INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Improving cognitive impairment classification by generative neural network-based feature augmentation

Bahman Mirheidari[1], Daniel Blackburn[2], Ronan O'Malley[2], Annalena Venneri[3], Traci Walker[4], Markus Reuber[4], and Heidi Christensen[1]

[1]Department of Computer Science, University of Sheffield, Sheffield, UK
[2]Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK
[3]Department of Human Communication Sciences, University of Sheffield, Sheffield, UK
[4]Academic Neurology Unit, University of Sheffield, Royal Hallamshire Hospital, Sheffield, UK
{b.mirheidari,heidi.christensen}@sheffield.ac.uk

Abstract

Early detection of cognitive impairment is of great clinical importance. Current cognitive tests assess language and speech abilities. Recently, we have developed a fully automated system to detect cognitive impairment from the analysis of conversations between a person and an intelligent virtual agent (IVA). Promising results have been achieved, however more data than is typically available in the medical domain is required to train more complex classifiers. Data augmentation using generative models has been demonstrated to be an effective approach. In this paper, we use a variational autoencoder to augment data at the feature-level as opposed to the speech signal-level. We investigate whether this suits some feature types (e.g., acoustic, linguistic) better than others. We evaluate the approach on IVA recordings of people with four different cognitive impairment conditions. F-scores of a four-way logistic regression (LR) classifier are improved for certain feature types. For a deep neural network (DNN) classifier, the improvement is seen for almost all feature types. The F-score of the LR classifier on the combined features increases from 55% to 60%, and for the DNN classifier from 49% to 62%. Further improvements are gained by feature selection: 88% and 80% F-scores for LR and DNN classifiers respectively.

Index Terms: Dementia detection, conversation analysis, speech recognition and segmentation, processing of pathological speech

1. Introduction

Dementia is a clinical syndrome affecting cognitive skills, memory, everyday functionalities, speech, language and communication. The number of people developing dementia is growing drastically. Around 850 thousand people are living with dementia in the UK and it is estimated that the figure will rise to 1.6 million by 2040 [1]. Now dementia is the leading cause of death in the country accounting for over 12 percent of total deaths [2]. The early diagnosis of dementia is of great clinical importance. The current test capable of identifying people with a high risk of dementia are expensive and invasive (positron emission tomography (PET): exposure to radiation; amyloid analysis of the cerebrospinal fluid (CSF): lumbar puncture). Thus there is a need for an automatic, easy-to-use, low-cost and accurate stratification tool.

Speech and language abilities are routinely assessed in current cognitive tests. Recently, research into automatic assessment methods based on speech as well people's interactive abilities have revealed promising cues for identifying cognitive decline. In particular, conversation analysis (CA) of patients and neurologists was shown to enable differentiation between patients with neurodegenerative disorder (ND) and functional memory disorder (FMD; exhibiting problems with the memory not caused by dementia) [3, 4]. However, the approach is expensive and difficult to scale up for routine clinical use; our recent work has focused on automating this interaction-based assessment approach.

To this end, we have developed a fully automatic system for identifying dementia based on analysing a person's speech and language as they speak to an intelligent virtual agent (IVA). The IVA asks a series of memory-probing questions similar to the *history taking* part of a normal face-to-face consultation. Initially, a number of features, routed in conversation analysis, were extracted and high accuracy levels were achieved when evaluating the system in a real memory clinic with patients diagnosed as having ND or FMD [5, 6]. We then expanded our data collection to include two more diagnostic classes (healthy controls (HC), and patients with mild cognitive impairment (MCI) [7, 8], and consequently changing the task from a binary decision for the classifier to a four-way classification. This naturally increases the difficulty due to the large overlap between symptoms. It also increases the need for training data to enable the training of more complex, data-hungry machine learning approaches - something which it is often difficult to obtain in sufficient quantity when working in medical domains. This paper addresses this issue by proposing a novel method for augmenting extracted features. We show that this enables us to increase performance and train deep learning based models. This approach would be applicable in similar sparse data domains.

