣病の予防を目的として、年に一度40歳から75歳までを対象に内臓脂肪肥満に着目した特定保険審査を実施している。これには身体計測や血液検査に加え、質問票を用いた問診も含まれる。この検査で生活習慣病の発症リスクが高く、生活習慣の改善で予防ができると判断された場合、医師や保健師による特定保健指導により、生活習慣の改善に取り組むことになる。今回は、全国土木建築国民健康保険組合加入者に対して、2013年3月から2017年3月の間に行われた特定保健師指導の電子的報告記録から抽出した28954件を、データセットとして用いた。

このデータには「年齢」や「身長」といった54種類の構造化データと、「既往歴」や「生活習慣病の改善状況」といった11種類の非構造化データが含まれている。表1に含まれている項目の一部を示す。

### 3.2 予測ラベルの作成

3.1節で示した通り、特定保険審査は生活習慣病の予防を目的とし、内臓脂肪型肥満に着目した健康診断となっている。そのため、今回は内臓脂肪の数値に関係すると考えられる腹囲と体重に注目した。なお、これら2つの項目については、次回健診時の数値も入力されていた。そのため、これらの数値を利用し、次回健診から腹囲が2cm以上減少しかつ、体重が2kg以上減少していた場合予後が良好であると判定することとした。

### 3.3 不均衡データの扱い

表2に示す通り予測対象となる予後が良好となっているデー

表 1 特定健診データに含まれる項目の一部。下線は入力文章の作成に用いた項目を表す。

| 種別      | 項目名                                                                                                                                                                                                                                    |
|---------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 構造化データ  | 患者 ID, 喫煙有無, 年齢, 性別, 生年月日, 歩行速度, 貧血, 飲酒頻度, 一年間の体重変化<br>30 分以上の運動習慣, 腹囲, BMI, 身長, 体重<br>収縮期・拡張期血圧, GOT(AST), <u>HbA1c</u> ,<br>LDL コレステロール, HDL コレステロール<br>中性脂肪, 空腹時血糖, 尿糖, 尿蛋白<br>血清クレアチニン, 次回健診時体重,<br>次回健診時腹囲, 診断回数, <u>ALT(GPT)</u> |
| 非構造化データ | メタボリックシンドローム判定, 保健指導レベル,<br>既往歴, 自覚症状, 自覚症状初見,<br>食べ方, 飲酒量, 保健指導時の会話記録,<br>飲酒習慣詳細, 睡眠習慣詳細, 生活習慣改善意思                                                                                                                                    |

タが 2698 件, 不良となっているデータが 28954 件となっており大きな偏りが生じている。このまま言語モデルを学習すると予後不良となっている多数派の影響を強く受けることでモデルの出力がほぼ予後不良となるように学習してしまう危険性がある。そのため今回はデータのダウンサンプリングを実施した。予後不良となっているデータからランダムに予後良好となっている件数と同じ 2698 件を抽出して用いることにした。これにより予後良好が 2698 件, 予後不良が 2698 件で同数となるデータセットを作成した。

## 4 手 法

### 4.1 モ デ ル

予後の良好・不良を予測する分類問題を解くに当たって, 今回は Bidirectional Encoder Representations from Transformers (BERT) [7] をベースとした 2 つの言語モデルを用いて実験を行った。1 つ目は 2019 年 9 月までの日本語版 Wikipedia に含まれる, 約 1700 万行の文書を用いて事前学習した bert-base-japanese-whole-word-masking (以下 Tohoku-BERT) [8], 2 つ目は東京大学附属病院の電子カルテに保存されていた約 1 億 2000 万行の文書をコーパスとして事前学習した UTH-BERT-BASE-512-WWM (以下 UTH-BERT) [9] を用いた。これらのモデルは両方とも日本語を学習したモデルであるが, 前者が汎用的な知識を獲得しているのに対し, 後者は医療分野に特化した知識を獲得している。今回使用するデータに含まれる自由記述部分には医師・保健師と患者による対話が収められた部分が多く存在している。そのため医療知識をあまり持たない患者向けに, 噛み砕いた説明になっていることが予想されたため, 医療特化のモデルだけではなく, 汎用的な知識を保持した日本語モデルも用いることとした。

構造化データを日本語に変換して結合した入力文

これは 4 回目の診察です。性別は男性。GPT(ALT) は正常。HbA1c の値は要注意。

非構造化データを日本語に変換して結合した入力文

先日はご多忙の中, お時間を頂きましてありがとうございました。〇〇〇〇〇〇です。足の痛みが和らいだこと, 痛風の治療と服薬は医師の指示どおり行えていることを伺い, 安心しました。今回のお電話では, ...

構造化・非構造化データを結合した入力文

これは 4 回目の診察です。性別は男性。GPT(ALT) は正常。HbA1c の値は要注意。先日はご多忙の中, お時間を頂きましてありがとうございました。〇〇〇〇〇〇です。足の痛みが和らいだこと, 痛風の治療と服薬は医師の指示どおり行えていることを伺い, 安心しました。今回のお電話では, ...

### 4.2 入力文章の作成

今回は言語モデルを用いて予測を行うことから, 入力 は自然言語による記述となるため, データセット内の特に自由記述部分の項目を主に扱うこととした。主な自由記述項目としては「生活改善指導時の発話」「継続配慮に関する申し送り」「飲料習慣」「睡眠習慣」「雑記」の 5 項目となっていたため, まずこの文章同士を繋げて一つの入力文章とした。加えて構造化データ部分に関しても予測結果に寄与すると考えられる項目が複数含まれていることから, これらの項目に関しても入力として用いることとした。図 2 はこのようにして作成された入力文章の一例である。特定保険審査データには 54 種類の構造化データが含まれていることは 3.1 節で紹介した通りである。これらの項目すべてを入力文章として変換して用いることは, 今回使用するモデルの入力トークン数の最大が 512 トークンとなっていることから困難であると考えた。そのためこの 54 項目のうち予測に寄与すると考えられる「性別」「ALT(GPT)」「HbA1c」「中性脂肪」「診断回数」の 5 項目を入力文章として追加することとした。この中で ALT(GPT) は肝臓脂肪に関連する数値である。HbA1c はヘモグロビンのうち糖化したものの割合を示しており, 主に糖尿病に関わる数値となっている。これらの数値を単純に文章化すると「性別は男性。ALT(GPT) は 50IU/L。HbA1c は 4.3 %。中性脂肪は 69mg/dl。診断回数は 2 回」といったな一方で, 「ALT(GPT)」「HbA1c」「中性脂肪」に関してはその数値が異常なのか判断がつかない。そのため, 今回はこの 3 項目に関しては基準値を調べ, ALT(GPT) の数値が 4 未満であれば「低い», HbA1c の数値が 5.8%であれば「正常», 中性脂肪の数値が 149mg/dl より大きければ「多い」という形で変換して文章を作成した。作成した文章とその数を表 2 に示す。また, 作成した文章のトークン数を調べた結果を図 3 に示す。これによりデータのほとんどがモデルの入力トークンの最大値である 512 トークン内に収まっていることを確認した。

