Interspeech 2018
2-6 September 2018, Hyderabad

Patient Privacy in Paralinguistic Tasks

Francisco Teixeira, Alberto Abad, Isabel Trancoso

INESC-ID / Instituto Superior Técnico, University of Lisbon, Portugal

francisco.s.teixeira@tecnico.ulisboa.pt

Abstract

Recent developments in cryptography and, in particular in Fully Homomorphic Encryption (FHE), have allowed for the development of new privacy preserving machine learning schemes. In this paper, we show how these schemes can be applied to the automatic assessment of speech affected by medical conditions, allowing for patient privacy in diagnosis and monitoring scenarios. More specifically, we present results for the assessment of the degree of Parkinsons Disease, the detection of a Cold, and both the detection and assessment of the degree of Depression. To this end, we use a neural network in which all operations are performed in an FHE context. This implies replacing the activation functions by linear and second degree polynomials, as only additions and multiplications are viable. Furthermore, to guarantee that the inputs of these activation functions fall within the convergence interval of the approximation, a batch normalization layer is introduced before each activation function. After training the network with unencrypted data, the resulting model is then employed in an encrypted version of the network, to produce encrypted predictions. Our tests show that the use of this framework yields results with little to no performance degradation, in comparison to the baselines produced for the same datasets.

Index Terms: computational paralinguistics, cryptography

1. Introduction

Privacy is one of the most important issues regarding technology nowadays. As the number and reach of Software as a Service (SaaS) applications grows, so does the concern on user privacy. Applications concerning speech and other biometric signals have access to a great deal of information that a person might not want revealed, not even to the entity they entrusted their personal data with. This is specially true in Medicine, where patient privacy is given the utmost importance. Secure Machine Learning is a growing field of research that aims to combine state-of-the-art machine learning algorithms with secure computation frameworks, such as cryptography. This way, data can be entrusted to a Machine Learning as a Service (MLaaS) provider, ensuring its protection from both the service provider and malicious third parties.

Privacy in speech processing is an interdisciplinary topic of research, that also receives growing attention. Earlier work by Pathak et al. [1] applied Secure Multi-Party Computation to speaker verification, using Gaussian Mixture Models. A promising line of work involved Secure Binary Embedding [2], a scheme based in nearest-neighbor search, using secure randomized vector embeddings, created through quantized random projections. This scheme allows information to be leaked if the corresponding SBE hashes of two vectors are close enough, while preserving information-theoretic security. It has been applied to several speech processing tasks such as speaker verification [3] and query-by-example speech search [4]. An extension of SBE, Secure Modular Hashing (SMH) [5], allows the user to control the extension of the information leakage, thus enabling a trade-off between accuracy and security. This method was first applied to a speech emotion recognition task in [6].

The most recent efforts on privacy preserving speech processing have followed the progress in secure machine learning, such as Cryptonets [7], described in the next section. In particular, an Encrypted Neural Network was applied to speech emotion recognition by [6].

In this paper we provide a proof-of-concept on how these privacy-preserving schemes can be used in medical applications concerning speech, for screening, monitoring and diagnosis purposes. More concretely, we apply Encrypted Neural Network schemes, to the detection and assessment of Cold, Depression and Parkinson's Disease. This selection was mainly motivated by the availability of corpora distributed in paralinguistic challenges, and corresponding baseline results.

The paper is organized as follows: Section 2 reviews the theoretical background necessary for this work. Our approach is introduced in Section 3. Section 4 describes the datasets used, and the experimental setup. Section 5 includes the results obtained, together with a critical analysis. Section 6 presents our main conclusions.

2. Background

2.1. Homomorphic Encryption

First proposed by Rivest et al. [8], Homomorphic Encryption (HE) is a type of encryption that allows for certain operations to be performed in the encrypted domain while preserving their results in the plaintext domain. In other words, if for example an addition or multiplication is performed on two encrypted values, the result of this operation is kept when the corresponding encrypted value is decrypted.

