2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Feature Fusion Strategies for End-to-End Evaluation of Cognitive Behavior Therapy Sessions

Zhuohao Chen[1], Nikolaos Flemotomos[1], Victor Ardulov[1], Torrey A. Creed[2],
Zac E. Imel[3], David C. Atkins[4] and Shrikanth Narayanan[1]

***Abstract*— Cognitive Behavioral Therapy (CBT) is a goal-oriented psychotherapy for mental health concerns implemented in a conversational setting. The quality of a CBT session is typically assessed by trained human raters who manually assign pre-defined session-level behavioral codes. In this paper, we develop an end-to-end pipeline that converts speech audio to diarized and transcribed text and extracts linguistic features to code the CBT sessions automatically. We investigate both word-level and utterance-level features and propose feature fusion strategies to combine them. The utterance level features include dialog act tags as well as behavioral codes drawn from another well-known talk psychotherapy called Motivational Interviewing (MI). We propose a novel method to augment the word-based features with the utterance level tags for subsequent CBT code estimation. Experiments show that our new fusion strategy outperforms all the studied features, both when used individually and when fused by direct concatenation. We also find that incorporating a sentence segmentation module can further improve the overall system given the preponderance of multi-utterance conversational turns in CBT sessions.**

***Index Terms*—Cognitive behavioral therapy, Motivational Interviewing, end-to-end evaluation, feature fusion strategies**

I. INTRODUCTION

In psychotherapy assessment, the quality of a session is generally evaluated through the process of behavioral coding in which experts manually identify and annotate behaviors of the participants [1]. However, this procedure is time-consuming, which makes it resource-heavy in terms of human capital and therefore often unfeasible in most treatment contexts. In recent years, researchers have developed automated behavioral coding algorithms using speech and language features for several clinical domains [2]–[4].

Cognitive Behavioral Therapy (CBT) is evidence-based psychotherapy predicated on the cognitive model involving shifts in the patient's thinking and behavioral patterns [5]. As a common type of talk therapy, CBT has been developed for many decades and become an effective treatment for a wide range of mental health conditions [6]. Extending upon this strong evidence base, recent research has explored whether combining CBT with other evidence-based psychotherapies might potentiate treatment outcome. For example, studies indicate that adding Motivational Interviewing (MI) as an adjunct to CBT may benefit patients by increasing motivation for and commitment to the intervention [7].

One of the early computational behavioral coding efforts for CBT is found in [8] which employed an end-to-end evaluation pipeline that overcomes the need of manual transcription and coding. This work formulated the CBT session quality evaluation as a classification task and compared the performance of various lexical features.

In this paper, we develop a new automated approach to assess CBT session quality and relate the MI and CBT in the computational behavioral modeling. Specifically, we utilize MI data to extract utterance-level features due to the similarities between MI and CBT and propose a novel fusion strategy. We experiment on both manual transcripts and automatically derived ones to show the superiority of the new fusion approach and the robustness of our automated evaluation system.

II. DATASETS

The CBT data, with accompanying audio-recorded sessions, used in this work come from the Beck Community Initiative [9]. The CBT quality is evaluated by the session-level behavioral codes based on Cognitive Therapy Rating Scale (CTRS) [10]. Each session receives 11 codes scored on a 7 point Likert scale ranging from 0 (poor) to 6 (excellent) for each evaluated dimension. We also compute the total CTRS by summing up the scores as an overall measurement of the quality of a session. Raters were doctoral-level experts who were required to demonstrate calibration prior to coding process to prevent rater drift, which resulted in high inter-rater reliability for the CTRS total score (ICC = 0.84) [9].

In this paper, we use 225 coded CBT sessions for experiments which manually transcribed with talk turns, speaker roles, and punctuation. The sessions were recorded at a 16kHz sampling rate and their lengths range from 10 to 90 minutes. We binarized the CTRS codes by assigning codes greater or equal to 4 as "high" and less than 4 as "low" since 4 is the primary anchor indicating the skill is fully present, but still with room for improvement [10]. The threshold indicative of CBT competence on the total CTRS is 40 [11]. The label distributions are shown in Table I.

