Interspeech 2018
2-6 September 2018, Hyderabad

Multi-Lingual Depression-Level Assessment from Conversational Speech Using Acoustic and Text Features

Yasin Ozkanca[1], *Cenk Demiroglu*[1], *Asli Besirli*[2], *Selime Celik*[2]

[1]Ozyegin University, Turkey
[2]Sisli Etfal Hospital, Turkey

yasin.ozkanca@ozu.edu.tr,
cenk.demiroglu@ozyegin.edu.tr,abesirli2006@yahoo.com,selimecelik2000@yahoo.com

Abstract

Depression is a common mental health problem around the world with a large burden on economies, well-being, hence productivity, of individuals. Its early diagnosis and treatment are critical to reduce the costs and even save lives. One key aspect to achieve that goal is to use voice technologies and monitor depression remotely and relatively inexpensively using automated agents. Although there has been efforts to automatically assess depression levels from audiovisual features, use of transcriptions along with the acoustic features has emerged as a more recent research venue. Moreover, difficulty in data collection and the limited amounts of data available for research are also challenges that are hampering the success of the algorithms. One of the novel contributions in this paper is to exploit the databases from multiple languages for feature selection. Since a large number of features can be extracted from speech, and given the small amounts of training data available, effective data selection is critical for success. Our proposed multi-lingual method was effective at selecting better features and significantly improved the depression assessment accuracy. We also use text-based features for assessment and propose a novel strategy to fuse the text- and speech-based classifiers which further boosted the performance.

Index Terms: Depression estimation, acoustic features, feature selection, multi-lingual applications

1. Introduction

Depression is a vital problem that affects a large portion of the population. It affects well-being and productivity of individuals as well as being heavy economic burden for the society [1]. Thus, inexpensive and accurate diagnosis with the help of technology is an increasingly important research challenge [2].

It has been shown that speech signal carries significant amount of information about mental health of the speakers [3, 4, 5]. In [6], phase distortion deviation that is used for voice quality examinations is found to be helpful for detecting depression. In [7], distortions in formant trajectories were used to detect depression. In [8], degradation in spectral variability was used. In [9], gender-dependent feature extraction was found to improve the detection performance.

Besides acoustics-only methods, there are also multi-modal approaches for detecting depression. In [10], face analysis and speech prosody are used for depression detection. Similarly, audio-visual features are used in [11, 12, 13, 14]. Retardations in motor control due to depression causes changes in coordination and timing of speech and face movements, which are used for audio-visual detection in [15].

Besides face features, text analysis of transcriptions have also been used as another mode of information [2]. In [16], transcription-derived features were used in addition to the speech features. Furthermore, sentiments analysis was performed on text and sentiment features were used to build an independent detector. Then, score fusion was used to combine acoustic and text-based system scores. Syntactic and semantic features were derived from transcriptions in [17] and shown to be effective indicators of depression.

In depression detection, another research challenge is to use speech data from other languages/cultures to train models. This approach is not only important for understanding universal cues of depression across different cultures/languages but also it allows use of data from other languages, which is important given the typically small amounts of data available in the public databases. In [18], prediction models built with a German database were shown to produce prediction scores in English that were correlated with self-assessment scores. In [19], combination of datasets in different languages was shown to yield high accuracy whereas if the train and test data are in different languages, performance was found to be lower.

Conversations with patients can be designed in a way to obtain data that is more indicative of depression, as opposed to a regular conversation. In [20], type of questions (positive and negative stimulus) during conversations have been shown to impact voice quality parameters in psychologically distressed subjects. Speech segments with higher articulation effort were found to be more informative for depression detection in [5].

This paper has two contributions. One of the contributions is novel algorithms for feature selection which was not explored as much in the literature. We propose a multi-lingual feature selection where Turkish and German databases were used together. Moreover, methods to improve redundancy and relevance computations in the case of data sparsity are proposed. The second contribution is a novel feature fusion technique where transcription-derived model predictions were used to adjust the predictions of the acoustic-only model when their predictions are highly conflicting. Significant improvements are obtained both for the Turkish and German databases using the proposed techniques.

2. Minimum redundancy maximum relevance (MRMR) feature selection

A large number of features can be derived from conversational speech to detect depression. However, building models with those features is challenging because of the curse of dimensionality especially given the typically small amounts of training data available in depression studies.

One way of reducing the dimensionality features is to use feature selection where features that are most relevant for the

10.21437/Interspeech.2018-2169

classification task and least correlated among themselves are selected. To that end, ”Minimum Redundancy Maximum Relevance” (MRMR) algorithm is commonly used [21, 22, 23].