表 2 カテゴリを一定の基準でグルーピングした際の基準と該当数

| 項目       |                        | 該当数   |
|----------|------------------------|-------|
| 予後       |                        |       |
|          | 良好                     | 26356 |
|          | 不良                     | 2598  |
| 性別       |                        |       |
|          | 男性                     | 27841 |
|          | 女性                     | 1113  |
| GPT(ALT) |                        |       |
|          | 低い ( $< 4IU/L$ )       | 1     |
|          | 正常                     | 23546 |
|          | 高い ( $44IU/L <$ )      | 5407  |
| HbA1c    |                        |       |
|          | 正常 ( $< 5.6 \%$ )      | 12532 |
|          | 要注意                    | 13418 |
|          | 極めて高い ( $6.4 \%$ $<$ ) | 3002  |
|          | 測定不能                   | 2     |
| 中性脂肪     |                        |       |
|          | 少ない ( $< 30mg/dl$ )    | 6     |
|          | 正常                     | 15175 |
|          | 多い ( $149mg/dl <$ )    | 13773 |

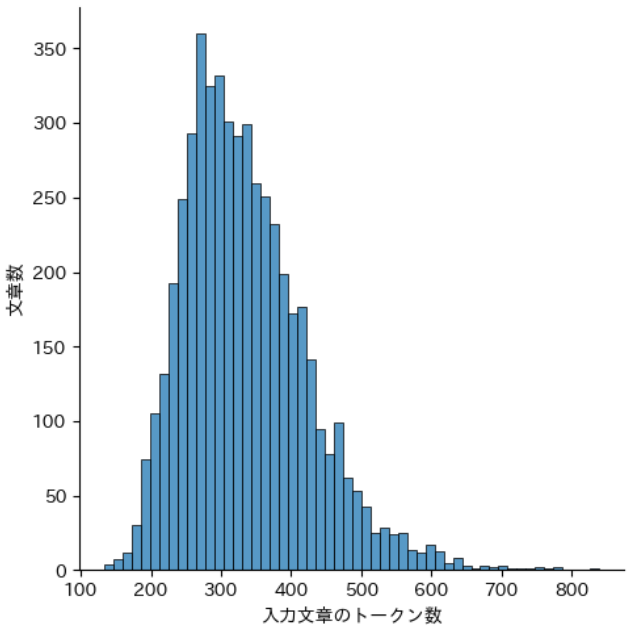


図 2 作成した入力文数とトークン数の分布

5 実 験

5.1 設 定

本研究では前回健診時からの患者予後が良好か不良かを判定する分類問題を設定した。データに関しては予後の数が不均衡になっていたため、これを解消するためにダウンサンプリング

を実施し、最終的に 5396 件のデータを用いて実験を行った。モデルに関しては一般テキストを用いて事前学習した Tohoku-BERT と、医療分野のテキストで事前学習した UTH-BERT を用いた。入力文章に関しては構造化データのみを文章化したもの、自由記述欄のみを結合して文章化したもの、これら 2 つの文章を結合させて一つの文章にしたものの 3 種類の入力文章を作成した。評価指標に関しては分類タスクということを考慮し F 値、AUROC、Accuracy の 3 種類を用いた。加えて予後判定時に用いる閾値を 0.1 ごとに区切って感度と特異度を計算した。感度とは実際に予測対象の疾病に罹患している人のうち予測結果が罹患しているとされた人の割合を示し、特異度とは予測対象の疾病に罹患していない人のうち予測結果が罹患していないとされた人の割合を示している。本研究の場合は、予後が良好となっている患者に対して過剰な医療リソースの投下を抑制するといった状態を想定している。そのため予後良好のラベルがつけられた人のうち予測結果が予後良好となっている人の割合を感度、予後不良のラベルがつけられた人のうち予測結果が予後不良となっている人の割合を特異度として計算している。実験のハイパーパラメータとしてはバッチサイズは 32、エポック数は 1000、アーリーストップは 50、学習率は  $2e-5$  とし、消費した GPU のメモリは約 29GB、約 100epoch で停止し、計算時間は約 2 3 時間ほど要した。

5.2 結 果

表 2 に実験結果を示す。まず入力データセット間における精度は、構造化データのみ、自由記述欄のみ、構造化データ+自由記述の順に全体として精度が向上することが確認できた。構造化データのみを文章化したものを入力として用いた場合は UTH-BERT のほうが性能が高く、自由記述欄のみを結合して文章化したものを入力として用いた場合は Tohoku-BERT のほうが精度が高くなった。検査値データが含まれた文章と自由記述欄を組み合わせで作成した文章の場合、F 値に関しては UTH-BERT のほうが精度が高く、AUROC と Accuracy に関しては Tohoku-BERT のほうが精度が高いという結果となった。

表 4 は感度・特異度の結果を表している。医療資源の最適化といった場面で過剰な医療リソースを投下しすぎないということを目的とした場合、閾値を 0.4 とした時の感度が Tohoku-BERT と UTH-BERT 共にが 85%ほどと十分な精度となっている。

図 4 は入力文章に Attention を表示して可視化した結果である。Attention が全体的にかかっており、ひと目見て注目すべき点がどこかを判定することは難しい。

5.3 考 察

5.3.1 モデルの特性について

モデル別の性能を比べてみると、検査値データのみを入力文章とした場合は UTH-BERT のほうが全体的な精度が高くなっている。一方で、自由記述部分のみを入力文章とした場合は Tohoku-BERT のほうが全体的な精度が高くなっている。UTH-BERT は医療テキストを用いて事前学習されていること

表 3 入力とモデル別に測定した性能の一覧。構造化データと非構造化データを組み合わせて作成した文章を入力に用いた際に一番精度が高くなる。

| 評価指標     | 構造化データのみ    |          | 非構造化データのみ   |          | 構造化+非構造化データ  |              |
|----------|-------------|----------|-------------|----------|--------------|--------------|
|          | Tohoku BERT | UTH-BERT | Tohoku BERT | UTH-BERT | Tohoku BERT  | UTH-BERT     |
| F 値      | 0.630       | 0.629    | 0.646       | 0.015    | 0.658        | <b>0.683</b> |
| AUROC    | 0.428       | 0.621    | 0.644       | 0.523    | <b>0.662</b> | 0.642        |
| Accuracy | 0.478       | 0.605    | 0.619       | 0.501    | <b>0.630</b> | 0.615        |

表 4 閾値をずらして感度と特異度を計算した結果

表 5 予後ラベル判別の閾値をずらして、感度と特異度を計算した結果

| 閾値  | 感度          |          | 特異度         |          |
|-----|-------------|----------|-------------|----------|
|     | Tohoku BERT | UTH-BERT | Tohoku BERT | UTH-BERT |
| 0.0 | 1           | 1        | 0           | 0        |
| 0.1 | 0.9769      | 0.9961   | 0.0961      | 0.0653   |
| 0.2 | 0.9269      | 0.9961   | 0.2461      | 0.1038   |
| 0.3 | 0.9038      | 0.9615   | 0.3         | 0.1807   |
| 0.4 | 0.8576      | 0.9      | 0.3846      | 0.3346   |
| 0.5 | 0.7115      | 0.8307   | 0.55        | 0.4      |
| 0.6 | 0.2461      | 0.5769   | 0.8384      | 0.5961   |
| 0.7 | 0.0038      | 0.1807   | 1           | 0.8884   |
| 0.8 | 0           | 0        | 1           | 1        |
| 0.9 | 0           | 0        | 1           | 1        |
| 1.0 | 0           | 0        | 1           | 1        |