III. APPROACH

Our evaluation approach includes two stages. In the first stage, we took the session recordings as inputs and used a speech processing pipeline to substitute manual transcription. In the second stage, we extracted the linguistic features

*This work was supported by the National Institutes of Mental Health (NIMH), grant number R56 MH118550.

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA sail.usc.edu

[2]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA tcreed@pennmedicine.upenn.edu

[3]Department of Educational Psychology, University of Utah, Salt Lake City, UT, USA zac.imel@utah.edu

[4]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA datkins@u.washington.edu

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

TABLE I

CBT BEHAVIOR CODES DEFINED BY THE CTRS MANUAL

Abbr.	CTRS Code	low/high
ag	agenda	131/94
at	application of cognitive-behavioral techniques	150/75
co	collaboration	111/114
fb	feedback	150/75
gd	guided discovery	146/79
hw	homework	165/60
ip	interpersonal effectiveness	47/178
cb	focusing on key cognitions and behaviors	122/103
pt	pacing and efficient use of time	135/90
sc	strategy for change	126/99
un	understanding	123/102
total	total score	134/91

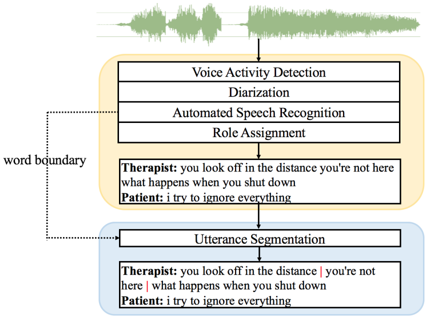

Voice Activity Detection
Diarization
Automated Speech Recognition
Role Assignment
word boundary
Therapist: you look off in the distance you're not here what happens when you shut down
Patient: i try to ignore everything
Utterance Segmentation
Therapist: you look off in the distance | you're not here | what happens when you shut down
Patient: i try to ignore everything

Fig. 1. Session Decoding Pipeline

from the therapist's transcripts to predict the binarized label of each code. The classification tasks are performed by a linear Support Vector Machine (SVM) with sample weights inversely proportional to their class frequencies.

A. Speech processing pipeline

To automatically transcribe the recorded sessions we adopted the speech pipeline described in [12] consisting of Voice Activity Detection (VAD), diarization, Automatic Speech Recognition (ASR) and role assignment presented by the yellow box in Fig. 1. The diarization error rate (including VAD errors) and ASR word error rate for the transcribed sessions are 21.47% and 44.01%, respectively. Error analysis revealed that the errors were highly inflated by the speech fillers (e.g., 'um', 'huh', etc). The role assignment module is trained to distinguish the therapist and patient in a counseling session and the annotation accuracy for the transcribed CBT sessions is 100% (225/225).

As shown in Fig. 1, the output of the speech pipeline in the yellow box is at the turn level without any punctuation. There might be multiple utterances within a turn, something which potentially affects the quality of utterance-level lexical features. Thus, we implemented an utterance segmentation module at the end of the pipeline. We made use of the word boundaries to split the text whenever the pause between consecutive words is more than 2 seconds, and then segmented the transcripts into utterances. The package we applied for utterance segmentation is an open source tool called "DeepSegment" [13] which achieves an F1 score of 0.7364 on the transcribed sessions.

B. Baseline mid-level features

We extract a number of different mid-level features from the transcribed text. The first set includes the term frequency - inverse document frequency (tf-idf) [14] transform of n-grams, while the second focuses on estimated Dialog Acts (DA) [15]. The tf-idfs and DAs were reported to achieve the best overall performance among the interpretable features in [8]. The third feature set considered here is inspired by utterance level codes drawn from MI. Under the hypothesis that there are shared characteristics between MI and CBT [7], we experimentally investigate the usefulness of MI based "features" in contributing to the quality assessment of CBT.