In the MRMR approach, for maximizing the relevance of selected features for the classification task, F-statistic is used.

$$F(g_i) = [\sum_k n_k(\bar{g_k} - \bar{g})^2/(K-1)]/\sigma^2, \tag{1}$$

$\bar{g_k}$ is the mean of the g_i, within the kth class. $\bar{g}$ is the global mean of whole feature set. The number of classes denoted by K and σ^2 is the pooled variance:

$$\sigma^2 = [\sum_k (n_k - 1)\sigma_k^2]/(n-K), \tag{2}$$

where for each class, n_k and the σ_k are size and the variance of those classes. Relevance of the feature set S is then defined as

$$maxV_F, \qquad V_F = \frac{1}{|S|}\sum_{i\in S} F(i). \tag{3}$$

Redundancy is defined using Pearson correlation for every possible feature combination:

$$minW_c, \qquad W_c = \frac{1}{|S|^2}\sum_{i,j} |c(i,j)|, \tag{4}$$

where absolute value of the correlation $c(i,j)$ is used. Finally, the optimization criteria for MRMR is

$$max(V_F - W_c). \tag{5}$$

3. Proposed feature selection algorithms

We propose several algorithms to improve the performance of the MRMR method for the depression detection problem where data is typically limited and, therefore, computation of F-statistic and correlation is unreliable.

3.1. Multi-lingual computation of relevance

The F-statistic computation in Eq.(1) assumes that there is enough data for each class to compute the mean of each class reliably, which is not the case when the number of classes is large and the data is limited. In the multi-lingual approach, the core idea is to use speech data collected from depression patients in other languages for relevance computation.

In order to increase the number of available samples for each class, hence improve the computation of relevance, we exploit the samples available in a different language for the same or neighboring classes with reduced weights assigned to the samples as the neighbors are further away on the depression scale. To that end, we have changed the computation of $\bar{g_k}$ and n_k. The weight parameter γ is defined as

$$\gamma_t = e^{-t^2}. \tag{6}$$

where t indicates how close the neighbors are on the depression scale. Number of samples in class k, $\hat{n}_k$ is adjusted using the parameter γ, the amount of adjustment depends on how much we need to satisfy the N_{min} constraint.

$$\hat{n}_k = \sum_{j=-J_k}^{+J_k} \gamma_j n_{k+j} \tag{7}$$

J is set such that $\hat{n}_k > N_{min}$. Thus, by including data from the same and neighboring classes in a different database, we ensure that there are at least N_{min} samples for each class in the target database. The adjusted mean of each class k, $\bar{g}_k{}'$, is then

$$\bar{g}_k^{'} = \frac{1}{\hat{n}_k}\sum_{j=-J_k}^{+J_k}\sum_{s=0}^{n_{k-j}-1} \gamma_j g_{k-j}(s)k-j(s) \tag{8}$$

where $g_{k-j}(s)$ is sample s in class $k-j$. Thus, the final equation to compute F-score becomes:

$$F(g_i) = [\sum_k \hat{n}_k(\bar{g_k}^{'} - \bar{g})^2/(K-1)]/\sigma^2, \tag{9}$$

3.2. Clustering approach

Even though the Beck depression scale is from 1 to 63 with a step size of 1, given randomness in the responses to Beck questionnaire, the resolution is expected to be lower than that. Hence, the difference between a person with a score of 3 or 4 may not be as significant to warrant different classes for those two cases especially given the very limited training data available.

In the clustering approach, we clustered the depression classes and reduced the number of classes in the MRMR training process to improve the feature selection performance by increasing the data available for each class. In this approach, data is split uniformly into N_{clus} classes.

3.3. Robust computation of redundancy (RCR)

Class labels are not required for the computation of redundancy, as shown in Eq.(4). Thus, large amounts of unlabeled, i.e. depression scores not available, speech data can be exploited for computing the redundancy. In this approach, we propose using unlabeled speech databases to compute redundancy when the amount of labeled data is limited.

4. Fusion with text-based features

4.1. Description of text-based features

We tagged sentiment (positive/negative/neutral) of the patients’ responses to interviewers questions. Even though sentiment tagging can be done automatically by using online services, here it is done manually. Total numbers of positive, negative, and neutral answers during the conversation are used as a three dimensional feature vector.

Using the timing information in the transcriptions, average length of the utterances and the rate of speech are also computed for each patient. Together with the 3 sentiment features, a total of 5 text-based features are thus extracted.

To further enrich the feature set, we also tagged the sentiment of the questions as positive, negative and neutral. Then, the 5 features described above are extracted separately for each question. For example, if the question is ”can you tell us a recent positive experience?” and the answer is ”I do not have any”, this is a negative answer to a positive question. Extracting the 5 features for the 3 question types, 15 dimensional feature vectors are obtained.