Data augmentation is a common method for increasing the number of samples in applications involving speech processing to alleviate problems with limited data [9]. There is an increasing number of studies applying generative models such as generative adversarial networks (GANs) for data augmentation in applications such as speech synthesis [10], speech recognition [11, 12], speech emotion recognition [13], speech enhancement [14], and speaker verification [15]. They almost all augment the data at the signal level. Recently we have investigated using three generative models to produce synthesized samples of *features* extracted from the IVA conversations[16]. The synthesized features were then added to the original features to provide more data to train a better classifier to distinguish between the four classes (FMD, ND, MCI and HC). We demonstrated how applying the approach on one fold of our dataset could improve the F-score of the DNN based classifier from 58% to 74%. We found that variational autoencoders (VAE) out-

http://dx.doi.org/10.21437/Interspeech.2020-2433

performed the two other generative models (conditional GAN (CGAN) and VAE combined with semi-supervised GAN (VAE-SGAN)). However, this preliminary work did not verify the approach across the full set of folds, nor did it analyse the susceptibility of different feature types to the proposed approach - does GAN-based augmentation of features provide different benefits depending on the vastly different feature types used, such as acoustic, linguistic and word vector-based features?

This paper describes this extension of previous work by performing a full 10-fold cross validation on our dataset and providing an in-depth feature-based analysis. Comparing to previous work, the ASR unit of our automatic system is also improved by applying transfer learning on a base model trained on the LIBRISPEECH dataset. This affects the extracted features from the conversations. In addition, new features are extracted from the conversational data to build more robust classifiers. Feature augmentation are done for the individual feature types as well as all features combined together to allow us to analyse which feature types benefit the most from the proposed feature augmentation approach.

2. Generative models for data augmentation

Machine learning models are either discriminative or generative. For the input features, x and its corresponding label, y, the discriminative models directly predict the probability of y given x ($P(y|x)$) by making boundaries from the features, while the generative models try to estimate the distribution of features and generate them (likelihood or probability of x: $P(x)$). The recently introduced GANs and variational autoencoders (VAEs) are examples of generative models. The GAN model ([17]) consists of two main components: the generator and the discriminator. The generator produces new samples and the discriminator (a binary classifier) authenticates the samples, evaluating whether they are genuine or not. The task of the generator is to create better samples to *deceive* the discriminator. Both the real and the synthesised samples are fed to the discriminator. The process of calling the generator and the discriminator are repeated until the discriminator cannot distinguish the real samples from the fake samples. Generally, the structure of the generator model is the reverse network of the discriminator. Stabilising the GAN models (i.e., deciding when to stop training) and low resolution of the produced features are the main challenges when using GAN models successfully. There have been different improvements to address these issues, including conditional GAN (CGAN [18]) and semi-supervised GANs (SGAN) (forcing the discriminator to produce the labels [19]).

Autoencoders (AEs) are generative models consisting of two components: the encoder and the decoder. The encoder encodes the input samples into a dimensionally reduced (compressed) representation, while the decoder reconstructs the samples from the representations. The AEs try to reconstruct a sample as similar as possible to the real sample. In the Variational AEs (VAE)s the compressed representations of the AE are normalised and the network tries to capture the distribution of the original samples, making better samples comparable to GANs [20]. GANs have been used in a number of medical applications (e.g., image segmentation task for the computed tomography (CT) cerebrospinal fluid and the fluid-attenuated inversion recovery magnetic resonance [21], classification of CT images of liver lesions [22]), cancer classification using gene data [23], disease classification on X-ray images [24], emotion recognition from electroencephalography signal, eye movement and directions [25]. However, very few studies have used a VAE as a generative model for data augmentation in the medical domain, where the data is often images and GANs are more suited. Here, our signal is speech and the aim is to explore the augmentation of the extracted features (acoustic, linguistic, CA, etc.) extracted from the audio conversations between the patients and the IVA. In this paper, we use a VAE as a generative model for feature augmentation, i.e., generating more data samples to train a robust classifier.

Table 1: *Datasets used for training the speaker diarisation and the ASRs. Len.:the total length in hours/mins, Utts.:number of utterances, Spks.:number of speakers, and Avg. Utts.:Average utterance length in seconds.*

Dataset (No)	Len.	Utts.	Spks.	Avg. Utts.
DR INTERVIEWS (295)	64.3h	39.2k	736	5.9s
IVA (93)	17.3h	5.6k	103	11.05s
LIBRISPEECH (281241)	961.1h	281.2k	5466	12.3s

3. Experimental setup

The data was collected between 2016 and 2019 using our in-house IVA system at the Department of Neurology, University of Sheffield, UK based at the Royal Hallamshire Hospital. A total of 93 participants were recorded of which 60 (15 FMD, 15 ND, 15 MCI and 15 ND) were chosen for the study. The rest were found to not have memory problems, however, we made use of that data for training the speaker diarisation and the ASR. For more details about the demographic information of the participants in the study, please refer to [8]. The data is divided into 10 folds to do k-fold cross-validation.