We extract all these features for the therapist side of the conversation only, because, as reported in [8], they perform robustly for the task of behavioral coding, and further fusing features of the two roles (i.e., therapist and patient) does not lead to substantial improvements. We compute the tf-idfs over unigrams. We additionally tag each utterance in a CBT session by one DA from the 7-class scheme described in Table II. We used a linear chain Conditional Random Field (CRF) model trained on the Switchboard-DAMSL dataset [16] which achieves 84.78% accuracy of the in-domain test set. For the DA-based feature representation we 1) count the utterances coded with each DA and normalize the counts with respect to the total number of utterances in each session; 2) count the words in the utterances tagged by each DA and normalize the count with respect to the total number of words in each session. Concatenating the two sets, we get a DA feature set of 7 × 2 = 14 dimensions.

To capture MI-like approaches used within a CBT session, we use specific utterance-level representations that describe MI relevant behaviors. In particular, we employ the set of Motivational Interviewing Skills [17] codes described in [3] and summarized in Table II. We cluster 'RES' and 'REC' into one class 'RE' since they are easily confused with each other [18]. We extract the MI relevant behavioral codes (MC, henceforth) the same way as in [18] which uses a neural architecture stacking an embedding layer, a bi-LSTM with attention layer and a dense layer. We train the model on the MI corpus used in [3] with train/validation/test split equal to 3/1/1 and the classification accuracy on the MI test set is 81.10%. The final MC-based feature representation is the same as the DA-based described previously. As observed in Table II, the DAs focus on the function of the dialog structure, while the MCs emphasize on the critical and causal elements deemed useful in the psychotherapy.

The tf-idfs are computed with regards to the occurrence of words in the sessions while the DAs and MCs are both annotations extracted at the utterance level. On this basis, we group the basic features into word-level features (tf-idfs) and utterance-level features (DAs, MCs).

IV. Feature Fusion Strategies

In this section, we discuss two feature fusion methods for combining the word-level and the utterance-level features.

A. Fusion by concatenation

The first fusion approach is straightforward, namely concatenation of the different feature sets. The hypothesis here is that the fused feature sets are complementary to each other so that they jointly carry richer information. Herein we combine the word-level feature tf-idfs with each of the utterance-level features (DAs, MCs) and denote the fused feature sets as tf-idfs + DAs and tf-idfs + MCs, respectively.

TABLE II

DETAILS OF DA AND MC (MI BEHAVIORAL CODES)

Coding Schemes	Codes
DA	Question, Statement, Agreement, Other Appreciation, Incomplete, Backchannel
MC	Facility (FA), Giving Information (GI),Reflection (RE), Closed Question (QUC), Open Question (QUO), MI Adherent (MIA), MI Non-Adherent (MIN)

B. *Augmenting words with utterance tags*

When we compute word-level features like tf-idfs and bag of words, contextual information is ignored. For example, the word ”homework” (an important element within CBT) in a question may denote that the therapist is checking if the patient has completed the given assignment, while in a reflection it might imply that the therapist is describing/confirming the assignment to/with the patient. The distribution of (just the) word ”homework” helps us evaluate how well a therapist incorporates the use of homework relevant to CBT. To incorporate the context in which they are used, we propose a fusion strategy of augmenting words with utterance level information.

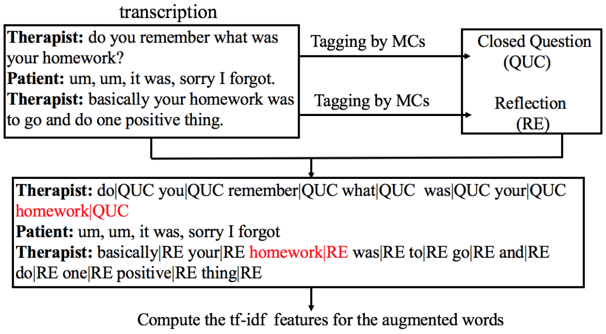

transcription
Therapist: do you remember what was your homework?
Patient: um, um, it was, sorry I forgot.
Therapist: basically your homework was to go and do one positive thing.
Tagging by MCs
Tagging by MCs
Closed Question (QUC)
Reflection (RE)
Therapist: do|QUC you|QUC remember|QUC what|QUC was|QUC your|QUC homework|QUC
Patient: um, um, it was, sorry I forgot
Therapist: basically|RE your|RE homework|RE was|RE to|RE go|RE and|RE do|RE one|RE positive|RE thing|RE
Compute the tf-idf features for the augmented words

Fig. 2. Word Augmentation. The features are extracted for the therapist side only. (MC: MI behavioral codes)

We show an example of the word augmentation we propose using MCs in Fig. 2. We first tag the therapist’s utterances by the model trained in Section III-B and then pad the words with the label of the utterance they belong to. In Fig. 2 the augmented tokens “homework|QUC” and “homework|RE” are viewed as different words for further analysis. Finally, we extract the tf-idfs based on the augmented words of the therapist to obtain the fused features.

