Multi-instance learning for bipolar disorder diagnosis using weakly labelled speech data

Zhao Ren, Jing Han, Nicholas Cummins, Qiuqiang Kong, Mark D. Plumbley, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Ren, Zhao, Jing Han, Nicholas Cummins, Qiuqiang Kong, Mark D. Plumbley, and Björn Schuller. 2019. "Multi-instance learning for bipolar disorder diagnosis using weakly labelled speech data." In *Proceedings of the 9th International Conference on Digital Public Health - DPH2019, November 20 - 23, 2019, Marseille, France*, 79–83. New York, NY: ACM Press. https://doi.org/10.1145/3357729.3357743.

Multi-instance Learning for Bipolar Disorder Diagnosis using Weakly Labelled Speech Data

Zhao Ren
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
zhao.ren@informatik.uni-augsburg.de

Jing Han
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
jing.han@informatik.uni-augsburg.de

Nicholas Cummins
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
nicholas.cummins@ieee.org

Qiuqiang Kong
Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
q.kong@surrey.ac.uk

Mark D. Plumbley
Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
m.plumbley@surrey.ac.uk

Björn W. Schuller*
GLAM – Group on Language, Audio & Music, Imperial College London, UK
schuller@ieee.org

ABSTRACT

While deep learning is undoubtedly the predominant learning technique across speech processing, it is still not widely used in health-based applications. The corpora available for health-style recognition problems are often small, both concerning the total amount of data available and the number of individuals present. The Bipolar Disorder corpus, used in the 2018 Audio/Visual Emotion Challenge, contains only 218 audio samples from 46 individuals. Herein, we present a multi-instance learning framework aimed at constructing more reliable deep learning-based models in such conditions. First, we segment the speech files into multiple chunks. However, the problem is that each of the individual chunks is weakly labelled, as they are annotated with the label of the corresponding speech file, but may not be indicative of that label. We then train the deep learning-based (ensemble) multi-instance learning model, aiming at solving such a weakly labelled problem. The presented results demonstrate that this approach can improve the accuracy of feedforward, recurrent, and convolutional neural nets on the 3-class mania classification tasks undertaken on the Bipolar Disorder corpus.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; • **Applied computing** → **Health care information systems**; **Health informatics**.

KEYWORDS

Bipolar Disorder, Weakly Labelled Data, Multi-instance Learning

*Björn Schuller is also with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany.

DPH' 19, November 20–23, 2019, Marseille, France

ACM ISBN 978-1-4503-7208-4/19/11…$15.00
https://doi.org/10.1145/3357729.3357743

ACM Reference Format:
Zhao Ren, Jing Han, Nicholas Cummins, Qiuqiang Kong, Mark D. Plumbley, and Björn W. Schuller. 2019. Multi-instance Learning for Bipolar Disorder Diagnosis using Weakly Labelled Speech Data. In *9th International Digital Public Health Conference (2019) (DPH' 19), November 20–23, 2019, Marseille, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3357729.3357743

1 INTRODUCTION

There has been an acceleration in recent years in the number of papers focused around the application of machine learning in the health domain [8]. A striking aspect of these papers is that, when reviewing the modelling techniques implemented, there is a clear lack of deep learning based approaches [8]. While deep learning has had great success in many speech-based learning tasks, [11, 18, 19], these tasks are commonly undertaken on large databases, thus enabling the learning of the many of millions of parameters in contemporary network structures. Speech-based health tasks, however, are commonly conducted with smaller datasets, bringing forth a need to seek novel and alternate approaches to facilitate the use of deep learning [8].

An example of a 'smaller' health corpus is the *Bipolar Disorder* (BD) corpus [6], which was recently made available for research purposes as part of the 2018 *Audio/Visual Emotion Challenge* (AVEC) [26]. The corpus contains only 218 speech samples from 46 individuals. The diagnosis of bipolar disorders, especially in primary care (general health) settings, is difficult [3]. Further, as early intervention strategies can have positive impacts, there is a need for tools which support an early and objective diagnosis [7]. The BD corpus was collected to aid efforts into identifying speech- and facial-based markers indicative of different mania level displayed by individuals with Bipolar – namely remission, hypomania, and mania. Note that the work presented herein focuses on using speech data only for this classification task of mania level.

