Semi-Supervised Learning with Generative Adversarial Networks for Pathological Speech Classification

Nam H. Trinh
ADAPT Centre, School of Computing
Dublin City University
Dublin, Ireland
nam.trinh@adaptcentre.ie

Darragh O'Brien
ADAPT Centre, School of Computing
Dublin City University
Dublin, Ireland
darragh.obrien@dcu.ie

***Abstract*—One application of deep learning in medical applications is the use of deep neural networks to classify human speech as healthy or pathological. In such applications, the audio signal is transformed into a spectrogram that captures its time-varying content and the latter "images" are fed into a classifier for classification. A challenge in applying this approach is the shortage of suitable speech data for training purposes. Labelled data acquisition requires significant human effort and/or time-consuming experiments. In this paper, we propose a semi-supervised learning approach that employs a Generative Adversarial Network (GAN) to alleviate the problem of insufficient training data. We compare the classification performance of a traditional classifier and our semi-supervised classifier. We observe that the GAN-based semi-supervised approach demonstrates a significant improvement in terms of accuracy and ROC curve when supplied an equivalent number of training samples.**

***Index Terms*—Generative Adversarial Network, pathological speech classification, semi-supervised learning**

I. Introduction

Deep learning applications in health care have attracted significant research effort over recent years. One such application is the use of neural networks to classify speech as healthy or pathological. A challenge to improvement in this area is the shortage of labelled data as the production of quality data requires significant human effort and expertise (e.g. experts to label a speech sample with an associated pathology) and time-consuming experiments (e.g. recordings of human speech that may give rise to ethical and privacy concerns).

Semi-supervised learning is a method of learning incorporating both labelled and unlabelled data into training models [1] [2]. While labelled data are expensive, unlabelled data are plentiful and inexpensive. Training with both labelled and unlabelled data enables the network to alleviate the data shortage problem and improve overall classification performance.

This paper explores a semi-supervised learning method employing a Generative Adversarial Network (GAN) for pathological speech classification. We demonstrate that the semi-supervised GAN outperforms the traditional Convolutional Neural Network (CNN) in terms of accuracy and area under the curve (AUC) when supplied with identical labelled training data. In Section 2, an overview of related work in the areas of pathological speech classification, GANs and semi-supervised learning is presented. In Section 3, we describe our approach to applying a GAN to semi-supervised learning for pathological speech classification. Section 4 describes our experimental settings and results. Section 5 concludes the paper.

II. Related Work

A typical pathological speech classification process (as depicted in Figure 1) consists of two main components: a feature extractor that applies speech signal processing techniques to compute salient features and a classifier that categorizes those features as indicative of healthy or pathological speech. For example, previous approaches employed a Support Vector Machine (SVM) as a classifier with Mel-Frequency Cepstral Coefficients (MFCCs) as input features. This approach has achieved an accuracy of 88% [3] on a dataset consisting of over 3750 healthy and Parkinson's speech samples.

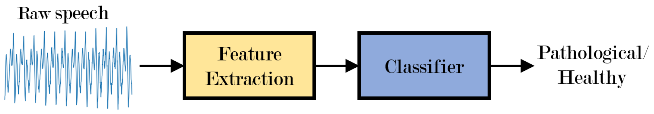

Fig. 1. A typical pathological speech classification model

With the emergence of deep learning algorithms, pathological speech classification models based on neural networks have also been proposed. For example, in [4], [5], [6] and [7], MFCCs serve as input vectors to a Multi Layer Perceptron (MLP). MLP drawbacks however, include overfitting and a potentially long training period due to the large number of model parameters. To address MLP-related issues, Convolutional Neural Network (CNN)-based models were proposed. Using a CNN-based approach (with a CaffeNet architecture), a state-of-the-art result of 98.77% accuracy was reported with the

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. Nam H. Trinh is funded by Science Foundation Ireland under grant No. 17/RC/PHD/3488.

Saarbrucken Voice Database consisting of 1616 pathological speech samples and 686 normal speech samples in the form of sustained vowel /a/ [8]. However, much work to date in the area of pathological speech classification has assumed that an adequate corpus of training samples (including both normal and pathological speech) is available to the model to be trained. In this paper we explore how a lack of training data can be mitigated through semi-supervised learning.