正解ラベル: 予後不良

予測ラベル: 予後不良

[CLS]これは2回目の診察です。性別は男性。GP# # T(AL# # T)は正常。H# # b# # A1Cの値が要注意。中性脂肪が多い。先# # 日はご# # 多# # 忙の中、お時間を頂# # きましてありがとうございました。〇〇〇〇〇〇です。間# # 食を継続して控えられていますね。菓子を食# # べたい時には、夜の摂取を避け# # て、少量を食# # べられるように工夫されていますね。無理をしすぎずに取り組みを継続されている事が、大変素晴らしいですね。夕食では、奥# # 様と別献# # 立を作られ、本当に良く頑# # 張# # られていますね。今は仕事# # が落ち# # 着# # かれ、睡眠・ストレス状況が前よりも改善# # しているとの事でしたね。ゆっくりと休養が# # 取れるようになられ、とても良い事ですね。お仕事# # が落ち着いて# # いるとの事なので、また折を見て歩行を始め# # てみましょう。6ヶ月間、本当にお# # 疲# # れ# # 様でした。今後も健康に気をつけお元気に# # 過# # ご# # ください。

図 3 入力文章に対する予測根拠部分の Attention 可視化。

から、検査値データに関する認識精度が高かったと考えられる。一方で、自由記述欄の内容は医師・保健師と患者の会話時の文章が多く含まれており、非医療従事者向けの語彙が多く含まれている。そのため、一般的な語彙を多く含む Tohoku-BERT の方が認識できる単語が多く含まれていることから、UTH-BERT よりも Tohoku-BERT の方が精度が高くなったと考えられる。検査値と自由記述文の両方を含む内容を入力とした場合、入力文に含まれる情報としては自由記述欄の情報の方が多く含まれるようになっている。そのため Tohoku-BERT の方が若干精度が高くなったと考えられる。

### 5.3.2 分類性能について

実験では入力文を検査値と自由記述文から作成したとき、AUROC が 0.662、Accuracy が 0.630 程度にとどまっている。医療現場で使用することを想定すると、精度改善の余地が十分に残されている。

まず、本研究では、使用するデータセットを大きくダウンサ

ンプリングしているため、予後不良ラベルが付与された未使用データが多く存在する。これを有効に利用するため、予後ラベルの出現割合が等しくなるようにリバランスしたデータセットを複数作成し、データセットごとにモデルを訓練した後、そのモデルの出力結果を多数決する方法が考えられる [10]。これにより、全データをモデルに投入することが可能となり、より広範囲な知識が獲得できると考えられる。一方で、今回の言語モデルのようにメモリを多く使用する場合、データセットごとのモデルを同時に学習するというのは難しい。そのため、一つのデータセットとモデルごとの結果を保存し、すべての計算が終わった後に集計処理を行うといった実装上の工夫が必要となる。

また不均衡データを、基準とする項目の個数が均衡するようにデータセットを改変して学習した後、不均衡データ全体を使って学習を行うと精度が向上するといった研究結果が存在する [11]。今回の実験では、データを均衡状態にしたため、作成したモデルを全体のデータを使ってもう一度学習することで、

精度が向上する可能性がある。

加えて、今回は、体重 2kg 以上減少かつ腹囲 2cm 以上減少した場合に予後良好とみなしたが、この条件が厳しかったという問題が考えられる。全データのうち、この条件を達成していたのは約 9%であったため、今回はこの数値に対して大きくダウンサンプリングを実施することになり、十分な精度を出すことができなかった。これに関しては、実際の現場で予後が良好と判断される人の割合と相違が無いのか調査する必要があると同時に、実験を実施する上では、条件を緩和する方法も考えられる。例えば今回の予後良好ラベルの設定を体重 1kg 以上減少、腹囲 1cm 以上減少とすることにより予後良好判定となるデータを増やすことができる。その状態での精度まず確認した上で、より厳しい条件の精度を確認するといった方法も考えられる。

### 5.3.3 予測確率の分布について

本研究ではデータをダウンサンプリングして予後の良好と不良が同数になったものをデータセットとして使用している。それに伴って閾値を 0.5 とした場合にモデルの出力も予後の良好と不良の出現割合がおよそ 1 対 1 になるようになっている。しかし現実には予後が良好となる人の割合は少なく、今回のデータセットの場合約 9%しか存在していない。仮に本研究で作成したモデルが実際の医療現場で使えるだけの精度が出た場合でも、実験に使ったデータセットの分布が現実の患者予後の分布と異なるため、今回の場合では予後良好となる確率が高くなってしまう。この場合予後が不良の患者を予後良好と誤診する場合は想定され、重大な医療過誤につながる恐れがある。そのため予測確率の補正が必要になってくる。この場合現場で使用するといった際にはデータセット上の分布と現実の分布の乖離に注意し、Isotonic Regression [12] などの手法で現実のデータに適応するように確率補正を行うという処理が必要となってくる。

### 5.3.4 説明可能性について

本研究では従来の手法では確認することが難しい予測結果の根拠という部分に着目した。そこで可視化結果に注目すると、全体的に Attention がかかっているようになっており、特に際立った根拠部分を見つけるとするのは難しい。今回は Attention 部分の可視化を行ったが、他には Integrated Gradients [13] といった手法が存在している。一方で Attention がかかっている部分が根拠として解釈してもよいのかという問題が存在しており、自然言語を用いた解釈性といった部分にはまだ検討の余地が残されている。

### 5.3.5 限界

特定健診の予後予測は経験を積んだ指導員にとっても難しい課題となっている。生活習慣の改善が求められるため、血糖値等の数値データよりも本人の家庭環境や改善への意思といったデータのほうが、予測する上では大きく寄与すると考えられる。例えば日々の仕事に追われておりほとんど家に帰宅しないといった人に対して、食生活の改善を促したとしても、それを実行するのが困難である場合が多いと考えられる。しかし従来の数値データを用いた手法ではこのコンテキストを把握することは難しく、そこに医師や保健師が診察を通じて記した自由記述

データを有効に活用する価値は多く残されている。一方で本研究の結果の通り精度や解釈性には多くの課題が残されており、今後どのように活用していくか検討の余地が残されている。

### 5.3.6 今後の展望

近年 ChatGPT に代表されるように、BERT よりも多くのパラメータを持つ言語モデルが登場してきた。医療分野に特化したものでは Med-PaLM [14] というモデルが登場している。モデル自体はまだ一般公開されているものではないが、既に米国の医師国家試験と同等レベルとされる MedQA タスクにおいて約 86.5%もの精度を達成したと報告されている。このような医療分野に特化したより大規模なモデルは日々開発させており、自然言語を用いた病院内データの利活用はより進んでいくと考えられる。

このような巨大な言語モデルが発達していくに連れて、人間がこれらを扱う方法についても変わってくる。本研究で用いた BERT のようなモデルは、一定のタスクを設定してそれに対してモデルを学習させて使用するといった形が一般的であった。そのため入力文章は文章、出力は数値といった形になるとその数値の算出過程を明らかにすることが難しいため解釈方法といった別の問題も出てきていた。一方で近年の巨大な言語モデルは、一つのモデルをあらゆる大量のデータで学習することで、そのモデルですべてのタスクを解くという方向に変わってきている。このようなモデルは入力、出力ともに文章となっているため、入力時の命令に「出力時には予測根拠も出力すること」といったようなプロンプトを入力しておくことでモデルの予測根拠のようなものを同時に出力することができる。また、ChatGPT のようにチャット形式で使用することで人間が補完しきれていない知識を取り出すことが可能となっている。そのため人間側が機械の出力結果を使用できる程度のタスクに落とし込むことさえできれば、解釈性といった問題を無視することができる。例えば出力に対する信頼性が低くとも、人間がその出力に対して裏取り調査を行った上で利用する等が考えられる。このような変化が起こる中で、言語モデルの進化に伴い人間側もその使用方法を改善することで導入の障壁を取り除くことができるのではと考えられる。

