INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Automatic Speech Recognition systems errors for objective sleepiness detection through voice

Vincent P. Martin[1], *Jean-Luc Rouas*[1], *Florian Boyer*[1,2], *Pierre Philip*[3]

[1]LaBRI, Univ. de Bordeaux, CNRS – UMR 5800, Bordeaux INP, Talence (France)
[2]Speech Lab, Airudit, Pessac (France)
[3]SANPSY, Univ. de Bordeaux, CNRS – USR 3413, CHU Bordeaux, Bordeaux (France)
{vincent.martin, rouas}@labri.fr, florian.boyer@airudit.com, pierre.philip@u-bordeaux.fr

Abstract

Chronic sleepiness, and specifically Excessive Daytime Sleepiness (EDS), impacts everyday life and increases the risks of accidents. Compared with traditional measures (EEG), the detection of objective EDS through voice benefits from its ease to be implemented in ecological conditions and to be sober in terms of data processing and costs. Contrary to previous works focusing on short-term sleepiness estimation, this study focuses on long-term sleepiness detection through voice. Using the Multiple Sleep Latency Test corpus, this study introduces new features based on Automatic Speech Recognition systems errors, in an attempt to replace hand-labeled reading mistakes features. We also introduce a selection feature pipeline inspired by clinical validation practices allowing ASR features to perform on par with the state-of-the-art systems on short-term sleepiness detection through voice (73.2% of UAR). Moreover, we give insights on the decision process during classification and the specificity of the system regarding the threshold delimiting the two sleepiness classes, Sleepy and Non-Sleepy.

Index Terms: Sleepiness, Excessive Daytime Sleepiness, Automatic Speech Recognition, Voice

1. Introduction

Chronic sleepiness, and specifically Excessive Daytime Sleepiness (EDS), impacts everyday life quality [1] and increases the accidental risks of people affected by it. Indeed, in case of antecedents of sleepiness at the wheel, people affected by EDS are between two and three times more likely to have a road accident [2]. Moreover, with a prevalence in the general population estimated between 10 % and 25 % [3], EDS is a public health concern.

EDS is usually measured either objectively (EEG) or subjectively (questionnaires). To measure objectively EDS, the gold-standard measure is the Multiple Sleep Latency Test (MSLT) [4]. This test consists of asking the patients to take a nap five times during a day and to measure through EEG (objectively) their sleep latency at each nap. Although this test is widely used, for example in narcolepsy diagnostic [5], it is very expensive: not only the patients are hospitalized during two nights and one entire day but the test requires qualified staff to monitor and interpret the EEG, and the equipment is costly and requires consumables [6].

Psychometric questionnaires, such as the Epworth Sleepiness Scale (ESS) [7], do not suffer the same drawbacks. Generally fillable in few minutes and requiring only a pen, they aim at measuring subjective EDS. However, what they exactly measure is still discussed and the obtained score does not necessarily correlate with objective measures [8, 9].

This study proposes a third modality: detection of EDS through voice. Indeed, when speakers are sleepy, their articulation and prosody are affected in a typical manner, allowing the estimation of their sleepiness [10]. This has the advantage to be implementable in various environments – including open environments outside laboratories, and it is not invasive and more comfortable for the patients, who do not have to wear electrodes. It does not require complex calibration or specific sensors and it is economical in terms of data processing, but also in terms of equipment. Moreover, this modality could be easily integrated into cutting-edge technologies such as virtual medical interviews [11].

If the detection of short-term sleepiness has already been the subject of two international challenges [12, 13], long-term sleepiness (i.e. EDS) has recently shown a rising interest. Based on the MSLT corpus (MSLTc) [14], two approaches have already been proposed to detect objective EDS through voice (binary classification between Sleepy – SL and Non-Sleepy – NSL). In [15], a system based on acoustic features explainable to physicians has achieved an Unweighted Average Recall (UAR) of 63.8%. Unfortunately, these performances are not sufficient to use the system in the medical field. Another approach relying on the hand-labeling of the mistakes the patients make during their reading has been proposed in [16] and achieved 86.2% of UAR [15]. Nevertheless, these features still represent a high human effort as they require time and specific qualifications to annotate the recordings.