3.1. Speaker diarisation and ASR

Table 1 shows information about the three datasets used:IVA (93 IVA-patient recordings), LIBRISPEECH (over 5 thousand people reading books) and DR INTERVIEWS (295 doctor-patient interviews) . The DR INTERVIEWS dataset is recorded at the same hospital and consists of interviews between doctors and patients with cognitive impairments, as well as conversation with people with seizure. It is only used for training the i-vector based diarisation module (the CALL_HOME recipe [26]). The LIBRISPEECH dataset was used to train a base TDNN acoustic model following Kaldi's Librispeech recipe[27]. The 10 fold cross-validation approach was used for training the diarisation and the ASR modules.

The language models for the ASRs were trained as four-grams with Turing smoothing interpolated with the four-gram language model from the LIBRISPEECH dataset (60% weight for the train set and 40% weight for the LIBRISPEECH language model). Then using the "transferring all layers" technique [28], the acoustic model was adapted to the acoustics of the IVA training set. In this approach, both the structure and the weights were transferred from the training of the LIBRISPEECH dataset. Then we ran one epoch of training on the training set data (the mini-batch size was 128 and the learning rate between 0.014 and 0.0014). One epoch was found to yield the best ASR performance. The diarisation error rate (**DER**) was **17.3%**, and the word error rate (**WER**) was **25.1%**(a significant improvement on previous work [16], where we achieved a WER of 38.3% by combining the DR INTERVIEWS dataset with the IVA dataset for training the acoustic model).

3.2. Extended features

In our previous research, we found that the CA-inspired features, as well as lexical and acoustic features, were useful in identifying dementia. In addition to the initial features (CA, acoustic (AC), lexical (LX), word vector with PCA (WV-PCA), verbal fluency test (FT), and MFCC acoustic features) introduced in [7, 16], two more sets of features were extracted: statistics (average, standard deviation, minimum, maximum and sum) of speech and silence durations per utterance (SilSP), and 384 acoustic features first proposed for the emotion detection challenge at Interspeech 2009 (IS09; 16 features and delta features of low-level descriptors (LLD) (zero-crossing-rate (ZCR), root mean square (RMS), frame energy, pitch frequency (F0), harmonics-to-noise ratio (HNR)) for each LLD [29]. The features were extracted for each utterance of the conversations and then averaged over the utterances. To extract the features the OpenSmile toolkit ([30]) was used. We also tried the Interspeech 2016 ComParE challenges features [31], but classification performance was inferior to IS09.

3.3. Generative model and classifiers

Previously we investigated three generative models [16] (CGAN, VAE and VAE combined with SGAN) and from our preliminary experiments, we found that VAE performed better than the others. Thus for this study, only the results of the VAE will be presented. The Keras Python library ([32]) back-ended by Tensorflow([33]) was used for training the generative models. The encoder part of the model consisted of two dense layers with 1024 and 512 neurons respectively, followed by a latent dense layer with 20 neurons which was then normalised using a lambda function. Also LeakyReLU, and Dropout layers were added in between the dense layers to avoid overfitting. The decoder had the reverse structure except for the lambda function. The MSE loss function and the Adam optimiser were used for training the network. Two classifiers were chosen for the experiments in this study: logistic regression (LR) classifier, and DNN classifier. The Sklearn python module [34] was used for training the LR classifier (with its default parameters with a fixed random state), and the Keras for training the DNN. The DNN model consisted of four layers of 1024 neurons and the last layer of 4 neurons (for the four classes). The activation for the four layers was LeakyReLU, and for the last layer was Softmax. The Categorical Cross Entropy was used for the loss function and the Adam function for the optimiser. To avoid the effect of random initialisation, the DNN classification was repeated 10 times in each experiment (the mode of the classes were used for calculating the performance of the classifier).

