2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA) | 978-1-6654-4337-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICMLA52953.2021.00239

Learning from Limited Data for Speech-based Traumatic Brain Injury (TBI) Detection

Apiwat Ditthapron
Computer Science Dept.
Worcester Polytechnic Institute (WPI)
Worcester, MA
aditthapron@wpi.edu

Emmanuel O. Agu
Computer Science Dept.
Worcester Polytechnic Institute (WPI)
Worcester, MA
emmanuel@wpi.edu

Adam C. Lammert
Biomedical Engineering Dept.
Worcester Polytechnic Institute (WPI)
Worcester, MA
alammert@wpi.edu

***Abstract*—Due to the high cost of collecting data especially for conditions that afflict a small percentage of the population, data is often scarce in healthcare. Inadequate data presents a challenge for training Deep Neural Networks (DNNs). Many Traumatic Brain Injury (TBI) patients require long periods of recovery with unexpected setbacks and possible rehospitalization, making continuous monitoring important. While speech-based TBI assessment has been found to be effective, available datasets are too small for DNNs. To solve the limited TBI speech data problem, we explore three Learning from Limited Data (LLD) methods (transfer, multi-task and meta-learning) that augment a small primary TBI dataset by learning from external non-TBI datasets to improve DNN performance. We found that all three LLD methods mitigate overfitting, improving binary TBI classification accuracy by 33.6%, 36.7%, and 26.4% respectively. External datasets with scripted speech improved the TBI detection accuracy of all three learning methods the most. Using a few-shot learning approach, we extrapolated results on real data to estimate the full trajectory of expected performance for various amounts of data.**

***Index Terms*—Traumatic Brain Injury (TBI), speech assessment, transfer learning, multi-task learning, meta-learning**

I. Introduction

Motivation: Over 1.4 million people suffer from Traumatic Brain Injury (TBI) in the US annually [1]. Patients of all TBI severities (mild, moderate or severe [1]) may have long-term sequelae and neuro-cognitive deficits that disrupt their lives and may require rehospitalization [2]. Consequently, periodic cognitive and neurological exams after injury are recommended. However, as these TBI assessments are costly and invasive, their use is often limited to severe TBI cases [3]. To increase access, non-invasive TBI assessment using sensed data such as speech and gait [4] have been proposed. Speech-based assessment tries to capture speech and language disorders that many TBI patients manifest in the form of poor speech production and comprehension [5]. Prior speech-based TBI assessment work [4] utilized traditional Machine Learning (ML) or DNNs [4] with handcrafted features but did not utilize DNNs to learn directly from raw data and thus did not fully exploit the power of DNNs.

The problem: Inadequate TBI speech data for training DNNs: Medical data is often scarce due to the high cost, need to involve medical experts, use of expensive instrumentation, patients' potential discomfort, and difficulty of recruiting adequate numbers of diverse patients [6]. Coelho [7] is currently the largest TBI speech dataset, containing speech discourse (story retelling, story generation, and conversation) from 55 subjects with TBI and 52 non-TBI control subjects, which is too small for training DNNs. In our pilot experiments, a Gated Recurrent Unit (GRU) with audio backbone DNN accurately classified TBI subjects in the Coelho dataset from healthy controls but overfit due to insufficient data.

This material is based on research sponsored by DARPA underagreement number FA8750-18-2-0077.

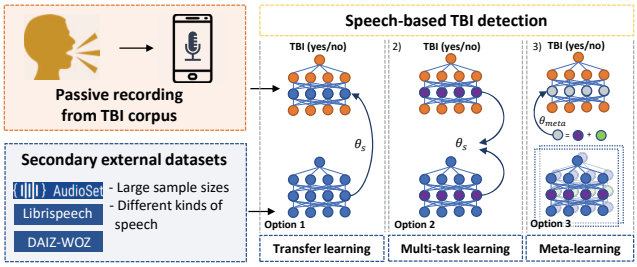

Fig. 1. Overall pipeline of LLD methods for TBI assessment

Our approach: To improve the performance of speech-based TBI assessement using DNNs with limited data, we investigate and compare three Learning from Limited Data (LLD) methods (*Multi-Task Learning* (MTL), *Transfer Learning* (TL) and *Meta-Learning (ML)*). LLD methods augment the primary TBI corpus with secondary non-TBI datasets (See Fig. 1). MTL jointly learns low-level speech features that are common to the source and target tasks. TL and ML both use low-level features learned from the source task as prior information to assist learning of the target task. These LLD methods have previously been adopted in other speech-based assessments [8]–[10] but not for TBI assessment. Moreover, the prior work did not investigate the effects of corpus size and similarity to the target task on prediction performance.