The objective of this study is twofold. First, in an attempt to automatize these annotations, we propose to study Automatic Speech Recognition (ASR) systems errors to automatically estimate EDS and eventually replace manually annotated reading errors. Second, we seek to validate these new features through a feature selection pipeline inspired by clinical validation of psychometric questionnaires.

This paper is organized as follows. Section 2 introduces the MSLTc and the objective EDS label used in this study. We introduce the features in Section 3 and the feature selection and classification pipeline in Section 4. Performances are presented in Section 5 and we discuss the selected features in Section 6. Finally, we conclude and draw future works in Section 7.

2. Corpus and EDS label

2.1. MSLT corpus

The corpus used in this study is the MSLT corpus (MSLTc), relying on the recordings of 106 patients of the Sleep Clinic of the Bordeaux University Hospital [14]. They undertake a Multiple Sleep Latency Test: every two hours between 9 am and 5 pm, the patients are asked to take a nap that has a maximum

http://dx.doi.org/10.21437/Interspeech.2021-291

duration of 35 minutes. Before each nap, the patients are asked to read out loud a 200 words text extracted from *Le Petit Prince* of Saint-Exupéry. As a consequence, each speaker of the corpus is recorded 5 times during the same day, with different texts and different emotional, fatigue, and circadian states. To ensure coherence of the speakers, we only keep the 93 patients of the corpus affected by different forms of Hypersomnia.

2.2. Objective EDS label

During each nap, the sleep latency – i.e. the time between the beginning of the test and the moment the patient falls asleep – is assessed by electroencephalography (EEG). This value is measured in minutes and ranges between 0 and 20 minutes: if the patients did not fall asleep during the first 20 minutes of the test, lights are switched on and the test is stopped. The objective EDS of the patient is measured by the average sleep latency across the nap: a mean MSLT sleep latency under 8 minutes is usually associated with objective EDS [4].

Table 1: *Distribution of the speakers across Sex and Sleepiness class. SL: Sleepy (MSLT $\leq$ 8 min.), NSL: Non-Sleepy (MSLT $>$ 8 min.)*

Sex	SL	NSL	TOTAL
Women	10	48	58
Men	11	24	35
TOTAL	**21**	**72**	**93**

3. Features

3.1. Acoustic features

The acoustic features used in this study are presented in detail in [17, 18]. They are twofold: on one side, statistics (length and ratio) of voiced and vocalic parts are automatically extracted from audio; on the other side acoustic features are computed on these parts (Harmonics and Formants amplitude and bandwidth, Harmonic to Noise ratio, ...).

3.2. Hand-labeled reading mistakes

These features rely on the hand-labeling of the errors the patients make during the reading of the texts. Elaborated with speech therapists, four errors have been studied for sleepiness detection: stumbling errors (”hesitations and breaks in the speech rhythm” [19]), paralexia (i.e. ”identification error of written words consisting in the production of a word instead of another” [19]), deletions of words and addition of words. Introduced in [16], these features have shown promising results in detecting objective EDS [15]. They are referred to as ”R. errors” in the following.

3.3. ASR errors

Annotating the previous errors is time-consuming and requires training to differentiate errors. In an attempt to automatize the labeling of reading errors, we measured the errors made by ASR systems. Indeed, when subjects are sleepy, their articulation and prosody are impaired [10] while the number of hesitations and repeats increases. This alteration of speech due to sleepiness may induce errors in ASR systems that could be used as biomarkers of sleepiness. Thanks to recent advances on end-to-end ASR systems allowing intermediate transcription units such as characters or tokens, it is now possible to transcribe not only words but also portions of words (Byte Pair Encoding – BPE).

In this study, we use an end-to-end system using RNN transducters with attention, based either on words, BPE, or characters, to transcribe words or BPE. The language model is trained on a word, BPE, or character version of the ESTER corpus [20]. A complete review of such systems and their performances is proposed in [21]. The end-to-end system achieving the best performances is the character-based one with a word-based RNN language model achieving 17.6% of Word Error Rate on the ESTER corpus.