## 6 おわりに

本研究では自然言語の持つ解釈性に着目し、特定健診データを使用して患者の予後予測を行った。データには構造化データも含まれていたため、これらを活用するべく 3 種類の文章を作成し、言語モデルの入力文章とた。予測ラベルに関しては次回健診時の体重と腹囲の変化から作成した。言語モデルは一般分野のテキストで事前学習済みの Tohoku-BERT と、医療分野のテキストで事前学習済みの UTH-BERT の 2 種類を使用した。これらを組み合わせた 6 種類の条件から予後の予測精度を調べた結果、約 63%程度の正解率となることが判明した。加えて医療分野で使用するためには説明可能性が重要になってくるため、今回はこの予測モデルの Attention を可視化することで根拠部分の判定が可能か調査を行った。その結果、現段階

では Attention のかかった部分を見ることで判断の根拠となる部分を判別することは難しかったが、一部診断回数には比較的 Attention が強くかかっていることが判明し改良次第では実応用可能なのではないかと考えられる。

## 7 謝 辞

本研究の内容の一部は戦略的創造研究推進事業（CREST）、及び JSPS 科研費 JP 19K10620 の助成を受けたものである。また本研究に使用したデータは、SOMPO ヘルスサポート株式会社および全国土木建築国民健康保険組合より提供を受けたものである。なお、本研究は京都大学医の倫理委員会（R0817-3）で承認を受けている。

## 文 献

- [1] 令和 4 年 (2022) 人口動態統計月報年計 (概数) の概況. p. 10, 2022.
- [2] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Transactions on Neural Networks and Learning Systems, Vol. 32, No. 11, pp. 4793–4813, 2021.
- [3] 荒木雅弘 新谷元司 吉川孝恒 川充. 健診データを用いた生活習慣病の発症予測. 2019.
- [4] 讃岐勝 我妻ゆき子 大場勇貴. 健康診断データを用いた疾患予測における解釈可能なモデルの構築. 2020.
- [5] Oliver Sander Thomas Lengauer André Altmann, Laura Toloşi. Permutation importance: a corrected feature importance measure. Bioinformatics, Vol. 26, No. 10, pp. 1340–1347, 2010.
- [6] Sutanay Choudhury Sindhu Tipirneni Pritam Mukherjee Colby Ham Suzanne Tamang Matthew Baker et al Agarwal, Khushbu. Preparing for the next pandemic via transfer learning from existing diseases with hierarchical multi-modal bert: a study on covid-19 outcome prediction. Scientific Reports, Vol. 12, , 2022.
- [7] Ming-Wei Chang Kenton Lee Devlin, Jacob and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.
- [8] M Suzuki. cl-tohoku/bert-japanese: Bert models for japanese text. Accessed: 3-Nov-2023.
- [9] Daisaku Shibata Emiko Shinohara Eiji Aramaki Kawazoe, Yoshimasa and Kazuhiko Ohe. A clinical specific bert developed using a huge japanese clinical text corpus. PLoS One, Vol. 16, No. 11, 2021.
- [10] Kevin Small Carla E. Brodley Wallace, Byron C. and Thomas A. Trikalinos. Class imbalance, redux. In 2011 IEEE 11th international conference on data mining, pp. 754–763, 2011.
- [11] Simon Wouters Ruben Cartuyvels Erfan Ghadery Malik, Farjad and Marie-Francine Moens. Two-phase training mitigates class imbalance for camera trap image classification with cnns. p. 10, 2021.
- [12] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. pp. 625–632, 2005.
- [13] Ankur Taly Sundararajan, Mukund and Qiqi Yan. Axiomatic attribution for deep networks. PMLR, pp. 3319–3328, 2017.
- [14] Tao Tu Juraj Gottweis Rory Sayres Ellery Wulczyn Le Hou Kevin Clark et al Singhal, Karan. Towards expert-level medical question answering with large language models. 2023.

# AI 音楽療法士（さつきさん）の開発研究

## 音楽療法のデジタルセラピューティクスにむけて

小杉 尚子<sup>†</sup> 石井健太郎<sup>†</sup> 児玉 直樹<sup>††</sup>

<sup>†</sup> 専修大学ネットワーク情報学部 〒214-8580 神奈川県川崎市多摩区東三田 2-1-1

<sup>††</sup> 新潟医療福祉大学医療技術学部 〒950-3198 新潟県新潟市北区島見町 1398 番地

E-mail: <sup>†</sup>naonaoan@gmail.com, <sup>††</sup>kenta@pc.fm.senshu-u.ac.jp, <sup>†††</sup>kodama@nuhw.ac.jp

**あらまし** 本論文では、認知症高齢者を対象とする、人工知能（AI）を用いた音楽療法士（AI 音楽療法士）の開発研究について述べる。音楽療法の認知症の「行動・心理症状」に効果があるため、介護負担の軽減に資することが期待されている。しかし、日本における音楽療法士の数は認知症高齢者の数に比べて非常に少なく、日本全国の認知症高齢者が誰でも気軽に音楽療法を受療できる状況ではない。そこで認知症高齢者がいつでもどこでも何回でも音楽療法を受療できるようにするために、パソコンの画面に現れるバーチャルな音楽療法士（AI 音楽療法士）「さつきさん」を開発している。本論文では、最新の「さつきさん」について報告すると共に、その「さつきさん」を、実際に認知症高齢者の介護に従事する介護スタッフに試用して頂いた際に得た意見をまとめ、本研究の方向性について考察する。わたしたちは、将来的には、AI 音楽療法士による音楽療法の受療者から多様なデータを収集し、受療者の状態の客観評価や認知症の進行予測なども可能な音楽療法のデジタルセラピューティクスを確立したいと考えている。

**キーワード** AI、音楽療法、認知症、デジタルセラピューティクス

### 1 はじめに

令和5年版高齢社会白書[4]によると、令和4年10月1日現在、65歳以上人口は3,624万人、75歳以上人口は1,936万人となり、総人口に占める割合（高齢化率）はそれぞれ29.0%、15.5%となったとのことである。令和4年版高齢社会白書[3]によると、令和3年10月1日現在、65歳以上人口は3,621万人、75歳以上人口は1,867万人で、総人口に占める割合（高齢化率）はそれぞれ28.9%、14.9%であったので、1年間で高齢化率はそれぞれ0.1%、0.6%上昇したことになる。また令和4年版高齢社会白書によると、65歳以上の要介護者数は増加しており、介護が必要となった主な原因としては、「認知症」が最も多く、18.1%を占めているとのことである。一方令和5年版高齢社会白書では、要介護者等と同居している主な介護者の年齢が60歳以上である割合が、男性では72.4%、女性では73.8%となっており、「老老介護」のケースが相当数存在していることや、介護を理由にした離職者の約3/4が女性であったことなどを報告しており、認知症高齢者数の増加と介護人材の不足が、重要な社会課題となっていることが分かる。

認知症とは、正常に発達した認知機能が、後天的な何らかの障害により、生理的老化の範囲を超えて広汎かつ慢性的に低下した状態のこと[1]で、記憶、言語、視空間認知などの認知機能の障害と、BPSD (Behavioral and Psychological Symptoms of Dementia) と呼ばれる徘徊や無気力などの「行動・心理症状」からなる[2]。認知症には、現状では有効な治療法がないため、認知症高齢者への対応はケアが中心となるが、認知症高齢者のケアは非常に困難であるため、介護者の負担を軽減するこ