Our main contributions are as follows.

1) Systematic comparison of the MTL, TL and ML LLD methods for DNN-based TBI assessment from speech.
2) Rigorous evaluation of improvements achieved by all LLD methods, while considering two backbone audio classification networks: 1) Wav2Vec [11], and 2) Sinc-

978-1-6654-4337-1/21/$31.00 ©2021 IEEE
DOI 10.1109/ICMLA52953.2021.00239

Net [12] and three external datasets for augmentation: 1) Google Audio Set (GA) [13] 2) Librispeech corpus (Libri) [14], and 3) Wizard-of-Oz (WOZ) interviews corpus [15]. These external source datasets were carefully selected to cover diverse size and levels of similarity to the source TBI dataset.

3) Using a few-shot learning approach, we extrapolated the results of LLD methods on real data from 1 to 50 TBI subjects to generate a complete performance trajectory and estimate the number of training examples required for each method to overcome overfitting. Few-short learning uses only few samples when modeling the target task. For instance, 5-shot learning uses only 5 samples from the target task to train the model.

II. BACKGROUND AND RELATED WORK

Transfer learning (TL): transferred low to mid-level audio representations ($\theta_{shallow}$) learned from non-TBI tasks (s) into our TBI detection model (t). $\theta_{shallow}$ of the pre-trained DNN ($f_{\theta_s}(f_{\theta_{shallow}})$), that was previously trained to minimize the loss function of source task $\mathcal{L}_s$, is transferred to the TBI detection model t, by initializing $f_{\theta_{shallow}}$ in $\min_{\theta_t} \mathcal{L}_t(f_{\theta_t}(f_{\theta_{shallow}}))$ from $\min_{\theta_s,\theta_{shallow}} \mathcal{L}_s(f_{\theta_s}(f_{\theta_{shallow}}))$. High-level features for TBI detection are learned in parameters θ_t, leaving $\theta_{shallow}$ unchanged. After the learning of f_{θ_t} is complete, $\theta_{shallow}$ and θ_t) are fine-tuned in $\min_{\theta_t,\theta_{shallow}} \mathcal{L}_t(f_{\theta_t}(f_{\theta_{shallow}}))$ with the weight $(f_{\theta_{shallow}})$.

Multi-Task Learning (MTL): optimizes $\theta_{shallow}$ across source and target tasks to prevent $f_{shallow}$ from the learning task-specific bias that causes overfitting. Parameters $\theta_{shallow}$ are shared among all tasks ($T_{1,2,3,\ldots,k}$) and jointly optimized to learn common data representations across all k tasks. The final prediction is made independently by an in-task network f_θ as in $\min_\theta \sum_{i=0}^{k} \mathcal{L}_i(f_{\theta_i}(f_{\theta_{shallow}}))$. MTL jointly learns common audio representations to decrease the total loss and increase the network's ability to discriminate classes that it previously confused for one another. MTL with multiple corpora has previously not been utilized for audio-based assessments. Prior audio-based assessment [9] only used MTL to jointly train multi-labels within a single dataset, which may be insufficient to overcome overfitting.

Meta-Learning (ML): We considered Model-Agnostic Meta-Learning (MAML) [16], a *learning to learn fast* Meta-Learning algorithm. MAML optimizes the global parameter θ by reducing the effort to learn (update θ) for each training task i as in $\theta \leftarrow \theta - \beta\nabla_\theta \sum_{i=0}^{k} \mathcal{L}_i\big(f(\theta - \alpha\nabla_\theta \mathcal{L}_i(f_\theta))\big)$. α and β are learning rates for in-task optimization and meta-update respectively. Non-TBI speech corpora were used to train MAML whereas the TBI corpus was used in the meta-testing step that contains the inner loss optimization with the same learning rate α. MAML has been used to adapt speech with a speech disorder [10] but not to mitigate overfitting in speech assessment, particularly for TBI.