Several Partially Homomorphic Encryption schemes allow for either additions or multiplications, such as the Paillier [9] and the El Gamal [10] cryptosystems, respectively, and although they can be of use in some cases, the restriction to a single mathematical operation makes them unsuitable for most applications. Fully Homomorphic Encryption (FHE) was designed to solve this limitation, when in 2009, Craig Gentry proposed a scheme in which both additions and multiplications were allowed and an arbitrary number of operations could be performed [11]. This scheme was however computationally unfeasible and Leveled Homomorphic Encryption (LHE) schemes rose as an alternative. LHE schemes take advantage of the fact that, in most applications, the user knows beforehand the number of arithmetic operations to be performed on encrypted values. Thus, as long as the number of operations does not exceed a previously set threshold, the scheme is both computationally feasible and provides correct results, however some limitations remain. In HE, operations increase the amount of noise in the encrypted values, and if a certain threshold is surpassed, it is impossible to recover their original value. LHE allows us to choose larger parameters that increase this noise threshold, but

10.21437/Interspeech.2018-2186

as these parameters increase, so does the computational complexity of the operations. Consequently there needs to be a trade-off between the number of operations to be computed in the encrypted domain, and the computational complexity of the application.

In this work we used SEAL's implementation of the Fan and Vercauteren(FV) scheme [12]. Specific details regarding this implementation, including a detailed security review of the scheme, can be found in its manual [13].

2.2. Encrypted Neural Networks

Neural networks have been shown to be especially suited for secure machine learning applications using FHE [7][14][15], as most operations can be replaced by additions and multiplications.

To comply with the restrictions FHE poses, some modifications are necessary. As stated in the previous section, a large number of operations translates into a high computational cost, therefore the number of hidden layers of the network needs to be reasonably small, to limit the amount of operations computed in the encrypted domain. Moreover, as HE only allows additions and multiplications to be computed, only polynomial functions can be computed, and thus activation functions have to be replaced by polynomials.

Considering multiplication has the highest toll on the noise budget (i.e. the amount of noise allowed before the noise threshold is reached), it is especially important to keep the number of multiplications to a minimum, which means that the degree of the polynomials replacing the activation functions needs to be as small as possible.

In view of the reasons stated above, it is necessary to find a suitable polynomial to replace the activation functions commonly present in neural networks. The REctified Linear Unit (*ReLU*) activation function is a widely used activation, and thus it has been the focus of most FHE neural network schemes, although other activation functions have also been considered, such as *tanh* and *sigmoid*.

The first approach to solve these constraints was proposed in Cryptonets [7], where the authors train a Convolutional Neural Network, with x^2 as a replacement for the *ReLU* activations. It is worth noting that in this method, the network is first trained with its original architecture, and the inference phase is done in a simplified network, for performance reasons.

In the work of Chabanne et al. [14], the authors go further than Cryptonets and approximate the *ReLU* function, using Least Squares, with second, fourth and sixth degree polynomials. These polynomials have a small interval of stability, as their approximation is made around 0, which motivated the authors to propose a key innovation, the introduction of a Batch Normalization (BN) Layer before each Activation layer. The use of the BN layer guarantees that the distribution of the inputs of each activation is close to the standard normal and that most inputs fall within the convergence interval of the approximation [16]. To avoid the unbounded derivative of polynomials, the authors train this network with regular *ReLUs*, which are replaced by their corresponding polynomial approximations during the prediction phase, although this problem only emerges in deep networks.

Another advance was made in CryptoDL [15], mainly concerning the polynomial approximations of the *ReLU*. This work suggests the use of modified Chebyshev Polynomials to approximate, not the *ReLU* directly, but its derivative, as the authors argue that the derivative of the activation is more important than the shape of the function itself. Furthermore, since the derivative of the *ReLU*, the Step function, is non-differentiable at 0, the Sigmoid function is used instead. This approach has the advantage of approximating the function in a selected interval, and not only around zero, which allows for wider convergence intervals. The resulting polynomial is then integrated, and used to train the network. Using this method, the authors were able to obtain the best results of the three works mentioned above, in the MNIST dataset [17].

In Dias et al. [6], the authors applied the scheme proposed by Chabanne et al. [14], to a speech emotion recognition task, using a small neural network.

In general, the training stage is still too computationally expensive to be performed in the encrypted domain. For this reason, most frameworks are trained with unencrypted data.

3. Encrypted Network Architecture

Since our purpose is to make a proof-of-concept on how HE might be used in medical applications, and considering the fact that the datasets that were used for training are relatively small, we used a shallow neural network, adapted with the methods described in the previous section.