4. Results

4.1. Classification

For the baseline (without feature augmentation), first, the eight types of features were extracted (we refer to them as the *original (*ORG. *) features*) and then passed to the LR and the DNN classifiers to distinguish between the four classes (HC, FMD, MCI, and ND). The feature values were normalised and scaled between 0 and 1. We also combined all features (COMB. ; a total number of 747 features) and repeated the classification tasks. The results of the classifiers in terms of F-score (F-s) are summarised in Table 2 and 3 (column 2) respectively. As can be seen, the results are different for the two classifiers. For the LR classifier, CA, IS09 and MFCC had better F-s than the others (53.9%, 50.7% and 41.8% respectively), while FT had the worst (22%; even less than the chance level of 25%). However, COMB. features achieved around 55% F-s (better than any of the individual feature types). For the DNN classifier, IS09, CA and WV-PCA had the three best F-s (43.9%, 42.2% and 40.2% respectively), and similarly FT reached the lowest F-s (26.7%; slightly better than chance level). Also COMB. features gained the highest F-s (49.3%). On the whole, using ORG. features, the LR classifier outperformed the DNN for both individual feature types and COMB. features.

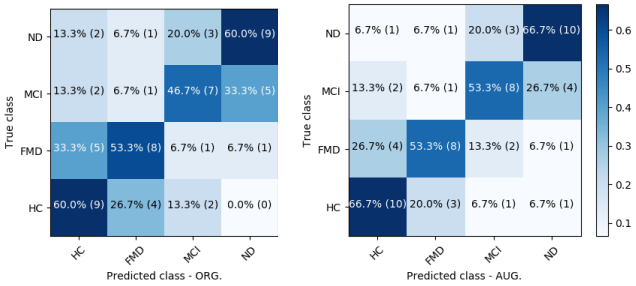

Figure 1: *Confusion matrix of the* COMB. *LR classifier.*

Table 2: *The LR classification results using the original features (*ORG. *) and augmented features (*AUG. *).*

Feature set	ORG. F-score%	AUG. F-score%
CA (20)	**53.9**	47.3
WV-PCA (7)	36.4	**38.6**
FT (6)	22.0	**40.5**
LX (60)	**39.0**	30.4
AC (60)	**36.0**	29.9
SilSP (10)	30.2	**33.8**
MFCC (200)	**41.8**	37.3
IS09 (384)	**50.7**	48.1
COMB. *(747)*	*54.9*	***59.8***

Then, the ORG. features were passed to the generative model and the augmentation processes were repeated 25[1] times. The augmented features were given to the LR and the DNN classifiers. The results of the classification are in Table 2 and 3 (column 3) respectively. Augmentation on the LR classifier only improved the F-s for some feature types, namely the FT, WV-PCA, and SilSP features (22% to 40.5%, 36.4% to 38.6%, and 30.2% to 33.8% respectively), however, overall, the COMB. features increased the F-s from 54.9% to 59.8%. The augmented features (AUG.), improved the DNN results significantly: all the individual feature types except for IS09 (with a slight drop from 43.9% to 43.8%), gained better F-s, and the COMB. features achieved the highest F-s of 62%.

To show the effect of feature augmentation on the classifiers for the individual classes, the confusion matrix (CM) of the classifiers were drawn. Figure 1 compares the CM of the LR classifier using COMB. ORG. features (left) to COMB. AUG. features (right). The LR ORG. classifier could correctly identify 60% (9/15) of the ND and HC groups, and 53% (8/15) of the FMD group. The most confusion was seen for the MCI class (33% wrongly identified as ND, 13.3% as MCI, and 6.7% as FMD). The LR AUG. classifier, however, could improve the correct identification of the three groups by 6.7% (ND and HC from

[1]We tried 50, 100 and 1000 times as well, however, increasing the number of repetition did not produce better results.

Table 3: *The DNN classification results using the original (*ORG. *) and augmented features (*AUG. *).*

Feature set	ORG. F-score %	AUG. F-score %
CA (20)	42.2	**45.9**
WV-PCA (7)	40.2	**40.8**
FT (6)	26.7	**29.4**
LX (60)	31.3	**41.2**
AC (60)	31.0	**33.4**
SilSP (10)	32.0	**37.5**
MFCC (200)	38.1	**42.7**
IS09 (384)	**43.9**	43.8
COMB. *(747)*	*49.3*	***62.0***

60% to 66.7%, and MCI from 46.7% to 53.3%), while the performance stayed the same for the FMD group. Figure 2 shows the CM for the DNN classifier using COMB. . ORG. features (left) vs. COMB. AUG. features (right). Similarly, MCI group had the highest confusion (26.7% as ND and FMD, 20% as HC) for the DNN classifier on all the original features. The augmentation process could improve correct identification of MCI by 26.7%, ND by 13%, and HC by 6.7% (FMD stayed the same). That is, the majority of the improvement seen by augmenting the features comes from improvements in performance for the class that had the lowest performance beforehand.