Similar to the previously-mentioned feature concatenation method, we fuse the augmented tf-idfs with each of the DAs and MCs and denote the fused feature sets as DA-tf-idfs and MC-tf-idfs, respectively.

V. EXPERIMENTAL RESULTS

We compute the tf-idfs, DA-tf-idfs and MC-tf-idfs using the TfidfVectorizer from the scikit-learn package [19]. We set the parameters max_df=0.95 and min_df=0.05 to ignore terms that appear in more than 95% or less than 5% of the documents and select the K best features based on cross-validation on the total CTRS using a univariate F-test. All the feature sets are z-normalized before being fed into the linear SVM classifier. A 5-fold cross-validation is conducted to report the F1 score of each CTRS code and the total CTRS.

TABLE III

F1 SCORES OF THE TASKS ON THE MANUAL TRANSCRIPTS.

	tf-idfs	DAs	MCs	tf-idfs +DAs	tf-idfs +MCs	DA-tf-idfs	MC-tf-idfs
ag	0.76	0.60	0.65	0.75	0.76	0.77	0.80
at	0.70	0.60	0.61	0.70	0.69	0.71	0.73
co	0.75	0.63	0.64	0.74	0.75	0.75	0.80
fb	0.75	0.56	0.63	0.75	0.76	0.74	0.76
gd	0.74	0.58	0.61	0.72	0.77	0.76	0.73
hw	0.66	0.55	0.61	0.66	0.70	0.70	0.68
ip	0.54	0.51	0.55	0.57	0.53	0.57	0.58
cb	0.73	0.55	0.69	0.75	0.75	0.77	0.77
pt	0.69	0.54	0.66	0.69	0.73	0.74	0.75
sc	0.73	0.60	0.61	0.74	0.76	0.76	0.76
un	0.74	0.56	0.58	0.74	0.76	0.73	0.75
avg	0.71	0.57	0.62	0.71	0.72	0.73	**0.74**
tot	0.78	0.62	0.67	0.77	0.78	0.81	**0.83**

TABLE IV

F1 SCORES OF THE TASKS ON THE AUTOMATICALLY DERIVED TRANSCRIPTS FROM THE SPEECH PIPELINE.

	tf-idfs	DAs	MCs	DA-tf-idfs	MC-tf-idfs
ag	0.75	0.62	0.64	0.77	0.78
at	0.69	0.59	0.61	0.73	0.73
co	0.73	0.61	0.66	0.74	0.75
fb	0.75	0.58	0.64	0.75	0.77
gd	0.68	0.57	0.60	0.72	0.70
hw	0.65	0.52	0.63	0.73	0.69
ip	0.55	0.53	0.53	0.52	0.53
cb	0.71	0.56	0.63	0.71	0.75
pt	0.66	0.46	0.64	0.68	0.72
sc	0.72	0.62	0.63	0.73	0.76
un	0.74	0.51	0.56	0.72	0.73
avg	0.69	0.56	0.62	0.71	**0.72**
tot	0.76	0.60	0.66	0.77	**0.80**

A. *Results on manual transcripts*

The results of the classification task on the manual transcripts are presented in Table III. From the reported results we find that among the basic feature sets, the tf-idfs achieve a substantially better performance compared to either DAs or MCs, which indicates that these utterance-level features cannot fully capture CBT-relevant information contained in the word-level features.