The network follows a similar structure to the one used by Dias et al. [6]. It is composed by two Fully Connected (FC) Layers, each followed by a Batch Normalization Layer and an Activation Layer, these are followed by a final output FC layer. For classification tasks, a Sigmoid Activation Layer is inserted after the last FC layer. Following the approaches of Cryptonets [7] and CryptoDL [15], we trained the network with the polynomial approximations of the activation functions.

During training, a dropout layer was inserted before the second and third FC layers. This serves as a regularizer, to help prevent the network from overfitting [18]. The BN layer already present in the architecture, also has a regularization effect [16], apart from assuring that the inputs of the activation layer are normally distributed.

3.1. Polynomial Approximations

As previously stated, the activation functions in the network need to be replaced by polynomials. We follow the approach suggested by CryptoDL, and use Chebyshev Polynomials, to approximate the ReLU through its derivative, or to be more precise, a function similar to its derivative, the Sigmoid.

We approximate this function, in the interval $[-120, 120]$, with a high degree Chebyshev polynomial, using Python's *numpy* package. The polynomial is then integrated, and it is necessary to compute its constant term. We want the polynomial to have a similar behaviour in the convergence interval as the ReLU function, this means that we not only want to have the smallest error between the real function and its approximation, but also want the approximation to have a value as close to zero as possible when $x < 0$. This can be achieved by minimizing the MSE error between the original function, and the polynomial, with a regularizer that penalizes negative values in the interval, as the one shown in Equation 1,

$$R = \sum_{n} max(-p(n) + c, 0)^2 \tag{1}$$

where $p(n)$ is the polynomial, and c is the constant to be optimized.

Using Python's *scipy.optimize*, with 10000 data points in the interval $[-50, 50]$, we were able to obtain the complete polynomial in Equation 2.

$$p(x) = 0.03664x^2 + 0.5x + 1.7056 \quad (2)$$

The approximation is limited to a second degree polynomial, to abide by the constraints stated in the previous sections.

For classification tasks, it is helpful to have a function constrained between 0 and 1 for the output. This is not possible using low degree polynomials, but it is possible to build a linear polynomial that is bounded between the same values in a given interval. To this end, we also approximated the Sigmoid function in the interval $[-10, 10]$, with a linear polynomial, obtaining:

$$p(x) = 0.004997x + 0.5 \quad (3)$$

4. Experimental Setup

In this section we briefly describe the datasets, features and methods used in our experiments.

4.1. Data

4.1.1. Cold: URTIC

The Upper Respiratory Tract Infection Corpus (URTIC) [19] was provided by the Institute of Safety Technology of the University of Wuppertal, Germany, for the Interspeech 2017 ComParE Challenge [20]. It consists of 630 subjects, performing both scripted and spontaneous speech tasks, recorded in quiet rooms. The 630 recordings are divided in 28,652 segments, varying between 3 and 10 seconds, amounting to a total of 45h. Both the Train and Development partitions include 210 subjects, of which 173 were healthy controls, and 37 were having a cold. These partitions include 9,505 and 9,596 segments, respectively. The labels included in this corpus indicate if the subjects have a cold, or if they are healthy, hence this dataset is used for a classification task.

4.1.2. Depression: DAIC-WOZ

The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ), is a subset of the DAIC database [21]. This database is composed by clinical interviews designed to support the diagnosis of psychological distress conditions. The DAIC-WOZ subset includes interviews conducted by a virtual interviewer, in both video and audio formats, containing 189 sessions, of which 106 are included in the training partition, and 34 in the development set. The training set includes 76 healthy controls and 30 subjects with depression. The development set has 12 subjects with depression, and 22 controls. This corpus comes with a set of labels classifying the interviewees in the PHQ-8 scale [22], as well as whether they are Depressed or not, allowing for both classification and regression tasks to be performed. The dataset also includes segmentation files that allow us to split the interviews into segments and to remove the interviewer's interventions.

4.1.3. Parkinson's Condition

In the assessment of Parkinson's Condition, we use the database provided for the ComParE Challenge of Interspeech 2015 [20]. This database is a subset of a Spanish corpus from Universidad de Antioquia [23], that includes 50 patients, 25 male and 25 female, recorded in noise controlled conditions. These subjects are classified in the UPDRS-III scale [24], ranging from 5 to 92. The recordings include 42 speech tasks, such as uttering isolated words, sentences, reading a text, a monologue and the rapid repetition of the syllables /pa-ta-ka/, /pa-ka-ta/ and /pe-ta-ka/. Of the 50 patients, 35 are included in the training set and the remaining 15 are included in the development set. This dataset is used for a regression task.