In line with the previous results on reading mistakes [15], we consider two types of errors in this study: insertions and substitutions. Each type of error is computed on tokens and on words, and we consider both the raw number of errors and their proportion over the total number of transcription units, leading to 8 features per system.

4. Classification pipeline

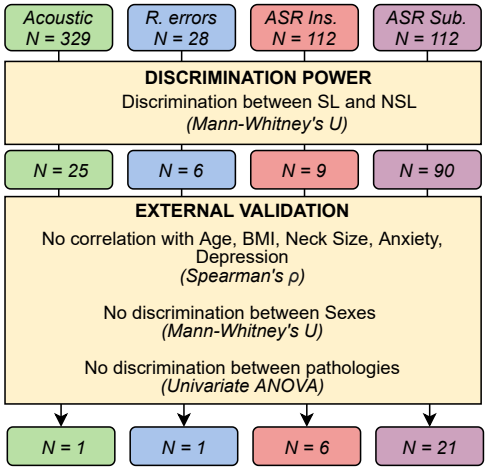


Figure 1: *Feature selection pipeline. R. errors: reading errors; ASR: Automatic Speech Recognition errors, Ins: insertions; Sub: substitutions. SL: Sleepy (MSLT $\leq$ 8 min.), NSL: Non-Sleepy (MSLT $>$ 8 min.). BMI: Body Mass Index*

4.1. Feature selection

For each speaker, each previously presented set of features is computed on each of the 5 naps of the MSLT, to which we aggregate the mean and the standard deviation across the naps, resulting in 7 measures for each feature and speaker. The proposed feature selection pipeline is represented in Figure 1. The selected features have to comply with two constraints:

- Their distribution for each sleepiness class, measured by a Mann-Whitney's U ($p < 0.05$) has to be statistically different (discrimination power of the features);
- They should not correlate (Spearman's ρ, $p > 0.05$) with age, Body Mass Index (BMI), neck size, anxiety or depression (measured by the Hospital Anxiety and Depression scale [22]); they should not discriminate sex (Mann Whitney's U, $p > 0.05$) or pathologies (Univariate ANOVA, $p > 0.05$).

This pipeline, inspired by the external validation of psychometric questionnaires, ensures that the selected features classify

only sleepiness, independently from the other measured speaker traits. Indeed, even if some of these factors could correlate with sleepiness, our aim is to train the classifier to learn the concept of sleepiness, not to learn a confounding factor correlating with sleepiness, that could still give good classification performances but make the interpretation of such results impossible.

Moreover, compared with performance-driven feature selection pipelines, this one works with few samples – statistical tests do not require large amounts of data – and is independent of the chosen metric: however the performances of the system are measured, the selected features remain the same.

4.2. Classification

To ensure generalization and avoid overlearning, the classification is carried out under Leave One Speaker Out Cross-Validation: each speaker is turn-by-turn isolated as a test speaker, while the classification system is trained on the others. Estimated and ground-truth sleepiness classes of the test speaker are stacked and the classification metrics are computed on this aggregation.

As the goal of this study is not to optimize the best possible classifier but to validate the use of new features for objective sleepiness estimation through voice, the previously selected features are aggregated (early-fusion), scaled, orthogonalized by a Principal Components Analysis (PCA) and classified by a logistic regression using the Python module sci-kit learn [23] with a *newton-cg* solver and a balanced class-weighting.

5. Results

5.1. Classification performances

Table 2: *Classification performances of the proposed pipeline. UAR: Unweighted Average Recall, F1-score: class-weighted average F1-score, AUC: Area Under the ROC Curve. R. errors: Reading errors, ASR: Automatic Speech Recognition system errors.*

	Features	UAR	F1	AUC
(a)	ASR	**73.2%**	**75.8%**	**74.8%**
(b)	R. errors	57.7%	73.4%	22.1%
(c)	Acoustic	59.5%	66.0%	60.1%
(d)	ASR + R. errors	71.8%	74.0%	74.5%
(e)	ASR + Acoustic	73.2%	75.9%	74.4%
(f)	R. errors + Acoustic	61.3%	70.2%	67.7%
(g)	All	73.9%	76.8%	74.6%

The obtained Unweighted Average Recall (UAR), weighed F1-score, and Area Under the ROC Curve (AUC) for the different feature combinations are presented in Table 2.