Recently, Generative Adversarial Networks (GANs) were proposed as a means of generating highly realistic images [9]. Since their introduction, an important application of GANs is in semi-supervised learning. In such a setting, a GAN's generator generates unlabelled data representing a real data distribution and a multi-class (rather than binary) discriminator classifies input data. Semi-supervised learning with GANs, introduced in [10] as a Categorical GAN or CatGAN, has shown significant improvements compared to traditional classifiers in image classification with several benchmark datasets. In [11], several features and training techniques were proposed to improve the performance of GANs for semi-supervised learning. In [12], the proposed semi-supervised GAN outperforms the traditional CNN-based approach with the same number of training samples with the MNIST dataset. In [13], a semi-supervised GAN also yields an accuracy gain with CIFAR10 and SHVN datasets. The proposed method in [14] is to train data generation and semi-supervised classification in parallel and achieves state-of-the-art results with several datasets. Besides the GAN-based approach, variational generative methods such as the Variational Auto-Encoder (VAE) are also employed in semi-supervised learning as an approximate Bayesian inference method to extract data density information for prediction [15]. In [16] a VAE is successfully employed for semi-supervised learning.

III. Methodology

In this section, we describe our method for modifying the traditional GAN architecture to suit the task of semi-supervised pathological speech classification.

The original GAN [9] is illustrated in Figure 2. GANs are generative models taking random noise as input and generating a real data distribution. A vanilla GAN consists of a discriminator and a generator. The generator takes random noise as input and generates new data samples. The discriminator's objective is to discriminate between real and generated samples (provided by the generator). The two networks compete with each other until an equilibrium point is reached where the discriminator cannot reliably discriminate between real and fake data.

Let D be the discriminator and G be the generator. The minimax game between D and G is modelled mathematically as follows:

$$V(G,D) = \min_{G} \max_{D} E_x[logD(x)] + E_z[1 - log(D(G(z)))] \tag{1}$$

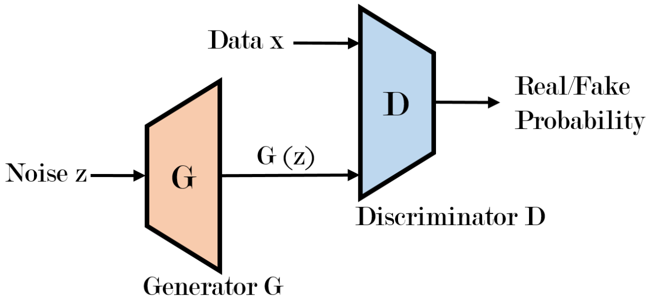

Fig. 2. The Original GAN [9]

where E_x is the expected value over all real data samples, $D(x)$ is the probability that a real data sample is categorized as real, E_z is the expected value over all noise samples, $G(z)$ is the generated output from the generator from input noise z. The objective of the training process is to train D to maximize the probability of classifying generated samples $G(z)$ as fake and to train G to convince D that generated samples, $G(z)$, are real. In other words, D is trained to maximize the loss function (1) while G is trained to minimize (1).

Semi-supervised GAN To alleviate the problem of a shortage of training data, unlabelled and labelled data are incorporated to enhance the decision boundary as depicted in Figure 3. By incorporating unlabelled data into the training process, the semi-supervised model attempts to shift the decision boundary to better cluster the data distribution [2]. This can be viewed as the model attempting to first cluster the data and subsequently, finding the decision boundary by assuming that unlabelled data points carry the same label as their neighbouring labelled data region.

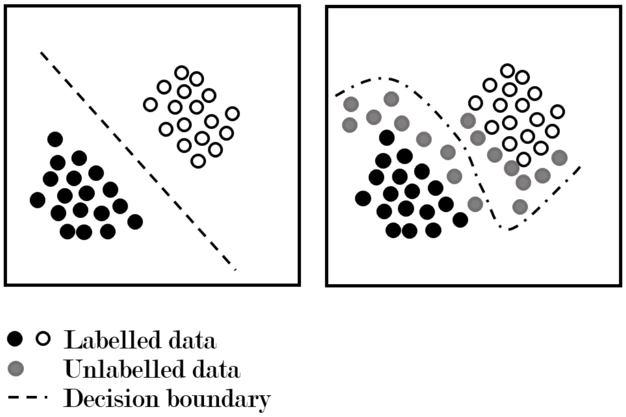

Fig. 3. Data points in supervised learning with limited amount of labelled data (left) and in semi-supervised learning with labelled data and unlabelled data (right)