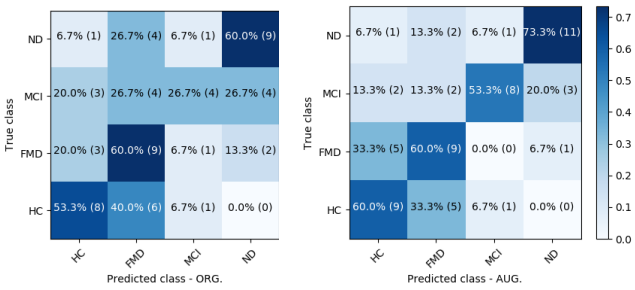

Figure 2: *Confusion matrix of the* COMB. *DNN classifier.*

4.2. Feature selection

The recursive feature elimination (RFE) and cross-validation feature selection (from the Sklearn python module [34]) were used[2]. The toolkit, however, provides the approach only for the conventional classifiers including the LR, not for DNNs. Using the RFE on the LR, the top features were extracted for both the original (RFE (ORG.)) and the augmented features RFE (AUG.) (105 and 19 features respectively). The top 105 RFE (ORG.) features contained features from across all of the individual feature types, however, for the RFE (AUG.), the top 19 contained only features from 5 features type groups: CA, LX, SilSP, MFCC, IS09. The full set of RFE (AUG.) features are listen in Table 4).

The classification tasks were repeated using RFE (ORG.) and RFE (AUG.) features and the results are summarised in Table 5. The LR classifier using RFE (ORG.) features achieved F-s 81% and with AUG. features improved to 88.3%, while the DNN gained F-s 77.5% and 80% respectively. Figure 3 shows the CM of the LR classifier on RFE (AUG.) that could identify correctly 93% of MCI (the highest improvement compared to

[2]We tried the univariate statistical tests (chi2, f-test, and mutual information test) and the results were (between 56% to 65% F-s).

Table 4: *The top RFE (*AUG. *) 19 features.*

Feature	Feature names
CA	Avg-NoOfTopicsChanged
LX	Sum-OtherPartOfSpeech
SilSP	Max-Silence
MFCC	Avg-MFCC-20, Avg-MFCC-40, Min-MFCC-6, Sum-MFCC-22
IS09	Linregc1-MFCC-4, linregc1-MFCC-6, Skewness-MFCC-10, Mean-MFCC-11, Linregc1-MFCC-12, Min-ZCR-A, MinPos-ZCR-A, Min-VoiceProb-A, Skewness-MFCC-Δ-1, Kurtosis-MFCC-Δ-8, Linregc2-MFCC-Δ-12, Mean-F0-Δ

Figure 1 left), as well as 87% of ND, FMD and HC groups.

Table 5: *Classification results using RFE (*ORG. *) and RFE (*AUG. *) features.*

Classifier	Feat.sel	F-score %
LR	RFE (ORG.)	81.0
LR	RFE (AUG.)	**88.3**
DNN	RFE (ORG.)	77.5
DNN	RFE (AUG.)	80.0

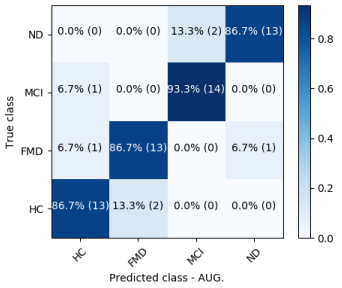

Figure 3: *Confusion matrix of RFE (*AUG. *) LR classifier.*

5. Conclusions

We have shown that feature based augmentation using VAE boosts the performance of both LR and DNN classifiers. Analysis showed that for the LR this improvment was only seen for some feature types (more conceptual features), for the DNN almost all individual feature types benefitted from the proposed feature augmentation method. Furthermore, augmentation for all features combined, gained the most improvements for the two classifiers. Augmentation even improved the F-score of the classifiers after applying a feature selection approach. In future work, we will focus on improving the DNN structure and the generative model, as well as the feature selection technique for the DNN classifier.