Backbone DNNs for speech-based assessment: binary TBI classification from speech, we explored including two backbone DNNs that are popular for various audio tasks, as shallow layers of our TBI detection model.

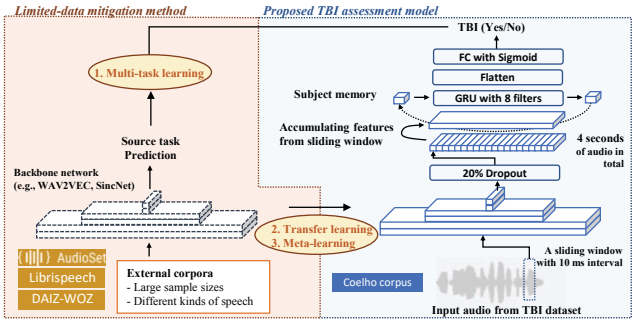

Fig. 2. Approach: LLD methods (MTL, TL, and ML) for TBI assessment.

Wav2Vec: is a semi-supervised model, originally used for Automatic Speaker Representation (ASR). It extracts acoustic features from the speech waveform [11] and consists of an encoder network with 5 CNN layers and a contextual network that combines the representations from the encoder network into a vector representing 210 ms of audio.

SincNet: applies a DNN directly on raw audio ($x[n]$), preserving the phase information of the signal [12]. f_1 and f_2 in a band-pass filter $y[n] = x[n] * [2f_2 sinc(2f_2 n) - 2f_1 sinc(2f_1 n)]$ were trained as a Finite Impulse Response (FIR) convolution filter on $x[n]$ that allows only signals within that band to pass through to subsequent CNN layers. SincNet outperformed a CNN in the speaker recognition task [12] and has been used as a feature extractor to detect neurodegenerative disease [17].

III. METHODOLOGY

An overview of our LLD approach is illustrated in Fig. 2.

A. Source dataset

Google Audio Set (GA) [13] is a large scene classification corpus containing 632 audio events with over 2 million human-labeled of 10-second sound clips extracted from YouTube videos. However, only a relatively small fraction of the clips contain human speech.

Librispeech corpus (Libri) [14] is a large corpus containing 1000 hours of audiobooks, frequently used to train DNNs for speaker recognition and automatic speech recognition. The speech includes male and female speakers but the speech is scripted and does not capture all variations of natural speech.

Wizard-of-Oz (WOZ) interviews corpus [15] is part of the Distress Analysis Interview Corpus (DAIC), which contains 50 hours of spontaneous speech recorded during 189 clinical interviews while virtual interviewer assessed depression. A Personal Health Questionnaire (PHQ-8) depression questionnaire administered to subjects was used as a target label.

B. Target dataset

The **Coelho corpus** [7] contains story generation and retelling, and conversation discourses from 55 TBI subjects with prior head injuries that had now closed, who were 55 native English speakers. Causes of the brain injuries included

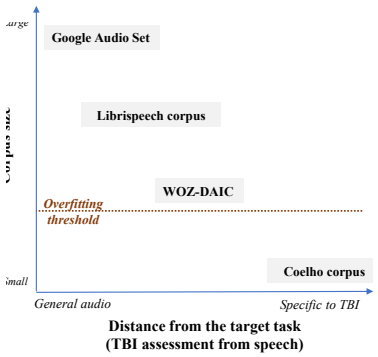

Fig. 3. Source dataset size vs similarity to target TBI task.

TABLE I
SUMMARY OF SOURCE AND TARGET DATASETS

Dataset	Task	Subject(M,F)	Size (hours)
GA [13]	Scene classification	-	5790
Libri [14]	Speaker recognition	125,126	1000
WOZ [15]	Depression assessment	189	34
Coelho [7]	TBI assessment	73,32	19

motor vehicle accidents, falls and being stuck by a car. As controls, speech from 52 native English speakers with no brain injury were also included.

The size of each corpus and its subjective domain distance to the TBI detection task are plotted in Fig. 3. GA and Libri are the two largest corpora, which are publicly available. The GA contains all types of audio whereas Libri contains speech recorded in a controlled environment as listed in Table I. WOZ-DAIC is a small dataset, containing spontaneous speech recorded for health assessment, which is similar in nature to the primary TBI dataset.