4.2. Fusion of acoustic- and text-based features

The fusion algorithm is designed based on the observation that acoustics-only system often overestimates in its prediction.

Those overestimations significantly impact the overall performance of the system and reduces its reliability. Text only result is 10.38, which over-performs the baseline audio result.

In our proposed approach, instead of performing commonly used score or feature fusion methods, we used a co-training algorithm to adjust the scores produced by the acoustics-only system. In this approach, any score above 30 is tagged as class-1 and any score below 18 is tagged as class-2. If the acoustic-only system generates a score that is above 30 or below 18 and if the text-only system also produces a score in the same range (agreement case), then the score from the acoustic-only system is used. If they are in disagreement, i.e., one of the system produces a score that is in class-1 and the other produces a score that is in class-2, the final score is adjusted by changing the acoustic-only result by getting it closer to opposite class. If the prediction of the acoustic system is p_{acou}, final prediction p_{fuse} is computed by adding or subtracting Γ from p_{acou}. Γ is determined using a grid search algorithm on the training data.

5. Experiment setup

5.1. Databases

The German database, distributed as part of the AVEC 2013 challenge [24], consists of conversations with 150 patients. [25]. Beck scores of the 100 patients are available whereas they are not available for the other 50 patients. The mean age of German database subjects is 31.5. The duration of the recordings range from 6 seconds to 4 minutes. Average BDI-II score is 15.1 and standard deviation is 12.3.

The Turkish database was collected at a hospital in Istanbul. It consists of 70 subjects. Mean age of the patients is 34. 14 of them are male and the rest is female. Beck scores of all subjects are available using the depression questionnaire, the Beck Depression Inventory-II (BDI-II) [26]. The average BDI-II score of the patients is 23.45 with a standard deviation of 11.01.

The Turkish database consists of interviews with the patients. Three types of questions were directed to the patients: neutral, positive and negative questions. For example, "Are you currently employed?" was a neutral question, "What made you happy lately?" is a positive question, and "What made you sad lately?" is a negative question. The interview consists of 16 questions. The mean length of the conversations is close to 5 minutes. The total length of the recordings is 6 hours. They were recorded using a headphone microphone connected to a built-in sound card of a laptop with a sampling rate of 48 kHz.

5.2. Acoustic features

The open-source toolkit OpenSMILE [27] was used for acoustic feature extraction. The AVEC 2013 feature extraction protocol was used. Feature vectors include 32 energy and spectral related low-level descriptors (LLDs) and their functionals [25]. 2268 dimensional features were extracted per speaker. Functionals were computed over 20 seconds time windows.

5.3. Baseline system

MRMR feature selection method was applied [28] to reduce the number of acoustic features. Support Vector Regression (SVR) was to used to model the relationship between the features and the depression scores. Because the amount of training data is small, leave-one-out method was used. Performance is measured using the root mean square error (RMSE) criterion.

6. Results and discussion

Two sets of experiments were conducted. In the first set, the proposed feature selection algorithms were tested and compared with the baseline MRMR algorithm both for the German and Turkish databases. The RCR algorithm proposed for redundancy computation in Section 3.3 was not used for the Turkish database since unlabeled data is not available in that database. In the second set, text-based features were tested only with the Turkish database since the transcriptions were not available for the German database.

The evaluation criteria for all experiments were root mean square error (RMSE), which is also used in the AVEC challenges [24, 25, 2, 29]. Significance of the results were tested using the t-test with $p < 0.05$. Support Vector Regression (SVR) was used in all systems.

6.1. Performance of the feature selection algorithms

Performance of the baseline and the proposed multi-lingual and RCR feature selection algorithms for the German database are shown in Table 1. In the multi-lingual approach, Turkish database was used to supplement additional features for each depression class in the German database when the number of samples is less than N_{min} as described in Section 3.1. Even though performance improved for $N_{min} = 3$, the improvement was not significant. Improvement with $N_{min} = 5$ was found to be significant only when the RCR algorithm was also used. RCR algorithm was not effective when it was used by itself.

Table 2 shows the results with the baseline and the multi-lingual feature selection algorithm for the Turkish database. Best result was 10.51 and the improvement compared to baseline was significant. RCR algorithm was not applied to Turkish due to lack of unlabeled data. The clustering algorithm proposed in Section 3.2 for the Turkish database was used with 2, 9, and 15 classes instead of the 45 distinct classes available in the databases. Results are shown in Table 3. Even though the system with 15 clusters significantly outperformed the baseline system, the improvement was not more than what was obtained with the multi-lingual MRMR approach.