Next we compare the results of the tf-idfs with tf-idfs + DAs and tf-idfs + MCs we conclude that directly concatenating the tf-idfs with utterance-level features does not lead to substantial improvements. Finally, we consider the proposed alternative fusion strategy. The performance of the DA-tf-idfs and MC-tf-idfs demonstrates that applying the new proposed fusion strategy to augment the words with the utterance tags, by either DAs or MCs, results in a better CBT relevant code prediction performance. Especially the MC-tf-idfs – which yield the best results among all the features sets – significantly improve the F1 score of the total CTRS and averaged F1 score over tf-idfs (with p-value $<$ 0.05 based on the combined 5×2cv F test [20]).

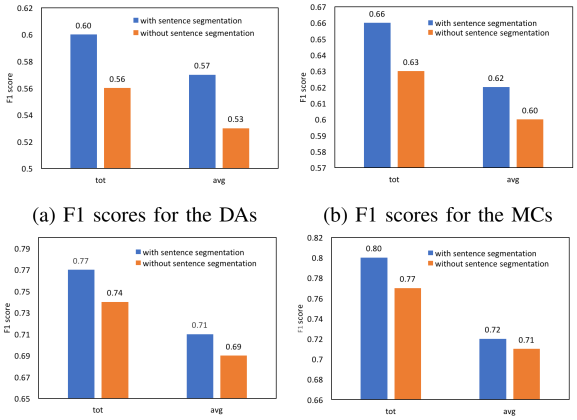
It is interesting to point out that the MCs always lead to better performance compared to DAs, no matter whether we try to predict the CTRS codes by the basic feature set or after fusing with the tf-idfs. This indicates that the behavioral codes defined in MI might exploit more useful therapy-relevant information, by encoding not only structural characteristics, but also more psychotherapy-based cues.

B. *Results of automatically-derived transcripts*

We next consider the end-to-end automated evaluation of CBT sessions using the transcripts generated by the speech processing pipeline described in Section III-A.

The experimental results are given in Table IV. Comparing the results with the ones in Table III, we observe that while the performance of the code prediction using the automatically derived transcripts is degraded compared to evaluating on manually-derived transcripts, the drop is relatively small. This modest performance degradation underscores both the robustness of this end-to-end speech processing system, and the room for further improvements. Again the tf-idf features achieve significantly better F1 scores than the DAs and MCs ($p < 0.01$) while DAs lead to the worst performance among the basic feature sets. The DA-tf-idfs and MC-tf-idfs both outperform the tf-idfs, which is consistent with the results in Table III. The MC-tf-idfs achieve the best overall metrics and F1 scores for the majority of the CTRS codes.

To demonstrate the effect of incorporating an utterance segmentation module, we experiment on the end-to-end evaluation tasks by removing this component from the pipeline. The comparison between the overall performances with and without the utterance segmentation is presented in Fig. 3. The results indicate that, for all the feature sets, removing the segmentation module leads to worse prediction outcomes. This confirms our hypothesis that multi-utterance turns need to be appropriately handled when we are employing utterance-specific representations.



(a) F1 scores for the DAs (b) F1 scores for the MCs

(c) F1 scores for the DA-tf-idfs (d) F1 scores for the MC-tf-idfs

Fig. 3. Comparison of the tasks performed with and without the utterance segmentation for different feature sets.

VI. CONCLUSIONS

We employed an end-to-end approach to assess CBT psychotherapy sessions automatically. The overall CBT session quality assessment was formulated as a binary classification task, and was performed using word-level and utterance-level linguistic feature sets and their fused combinations. In particular, we introduce utterance-level MI codes as one of the feature sets. A new feature fusion strategy was proposed where we augmented the words of an utterance with an utterance-level tag. The experimental results showed that our end-to-end automated approach was robust and the final performance was comparable to using manual transcripts. The best performance was achieved by the fused features of the tf-idfs and MI codes obtained with the new fusion strategy. Additionally, we confirmed the importance of including an utterance segmentation module into the pipeline.

REFERENCES

[1] R. Bakeman, “Behavioral observation and coding,” *Handbook of Research Methods in Social and Personality Psychology*, p. 138, 2000.

[2] B. Shiner, L. W. D’Avolio, T. M. Nguyen, M. H. Zayed, B. V. Watts, and L. Fiore, “Automated classification of psychotherapy note text: implications for quality assessment in ptsd care,” *Journal of evaluation in clinical practice*, vol. 18, no. 3, pp. 698–701, 2012.