Data pre-processing: The conversational speech produced by a target was extracted using onset time in the Coelho corpus transcript. All corpora were normalized using Vocal Tract Length Normalization (VTLN) [18] to reduce inter-speaker variations caused by differences in the vocal tract lengths of various genders and ages, followed by a min-max normalization ($\frac{x-x_{min}}{x_{max}-x_{min}}$). The frequency warping factor in VTLN was determined using the phoneme classification task [19], and applied independently on each subject.

C. GRU for TBI classification

GRUs have previously been used on sequential speech data for assessing depression [20] but not TBI. We applied the GRU model to multi-scale acoustic representations extracted from 4-second segments of audio by each audio backbone DNN. As input to each backbone model, a sliding window of the length specified in Table II with 10 ms interval was applied to format the input shape of the raw audio to match the backbone DNN. As the proposed method operates over a sequence of acoustic features, the concatenated outputs from the backbone model are combined representations that are input to the TBI cascaded GRU model. The cascaded GRU model is comprised of one GRU layer followed by a parametric ReLU activation function and a dropout layer with a dropout rate of 0.2. Since overfitting is the primary concern of this study and the LLD methods are only applied on the backbone network, we employed a GRU with a small filter size of 8. We considered the cell gate unit of the GRU as subject-dependent prior information that provides temporal information of acoustic speech. The final binary (Yes/No) TBI classification is made by a Fully Connected (FC) layer with a Sigmoid activation. Fig. 2 presents an overview of our approach.

TABLE II
INPUT AND OUTPUT DIMENSIONS OF THE BACKBONE MODELS

Model	Original task	Input audio	Bottleneck feature size
Wav2Vec	Speech embedding	210 ms	512
SincNet	Speaker recognition	200 ms	2048

D. Implementation

While trained differently, TL, MTL and ML were all implemented[1] using the PyTorch library ver. 1.4 [21] on NVIDIA Tesla P100 and V100 GPUs. The implementation of MAML was derived from MAML++ [22]. Subject-wise, 10-fold cross-validation was used during evaluation to avoid any bias with a multi-label stratification method [23]. Gender, age and TBI severity distributions were maintained in each fold. During each cross-validation run, two subjects from each TBI class were randomly selected as a validation set.

IV. EVALUATION

A. Evaluation Metrics

Evaluation metrics include Balanced Accuracy (BA) $=$ $(\text{Sensitivity}+\text{Specificity})/2$, F1 score (F1) and Area under the Curve ROC Curve (AUC-ROC). The estimation error of each metric is reported as standard error. We tuned hyperparameters using grid search and selected the optimal model based on classification BA using subject-wise 10-fold cross-validation.

B. Experiments

1) Single Task Learning (STL) and conventional machine learning: This evaluation included STL that learns the TBI detection task without external datasets as a baseline and compared both validation loss and TBI prediction accuracy. Parameters in the backbone network and GRU were initialized randomly and trained using only speech from TBI corpora. For traditional machine learning, we explored two feature sets previously used to detect TBI and other speech disorders: 1) COMPARE [24] and 2) Bag-Of-Audio-Word (BOAW) [25] features. TBI classification was performed by Support Vector Machine (SVM), Random Forest (RF) and Multi-Layer Perceptron (MLP) classifiers with hyperparameter fine-tuning.

2) Few-short learning: We evaluated the learning method using data from 1, 5, 10, 25, 30, 40 and 50 (all) subjects in the Coelho corpus. Extrapolations to 75, 100 and 125 subjects were projected using a linear spline $S_{linear,k}(x) = y_k + \frac{y_{k+1}-y_k}{x_{k+1}-x_k}(x - x_k)$.