とが重要である。

認知症高齢者の介護を困難にするのは徘徊、不安、抑うつなどのBPSDであるが、これは適切なケアを行うことで発症を抑制したり、症状を軽減したりすることができるため、非薬物療法で対応することが推奨されている。認知症疾患診療ガイドライン2017には、非薬物療法として運動療法、回想法、音楽療法などが挙げられている。

音楽療法は認定資格を持つ音楽療法士によって実施されるが、日本で活動する音楽療法士の数は2,461名（2021年3月末日現在）[5]と少なく、その4〜5割が関東域で活動している可能性が高いため、絶対数不足と首都圏偏在が問題になっている。また音楽療法士には女性が多いため、出産・子育てや介護などが音楽療法士として経験を重ねた量を磨く時期に重なる場合も多く、そのような時期には十分な臨床実践が難しくなるという問題も抱えている。これは結果的に、音楽療法士は自身の成長機会を失い、療養者も音楽療法の受療機会を失うという、音楽療法士と療養者の両者の問題になっている。

そこで私達は、認知症高齢者が「いつでもどこでも何回でも」気軽に音楽療法を受療できる社会の確立を目指して、AI(人工知能)を用いた音楽療法士（以下、AI 音楽療法士）の開発研究を推進している[8]。私達が開発しているAI 音楽療法士は「さつきさん」という名前で、認知症高齢者の「発話サポート」と「歌唱伴奏」に特化したAI 音楽療法士である。本論文では「さつきさん」の研究開発の最新の進捗について述べると共に、それを認知症介護従事者に試用してもらってインタビューしたのでその結果について報告する。そしてそれに基づいて、今後の展開について議論する。本論文の構成は以下である。第2章では音楽療法とデジタルセラピューティクスについて説明する。

第3章でAI音楽療法士について説明し、第4章で実験と評価結果を報告し考察を述べる。最後に第5章で本論文をまとめる。

## 2 音楽療法とDTx(デジタル・セラピューティクス)

音楽療法とは「音楽の持つ生理的、心理的、社会的働きを用いて、心身の障害の回復、機能の維持改善、生活の質の向上、行動の変容などに向けて、音楽を意図的、計画的に使用すること」である[6]。週に1回から月に1,2回の頻度で、歌唱(合唱)や楽器演奏などを通して、受療者のストレス発散や不安の軽減[11]、認知機能の改善[15],[16],[17]などを支援する。主な対象者は、精神疾患患者や高齢者などの投薬や手術だけでは状態の改善が難しい方々である。日本では5~10名を対象に、1時間前後のプログラムが行われることが多いが、受療者の主訴や人数等に応じて柔軟に調整可能である[18],[19],[20]。

音楽療法は、第1章でも述べた通り、認知症に対する有効性が既知であるが、日本では音楽療法士が不足しているため、誰でも気軽に受療できる状況ではない。そこで著者らは、インターネットを利用して音楽療法を提供・受療する遠隔音楽療法システム(Music-Telepy)を開発した[9]。Music-Telepyは、音楽療法士と認知症高齢者施設を光回線で結び、パソコン、Webキャスティング・ミキサー、マイク、スピーカ、モニタを用いて音楽療法を実施するもので、特に音楽療法士のいない離島や冬場の外出が困難になる豪雪地では大変好評だった[10]。そして著者らは、このような臨床実践・臨床研究を通して、遠隔音楽療法が、認知症高齢者に対して、対面方式の音楽療法とほぼ同等の臨床効果があることと、認知症高齢者が、モニタに映し出された音楽療法士と自然に会話できることも確認した[11]。

「デジタル・セラピューティクス(Digital Therapeutics: DTx)」は、日本語では「治療用アプリ」と訳される。医療現場におけるICT技術の活用は、主に診断の場で進められてきた。しかし昨今、ICT技術の新しい活用場として注目を浴びているのが「治療の場」である。日本では2020年8月21日に禁煙を支援するためのニコチン依存症治療用アプリ「CureApp SC」と呼気中の一酸化炭素濃度を計測する「ポータブルCOチェッカー」が薬事承認を得て注目を集めた[12]。アメリカでは約10年前にアメリカ食品医薬品局が糖尿病の治療補助アプリ「BlueStar」を承認し、医療保険の適用が認められている。現代社会では、医療技術の進化に伴う超高齢化や、重大事故などによる死亡が減少している。これ自体は好ましいことであるが、超高齢者や重大事故などで障害を負った方々が、その人らしく生きる・生活するためには、病院での治療に加えて様々な医療や介護サポートが必要であることは想像に難くない。また社会のグローバル化・複雑化等で精神疾患や慢性疾患を抱えながら社会生活を送る人々が増えている。そのような傷病では「行動変容」や「自己管理」が治療の過程で欠かせないため、いつでもどこでも利用できるDTxに期待が高まるのは必然である。さらにDTxではセラピストの力量に左右されない均質なセラピーや、患者の応答内容・表情や声のトーン等の様々な情報を利用・学習することで、患者の状態のタイムリーな把握、

治療効果の客観評価、個々の患者に合わせたアプローチや、病気・症状の進行予測、それに基づいた予防措置の提案等も可能になると考えられるため、今後益々関心が高まることは間違いない。

## 3 AI音楽療法士

著者らは音楽療法のDTx(第2章参照)としてAI音楽療法士を研究開発している。

AIによって代替できない職種として常に挙げられるものの1つに「セラピスト(Therapist, 療法士)」があり、AI療法士の開発は容易ではないと考えられている。音楽療法士(ミュージック・セラピスト)は、音楽を使って対象者に寄り添う専門職であるため、これを完全にAIで実現するには相当困難であると考えられる。しかし音楽療法士の業務は多様で多岐に渡るため、AIが実施可能、むしろAIの方が得意な業務も存在することは想像に難くない。そこで我々は、音楽療法を「AI音楽療法士と生身の音楽療法士の協業」と捉え直し、AI音楽療法士が実施可能な業務を厳選してその完成度を高め、生身の音楽療法士との協業の形を新しく定義するという方針で研究を進めている。

### 3.1 概要

著者らはAI音楽療法士が実施可能な業務として、私達は「会話」と「歌唱伴奏」を選定した。認知症高齢者に対する音楽療法では、歌唱、体操、楽器演奏などが行われることが多く、それらの活動の間や、セッションの最初と最後では、音楽療法士との楽しい会話が人気だからである。

特に「会話」については認知症高齢者の「発話サポート」にフォーカスし、高齢者に安心して楽しく好きなおしゃべりしてもらえる環境構築を目指している。そのため、「～と言えば?」や「～の思い出は?」等の「問いかけ」や、「〇〇が大変だったのですね」や「私も〇〇が好きです」等の「復唱」を多用する、シナリオベースの会話のデザインを進めている。これは「イライザ効果」[26]として知られる、「復唱と簡単な質問を繰り返すだけで、あたかも自分が理解されている・受け容れられていると勘違いし、ユーザは対話に没頭した」ことを参考にしたもので、私達もこのような対話デザインを採用することで、話好きな高齢者の発話を促進し、対話による満足感を高め、AI音楽療法士の定期的かつ長期的な利用に繋がれるのではないかと考えている。AI音楽療法士は音楽療法のデジタルセラピューティクスとして開発するため、利用者の定期的かつ長期的な利用は療養データの収集には欠かせない重要な特性である。