A GAN-based approach for semi-supervised learning (as illustrated in Figure 4) incorporates generated data from the GAN's generator as unlabelled data and feeds these data into the discriminator. In this work, we modify the discriminator to not only classify a data sample as real or fake as in the original GAN but to also classify that sample as healthy or pathological. We modify the discriminator's architecture by adding an additional output layer in parallel with the output layer for real/fake classification to classify speech data as pathological or healthy. A detailed description of this implementation is presented in Section IV.

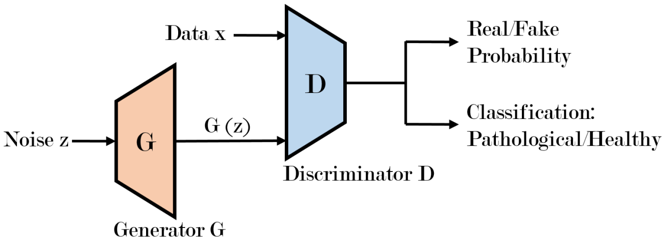

Fig. 4. The proposed Semi-Supervised GAN

IV. Experiments and Results

A. Experimental Design

To validate our semi-supervised learning method, we compare the performance of a semi-supervised GAN with a traditional CNN (with the latter having the same architecture as the discriminator in the proposed GAN) in a pathological speech classification task.

Dataset: To validate the proposed method, we train and test our model with the Spanish Parkinsons Disease Dataset (SPDD) [17]. SPDD consists of speech samples from 50 Parkinsons disease patients and 50 healthy controls, 25 men and 25 women per group. All subjects are Colombian native Spanish speakers. Several types of speech recordings are included:

- sustained vowels including /a/, /u/, /i/, /e/ and /o/ in Spanish,
- some specific words and phonemes,
- three sets of different words,
- conversational speech.

We use speech data extracted from the sustained /a/ vowel recordings at 44100 Hz in the experiments described below.

Speech Spectrogram Extraction: We use the librosa [18] speech processing framework to extract spectrograms from the speech signal using the Short-time Fourier Transform method with 128 frequency components. The resulting feature vectors of shape $(128, 96)$ are then zero-padded to $(128, 128)$ square vectors before being fed into our models.

Semi-Supervised GAN: Our Semi-Supervised GAN includes a discriminator and a generator as shown in Figure 4. The GAN's architecture is inspired by that of the DCGAN [19].

The discriminator is built with the architecture shown in Figure 5. The input to the discriminator has shape $(128, 128, 1)$. We use 2D convolutional layers with numbers of filters equal to $32, 64, 128, 256$ and 512. After each convolutional layer, we apply LeakyReLU with an alpha of 0.2, a drop-out layer with a rate of 0.25 and a batch normalization layer with a momentum of 0.8. The outputs of each of these layers are then flattened. Flattened data are then fed in two directions: to a discriminator for classification as fake or real and to a second discriminator for classification as pathological or healthy. For pathological speech classification, the final output layer is a single neuron with a sigmoid activation function for binary classification. For real/fake discrimination, we create a custom layer to calculate the probability of data being real. The output layer of this discriminator also consists of a single neuron with a sigmoid activation function for binary classification.

The generator is built according to the architecture shown in Figure 6. The shape of the random noise input to the generator is $(16, 16, 64)$ after being reshaped from $(16384, 1)$. Upsampling layers are subsequently employed to increase the dimensions of the data from $(16, 16)$ to $(128, 128)$. After each upsampling layer, we apply convolutional layers with a stride of 2 and a batch normalization layer with a momentum of 0.8. The generated spectrograms have a shape of $(128, 128, 1)$.

Traditional CNN: For the traditional CNN, we use the same architecture as for the discriminator in the GAN in order to ensure the performance of the semi-supervised GAN and that of the traditional CNN are directly comparable.