[1]Python code: https://github.com/aditthapron/learning-methods-for-speech-based-TBI-detection

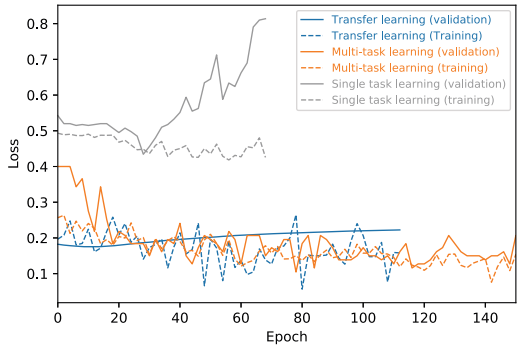

Fig. 4. Training and validation losses using MTL, TL and ML methods

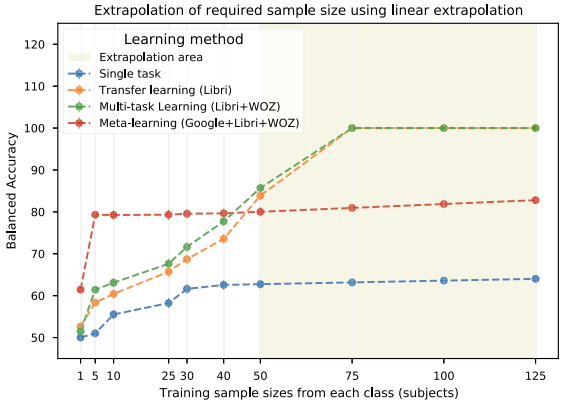

Fig. 5. TBI detection BA for various training sample sizes using linear fit

V. Results and Discussion

Comparison of LLD methods: As shown in Table III, MTL jointly trained with Libri and WOZ corpora on the SincNet model achieves the best TBI classification performance. Also, the backbone network used during training affects performance since TL outperforms MTL and MAML when the Wav2Vec model is used. Wav2Vec contains fewer parameters than the SincNet, preventing it from overfitting during fine-tuning of TL. Although MAML used the same source dataset combinations as MTL, it gains more of an advantage when more secondary tasks are considered. TL and MTL both mitigate overfitting but with different behaviors (See Fig. 4). The loss plot does not include MAML because it computes the loss iteratively on each secondary task, excluding the TBI dataset until the meta-testing step. TL begins to overfit around epoch 16. However, the gaps between validation and training losses are smaller than for STL. TL outperforms STL as it utilizes pre-trained weights from non-TBI datasets. However, overfitting still occurs as fine-tuning uses limited data to update all parameters. MTL mitigates overfitting better than TL as the training and validation losses attained the same error level throughout the training. However, MTL consumes more computing resources as it has to optimize multiple networks and corpora simultaneously, while TL and MAML optimize only one network using one corpus at a time.

Comparison of backbone DNNs: SincNet outperforms other backbone models for all LLD methods because it can track the formant feature, which is important for TBI classification [4]. Wav2Vec also performed well because it captures local phoneme features that may benefit word recognition.

Comparison of external corpora: In all experiments, training with Libri yielded better performance than other datasets. Even though Libri is neither the largest corpus nor the most similar to the TBI corpus, it contains scripted speeches from professional speakers collected in a controlled environment. Increasing the number of source datasets improves the TBI detection performance of MTL and ML. However, using all datasets does not provide the highest BA, possibly because GA contains a wider range of audio recording types and is not specific to the speech domain. A combination of Libri and WOZ containing scripted and spontaneous speech trained on speaker identification and depression assessment tasks aids the TBI classification model in learning common audio representations and preventing overfitting.

Few-shot learning: As shown in Fig. 5. in both 1-shot and 5-shot learning, MAML achieves the lowest TBI detection accuracy of all 3 LLD methods. Meta-learning is an inner-outer iteration method where the outer iteration learns a meta-weight vector and requires less optimization in the inner iteration. This optimization mechanism makes few-shot learning (few inner-loop iterations) possible and provides higher detection accuracy than the other learning methods that do not limit the optimization step or adaptation process on the target task. During the meta-testing step, MAML has access to only a subset of subjects (x-axis of Fig.5), which are used to train the meta-weight. This makes MAML a promising LLD method for training DNNs. Otherwise, MTL and TL are preferred methods. In the extrapolations of TL and MTL, we limited the BAC to 100%, where a plateau exists. STL and MAML do not improve much with more data.