「歌唱」は音楽療法において高齢者に人気の高い最頻活動であるため必須だと考えている。将来的には発話音声もAIを用いて分析し、対象者の歌声のキーやテンポを推定しながら[13],[14]、目の前の高齢者の歌唱に合わせた伴奏をすることを目指している。現段階では音楽療法士に指定されたキーとテンポで、「幸せなら手をたたこう」の伴奏音源を再生している。本研究で開



図 1 AI 音楽療法士が高齢者に音楽療法を提供しているイメージ

発する AI 音楽療法士のイメージを図 1 に示す。

### 3.2 システム構成

AI 音楽療法士のシステム構成を図 2 に示す。

システムは、1) 音声認識モジュール、2) 発話伴奏モジュール、3) パーチャルエージェントモジュール、4) 対話シナリオ管理モジュールからなる。

音声認識モジュールは、対話シナリオ管理モジュールからのトリガーにより、「療対象者」の発話音声进行認識し、テキストへの変換を行う。音声からテキストへの変換には外部の音声認識エンジンを利用しており、変換されたテキストは対話シナリオ管理モジュールに転送される。

発話伴奏モジュールは、対話シナリオ管理モジュールからのトリガーにより、「さつきさん」の発話音声と音楽の伴奏を再生する。再生する発話音声と伴奏はファイルとしてシステムの実行環境に予め用意してあり、システムの実行時には指定のファイルが再生される仕組みである。今後は、システムの実行時に音声合成や伴奏音源の微調整を行いながら再生することも検討している。

パーチャルエージェントモジュールは、対話シナリオ管理モジュールからのトリガーにより、「さつきさん」の身体動作を再生する。再生する身体動作はファイルとしてシステムの実行環境に予め用意してあり、システムの実行時には指定のファイルが再生される仕組みである。今後は、まばたきなどの不随意動作をシステムの実行時に自動で追加する仕組みを導入することも検討している。

対話シナリオ管理モジュールは、音声認識テキストに基づき、対話の状態と次の話題を決定する。決定した話題から、「さつきさん」に何と発話させるのか、あるいは、音楽の伴奏を再生するのも決定し、音声認識モジュール・発話伴奏モジュール・パーチャルエージェントモジュールにそれぞれのトリガーをかける。

パーチャルエージェントモジュールの画面例を図 3 に示す。

画面に現れる女性が AI 音楽療法士の「さつきさん」である。

### 3.3 実装

本節では、「さつきさん」の実装について説明する。システムは JavaScript によるウェブアプリとして実装されており、PC・タブレット・スマートフォンのウェブブラウザから利用可能である。

音声認識モジュールは、まず音声区間検出 (Voice Activity Detection; VAD) を行い、マイク入力中の音声が入力されている区間を切り出してから、外部の音声認識エンジンによる音声認識を実行する。VAD には「Silero VAD [21]」の JavaScript 実装 [22] を利用し、音声認識エンジンには OpenAI が開発した「Whisper [23]」を利用している。

発話伴奏モジュールは、指定の発話音声ファイルや伴奏ファイルを再生する機能を持つ。発話音声ファイルは、第 3.4 節に後述する SSML 形式のシナリオテキストを、Google Cloud の Text-to-Speech [24] を用いて音声合成して作成している。MIDI ファイルは、音源としては MIDI データの利用を想定しているが、現段階では対象となる楽曲の MIDI 音源から、キーの異なる音源を MP3 データの形式で予め複数作成し、伴奏再生時にキーを指定して、該当する伴奏音源を再生することができるように実装されている。

パーチャルエージェントモジュールは、「さつきさん」や仮想空間周囲の物体を表示するためのモジュールで、three.js [25] を利用している。「さつきさん」を含む仮想空間のオブジェクトは glTF 形式のファイルとして予め作成してあり、身体動作用のモーションも同じ glTF ファイルに含めている。パーチャルエージェントモジュールは、指定の身体動作のモーションを再生する機能を持ち、対話シナリオ管理モジュールからのトリガーによりモーションを再生する。

対話シナリオ管理モジュールは、以上の各モジュールを制御する JavaScript プログラムとして実装されている。

### 3.4 シナリオ

さつきさんのシナリオは、音声合成マークアップ言語 SSML (Speech Synthesis Markup Language) で記載している。以下に実際のシナリオの一部を示す。さつきさんは、(1)～(5) の各発話の後、対象者からの回答を待ち、回答を得た後は、音声認識を行って次の発話に進む。

(1) こんにちは。

(2) AI 音楽療法士の <s>「さつき」</s> です。5 月生まれなので名前は「さつき」と言います。今日、<s> 音楽 </s> を一緒にさせていただきますね。<s> よろしくお願ひします。</s> さて、今日のご気分はいかがですか？

(3) そうですか。<break time="2s"/> 私は朝ごはんをたくさん食べましたので <s> とても元気 </s> です。今日は、そちらは、どんなお天気ですか？

(4) そうですか。<break time="2s"/> 晴れ、曇、雨などありますが、どんなお天気が好きですか？

(5) そうなんですね。<break time="1s"/> 私は晴れの

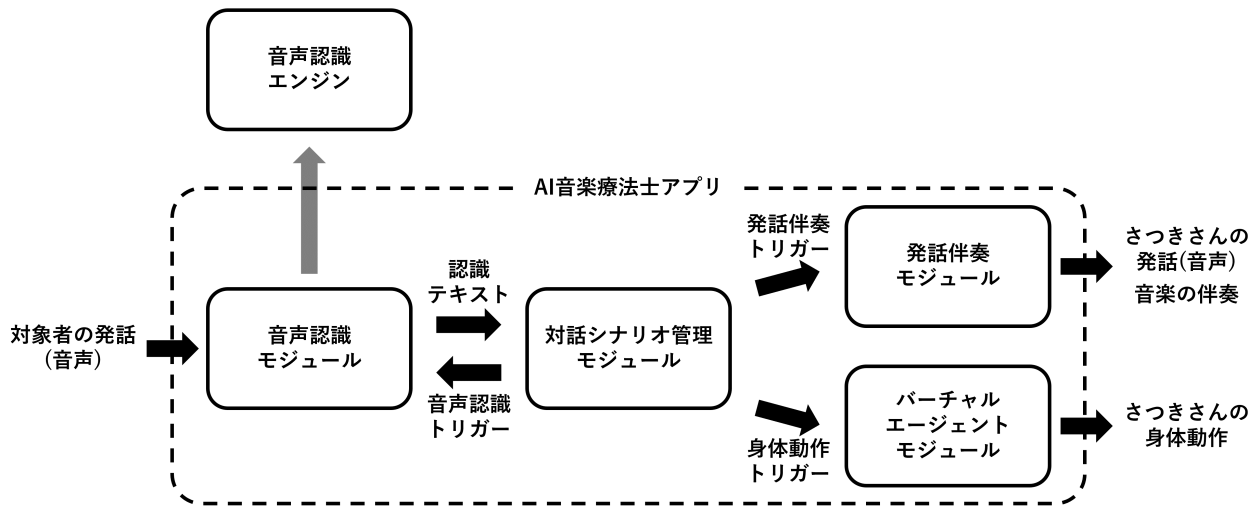


図2 システム構成図

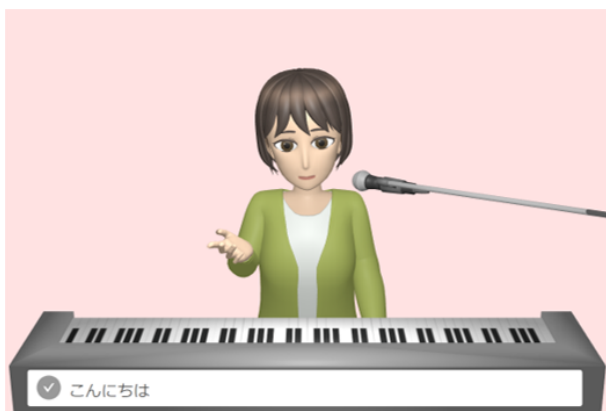


図3 AI音楽療法士の「さつきさん」

日が好きです。青空の中にふわふわの白い雲を見つけると、嬉しくなり、わたあめが食べたくなります。

## 4 実験と評価、および考察

「さつきさん」を、認知症高齢者を対象とした介護施設で従事する5名の介護スタッフに見てもらい、その動作や使用感についてヒアリングを行った。本章ではヒアリングで得られた指摘事項をまとめ、各指摘事項に対する考察と今後の進め方について述べる。