The best performances are obtained by the system (g), aggregating the three sets of features: this system achieves 73.9% of UAR, 76.8% of weighted F1-score, and 74.6% of AUC. In this system, the selected reading error is the number of additions on the fourth nap, and the selected acoustic features are the bandwidth of the first Formant on the first nap. The selected ASR errors are detailed below.

However, the ASR features alone (system (a)) achieve classification performances that are only slightly below (73.2% of UAR, 75.8% of F1-score, and 74.8% of AUC): the acoustic and reading errors seem to have little importance on the classification. Regarding the cost-benefices balance of the hand-labeled reading mistakes in comparison with the small performance enhancement they are the cause of when combined with ASR and acoustic features (0.7% of absolute improvement regarding UAR, 1% regarding the weighted F1-score), we choose to discard these features. As the combination between ASR and acoustic features (system (c)) achieves poorer results than ASR features alone, we choose to focus on system (a), based only on ASR features.

5.2. Performances of the selected system

The ROC curve of the systems (a) and (g) and the confusion matrix of the system (a) are respectively represented in Figures 2a) and 2b). The ROC curve confirms the close similarity of performances of systems (a) and (g) and consolidates the choice to focus on system (a).

Moreover, inspired by a previous study [18], we represented in Figure 2c) the performances of the system (a) depending on the threshold to distinguish SL from NSL. This graph represents the specificity of the selected features to the phenomena they aim at measuring. As intended, the best performances are obtained for a threshold of 8 minutes. Moreover, excepting a crook for 7.5 minutes, these features achieve performances higher than 70% for thresholds between 7.0 minutes and 9.5 minutes, allowing physicians to select the severity of objective EDS they want to detect.

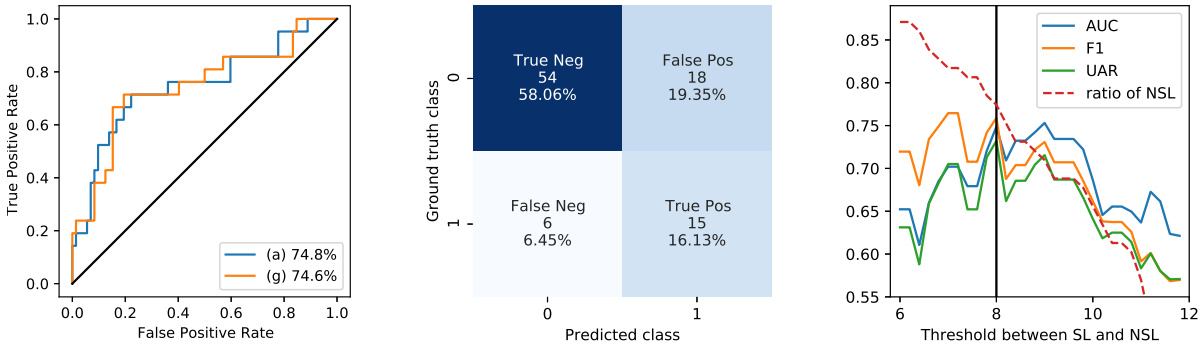
6. Discussions

6.1. PCA Analysis

Along the cross-validation process, the parameters of the PCA and the weights of the logistic regression are averaged. Figure 3 represents the four different PCA dimensions and their averaged corresponding weights in the classification. The most important PCA component in the decision of the classifier is the fourth dimension, relying on the difference between the number of insertions during the third nap and the standard deviation of the substitutions (mean coefficient of the logistic regression: $\alpha_4 = 0.85$). The sum of the same features directs the third dimension of the PCA, which has a weight of $\alpha_3 = -0.22$ in the classifier decision. The insertions errors measured at the third nap are made by a character-based ASR system with a word-based language model, whereas the involved substitutions are made by a BPE-based ASR system without a language model. The second most important dimension is directed by the substitutions measured on second, fourth, and fifth naps, and the mean value across the naps ($\alpha_1 = 0.29$). These errors come from 7 ASR systems based on different units, with and without language models. Finally, the least important PCA dimension relies on the number of insertions on the first and the third naps ($\alpha_2 = -0.05$). Contrary to the insertions of the third nap, the selected insertions of the first nap are made by a BPE-based ASR system without a language model.