VI. Acknowledgement

This material is based on research funded by DARPA under agreement number FA8750-18-2-0077. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

VII. Conclusion

Training DNNs in medical applications such as TBI detection is challenging due as data is often scarce. This study explored three LLD methods (TL, MTL, and MAML) for mitigating overfitting and improving the performance of TBI detection from speech using limited data. The best external in terms of corpus size and similarity to the target task were also investigated. All three LLD methods improved

TABLE III
TBI CLASSIFICATION RESULTS USING LEARNING METHODS FOR LIMITED DATA LEARNING

Model	Source dataset	TL			MTL			MAML		
		BAC	F1 score	AUC	BAC	F1 score	AUC	BAC	F1 score	AUC
SincNet	GA	67.34 (1.12)	70.86 (1.34)	70.75 (1.44)	66.13 (1.27)	74.75 (1.31)	71.14 (1.52)	66.27 (1.34)	71.11 (1.49)	70.51 (1.40)
	Libri	**83.87 (1.64)**	**85.09 (2.58)**	**87.44 (1.20)**	84.22 (1.00)	86.51 (1.59)	88.68 (1.61)	72.65 (1.32)	74.19 (1.47)	74.80 (1.26)
	WOZ	72.15 (1.34)	75.86 (1.82)	81.41 (1.78)	80.62 (1.49)	82.77 (1.67)	84.27 (1.69)	70.69 (1.35)	74.37 (1.33)	71.65 (1.30)
	GA+ Libri	-	-	-	79.41 (1.32)	82.42 (1.44)	76.12 (1.65)	77.40 (1.35)	80.40 (1.62)	80.96 (1.35)
	GA + WOZ	-	-	-	78.92 (1.25)	80.02 (1.33)	76.25 (1.57)	73.89 (1.32)	77.63 (1.29)	78.56 (1.53)
	Libri + WOZ	-	-	-	**85.70 (1.18)**	**87.12 (1.51)**	**90.13 (1.43)**	79.05 (1.16)	81.09 (1.45)	81.25 (1.87)
	GA+ Libri + WOZ	-	-	-	81.11 (1.62)	84.52 (1.28)	85.04 (1.77)	**79.32 (1.04)**	**82.36 (1.53)**	**81.41 (2.00)**
Wav2Vec	GA	58.84 (1.30)	76.10 (1.98)	61.10 (0.96)	67.42 (1.23)	74.26 (1.44)	69.03 (1.68)	64.28 (1.03)	70.25 (1.20)	69.53 (1.30)
	Libri	**76.13 (1.19)**	**77.13 (2.19)**	**79.70 (1.33)**	72.00 (1.35)	81.32 (1.20)	77.65 (1.52)	66.82 (0.94)	72.64 (0.88)	71.34 (0.99)
	WOZ	75.22 (1.29)	75.90 (1.61)	78.37 (1.68)	70.25 (1.34)	80.89 (1.32)	73.31 (1.43)	66.15 (1.04)	71.28 (1.11)	71.52 (1.24)
	GA + Libri	-	-	-	71.42 (1.18)	81.02 (1.30)	75.60 (1.44)	67.50 (1.35)	74.87 (1.31)	74.87 (1.34)
	GA + WOZ	-	-	-	70.27 (1.53)	79.88 (1.61)	72.68 (1.67)	66.72 (1.31)	71.08 (1.23)	71.25 (1.29)
	Libri + WOZ	-	-	-	**73.16 (1.35)**	**82.66 (1.31)**	**80.52 (1.60)**	69.05 (1.22)	**74.35 (1.19)**	77.32 (1.53)
	GA + Libri + WOZ	-	-	-	70.57 (1.37)	80.65 (1.51)	78.53 (1.52)	**69.29 (1.29)**	**74.38 (1.39)**	**78.75 (1.43)**

TABLE IV
BASELINE FOR TBI CLASSIFICATION

Method		BAC	F1	AUC
STL	SincNet	62.74 (1.22)	69.26 (1.25)	65.62 (0.87)
	Wav2Vec	61.99 (1.14)	65.60 (0.98)	68.35 (1.11)
COMPARE	SVM	56.37 (1.46)	60.14 (1.14)	59.62 (1.20)
	RF	52.18 (0.96)	57.66 (1.30)	56.27 (1.23)
	MLP	52.85 (1.21)	54.32 (1.43)	54.12 (1.15)
BOAW	SVM	**66.05 (1.36)**	**74.44 (1.21)**	**71.41 (2.00)**
	RF	50.14 (1.46)	53.72 (1.72)	52.17 (1.15)
	MLP	62.97 (1.23)	66.74 (1.20)	65.28 (1.49)

TBI classification accuracy by up to 34%. MTL yielded the most improvement in TBI classification accuracy with no overfitting. In few-shot scenarios where audio is very scarce, MAML is a promising method as it required data from only 5 subjects to achieve 78% BA. Also, using a scripted speech dataset as an external dataset yielded the best TBI detection performance.