### 4.1 質問項目

以下に今回のヒアリングの項目を列記する。

- (1) 「さつきさん」との会話は楽しかったですか? 「さつきさん」との会話についてご意見をお聞かせください。
- (2) 「さつきさん」とどんなことを話してみたいですか?
- (3) 「さつきさん」の伴奏による歌唱について、ご意見をお聞かせください。
- (4) 「さつきさん」のどんなところが良いと思いましたか? 特に認知症高齢者にとって良いと思うのはどのようなことですか?
- (5) 「さつきさん」のどんなところが問題だと思いましたか? それに対する改善策としては何が考えられると思いますか?

か? それに対する改善策としては何が考えられると思いますか?

(6) 「さつきさん」に今後、どのようなことを期待しますか?

(7) 将来的にどのような利用方法が良いと思いますか? 「さつきさん」をどのように利用してみたいですか?

(8) その他、ご意見・ご要望などございましたら、なんでも自由にお答えください。

これらの質問を通して得られた評価を以下にまとめる。

### 4.2 さつきさんとの会話について

さつきさんとの会話について、複数の人から得た評価は(1) もっと言ったことに反応して欲しい、(2) 名前を呼んでくれると良い、の2点である。言ったことに反応したり、名前を呼ぶためには、音声認識結果を形態素解析するなどして、受療者の発話から、主要な単語や名前を切り出す・認識する必要がある。特に、受療者を名前で呼ぶことは音楽療法の現場でも大切に考えられていることなので、できるだけ早く対象者の名前を聞いて復唱できるように修正したい。その上で、今後は音楽療法の現場で交わされる会話データを収集して、反応パターンのバリエーションを増やし、シナリオを充実させていきたいと考える。

さつきさんとの会話について、各介護スタッフから得られた良い評価は以下である。

- (1) 自分の言ったことを広げてくれる感じが良いと思う。
  - (2) 身近な質問を選んでいるところは良い。
  - (3) 私はこういうのが好きですよ、などと言っていたのは親しみやすく良かった。
- 身近な話題を選んでいることや、さつきさんからの質問に対する回答を受けて、さつきさんが自分のことを話すというやりとりも好意的に捉えられたようである。

一方、各介護スタッフから得られた指摘事項は以下である。

- (1) 感情が足りない。事務的にしゃべっているように感じる。
- (2) 共感するという感覚が足りない感じがする。「私もです」って言ってもらえると良い。

(3) 一方的に聞かれるだけではなくて、こちらからも話せるようになると良い。話している人が、まだ心を許せない感じがしてしまう。

(4) もっとコミュニケーションの取り方に時間をかけた方が良い。

これらの指摘に共通するキーワードは「共感」だと考えられる。もっと「寄り添ってくれている」という印象を与えられるような会話ができるように改良したい。

なお、「音楽のことを話題できると良いと思う」「だじゃれとか、ただの会話というよりも、笑いが出ちゃう感じだと楽しいと思う」といった、会話の内容に関するアドバイスも得ることができたので、これらの指摘はシナリオの改善に役立てたいと考えている。

#### 4.3 さつきさんの声、見た目、動きについて

さつきさんの声や外見等について、複数の人から得た評価は「表情が乏しい」だった。認知症を含めて、脳に障害がある場合は、認知機能の中の情報処理機能が低下している可能性があるため、生身の人間の複雑な表情の読み取りが困難なケースがある。そのため、さつきさんの表情があまり豊かでは無い方が良いという可能性が考えられる。しかし介護スタッフによると、認知症高齢者は非常によくスタッフや周囲の人の表情を見ている、とのことだったので、今後はどの程度の表情の豊かさが適切であるか調査が必要だと考えている。

さつきさんの声については、「聞き取りやすい声だった」という良い評価も、「もっと抑揚があった方が良い」「語尾を伸ばしたりすると親しみやすいかもしれない」などといった改善点も指摘された。抑揚は聞き取りやすさに関連するので調整を進めていきたい。

一方、さつきさんの外見や動作については、上述の「表情が乏しい」以外は、「目ははっきりしていて良い」「拍手する動作が良い」「かわいくて髪型も良いと思う」など、総じて評価が高かった。今後は動作のバリエーションを増やしていきたいと考えている。

#### 4.4 さつきさんの伴奏による歌唱について

さつきさんの伴奏による歌唱については、さつきさんも一緒に歌ってくれた方が良い、という指摘があった。ボーカロイド技術の利用を検討していきたい。

それ以外は、

- (1) こちらのリクエストに応じてもらえると良い
- (2) 地域ごとの季節の歌を聞ける・歌えると良い
- (3) 歌詞も画面に出ると良いと思う
- (4) これから、何が始まるんだろう続いていくんだろうといった、どきどき感とかわくわく感が足りない。今日は唱歌にしましょうとか、ちょっと期待感を持てるようなのがいいと思う。

(5) 歌い終わったら「上手ですよ」と言ってくれると良いなど、曲目や歌詞の表示、歌う前と歌い終わった後のコミュニケーションに関する指摘事項だった。

歌唱は音楽療法セッションの中でも主要なアクティビティであるため、今後、様々な改善点やアイデアが出てくる可能性が高いと考えられる。ぜひ AI 音楽療法士ならではの歌唱アクティビティの提供に繋がられるように手を広げていきたい。

#### 4.5 さつきさんへの更なる期待について

さつきさんへの更なる期待については、主に (1) さつきさん自身の発展、(2) さつきさんの利用方法、(3) さらに将来的な期待、の3つに分類できる。

##### 4.5.1 さつきさん自身の発展について

「〇〇さんいいですね、お口開いてますね、いい笑顔ですねのような、さつきさんがこっちを見ている感じが出るといいと思う」や、「季節感があると良い。クリスマスだったら着るものを変えてみるなど。」といった意見を頂いた。

認知症高齢者の介護施設でのアクティビティでは、「見てくれる」と高齢者が感じられるようにすることに対する要望は高いので、妥当な指摘だと考える。表情認識技術の導入を検討したい。また、認知症高齢者は自力での外出が困難なため、屋内で過ごす時間が長く、季節感を感じるのが難しい状況である。音楽療法士は季節感のある曲を選ぶと共に、季節感のある服装をしたり、小道具を用意したりして、屋内で過ごす時間の長い療養者に、季節感を感じてもらえるように配慮するので、さつきさんもそのようなことができるように改良していきたい。