Hyperparameter Configuration: We train both models with 10000 epochs, with a batch size of 32, with the Adam optimizer [20] and with a learning rate of 0.00002. For loss functions, we use binary cross entropy for both networks. Accuracy is chosen as the evaluation metric as the number of healthy and pathological samples are balanced. We separate a total of 4000 speech spectrograms into 3200 samples for training and 800 samples for testing. Across each experiment we reduce the number of samples for training as follows: $3200, 2400, 1600, 1000$ and 800 samples. The performance of both networks are then compared for each number of training samples.

B. Results

1) Generative Results: A spectrogram generated by the proposed GAN is shown in Figure 7.

2) Classification Results: We compare the performance of the traditional CNN against the semi-supervised GAN based on two evaluation metrics: classification accuracy and area under curve (AUC).

The classification accuracy of the traditional CNN and the semi-supervised GAN using decreasing numbers of training samples are shown in Table I. The accuracies achieved by both models decrease as the number of samples is reduced from 3200 to 800. Comparing the two models, when trained with the same number of training samples, we observe a significant improvement in accuracy with the semi-supervised GAN. Our result confirms that semi-supervised learning with the GAN-based approach considerably improves the overall accuracy and alleviates the problem of insufficient training data.

To further assess our approach, we plot the Receiver Operator Characteristic (ROC) curves and measure the Area Under the ROC curve (AUC) for both models. The ROC curve captures the performance of a classifier by capturing the relationship between the false positive rate and the true

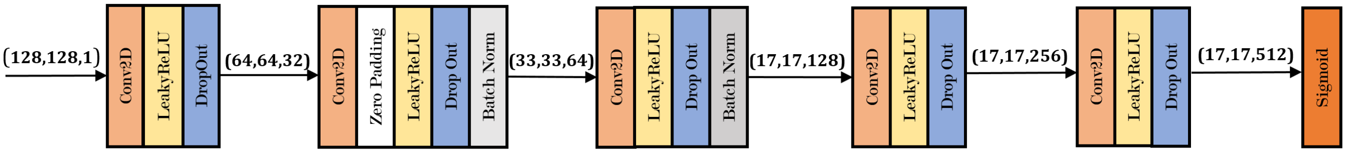

Fig. 5. The Discriminator's architecture

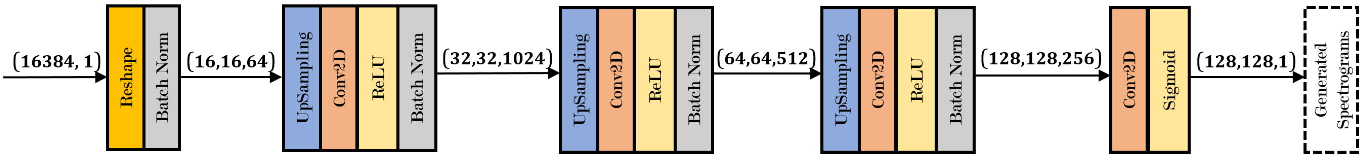

Fig. 6. The Generator's architecture

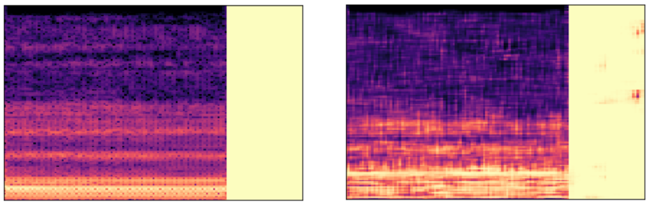

Fig. 7. An original (real) spectrogram (left) and a generated spectrogram by the proposed GAN (right)

TABLE I
ACHIEVED ACCURACY(%) WITH SPDD

Number of training samples	Traditional CNN	Semi-supervised GAN
3200	91.35	**96.63**
2400	90.00	**95.63**
1600	88.00	**90.75**
1000	86.50	**90.50**
800	79.62	**88.25**

positive rate along the x-axis and y-axis respectively using different threshold values. The ROC curves generated from 800 training data samples are shown in Figure 8.

The AUC of the semi-supervised GAN versus that of the traditional CNN is summarised in Table II. We observe a similar trend as for classification accuracy, i.e. the AUC of the semi-supervised GAN is higher than that of the traditional CNN for similar numbers of training samples. This result provide further evidence that our GAN-based semi-supervised approach outperforms our traditional CNN for this pathological speech classification task.