REFERENCES

[1] M. Faul, M. M. Wald, L. Xu, and V. G. Coronado, "Traumatic brain injury in the united states; emergency department visits, hospitalizations, and deaths, 2002-2006," CDC, 2010.

[2] F. M. Hammond, J. D. Corrigan, J. M. Ketchum, T. A. Novack, J. Bogner, M. N. Dahdah, and G. G. Whiteneck, "Prevalence of medical and psychiatric comorbidities following traumatic brain injury," The Journal of head trauma rehabilitation, vol. 34, no. 4, pp. E1–E10, 2019.

[3] M. A. Lindberg, S. A. Kiser, and E. M. M. Martin, "Mild tbi/concussion clinical tools for providers used within the department of defense and defense health agency," Federal practitioner, vol. 37, no. 9, p. 410, 2020.

[4] T. Talkar, S. Yuditskaya, J. R. Williamson, A. Lammert, H. Rao, D. Hannon, A. O'Brien, G. Vergara-Diaz, R. DeLaura, D. Sturim et al., "Detection of subclinical mild traumatic brain injury (mtbi) through speech and gait," INTERSPEECH, pp. 135–139, 2020.

[5] R. S. Norman, C. A. Jaramillo, M. Amuan, M. A. Wells, B. C. Eapen, and M. J. Pugh, "Traumatic brain injury in veterans of the wars in iraq and afghanistan: Communication disorders stratified by severity of brain injury," Brain injury, vol. 27, no. 13-14, pp. 1623–1630, 2013.

[6] C. H. Lee and H.-J. Yoon, "Medical big data: promise and challenges," Kidney research and clinical practice, vol. 36, no. 1, p. 3, 2017.

[7] C. A. Coelho, K. M. Youse, and K. N. Le, "Conversational discourse in closed-head-injured and non-brain-injured adults," Aphasiology, vol. 16, no. 4-6, pp. 659–672, 2002.

[8] K. Feng and T. Chaspari, "A review of generalizable transfer learning in automatic emotion recognition," Frontiers in CS, vol. 2, p. 9, 2020.

[9] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning." in INTERSPEECH, 2019, pp. 2803–2807.

[10] N. R. Koluguri, M. Kumar, S. H. Kim, C. Lord, and S. Narayanan, "Meta-learning for robust child-adult classification from speech," in IEEE ICASSP, 2020, pp. 8094–8098.

[11] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition." in Interspeech, 2019.

[12] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in IEEE SLT, 2018, pp. 1021–1028.

[13] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in IEEE ICASSP), 2017, pp. 776–80.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in IEEE ICASSP, 2015, pp. 5206–5210.

[15] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella et al., "The distress analysis interview corpus of human and computer interviews." in LREC, 2014, pp. 3123–3128.

[16] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," arXiv:1703.03400, 2017.

[17] Y. Pan, B. Mirheidari, Z. Tu, R. O'Malley, T. Walker, A. Venneri, M. Reuber, D. Blackburn, and H. Christensen, "Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification." in INTERSPEECH, 2020, pp. 4806–4810.

[18] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in IEEE ICASSP, vol. 1, 1996, pp. 353–356.

[19] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in IEEE ASRU, 2011, pp. 24–29.

[20] Y. Zhang, W. Hu, and Q. Wu, "Autoencoder based on cepstrum separation to detect depression from speech," in ICTEE, 2020, pp. 508–510.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, and Gimelshein, "Pytorch: An imperative style, high-performance deep learning library," in NIPS, 2019, pp. 8024–8035.

[22] A. Antoniou, H. Edwards, and A. Storkey, "How to train your maml," arXiv:1810.09502, 2018.

[23] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in ECML PKDD. Springer, 2011, pp. 145–158.

[24] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi et al., "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in INTERSPEECH, 2013.

[25] C. Kohlschein, M. Schmitt, B. Schüller, S. Jeschke, and C. J. Werner, "A machine learning based system for the automatic evaluation of aphasia speech," in Healthcom. IEEE, 2017, pp. 1–6.