Dim	Baseline	$N_{min} = 3$	$N_{min} = 5$
3	12.30	**10.51** ($p = 0.05$)	11.26
4	12.45	10.85	**10.74** ($p = 0.37$)
5	12.56	10.58	11.23
10	12.45	10.82	12.13
15	12.08	11.12	12.00
20	12.87	11.91	11.46
40	13.28	12.67	11.98
80	11.58	12.28	13.06
100	11.75	11.95	13.08
200	11.32	11.55	12.14
400	11.42	11.72	12.00
800	11.31	11.39	11.35

Table 2: *Performance of the multi-Lingual MRMR Methods for the Turkish database when the minimum occurrence threshold N_{min} is set 3 and 5. Best results are shown in bold together with their statistical significance using t-test.*

6.2. Performance of score fusion

Table 4 shows results when speech-based features were fused with text-based features using the proposed approach described

Dim	Baseline	$N_{min} = 5$	$N_{min} = 5$ and RCR	$N_{min} = 3$	$N_{min} = 3$ and RCR	RCR
10	9.90	9.97	9.99	10.37	12.39	10.02
15	9.81	10.21	**9.43** $p(0.01)$	10.13	12.12	10.08
20	9.86	10.32	9.52	9.84	10.68	9.74
40	10.25	10.35	10.45	9.73	11.36	10.22
80	10.69	9.93	9.88	**9.42** $p(0.47)$	10.93	10.06
100	10.48	9.93	9.74	9.50	10.54	10.17
200	10.12	10.00	10.38	9.69	10.28	10.44
400	10.14	9.79	10.21	9.58	10.29	10.13
800	10.08	9.86	10.11	9.91	10.11	9.89
1000	10.02	9.85	10.14	9.79	10.16	9.98

Table 1: *Performance of the multi-lingual MRMR Methods for the German database when the minimum occurrence threshold N_{min} is set 3 and 5. Results are shown both when the RCR algorithm is used and not used. Best results are shown in bold together with their statistical significance using t-test.*

Dim	Baseline	2-Cluster	9-Cluster	15-Cluster
5	12.56	11.35	13.14	11.99
10	12.45	10.95	13.42	12.25
15	12.08	11.13	13.07	11.75
20	12.87	11.74	13.23	12.95
40	13.28	12.33	13.73	12.06
80	11.58	12.72	13.33	**10.83**
100	11.75	13.22	13.09	10.97
200	11.32	11.72	12.66	11.50
400	11.42	11.83	12.00	11.40
800	11.31	11.62	11.70	11.64

Table 3: *Results with feature selection using the clustering approach with 2, 9, and 15 clusters. Turkish database is used. Statistically significant ($p < 0.05$) improvement is shown in bold.*

in Section 4.2. Fusion algorithm significantly improved the performance (p-value=0.01) compared to the baseline case by reducing the error by more than 10%. Spread of the prediction errors is substantially reduced after fusion as shown in Fig. 1. Fourth column in Table 4 shows the results for the multi-lingual feature selection for Turkish when $N_{min} = 3$. Even though that approach worked well compared to the baseline when fusion was used, improvement with it compared to the base-fusion was not found to be significant. Fifth column shows the best result obtained with the clustering approach together with the fusion method. That algorithm not only outperformed the baseline but also outperformed the base-fusion algorithm significantly.

7. Conclusion and future work

We investigated using multi-lingual databases for feature selection in the context of depression assessment, which was found to be effective. This result is significant not only because it is a step towards using larger multi-lingual databases for depression detection, but also it indicates that there are similarities between two entirely different languages in the way they manifest depression. As a second contribution, we proposed novel features derived from transcriptions and fused them with the acoustic features in way that significantly improved the performance.

In future work, we will add more languages to our database and continue to improve the feature selection process. Moreover, we believe that our text features are also language-independent and we will investigate fusion algorithms in a multi-lingual setting.