6.2. Measures of the features

Regarding the selected features, the insertions seem to be relevant precisely on the first and third naps, excluding the others. When studying the selected features after the first step of the validation process, the insertions of the first, third but also fourth nap and their standard deviation across the naps discriminate objective EDS. However, insertions during the fourth nap correlate significantly with the BMI (Spearman's ρ, $p < 0.05$) and their standard deviation across naps discriminate patholo-



(a) (b) (c)

Figure 2: ***(a)*** *ROC of the system (a) and (g) and their corresponding AUC.* ***(b)*** *Confusion matrix of the system (a)* ***(c)*** *Performances of the system (a) depending on the threshold between Sleepy – SL – and Non-Sleepy – NSL – classes*

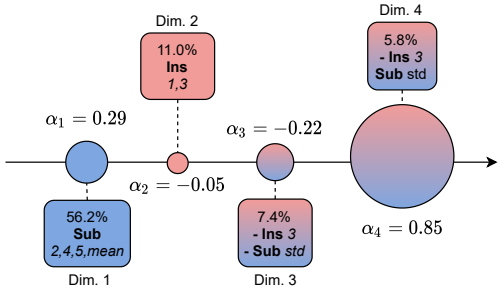


Figure 3: *PCA components and their associated weight (α_i) in the logistic regression for the system (a). For each box, from top to bottom: mean ratio of explained variance in the PCA, feature (in bold), nap during which it is measured (in italic).*
Red background: Insertions, Blue background: Substitutions. - : negative PCA weight; Sub: substitutions; Ins: insertions. Dim: Dimension

gies (Univariate ANOVA, $p < 0.05$). This could be explained by two phenomena. First, the texts are different for each iteration of the MSLT procedure. The ASR systems having an inherent error rate depending on the content of the processed text reading, it may be possible that the third text could be the only one on which the link between insertions and objective EDS is distinguished independently from the other speaker traits. Second, the speakers are recorded five times during the day, in different emotional, fatigue, or circadian states, filtering the expression of the speakers' traits. Indeed, the first recording is made at 9 am after breakfast and patients lunch few minutes before the third nap: the state induced by taking a meal could favor voice phenomena inducing the ASR system to produce insertions errors linked with sleepiness, but discriminating it from other traits.

7. Conclusion and perspectives

After validating the usefulness of reading mistakes analysis for sleepiness detection, we wanted to fully automatize the process by removing any human intervention in the labeling of errors. We thus proposed to use the outputs of several ASR systems based on end-to-end architectures using different target units (words, characters, and BPE). The analysis of the errors produced by the systems when compared to the reference transcription of the texts allowed us to reach a satisfying 73.2% of UAR in objective sleepiness classification with a fully automated pipeline, compared to 55.7% UAR for manually annotated reading mistakes. As a perspective, we will focus on improving the feature selection process using for example decorrelation techniques. Furthermore, we will validate the proposed pipeline for the prediction of other objective and subjective sleepiness measurements.

8. References

[1] M. M. Ohayon, C. F. Reynolds, and Y. Dauvilliers, "Excessive sleep duration and quality of life: Excessive Sleep in USA," *Annals of Neurology*, vol. 73, no. 6, pp. 785–794, Jun. 2013. [Online]. Available: http://doi.wiley.com/10.1002/ana.23818

[2] P. Philip, P. Sagaspe, E. Lagarde, D. Leger, M. M. Ohayon, B. Bioulac, J. Boussuge, and J. Taillard, "Sleep disorders and accidental risk in a large group of regular registered highway drivers," *Sleep Medicine*, vol. 11, no. 10, pp. 973–979, Dec. 2010.

[3] T. B. Young, "Epidemiology of daytime sleepiness: definitions, symptomatology, and prevalence," *The Journal of Clinical Psychiatry*, vol. 65 Suppl 16, pp. 12–16, 2004.