4.2. Feature Extraction and Pre-processing

For both Cold and Depression tasks, the *extended Minimalistic Acoustic Parameter Set* (eGeMAPS) [25] was used. This feature set was designed to serve as a baseline for paralinguistic tasks. It is composed by 88 paralinguistic features, including information on frequency, energy, spectral and cepstral characteristics.

For the Parkinson's Condition task, as eGeMAPS did not obtain very significant results, a more specific feature set was used. This set, developed by Pompili et al. [26], includes common features used in the assessment of Parkinson's Disease. It is based on GeMAPS [25], containing 36 GeMAPS based features, alongside with 78 MFCC based features, resulting in a 114-dimensional feature vector. Both feature sets were extracted from the audio files using openSMILE [27].

In every feature vector, the features were zero-centered and scaled using the mean and standard deviation computed from the training data of their respective datasets.

4.3. Neural Network Training

For our experiments, we implemented the model described in Section 3 in Keras [28], with both polynomial and regular activation functions. In these models, the first and second fully connected layers have 120 and 50 hidden nodes, respectively, while the output FC layer has only one, for both classification and regression tasks.

All four models, classification for Cold and Depression, and regression for Depression and Parkinson's Disease, were trained with a learning rate of 0.02, 100 epochs and a weight decay of 0.005, using RMSProp, an adaptive learning rate backpropagation algorithm implemented in Keras. As loss functions we used Binary Cross Entropy (BCE) for classification and Mean Squared Error (MSE) for regression.

As was previously stated a dropout layer was included before the second and third FC layers, to prevent overfitting during training. The values of the probability of dropout used were found through random search: 0.3746 and 0.5838 for the Cold; 0.092 and 0.209 for Depression; 0.877 and 0.246 for Parkinson's Condition.

When training for classification, it was noted that the Cold and Depression training partitions had a large misrepresentation of the subjects that presented the respective condition. To balance this, weights were attributed to each class. For Depression a weight of 0.8 was attributed to positive samples, and 0.2 was given to negative samples. In Cold, the difference was larger, thus we gave weights of 0.9 and 0.1 to the positive and negative samples, respectively.

4.4. Encryption Parameters

To make predictions in an encrypted setting, the parameters of the models trained with Keras were applied to an implementation of the same network using encrypted operations, implemented in C++ with SEAL [13]. This library also requires a set of encryption parameters to be selected. These are related with the security of the encryption and the number of operations allowed on the encrypted domain. We selected a polynomial modulus of 4,096 and a plaintext modulus of 2^{30}.

5. Results

In this section we provide the results obtained in our tests. For each dataset we present the values corresponding to two neu-

Table 1: *The baseline and the results obtained for Cold classification.*

Method	UAR(%)	F1 Score	Precision	Recall
Baseline	66.1	-	-	-
NN	66.9	.279 (.687)	.169 (.959)	.803 (.535)
ENN	66.7	.278 (.687)	.168 (.958)	.799 (.535)

ral networks: an unencrypted neural network (NN) trained with normal activation functions, and an Encrypted Neural Network (ENN), trained with polynomial approximations and performing encrypted predictions. All results presented correspond to the Development partition of the datasets, at the segment level. When referring to the F1 scores, Precision and Recall, the value in brackets always corresponds to the negative (0) class, while the value outside the brackets corresponds to the positive (1) class. In the Depression task, a positive sample means that the subject is Depressed, and in the Cold task that the subject has a cold.

5.1. Cold

In Table 1, it is possible to observe the results regarding the Cold classification task and the baseline stated for the Interspeech 2017 ComParE Challenge [29]. In this challenge, UAR was chosen to be the metric, but we also present the F1 scores, together with the Precision and Recall for both classes, for a better understanding of the performance of our models. Both the model with the original activation functions and the encrypted model performed above the baseline. In this case, there is a very small performance degradation from the unencrypted NN to the ENN. Most likely, this difference is not due to the ReLU approximation, but because of the output Sigmoid, which, in the NN case, is a bounded function, and in the ENN is a linear polynomial.