##### 4.5.2 さつきさんの利用方法について

さつきさんの利用方法については多数の意見を頂いた。

- (1) 利用者さんにとっては、(音楽療法の)機会が増えるのは良いこと
- (2) 施設など、複数の利用者さんがいるところで会話できるようになると良い
- (3) 日々のデイサービスの中で、毎日の朝の会とかで活用できるとよい。
- (4) スタッフが記録を残したりしている間に AI が利用者さんの相手をしてくれるとよい
- (5) 独居の高齢者の方が受け入れやすいと思う

音楽療法に接する機会が増えることで、毎日利用することが想定されたり、忙しい介護スタッフをサポートする形での利用など、こちらが思いつかない利用方法を聞くことができた。

複数の高齢者に同時に対応するような、高齢者のグループでの利用はかなり難しいと考えられるので、まずは受療者1人を対象とする設定で完成度を高めていきたいと考えている。その上で、人数分のパソコンがあれば、さつきさんがそれぞれの高齢者に対応するので、そのような形での利用であれば施設などでもサービス展開することが可能と考えられる。

##### 4.5.3 さらに将来的な期待について

「違う人がでてくると、利用者さんにも刺激があってよい。」「利用者さんの反応の記録もさつきさんができるようになると良い」といった意見を聞くことができた。特に後者は、私達が目指している音楽療法のデジタルセラピューティクスに関連するものなので、研究の方向性が適切であることを確認することができた。

## 5 ま と め

本論文では、AI 音楽療法士の「さつきさん」の開発研究について説明した。AI 療法士の開発は容易ではないと考えられているが、我々は、音楽療法を「AI 音楽療法士と生身の音楽療法士の協業」と捉え、AI 音楽療法士が実施可能な業務を厳選してその完成度を高め、生身の音楽療法士との協業の形を新しく定義するという方針で研究を進めている。私達は AI 音楽療法士が実施可能な業務として、発話促進に特化した対話と歌唱伴奏に絞り込んだ。今後は 4 章に記載した認知症高齢者の介護に従事する介護スタッフによる評価に対する対応を中心に開発を進めると共に、さつきさんの利用データを収集し、認知症の重症度の客観評価や進行予測などに活かすための環境整備も進め、デジタルセラピューティクスとしての完成度を追求する予定である。そして、さつきさんと生身の音楽療法士との新しい協業の形を考案していく予定である。「さつきさん」により「会話を飢えている」と言われる高齢者が、いつでもどこでも何回でも気軽にさつきさんとの会話と歌唱を楽しみ、それによって認知症高齢者の BPSD が軽減し、それが介護者の負担軽減に繋がって、高齢者と介護者双方の QOL が向上することが期待される。

## 謝 辞

AI 音楽療法士の開発に協力して下さった株式会社 dott の清水俊之介氏、AI 音楽療法士のシナリオ作成に協力して下さった NAT 音楽療理事務所の認定音楽療法士の相川直子さんに感謝いたします。評価実験にご協力いただいた、「福祉ネットワーク・やえやま」の當山房子先生と、「あかゆら」の介護スタッフの皆さまに感謝いたします。本研究は JSPS 科研費 JP20K11059, JP23H03407、新潟医療福祉大学学長裁量研究費、AMED の課題番号:JP21dk0207056、専修大学情報科学研究所共同研究助成の助成を受けました。

## 文 献

- [1] 「臨床精神医学」編集委員会編集, 精神科臨床評価検査法マニュアル, 2004
- [2] 日本神経学会監修, 認知症疾患臨床ガイドライン 2017
- [3] 内閣府, 令和 4 年版高齢社会白書, [http://www8.cao.go.jp/kourei/whitepaper/w-2022/zenbun/04pdf\\_index.html](http://www8.cao.go.jp/kourei/whitepaper/w-2022/zenbun/04pdf_index.html)
- [4] 内閣府, 令和 5 年版高齢社会白書, [http://www8.cao.go.jp/kourei/whitepaper/w-2023/zenbun/05pdf\\_index.html](http://www8.cao.go.jp/kourei/whitepaper/w-2023/zenbun/05pdf_index.html)
- [5] 一般社団法人 日本音楽療法学会 入会のご案内, <https://www.jmta.jp/publish/pdf/information.pdf>
- [6] 一般社団法人 日本音楽療法学会, <https://www.jmta.jp>
- [7] 認知症ちえのわ net, <https://chienowa-net.com/>
- [8] 小杉尚子, 石井健太郎, 児玉直樹, AI 音楽療法士の開発研究-音楽療法のデジタルセラピューティクスにむけて-, 第 15 回データ工学と情報マネジメントに関するフォーラム, 2023
- [9] 小杉尚子, 児玉直樹, 清水幸子, 数井裕光, 遠隔音楽療法サービスの開発研究, 情報処理学会論文誌, 60(1), 193-202, 2019
- [10] 琉球新報, 音楽療法遠くでも OK ネット活用最新機器で音ずれず, <https://ryukyushimpo.jp/news/preentry-210864.html>
- [11] 小杉尚子, 児玉直樹, 清水幸子, 数井裕光, 認知症高齢者に対する遠隔音楽療法の効果, 日本遠隔医療学会雑誌, 15(2), 145-148, 2019, 優秀論文賞受賞
- [12] 承認番号:30200BZX00271000, 管理医療機器 禁煙治療補助システム (JMDN コード:71087002)
- [13] 小杉尚子, 小島明, 片岡良治, 串間和彦, 大規模音楽データベースのハミング検索システム, 情報処理学会論文誌 43(2), 287-298, 2002
- [14] 小杉尚子, 櫻井保志, 山室雅司, 串間和彦, SoundCompass: ハミングによる音楽検索システム, 情報処理学会論文誌, 45(1), 333-345, 2004
- [15] Kosugi N, Oshiyama C, Kodama N, Niwa S, Incorporating music therapy into cognitive remediation to improve both cognitive dysfunction and negative symptoms in schizophrenia, Schizophrenia Research, 204, 423-424, 2018
- [16] Kosugi N, Oshiyama C, Kodama N, Niwa S, Introduction of Music Therapy Incorporated into Cognitive Remediation: A New Approach to Cognitive Dysfunction in Psychiatric Disorders and a Preliminary Report on Its Effects in Schizophrenia, Open Journal of Psychiatry, 9(1), 23-38, 2019
- [17] Kosugi N, Oshiyama C, Kodama N, Niwa S, Predictability of a favorable outcome in schizophrenia associated with positive effects of “music therapy incorporated into cognitive remediation”, Schizophrenia Research, 238, 52-53, 2021
- [18] 日野原重明監修, 篠田知璋・加藤美知子編集, 標準音楽療法入門 (上) 理論編, 春秋社, 1998
- [19] 日野原重明監修, 篠田知璋・加藤美知子編集, 標準音楽療法入門 (下) 実践編, 春秋社, 1998
- [20] 遠山文吉, 阪上正巳, 岡崎香奈, 中野万里子, 屋部操, 国立音楽大学音楽研究所音楽療法研究部門編著, 音楽療法の現在, 株式会社人間と歴史社, 2007
- [21] <https://github.com/snakers4/silero-vad>
- [22] <https://github.com/ricky0123/vad>
- [23] <https://openai.com/blog/whisper/>
- [24] <https://cloud.google.com/text-to-speech>
- [25] <https://threejs.org>
- [26] 東中竜一郎, 稲葉通将, 水上雅博, Python でつくる対話システム, 株式会社オーム社, 2020