V. CONCLUSION

In this paper, we explore an approach employing semi-supervised learning with a GAN for pathological speech classification. We conduct experiments on the SPDD and compare the classification performance of a traditional CNN with that of a GAN-based semi-supervised approach under the same training conditions. Classifiers are compared using two evaluation metrics: classification accuracy and Area Under Curve (AUC). Our results indicate that for our speech classification task the semi-supervised approach both outperforms the traditional CNN in terms of higher classification accuracy and AUC and has the potential to alleviate the data shortage problem associated with speech pathology classification.

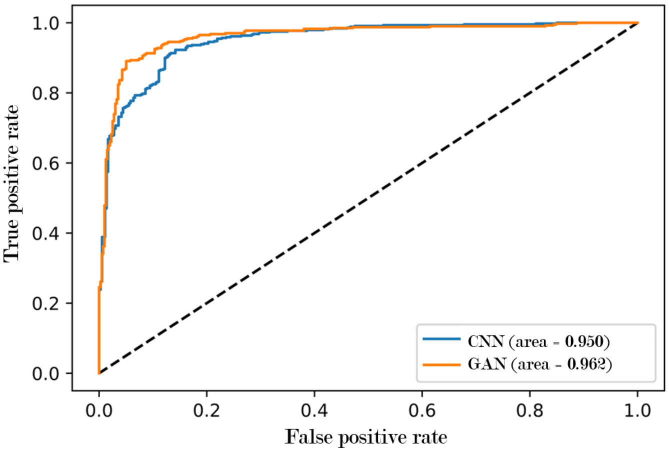

Fig. 8. ROC Curve of the traditional GAN and the semi-supervised GAN with 800 training data samples

TABLE II
AREA UNDER CURVE (AUC) WITH SPDD

Number of training samples	Traditional CNN	Semi-supervised GAN
3200	0.968	**0.995**
2400	0.975	**0.995**
1600	0.959	**0.982**
1000	0.936	**0.974**
800	0.950	**0.962**

ACKNOWLEDGEMENT

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional

Development Fund. Nam H. Trinh is funded by Science Foundation Ireland under grant No. 17/RC/PHD/3488.

References

[1] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[2] X. J. Zhu, "Semi-supervised learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2005.

[3] A. H. Poorjam, M. A. Little, J. R. Jensen, and M. G. Christensen, "A parametric approach for classification of distortions in pathological voices," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 286–290.

[4] J. Moon and S. Kim, "An approach on a combination of higher-order statistics and higher-order differential energy operator for detecting pathological voice with machine learning," in *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct 2018, pp. 46–51.

[5] Smitha, S. Shetty, S. Hegde, and T. Dodderi, "Classification of healthy and pathological voices using mfcc and ann," in *2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, Feb 2018, pp. 1–5.

[6] S.-H. Fang, Y. Tsao, M.-J. Hsiao, J.-Y. Chen, Y.-H. Lai, F.-C. Lin, and C.-T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, 2018.

[7] S. E. Shia and T. Jayasree, "Detection of pathological voices using discrete wavelet transform and artificial neural networks," in *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, March 2017, pp. 1–6.

[8] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41 034–41 041, 2018.

[9] 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.

[10] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *arXiv preprint arXiv:1511.06390*, 2015.

[11] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems*, 2016, pp. 2234–2242.

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

[13] A. Kumar, P. Sattigeri, and T. Fletcher, "Semi-supervised learning with gans: Manifold invariance with improved inference," in *Advances in Neural Information Processing Systems*, 2017, pp. 5534–5544.

[14] Z. Dai, Z. Yang, F. Yang, W. W. Cohen, and R. R. Salakhutdinov, "Good semi-supervised learning that requires a bad gan," in *Advances in neural information processing systems*, 2017, pp. 6510–6520.

[15] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[16] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther, "Auxiliary deep generative models," *arXiv preprint arXiv:1602.05473*, 2016.

[17] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. V. Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease." in *In Proc. Of the International Confer- ence on Language Resources and Evaluation (lrec)*, Reykjavik, Iceland, 2014, pp. 342–347.

[18] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[19] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.