Despite having small amounts of data, in terms of the number of speakers and the overall number of samples, it is still possible to employ deep learning in speech-health applications. Techniques such as transfer learning [4], data augmentation [10], and representation learning [14] have all been used in this domain. One such approach,

which to the best of our knowledge is yet to be applied in the speech-health domain, is *multi-instance learning*. Multi-instance learning is the training of a machine learning model with a 'bag' of instances, rather than a single feature vector [13]. Multi-instance learning has been used in a diverse range of tasks including speech-based interest recognition [28], and audio event detection [17].

The work presented in this paper uses chunking to artificially increase the number of data instances for the training of a *Deep Neural Network* (DNN) based classifier. However, while we assign each chunk to a label corresponding to the clip it was partitioned from, we considered all chunks as being *weakly labelled*. We make this assumption on the basis that while some chunks will be indicative of their corresponding labels, we cannot guarantee that this is the case across all chunks. We, therefore, explore the benefits of using different multi-instance learning techniques [1], in combination with multiple DNN-based classifiers, to overcome this weakly labelled problem. The aim is to construct DNN models, under the framework of multi-instance learning, such that we can train reliable DNN models on datasets with smaller amounts of data. We demonstrate the effectiveness of this approach on the BD corpus.

Only a small number of papers have used deep learning on the BD corpus to date. A transfer learning approach is presented in [35]. The authors of [35] trained a *Long Short-Term Memory Recurrent Neural Network* (LSTM-RNN) [15] on affective data from the RECOLA dataset [27], and used it to extract 'emotion' based features from the BD corpus. A data-driven deep learning approach was presented in [12]. This approach consisted of an Inception Net [31] combined with an LSTM-RNN designed to learn multi-resolution features from a *Mel Frequency Cepstral Coefficients* (MFCCs) feature space. Interestingly, other than the AVEC 2018 baseline system [26], this approach did not outperform the more conventional machine learning systems presented in the challenge [30, 33].

The rest of this paper is laid out as follows. First, our multi-instance learning framework is presented in Section 2. We then present our key experimental settings, including an overview of the BD corpus in Section 3. Our experimental results are given in Section 4, and finally our conclusions and future work in Section 5.

2 METHODOLOGY

Compared with more standard speech recognition tasks, based on strongly labelled training data, the bipolar disorder classification task with weakly labelled data contains only clip-level labels. In this section, we first describe the three deep neural networks utilised in this work. We then introduce the multi-instance learning approach which unifies the predictions from the weakly labelled data to formulate the overall clip-level labels.

2.1 Deep Neural Networks

We apply three common deep learning methods in the multi-instance learning framework: feed-forward DNNs (herein denoted as DNNs), Gated Recurrent Neural Networks (GRNNs), and Convolutional Neural Networks (CNNs). These three models have achieved success in similar audio classification tasks [23, 25], and health-related speech processing tasks [9, 34]. All of the three networks have a fully connected layer and a softmax layer for the final prediction (Table 1). Aside from these aspects, the networks differ slightly. The DNN models are constructed using two fully connected layers. The GRNN models include three Gated Recurrent Unit (GRU) layers [5]; we choose the label of the last time step in GRNNs as the final prediction. Based on previous work [22, 23], the CNN structure contains three convolutional layers with a kernel size of (3, 3), a local max pooling layer with a kernel size of (2, 2), and a global max pooling layer. We have previously observed that the global max pooling tends to result in accurate classification by reducing the dimensionality of the output of the final convolutional layer [23]. The structures and parameters of these networks were set empirically to fit the training data of BD corpus.

Table 1: A comparison of the structures of three deep neural networks examined in our proposed mutli-instance learning framework. Note, 'fc' is a fully connected layer, and 'conv' denotes a convolutional layer. The value following each layer is the number of output neurons.

DNN	RNN	CNN
fc-1024	GRU-256	conv-64
fc-1024	GRU-256	conv-128; local max-pool
	GRU-256	conv-256; global max-pool
a fully connected layer with three neurons		
a softmax layer of probabilities for three classes		

Differing from using conventional machine learning methods to process the complete speech samples, we apply the deep neural networks to learn high-level representations from smaller speech chunks. The classification procedure has two potential benefits. First, segmenting speech clips into smaller chunks enables the use of deep learning which relies on big data. Further, deep neural networks can better fit the data by learning non-linear functions more effectively than conventional machine learning classifiers.