5.2. Depression

Tables 2 and 3 show the results regarding classification and regression for the Depression task. The metrics used correspond to the ones presented in the AVEC 2016 Challenge [30]. In this challenge, the metrics used for classification were the F1 score, Precision and Recall, for both classes. For regression, the metrics were the RMSE and MAE. We also include UAR as a metric, for coherence with the Cold task. In the case of classification, when comparing the results of the NN and ENN, there is just a slight performance degradation due to the polynomial approximations. For the regression task, the ENN performs better than the NN for both the Root Mean Square Error (RMSE) and the Mean Average Error (MAE). The baseline results presented for the AVEC 2016 Challenge are reported at the interview level, combining frame level results with majority voting and a simple average. For fair comparison with the other tasks, our results are reported at the segment level, instead of the interview level. Nevertheless, the baseline provided for the challenge yields a UAR of 69.9% for classification, and for regression a RMSE of 6.74 and a MAE of 5.36. Using majority voting, the ENN obtained a UAR close to that of the baseline, achieving 67.9%. This discrepancy may be due to the fact that AVEC's baseline uses features from COVAREP [31], whereas our experiment was conducted using eGeMAPS.

Table 2: *The baseline and the results obtained for Depression classification.*

Method	UAR(%)	F1 Score	Precision	Recall
NN	60.6	.586 (.515)	.454 (.782)	.827 (.384)
ENN	60.2	.541 (.642)	.480 (.713)	.621 (.584)

Table 3: *Baseline and results obtained for Depression severity.*

Method	RMSE	MAE
NN	7.43	5.80
ENN	6.77	5.64

5.3. Parkinson's Condition

The results for Parkinson's Condition can be observed in Table 4. Here we chose to include, not only the metric proposed in Interspeech 2015 [20], Spearman's Correlation Coefficient ρ, but also the RMSE and MAE. The reason for this was to have the same metrics present in both regression tasks. The NN performs better than the baseline with regard to Spearman's Coefficient, but the ENN does not. Considering the RMSE and MAE, there is not a relevant difference in the results of both models, however both are lower for the ENN.

Table 4: *Baseline and results obtained for Parkinson's Condition.*

Method	RMSE	MAE	ρ
Baseline	-	-	0.492
NN	16.6	12.6	0.507
ENN	16.0	12.5	0.450

6. Conclusions

This work contributes with a proof-of-concept on how paralinguistic health-related tasks can be made secure through the use of Fully Homomorphic Encryption.

The same generic architecture was adopted for the three tasks, without any customization effort, which results in almost no gains over the baselines. Health-related tasks are still typically characterized by limited amounts of training data, which in turn, limits the improvements potentially obtainable with state-of-the-art machine learning techniques, using speech as a single modality, without any speaker clustering. However, the slight difference between the results obtained by encrypted neural networks and their non-encrypted counterparts showed the validity of our secure approach. Nevertheless, the limited amount of data does not allow a thorough analysis of performance degradation in deeper networks.

As future work, we plan to investigate other promising solutions, such as using deeper neural networks, and adapting end-to-end architectures to the restrictions of FHE. Secure training is also an open problem, that if solved can contribute to the increase in size of existing databases, allowing for better models to be trained for real world applications.

7. Acknowledgements

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013.

8. References

[1] M. A. Pathak and B. Raj, "Privacy-Preserving Speaker Verification and Identification Using Gaussian Mixture Models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 397–406, Feb 2013.

[2] P. Boufounos and S. Rane, "Secure Binary Embeddings for Privacy Preserving Nearest Neighbors," in *Information Forensics and Security (WIFS), 2011 IEEE International Workshop*. IEEE, 2011, pp. 1–6.

[3] J. Portêlo, A. Abad, B. Raj, and I. Trancoso, "Secure Binary Embeddings of Front-end Factor Analysis for Privacy Preserving Speaker Verification," in *INTERSPEECH*, 2013, pp. 2494–2498.

[4] J. Portêlo, A. Abad, B. Raj, and I. Trancoso, "Privacy-preserving Query-by-Example Speech Search," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference*. IEEE, 2015, pp. 1797–1801.

[5] A. Jiménez, B. Raj, J. Portêlo, and I. Trancoso, "Secure Modular Hashing," in *Information Forensics and Security (WIFS), 2015 IEEE International Workshop*. IEEE, 2015, pp. 1–6.

[6] M. Dias, A. Abad, and I. Trancoso, "Exploring Hashing and Cryptonet based Approaches for Privacy-preserving Speech Emotion Recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference*. IEEE, 2018.