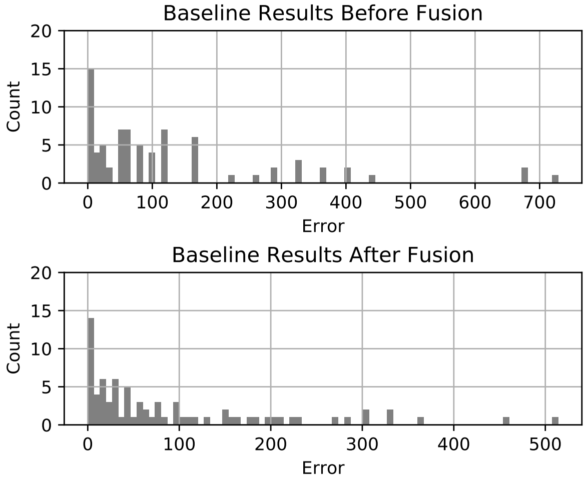

Figure 1: *Distribution of squared errors for the baseline MRMR case is shown in the top figure. The bottom figure shows the squared error distribution for the Baseline MRMR after fusion.*

Dim	Baseline	Base-fusion	Fusion	Fusion
			($N_{min} = 3$)	(15 Clus.)
3	12.30	10.87	**9.59**	11.15
4	12.45	10.81	9.94	10.81
5	12.56	10.79	9.66	10.62
10	12.45	11.07	9.92	10.67
15	12.08	10.53	10.06	10.13
20	12.87	11.02	10.35	11.18
40	13.28	11.51	11.23	10.41
80	11.58	10.19	10.71	**9.59**
100	11.75	10.33	10.30	9.93
200	11.32	10.15	10.17	10.40
400	11.42	10.08	10.04	10.25
800	11.31	**10.03**	10.10	10.35

Table 4: *Results after fusing with text classification. Turkish database was used. Baseline acoustic system predictions are used in base-fusion. Bold results show cases where the improvement is significant compared to the baseline case but not to the base-fusion case. In the underlined bold case, improvement is significant both compared to the baseline system and the base-fusion system.*

8. References

[1] A. Halfin, "Depression: the benefits of early and appropriate treatment." *The American journal of managed care*, vol. 13, no. 4 Suppl, pp. S92–7, 2007.

[2] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[3] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[4] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[5] B. Stasak, J. Epps, and R. Goecke, "Elicitation design for acoustic depression classification: An investigation of articulation effort, linguistic complexity, and word affect," in *Proc. Interspeech 2017*, 2017, pp. 834–838. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1223

[6] O. Simantiraki, P. Charonyktakis, A. Pampouchidou, M. Tsiknakis, and M. Cooke, "Glottal source features for automatic speech-based depression assessment," in *Proc. Interspeech 2017*, 2017, pp. 2700–2704. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1251

[7] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, "Classification of depression state based on articulatory precision." in *Interspeech*, 2013, pp. 2172–2176.

[8] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Probabilistic acoustic volume analysis for speech affected by depression." in *Interspeech*, 2014, pp. 1238–1242.

[9] B. Vlasenko, H. Sagha, N. Cummins, and B. Schuller, "Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition," in *Proc. Interspeech 2017*, 2017, pp. 3266–3270. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-887

[10] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–7.

[11] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," *depression*, vol. 1, no. 1, 2014.

[12] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression estimation using audiovisual features and fisher vector encoding," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 87–91.

[13] R. Gupta and S. S. Narayanan, "Predicting affective dimensions based on self assessed depression severity." in *INTERSPEECH*, 2016, pp. 1427–1431.

[14] R. Gupta, S. Sahu, C. Espy-Wilson, and S. S. Narayanan, "An affect prediction approach through depression severity parameter incorporation in neural networks," in *Proc. Interspeech 2017*, 2017, pp. 3122–3126. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-120

[15] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 65–72.

[16] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 33–40.

[17] M. R. Morales and R. Levitan, "Speech vs. text: A comparative analysis of features for depression detection systems," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 136–143.

[18] V. Mitra, E. Shriberg, D. Vergyri, B. Knoth, and R. M. Salomon, "Cross-corpus depression prediction from speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4769–4773.

[19] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. Cohn, "Cross-cultural depression recognition from vocal biomarkers," *Interspeech 2016*, pp. 1943–1947, 2016.

[20] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd." in *Interspeech*, 2013, pp. 847–851.

[21] B.-Q. Li, L.-L. Hu, L. Chen, K.-Y. Feng, Y.-D. Cai, and K.-C. Chou, "Prediction of protein domain with mrmr feature selection and analysis," *PLoS One*, vol. 7, no. 6, p. e39308, 2012.

[22] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, and Y. Li, "Prediction of lysine ubiquitination with mrmr feature selection and analysis," *Amino acids*, vol. 42, no. 4, pp. 1387–1395, 2012.

[23] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by svm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2297–2307, 2010.

[24] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10.

[25] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 3–10.

[26] A. T. Beck, R. A. Steer, and G. K. Brown, "Beck depression inventory-ii," *San Antonio*, vol. 78, no. 2, pp. 490–8, 1996.

[27] 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*. ACM, 2010, pp. 1459–1462.

[28] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[29] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 3–9.