However, within this framework, the label of each chunk is unknown, and hence must be assumed weakly labelled. This weakly labelled problem results in two main difficulties: (i) how to give each chunk in a clip a label; and (ii), how to predict a single label for an entire clip from the varied chunk-level predictions. Next, we introduce our methodology to overcome these two difficulties via a multi-instance learning framework.

2.2 Multi-instance Learning

To solve the weakly labelled data problem, we now introduce two stages of multi-instance learning framework. The first is known as instance-level classification, and the other is bag-level classification.

2.2.1 Instance-level Classification. When using *strongly labelled* data, the data set $\mathcal{L}$ contains a set of pairs $\{(\boldsymbol{x}_i, y_i) \mid i = 1, ..., m\}$, where m denotes the total number of speech samples, $\boldsymbol{x}_i$ is a d-dimensional feature vector, $\boldsymbol{x}_i \in \mathbb{R}^d$, and y_i is the label for the i-th speech sample, $y_i \in \{0, 1\}^C$ where C is the number of classes. Multi-instance learning, on the other hand, aims to solve the problems associated with weakly labelled data. Multi-instance learning is defined in terms of *bags*, where single *bag* consists of several *instances*. The number of instances can vary between bags. In such a framework, the data set $\mathcal{L}$ now consists of a variety of pairs $\{(B_i, y_i) \mid i = 1, ..., m\}$, where the bag B_i is labelled as y_i. A bag B_i contains feature vectors of multiple instances, $B_i = \{\boldsymbol{x}_i^j \mid j = 1, ..., n_i\}$, where n_i means the total number of

instances in the bag B_i. In the training procedure, as the data is weakly labelled, the label of each instance $\boldsymbol{x}_i^j$ is assumed to be consistent with the label of bag B_i.

In our multi-instance learning framework for the level of mania classification, we consider each speech clip as a bag, containing several chunks, herein named as instances. In the instance space, constructed by speech chunks, an instance-based classifier $f(\boldsymbol{x}) \in \{0,1\}^C$ can be constructed, from training data, using the aforementioned deep learning methods (Section 2.1). Further, given a new bag B, there is a need to construct a mapping between the objective bag-level classifier $F(B) \in \{0,1\}^C$ and $f(\boldsymbol{x})$. The classifier $F(B)$ can be expressed by: $F(B) = g_{\boldsymbol{x}\in B}(f(\boldsymbol{x}))$, where $g(f(\boldsymbol{x}))$ denotes a transfer function based on $f(\boldsymbol{x})$.

2.2.2 Bag-level Classification. In order to define the transfer function mentioned above, four different assumptions can be made. The first is the *standard assumption* in multi-instance learning, and the others are the *vocabulary-based assumption*, the *collective assumption*, and the *weighted collective assumption* [1].

Standard Assumption: To obtain a bag-level classifier $F(B)$, the standard assumption assumes that a speech clip is labelled as positive for a specified mania level when the patient's state at this level occurs at more than one chunk. On the other hand, a speech clip is labelled as negative when the patient's state at the specified level does not occur at any chunk. Under this standard assumption, we can then define a bag-level classifier using the *max rule*,

$$F(B) = \max_{\boldsymbol{x}\in B} f(\boldsymbol{x}). \tag{1}$$

The standard assumption has been applied in many circumstances [29, 32]. However, it has an underlying issue, in that for certain instances it can make the bag positive when the real label is negative. It over- or underestimates the contribution of other instances. Thus, more accurate approaches are used in our work to consider the information from all instances to make a decision.

Vocabulary-based Assumption: The simplest solution to the problem from the standard assumption is to represent the statistic information of each bag using a vocabulary. We apply a *histogram-based method* to construct the vocabulary [1]. To do this, we compute a function which maps a bag B into a histogram $H = h_1, ..., h_C$, where the k-th histogram h_k denotes the number of instances which fall into class k. The mapping function can, therefore, be defined as $M(B, H) = (h_1, ..., h_C)$. Finally, the label of bag B is predicted as the class which holds the maximum histogram count.