[3] B. Xiao, D. Can, J. Gibson, Z. E. Imel, D. C. Atkins, P. G. Georgiou, and S. S. Narayanan, “Behavioral coding of therapist language in addiction counseling using recurrent neural networks.” in *Interspeech*, 2016, pp. 908–912.

[4] V. Ardulov, M. Mendlen, M. Kumar, N. Anand, S. Williams, T. Lyon, and S. Narayanan, “Multimodal interaction modeling of child forensic interviewing,” in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 179–185.

[5] J. S. Beck and A. T. Beck, *Cognitive therapy: Basics and beyond*. Guilford press New York, 1995, no. Sirsi) i9780898628470.

[6] S. G. Hofmann, A. Asnaani, I. J. Vonk, A. T. Sawyer, and A. Fang, “The efficacy of cognitive behavioral therapy: A review of meta-analyses,” *Cognitive therapy and research*, vol. 36, no. 5, pp. 427–440, 2012.

[7] C. L. Randall and D. W. McNeil, “Motivational interviewing as an adjunct to cognitive behavior therapy for anxiety disorders: A critical review of the literature,” *Cognitive and behavioral practice*, vol. 24, no. 3, pp. 296–311, 2017.

[8] N. Flemotomos, V. R. Martinez, J. Gibson, D. C. Atkins, T. Creed, and S. Narayanan, “Language features for automated evaluation of cognitive behavior psychotherapy sessions.” in *Interspeech*, 2018, pp. 1908–1912.

[9] T. A. Creed, S. A. Frankel, R. E. German, K. L. Green, S. Jager-Hyman, K. P. Taylor, A. D. Adler, C. B. Wolk, S. W. Stirman, S. H. Waltman, *et al.*, “Implementation of transdiagnostic cognitive therapy in community behavioral health: The beck community initiative.” *Journal of consulting and clinical psychology*, vol. 84, no. 12, p. 1116, 2016.

[10] J. Young and A. T. Beck, *Cognitive therapy scale: Rating manual*. Bala Cynwyd, PA: Beck Institute for Cognitive Behavior Therapy, 1980, vol. 36.

[11] B. F. Shaw, I. Elkin, J. Yamaguchi, M. Olmsted, T. M. Vallis, K. S. Dobson, A. Lowery, S. M. Sotsky, J. T. Watkins, and S. D. Imber, “Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression.” *Journal of Consulting and Clinical Psychology*, vol. 67, no. 6, p. 837, 1999.

[12] V. R. Martinez, N. Flemotomos, V. Ardulov, K. Somandepalli, S. B. Goldberg, Z. E. Imel, D. C. Atkins, and S. Narayanan, “Identifying therapist and client personae for therapeutic alliance estimation,” *Proc. Interspeech 2019*, pp. 1901–1905, 2019.

[13] “Deepsegment: A sentence segmenter that actually works!” [Online]. Available: https://github.com/bedapudi6788/deepsegment/

[14] M. Dillon, “Introduction to modern information retrieval: G. salton and m. mcgill. new york: Mcgraw-hill (1983),” 1983.

[15] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, “Dialogue act modeling for automatic tagging and recognition of conversational speech,” *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[16] D. Can, D. C. Atkins, and S. S. Narayanan, “A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[17] J. Houck, T. Moyers, W. Miller, L. Glynn, and K. Hallgren, “Motivational interviewing skill code (misc) version 2.5,” *(Available from http://casaa.unm .edu/download/misc25.pdf)*, 2010.

[18] Z. Chen, K. Singla, J. Gibson, D. Can, Z. E. Imel, D. C. Atkins, P. Georgiou, and S. Narayanan, “Improving the prediction of therapist behaviors in addiction counseling by exploiting class confusions,” in *ICASSP*, 2019, pp. 6605–6609.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[20] E. Alpaydm, “Combined 5× 2 cv f test for comparing supervised classification learning algorithms,” *Neural computation*, vol. 11, no. 8, pp. 1885–1892, 1999.