[4] D. Arand, M. Bonnet, T. Hurwitz, M. Mitler, R. Rosa, and R. B. Sangal, "The Clinical Use of the MSLT and MWT," *SLEEP*, vol. 28, no. 1, pp. 123–144, 2005.

[5] M. S. Aldrich, R. D. Chervin, and B. A. Malow, "Value of the multiple sleep latency test (MSLT) for the diagnosis of narcolepsy," *Sleep*, vol. 20, no. 8, pp. 620–629, 1997.

[6] A. S. Association, "Multiple Sleep Latency Test," Mar. 2021. [Online]. Available: https://www.sleepassociation.org/sleep-treatments/multiple-sleep-latency-test/

[7] M. W. Johns, "A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale," *Sleep*, vol. 14, no. 6, pp. 540–545, 1991.

[8] R. Sangal, "Subjective sleepiness ratings (Epworth sleepiness scale) do not reflect the same parameter of sleepiness as objective sleepiness (maintenance of wakefulness test) in patients with narcolepsy," *Clinical Neurophysiology*, vol. 110, no. 12, pp. 2131–2135, 1999.

[9] E. Evangelista, A. L. Rassu, L. Barateau, R. Lopez, S. Chenini, I. Jaussent, and Y. Dauvilliers, "Characteristics associated with hypersomnia and excessive daytime sleepiness identified by extended polysomnography recording," *Sleep*, Nov. 2020. [Online]. Available: https://academic.oup.com/sleep/advance-article/doi/10.1093/sleep/zsaa264/6010320

[10] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, and B. Schuller, "Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech," *Neurocomputing*, vol. 84, pp. 65–75, 2011.

[11] P. Philip, L. Dupuy, M. Auriacombe, F. Serre, E. de Sevin, A. Sauteraud, and J.-A. Micoulaud-Franchi, "Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients," *npj Digital Medicine*, vol. 3, no. 1, p. 2, 2020.

[12] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Interspeech 2011*, 2011, pp. 3201–3204.

[13] B. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychocz, R. Vollman, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. Warlaumont, L. Yankowitz, E. Nöth, S. Amiriparian, S. Hantke, and M. Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *Interspeech 2019*, 2019.

[14] V. P. Martin, J.-L. Rouas, J.-A. Micoulaud-Franchi, and P. Philip, "The Objective and Subjective Sleepiness Voice Corpora," in *12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020, p. 6525-6533. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.803.pdf

[15] V. P. Martin, J.-L. Rouas, and P. Philip, "Détection de la somnolence dans la voix : nouveaux marqueurs et nouvelles stratégies," *Traitement Automatique des Langues*, vol. 61, no. 2, pp. 67–90, 2020.

[16] V. P. Martin, G. Chapouthier, M. Rieant, J.-L. Rouas, and P. Philip, "Using reading mistakes as features for sleepiness detection in speech," in *10th International Conference on Speech Prosody 2020*, Tokyo, Japan, 2020, pp. 985–989. [Online]. Available: http://dx.doi.org/10.21437/SpeechProsody.2020-201

[17] J.-L. Rouas, T. Shochi, M. Guerry, and A. Rilliard, "Categorisation of spoken social affects in Japanese: human vs. machine," in *ICPhS*, 2019.

[18] V. P. Martin, J.-L. Rouas, P. Thivel, and J. Krajewski, "Sleepiness detection on read speech using simple features," in *10th Conference on Speech Technology and Human-Computer Dialogue*, Timisoara, Romania, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8906577

[19] F. Brin, C. Courrier, E. Lederle, and V. Masy, *Dictionnaire d'orthophonie - 4ème édition*, orthoedition ed., Sep. 2018.

[20] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts," in *Interspeech 2009*, 2009, pp. 2583–2586.

[21] F. Boyer and J.-L. Rouas, "End-to-End Speech Recognition: A review for the French Language," *[Unpublished]*, 2019, arXiv: 1910.08502. [Online]. Available: http://arxiv.org/abs/1910.08502

[22] A. S. Zigmond and R. P. Snaith, "The hospital anxiety and depression scale," *Acta Psychiatrica Scandinavica*, vol. 67, no. 6, pp. 361–370, 1983.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.