Collective Assumption: The collective assumption states that all instances in a bag have *some* form of contribution to the bag's label. The collective assumption can be realised by using the *mean rule* [16]. The *mean rule* assumes all instances in a bag have equal contribution to the bag's label. Given a new bag B, the bag-level classifier can then be obtained by:

$$F(B) = \frac{1}{|B|} \sum_{\boldsymbol{x}\in B} f(\boldsymbol{x}), \tag{2}$$

where $|B|$ is the normalisation constant for that final evaluation. The *mean rule* can provide good results in many applications [9, 23]. However, the instances can also contribute at a variety of levels to the bag's label in the real data.

Table 2: The distribution of speech clips(chunks) for the three classes in the three data sets. The speech clips are the official dataset of AVEC 2018 [26]. The chunks are obtained by chunking speech samples by a sliding window with length of one frame.

Mania level	Training	Development	Test
Remission	25(121 665)	18(80 234)	18(80 234)
Hypomania	38(314 463)	21(117 027)	18(117 027)
Mania	41(355 323)	21(133 198)	18(133 198)
Sum	104(791 451)	60(330 459)	54(388 386)

Weighted Collective Assumption: As an alternative implementation of the *mean rule*, the *weighted rule* under the weighted collective assumption computes a weight value for each instance in a bag. The *weighted rule* can be defined as,

$$F(B) = \frac{1}{\sum_{\boldsymbol{x}\in B} w(\boldsymbol{x})} \sum_{\boldsymbol{x}\in B} w(\boldsymbol{x}) f(\boldsymbol{x}), \tag{3}$$

where $w(\boldsymbol{x})$ denotes the weight of the classifier $f(\boldsymbol{x})$. To compute the weight values, we propose the use of the *Margin Sampling Value* (MSV) of the predicted class for each instance. We define the MSV for an instance $\boldsymbol{x}$ using:

$$C(\boldsymbol{x}) = ||P_0 - P_1||, \tag{4}$$

where the P_0 and P_1 are the first and second maximum posterior probabilities respectively. MSV has been used for decision fusion [24] as metrics directly and in a cooperative learning framework as a query function [36]. To the best of our knowledge, MSV has not been applied to determine the weight value in multi-instance learning. Finally, to ensure the sum of all MSVs in a bag is equal to one, we apply a softmax function on the MSVs:

$$C(\boldsymbol{x}) = \frac{e^{C(\boldsymbol{x})}}{\sum_{\boldsymbol{x}\in B} e^{C(\boldsymbol{x})}}. \tag{5}$$

3 KEY EXPERIMENTAL SETTINGS

3.1 Database

The BD corpus, as used in the AVEC 2018 [26], contains recordings from 46 Turkish speaking individuals (16 females, 30 males) who have bipolar disorder. Both audio and video data were recorded while the participants completed various speaking tasks. At each recording, the patients' level of mania, a core feature of bipolar, was determined using the *Young Mania Rating Scale* (YMRS). According to their corresponding YMRS scores, the recordings are grouped into one of three levels: remission (YMRS <= 7), hypomania (7 < YMRS < 20), and mania (YMRS >= 20), leading to a three-class classification task. As part of AVEC 2018, the data set was partitioned into training/development/test sets (Table 2). For further information on the corpus, the interested reader is referred to [6, 26]. In this work, we analyse the speech data with sampling rate of 16 kHz.

3.2 Ensemble Learning

Ensemble approaches have been repeatedly shown to improve the performance of multiple single (weak) classifiers [2, 37]. As an extension to our framework, we apply the approach of *bagging* to ensemble the classifiers learnt through multi-instance learning [37].

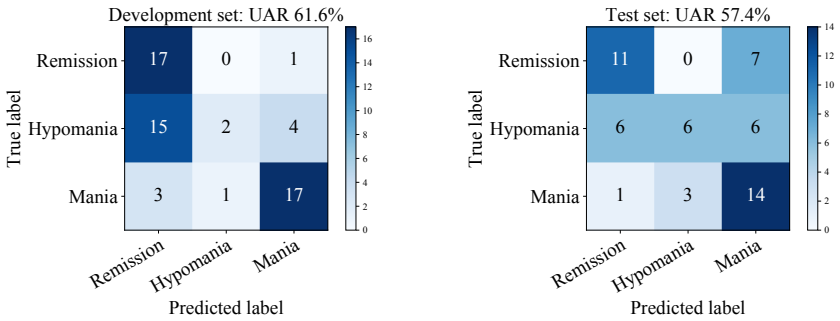

Figure 1: Confusion matrices of the development and test set from our best result on the BD corpus.