[7] R. Gilad-Bachrach, N. Dowlin, and K. Laine et al., "CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy." in *ICML*, ser. JMLR Workshop and Conference Proceedings, vol. 48, 2016, pp. 201–210.

[8] R. L. Rivest, L. Adleman, and M. L. Dertouzos, "On Data Banks and Privacy Homomorphisms," *Foundations of Secure Computation, Academia Press*, pp. 169–179, 1978.

[9] P. Paillier, "Public-Key Cryptosystems Based on Composite Degree Residuosity Classes," in *Advances in Cryptology - EUROCRYPT 99*, ser. Lecture Notes in Computer Science, vol. 1592. Springer, 1999, pp. 223–238.

[10] T. Elgamal, "A Public Key Cryptosystem and a Signature Scheme based on Discrete Logarithms," *IEEE Transactions on Information Theory*, vol. 31, no. 4, pp. 469–472, 1985.

[11] C. Gentry, "A Fully Homomorphic Encryption Scheme," Ph.D. dissertation, Stanford University, 2009.

[12] J. Fan and F. Vercauteren, "Somewhat Practical Fully Homomorphic Encryption." *IACR Cryptology ePrint Archive*, vol. 2012, p. 144, 2012, informal publication.

[13] K. Laine, H. Chen, and R. Player, "Simple Encrypted Arithmetic Library - SEAL v2.3.0," Microsoft, Tech. Rep., December 2017. [Online]. Available: https://www.microsoft.com/en-us/research/publication/simple-encrypted-arithmetic-library-v2-3-0/

[14] H. Chabanne, A. de Wargny, J. Milgram, and C. Morel et al., "Privacy-Preserving Classification on Deep Neural Network." *IACR Cryptology ePrint Archive*, vol. 2017, p. 35, 2017.

[15] E. Hesamifard, H. Takabi, and M. Ghasemi, "CryptoDL: Deep Neural Networks over Encrypted Data." *CoRR*, vol. abs/1711.05189, 2017.

[16] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." *CoRR*, vol. abs/1502.03167, 2015.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[18] N. Srivastava, G. E. Hinton, A. Krizhevsky, and I. Sutskever et al., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[19] J. Krajewski, S. Schnieder, and A. Batliner, "Description of the Upper Respiratory Tract Infection Corpus (URTIC)." in *INTERSPEECH*. ISCA, 2017.

[20] B. W. Schuller, S. Steidl, A. Batliner, and S. Hantke et al., "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition." in *INTERSPEECH*. ISCA, 2015, pp. 478–482.

[21] J. Gratch, R. Artstein, G. M. Lucas, and G. Stratou et al., "The Distress Analysis Interview Corpus of human and computer interviews." in *LREC*. European Language Resources Association (ELRA), 2014, pp. 3123–3128.

[22] K. Kroenke, T. W. Strine, R. L. Spitzer, and J. B. Williams et al., "The PHQ-8 as a measure of current depression in the general population," *J Affect Disord*, vol. 114, no. 1-3, pp. 163–173, Apr 2009.

[23] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. V. Bonilla, and M. C. Gonzalez-Rátiva et al., "New Spanish Speech Corpus Database for the Analysis of People Suffering from Parkinson's Disease." in *LREC*. European Language Resources Association (ELRA), 2014, pp. 342–347.

[24] C. G. Goetz, B. C. Tilley, and S. R. Shaftman et al., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results," *Mov. Disord.*, vol. 23, no. 15, pp. 2129–2170, Nov 2008.

[25] F. Eyben, K. Scherer, B. Schuller, and J. Sundberg et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 4 2016, open access.

[26] A. Pompili, A. Abad, P. Romano, and I. P. Martins et al., "Automatic Detection of Parkinson's Disease: An Experimental Analysis of Common Speech Production Tasks Used for Diagnosis." in *TSD*, ser. Lecture Notes in Computer Science, vol. 10415. Springer, 2017, pp. 411–419.

[27] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich Open-source Multimedia Feature Extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838.

[28] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.

[29] B. Schuller, S. Steidl, A. Batliner, and E. Bergelson et al., "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017.

[30] M. F. Valstar, J. Gratch, B. W. Schuller, and F. R. et al., "AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge," *CoRR*, vol. abs/1605.01600, 2016.

[31] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - A Collaborative Voice Analysis Repository for Speech Technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 960–964.