6. Acknowledgements

This research has been partly supported under the European Union's H2020 Marie Skłodowska-Curie programme TAPAS (Training Network for Pathological Speech processing; Grant Agreement No. 766287)

7. References

[1] Alzheimers.org.uk, "Facts for the media," 2020, accessed on April 23, 2020. [Online]. Available: https://www.alzheimers.org.uk/about-us/news-and-media/facts-media

[2] Dementia Statistics, "Deaths due to dementia," 2018, accessed on October 12, 2019. [Online]. Available: https://www.dementiastatistics.org/statistics/deaths-due-to-dementia

[3] C. Elsey, P. Drew, D. Jones, D. Blackburn, S. Wakefield, K. Harkness, A. Venneri, and M. Reuber, "Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics," *Patient Education and Counseling*, vol. 98, pp. 1071–1077, 2015.

[4] D. Jones, P. Drew, C. Elsey, D. Blackburn, S. Wakefield, K. Harkness, and M. Reuber, "Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders," *Aging & Mental Health*, vol. 7863, pp. 1–10, 2015.

[5] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "An avatar-based system for identifying individuals likely to develop dementia," *Proc. Interspeech*, pp. 3147–3151, 2017.

[6] B. Mirheidari, D. Blackburn, A. Venneri, M. Reuber, T. Walker, and H. Christensen, "Detecting signs of dementia using word vector representations," in *Proc. Interspeech*. ISCA, 2018.

[7] B. Mirheidari, D. Blackburn, R. O'Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Computational cognitive assessment: Investigating the use of an intelligent virtual agent for the detection of early signs of dementia," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2732–2736.

[8] R. O'Malley, B. Mirheidari, K. Harkness, M. Reuber, A. Venneri, T. Walker, H. Christensen, and D. Blackburn, "A fully automated cognitive screening tool based on assessment of speech and language," *Journal of Neurology, Neurosurgery & Psychiatry*, 2020, in preparation.

[9] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[10] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4910–4914.

[11] H. Hu, T. Tan, and Y. Qian, "Generative adversarial networks based data augmentation for noise robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5044–5048.

[12] P. Sheng, Z. Yang, H. Hu, T. Tan, and Y. Qian, "Data augmentation using conditional generative adversarial networks for robust speech recognition," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 121–125.

[13] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using gans for speech emotion recognition," *Proc. Interspeech 2019*, pp. 171–175, 2019.

[14] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[15] D. Michelsanti and Z. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *arXiv preprint arXiv:1709.01703*, 2017.

[16] B. Mirheidari, Y. Pan, D. Blackburn, R. O'Malley, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Data augmentation using generative networks to identify dementia," *arXiv preprint arXiv:2004.05989*, 2020.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[18] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[19] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[20] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[21] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, M. V. Dickie, D. Alexander H., J. Wardlaw, and D. Rueckert, "Gan augmentation: augmenting training data using generative adversarial networks," *arXiv preprint arXiv:1810.10863*, 2018.

[22] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 289–293.

[23] P. Chaudhari, H. Agrawal, and K. Kotecha, "Data augmentation using mg-gan for improved cancer classification on gene expression data," *Soft Computing*, pp. 1–11, 2019.

[24] D. Bhattacharya, S. Banerjee, S. Bhattacharya, B. U. Shankar, and S. Mitra, "Gan-based novel approach for data augmentation with improved disease classification," in *Advancement of Machine Intelligence in Interactive Medical Image Analysis*. Springer, 2020, pp. 229–239.

[25] Y. Luo, L.-Z. Zhu, and B.-L. Lu, "A gan-based data augmentation method for multimodal emotion recognition," in *International Symposium on Neural Networks*. Springer, 2019, pp. 141–150.

[26] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision. IEEE 11th International Conference*, 2007, pp. 1–8.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[28] V. Manohar, D. Povey, and S. Khudanpur, "Jhu kaldi system for arabic mgb-3 asr challenge using diarization, audio-transcript alignment and transfer learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 346–352.

[29] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[31] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Interspeech 2016*, 2016, pp. 2001–2005.

[32] F. Chollet *et al.*, "Keras: Deep learning library for theano and tensorflow," *URL: https://keras. io/k*, vol. 7, p. 8, 2015.

[33] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[34] F. Pedregosa and G. Varoquaux, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.