Table 3: The results of different methods of bag-level classification. The form of results is represented as UAR[%] on test set (development set). The results are presented as two parts: the multi-instance learning from single iteration, and the ensemble multi-instance learning from multiple iterations.

UAR[%]	Max	Histogram	Mean	Weighted
Multi-instance Learning				
CNN	51.9(51.9)	53.7(56.9)	53.7(55.3)	55.6(56.9)
DNN	48.1(44.7)	53.7(58.2)	53.7(56.6)	51.9(58.5)
RNN	44.4(53.4)	50.0(56.3)	50.0(56.3)	53.7(57.7)
Ensemble Multi-instance Learning				
CNN	**46.3(55.3)**	46.3(55.3)	50.0(54.0)	51.9(56.6)
DNN	42.6(43.1)	**51.9(61.6)**	**53.7(61.6)**	**57.4(61.6)**
RNN	46.3(48.7)	51.9(60.1)	51.9(56.9)	55.6(58.5)

Table 4: Comparison of the state-of-the-art methods and our deep neural networks-based multi-instance learning framework. The compared methods can focus on either Audio (A) data, or both Audio and Video data (A+V). Our multi-instance learning approach is only applied on audio data.

UAR[%]	*Dev*	*Test*
SVMs (A) [26]	55.0	50.0
GEWELMs (A) [30]	55.0	48.2
Multistream (A+V) [35]	78.3	40.7
IncepLSTM (A+V) [12]	65.1	–
Hierarchical recall model(A+V) [33]	86.8	57.4
Multi-instance learning (A only)	**61.6**	**57.4**

First, the training set is used to form a set of deep neural networks, resulting in a set of classifiers $\{S_1, ..., S_t\}$ obtained from t training iterations. We then bag the outputs of all of these classifiers using the bag-level classification methods introduced earlier (Section 2.2.2).

3.3 Experimental Set-up

First, the speech clips are segmented into a set of chunks by a sliding window of width {1, 2, 4, 8, 16, 32} frames and a step size of 1 frame in order to obtain a similar number of chunks from different window lengths. Therefore, the chunks obtained by a window with a length of 1 frame are distributed as Table 2. We then extract MFCCs with the bands 1–20 empirically as the low-level descriptors for each chunk. MFCCs have achieved success in many applications, such as speech-based emotion recognition [21], and audio classification [20]. The three deep learning algorithms are trained for 6 000 iterations each with a batch size of 128, an initial learning rate of 0.001, using Adam optimisation. To counter over-fitting, and stabilise the training procedure, the learning rate is decreased by a factor of 0.9 at every 50 iteration steps. The set-up of parameters of deep neural networks were empirical.

4 RESULTS AND DISCUSSIONS

In our experiments we compared two methods, *multi-instance learning* and *ensemble multi-instance learning*. The classification metric is Unweighted Average Recall (UAR), which was the metric in AVEC 2018. The baseline was obtained using Support Vector Machines (SVMs) for classification that were trained and tested with a single feature representation extracted from the entire speech sample [26].

For the *multi-instance learning*, we choose the best results obtained from the chunks with different window lengths of {1, 2, .., 32} (Section 3.3) for each setup of network and bag-level classification method. Also, each deep learning method was stopped at the 5000-th training iteration. From these experiments we observed that the bag-level classification approaches of *histogram*, *mean*, and *weighted* perform better than *max* (Table 3). This observation is consistent with the analysis presented in Section 2.2.2, , the *max rule* will perform the weakest as it does not consider the contribution of all instances in a bag.

In the *ensemble multi-instance learning*, the ten models are formed from {5 000, 5 100, ..., 5 900} training iterations. Classification performances are improved by using the ensemble extension (Table 3). This improvement is observable across the four bag-level classification methods, and in the three deep learning frameworks, especially DNNs and RNNs.

The best result obtained by our proposed method is highly competitive with those obtained by state-of-the-art methods (Table 4). Except for [35] which also segmented the speech files, these methods fed features obtained from the full audio/video records into SVMs [26], Greedy Ensembles of Weighted Extreme Learning Machines (GEWELMs) [30], Inception LSTM (IncepLSTM) [12], or a hierarchical recall decision tree model [33]. Our multi-instance learning method performs better than all other audio-based methods, and better than, or comparable with, the methods using both speech and visual information. Notably, our algorithm improves the performance significantly more than the multistream approach by [35] on the test set ($p < .05$ in a one-tailed z-test).

Finally, when viewing two confusion matrices of our best results on the development and test sets from ensemble multi-instance learning with the *weighted rule*, it can be observed that the classes of *remission* and *mania* are classified better than the class of *hypomania* (Figure 1). We speculate that the reason might be that the YMRS of the class *hypomamia* sits between the other two classes (aforementioned in Section 3.1), meaning these samples are more challenging to be classified.

5 CONCLUSIONS

We proposed a multi-instance learning framework based on deep neural networks to process the weakly labelled speech data for classifying the level of mania in bipolar patients. In the instance-level classification, the speech clips were segmented into a set of chunks as the input of deep neural networks. Then, the labels of speech clips were predicted using the methods in bag-level classification. The best result was obtained using an ensemble multi-instance learning framework with the *weighted rule*. This model performed better than, or matched to, the state-of-the-art methods on the test partition of the Bipolar Disorder corpus from the 2018 Audio/Visual Emotion Challenge.

In future efforts, we will consider training the neural networks at the bag-level instead of the instance-level. Further, attention based deep learning models will be investigated to better evaluate the contribution of instances in the bag-level assumption.

6 ACKNOWLEDGEMENTS

We acknowledge support by the EU's Horizon H2020 Marie Skłodowska-Curie grant agreement No. 766287 (TAPAS), the EPSRC grant EP/N014111/1 "Making Sense of Sounds", a Research Scholarship from the China Scholarship Council (CSC) No. 201406150082, and the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 115902 supported by Horizon 2020 and EFPIA.

REFERENCES

[1] Jaume Amores. 2013. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence* 201 (2013), 81–105.
[2] Alberto Cano. 2017. An ensemble approach to multi-view multi-instance learning. *Knowledge-Based Systems* 136 (2017), 46–57.
[3] Mauro Giovanni Carta and J. Angst. 2016. Screening for bipolar disorders: A public health issue. *Journal of Affective Disorders* 205 (2016), 139–143.
[4] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Jianhua Tao. 2015. Multi task sequence learning for depression scale prediction from video. In *Proc. ACII*. Xi'an, China, 526–531.
[5] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proc. NIPS*. Montreal, Canada. 9 pages.
[6] Elvan Ciftci, Heysem Kaya, Hüseyin Gülec, and Albert Ali Salah. 2018. The Turkish audio-visual bipolar disorder corpus. In *Proc. ACII Asia*. Beijing, China.
[7] Philippe Conus, Craig Macneil, and Patrick D. McGorry. 2014. Public health significance of bipolar disorder: Implications for early intervention and prevention. *Bipolar Disorders* 16, 5 (2014), 548–556.
[8] Nicholas Cummins, Alice Baird, and Björn Schuller. 2018. The increasing impact of deep learning on speech analysis for health: Challenges and opportunities. *Methods, Special Issue on on Translational data analytics and health informatics* 151 (2018), 41–54.
[9] Jun Deng, Nicholas Cummins, Jing Han, Xinzhou Xu, Zhao Ren, Vedhas Pandit, Zixing Zhang, and Björn Schuller. 2016. The University of Passau open emotion recognition system for the multimodal emotion challenge. In *Proc. CCPR*. Chengdu, P. R. China, 652–666.
[10] Jun Deng, Nicholas Cummins, Maximilian Schmitt, Kun Qian, Fabien Ringeval, and Björn Schuller. 2017. Speech-based Diagnosis of Autism Spectrum Condition by Generative Adversarial Network Representations. In *Proc. DH*. London, UK, 53–57.
[11] Li Deng. 2016. Deep learning: From speech recognition to language and multimodal processing. *APSIPA Transactions on Signal and Information Processing* 5 (2016), 1–15.
[12] Zhengyin Du, Weixin Li, Di Huang, and Yunhong Wang. 2018. Bipolar disorder recognition via multi-scale discriminative audio temporal representation. In *Proc. AVEC*. Seoul, South Korea, 23–30.
[13] James Foulds and Eibe Frank. 2010. A review of multi-instance learning assumptions. *The Knowledge Engineering Review* 25, 1 (2010), 1âĂŞ–25.
[14] Gábor Gosztolya, Róbert Busa-Fekete, Tamás Grósz, and László Tóth. [n.d.]. DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification. In *Proc. INTERSPEECH*. Stockholm, Sweden.
[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[16] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D Plumbley. 2018. Audio set classification with attention model: A probabilistic perspective. In *Proc. ICASSP*. Calgary, Canada, 316–320.
[17] Anurag Kumar and Bhiksha Raj. 2016. Audio event detection using weakly labeled data. In *Proc. ACM Multimedia*. Amsterdam, Netherlands, 1038–1047.
[18] Zhen-Hua Ling et al. 2015. Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine* 32, 3 (2015), 35–52.
[19] Ali Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. 2019. Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access* 7 (2019), 19143–19165.
[20] Alain Rakotomamonjy and Gilles Gasso. 2015. Histogram of gradients of time–frequency representations for audio scene classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 1 (2015), 142–153.
[21] Arti Rawat and Pawan Kumar Mishra. 2015. Emotion recognition through speech using neural network. *International Journal of Advanced Research in Computer Science and Software Engineering* 5, 5 (2015), 422–428.
[22] Zhao Ren, Qiuqiang Kong, Jing Han, Mark Plumbley, and Björn Schuller. 2019. Attention-based atrous convolutional neural networks: Visualisation and understanding perspectives of acoustic scenes. In *Proc. ICASSP*. Brighton, UK, 56–60.
[23] Zhao Ren, Qiuqiang Kong, Kun Qian, Mark Plumbley, and Björn Schuller. 2018. Attention-based convolutional neural networks for acoustic scene classification. In *Proc. DCASE*. Surrey, UK, 39–43.
[24] Zhao Ren, Vedhas Pandit, Kun Qian, Zijiang Yang, Zixing Zhang, and Björn Schuller. 2017. Deep sequential image features on acoustic scene classification. In *Proc. DCASE*. Munich, Germany, 113–117.
[25] Zhao Ren, Kun Qian, Yebin Wang, Zixing Zhang, Vedhas Pandit, Alice Baird, and Björn Schuller. 2018. Deep scalogram representations for acoustic scene classification. *IEEE/CAA Journal of Automatica Sinica* 5, 3 (2018), 662–669.
[26] Fabien Ringeval et al. 2018. AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proc. AVEC*. Seoul, South Korea, 3–13.
[27] Fabien Ringeval, Andreas Sonderegger, Jürgen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proc. EmoSPACE*. Shanghai, P. R. China.
[28] Björn Schuller and Gerhard Rigoll. 2009. Recognising interest in conversational speech-comparing bag of frames and supra-segmental features. In *Proc. INTERSPEECH*. Brighton, UK, 1999–2002.
[29] P Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine. 2019. Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications* 117 (2019), 103–111.
[30] Zafi Sherhan Syed, Kirill Sidorov, and David Marshall. 2018. Automated screening for bipolar disorder from audio/visual modalities. In *Proc. AVEC*. Seoul, South Korea, 39–45.
[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proc. CVPR*. Boston, MA, 1–9.
[32] Junxiang Wang and Liang Zhao. 2018. Multi-instance domain adaptation for vaccine adverse event detection. In *Proc. WWW*. Lyon, France, 97–106.
[33] Xiaofen Xing, Bolun Cai, Yinhu Zhao, Shuzhen Li, Zhiwei He, and Weiquan Fan. 2018. Multi-modality hierarchical recall based on GBDTs for bipolar disorder classification. In *Proc. AVEC*. Seoul, South Korea, 31–37.
[34] Le Yang, Dongmei Jiang, Wenjing Han, and Hichem Sahli. 2017. DCNN and DNN based multi-modal depression recognition. In *Proc. ACII*. San Antonio, TX, 484–489.
[35] Le Yang, Yan Li, Haifeng Chen, Dongmei Jiang, Meshia Cédric Oveneke, and Hichem Sahli. 2018. Bipolar disorder recognition with histogram features of arousal and body gestures. In *Proc. AVEC*. Seoul, South Korea, 15–21.
[36] Zixing Zhang, Eduardo Coutinho, Jun Deng, and Björn Schuller. 2015. Cooperative learning and its application to emotion recognition from speech. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 23, 1 (2015), 115–126.
[37] Zhi-Hua Zhou and Min-Ling Zhang. 2003. Ensembles of multi-instance learners. In *Proc. ECML*. Cavtat-Dubrovnik, Croatia, 492–502.