DL4DED: Deep Learning for Depressive Episode Detection on Mobile Devices

Afef Mdhaffar^{1,2(✉)}, Fedi Cherif³, Yousri Kessentini¹, Manel Maalej⁴,
Jihen Ben Thabet⁴, Mohamed Maalej⁴, Mohamed Jmaiel^{1,2},
and Bernd Freisleben⁵

¹ Digital Research Center of Sfax, Sfax, Tunisia
afef.mdhaffar@gmail.com

² National School of Engineering of Sfax, University of Sfax, Sfax, Tunisia

³ Geeks Data Consulting, Tunis, Tunisia

⁴ Department of Psychiatry “C”, Hedi Chaker University Hospital,
Faculty of Medicine of Sfax, University of Sfax, Sfax, Tunisia

⁵ Department of Mathematics and Computer Science, University of Marburg,
Marburg, Germany

Abstract. This paper presents a deep learning approach for depressive episode detection on mobile devices, called DL4DED. It is based on a convolutional neural network and a long short-term memory network to identify the status of a patient’s voice extracted from spontaneous phone calls. To run DL4DED on mobile devices, two neural network model compression techniques are used: quantization and pruning. DL4DED protects data privacy, since it can be executed on a patient’s smartphone. Our proposal is validated on the DAIC-WOZ database. The obtained results show that the accuracy of DL4DED with model compression is only slightly lower than the accuracy of DL4DED without model compression. Furthermore, our experiments indicate that the power consumption of DL4DED is reasonably low.

Keywords: Deep learning · Mobile application · Depression · Pruning · Quantization · Optimization

1 Introduction

According to the World Health Organization (WHO), 800 000 people die every year due to suicide¹, which is the second leading cause of death in people aged between 15 and 29 years. However, many suicides can be prevented. Indeed, mental disorders, such as depression, contribute to many of them around the world. Hence, early detection and appropriate management are key to ensuring that people receive the care they need². Therefore, patients suffering from Major

¹ <https://www.who.int/news-room/fact-sheets/detail/depression>.

² https://apps.who.int/iris/bitstream/handle/10665/131056/9789241564779_eng.pdf.

Depressive Disorder (MDD) and Bipolar Disorders (BD) should be continuously monitored, since they might experience frequent depressive episodes.

Several mental health monitoring approaches using mobile devices have been proposed. Most of them [1–3, 5–11, 13, 14] are based on (1) collecting and analyzing smartphone features such as activity, localization, and phone calls, and (2) launching interactive questionnaires such as PHQ-9³ and BDI⁴. “Active” monitoring approaches (i.e., requiring a patient’s intervention) are less used and less effective than “passive” ones (i.e., not requiring a patient’s intervention) in practice. Actually, patients do not like to answer questionnaires every day, and even if they answer them, we cannot be sure about the honesty of their answers.

Passive monitoring approaches [3, 5–8, 10, 11, 13, 14] can be divided into two categories. The first category tries to use most of the smartphone’s sensors, such as accelerometer and camera, to collect related features [5–7, 11, 14]. These approaches require special environmental conditions to ensure the precision of collected data. For instance, Maxhuni et al. [11] require the smartphone to be held in a special position to accurately capture the patient’s activity. This makes these approaches difficult to use in practice. The second category of approaches [3, 8, 10, 13] focuses on analyzing “only” voice sounds, recorded from phone calls. Indeed, acoustic changes in speech allow us to detect depressive episodes [8]. Most of these approaches are based on the use of complex algorithms for speech analysis, such as deep learning approaches. Typically, these approaches require significant computational resources. Therefore, this category of approaches sends recorded phone calls to an external server for analysis, since running complex algorithms on smartphones is still a challenging task [12]. This reduces the chances of using these mobile applications, because they do not protect data privacy. Actually, many patients do not accept that external servers process their recorded phone calls.

In this paper, we present a novel deep learning approach that can be loaded and executed on a patient’s smartphone to allow real-time detection of depressive episodes. Our proposal, called DL4DED (Deep Learning for Depressive Episode Detection), is based on optimizing and compressing a deep learning model to be integrated in our mobile application. DL4DED (1) records phone calls, (2) executes our deep learning model, and (3) triggers alerts if a depressive episode is detected. DL4DED does not send recorded phone calls to an external server and consequently preserves data privacy. To evaluate our approach, two groups of experiments have been conducted. The first one illustrates the efficiency of DL4DED in terms of accuracy. The second one demonstrates that the power consumption of DL4DED is reasonable when compared to baseline approaches.

The paper is organized as follows. Section 2 discusses related work. DL4DED and its implementation are detailed in Sects. 3 and 4, respectively. Section 5 presents experiments and results. Section 6 concludes the paper and outlines areas for future research.

³ Patient Health Questionnaire (PHQ-9).

⁴ Beck’s Depression Inventory.

2 Related Work

A comprehensive survey reviewing existing mobile applications for mental health diseases (e.g., BD, MDD, schizophrenia) has been published by Cornet et al. [1]. This survey explains different parameters that have been used throughout the years to detect mental health diseases. These parameters are usually extracted from smartphone sensors, such as microphone (i.e., phone calls), accelerometer (i.e., movement data), and GPS (i.e., localization). The authors also describe the used analysis algorithms, where the most advanced ones are based on deep learning networks. Moreover, the authors discuss data privacy issues related to all reviewed approaches, since all extracted smartphone features are sent to an external server for analysis.

Haque et al. [6] present a deep learning model to measure the severity of depression symptoms. The proposed approach is based on analyzing 3D facial expression and spoken language. It has been evaluated on the DAIC-WOZ dataset [4]. The proposed approach requires an active intervention from the patient, which makes it unusable in some cases. Actually, it is hard to ask a patient to daily record videos, due to privacy issues. In contrast, our approach is passive, analyzes spontaneous phone calls, and preserves data privacy.

Su et al. [14] present a diagnosis assistance system that is based on deep learning and fusion techniques. The proposed system makes use of two deep learning models. The first one is used for speech classification, while the second one processes facial expression. This work is aimed at assisting doctors to avoid misdiagnosis in the case of BD. Actually, BD is usually confused with MDD. The results are quite promising, but the system architecture cannot not be applied in a spontaneous manner. Indeed, the system requires special environmental conditions to be applied, such as a camera.

Huang et al. [7] propose an attention-based convolutional neural network (CNN) and long short-term memory (LSTM) approach for distinguishing between MDD and BD. The proposed approach identifies mood disorders on the basis of responses given to 6 video sequences. Consequently, it can be applied to depressive episode detection. The analysis includes speech processing and facial recognition to extract emotional state. In contrast to our approach, this work cannot be applied in a spontaneous manner and does not protect data privacy. Actually, recording responses to 6 particular videos cannot be expected on a daily basis and without the intervention of a psychiatrist.

Grunerbl et al. [5] present a smartphone-based approach for mood status identification in BD. The proposed approach makes use of smartphone features (e.g., location, phone call sounds, phone light) to detect depressive and maniac episodes. Features are extracted manually, and data privacy is not protected.

Maxhuni et al. [11] define an analysis approach that identifies manic and depressive episodes on the basis of activity data and phone calls. To monitor activities, the authors mention that the phone has to be held in a special position. Moreover, phone calls are recorded, stored in the phone's memory card and sent to another server for analysis. This means that this approach does not protect data privacy and requires special environmental conditions to generate accurate results.

Khorram et al. [10] propose a machine learning (ML) approach for depression detection in BD. The proposed approach is based on the analysis of acoustic features of the patient's voice. The paper shows the effectiveness of speech features in detecting depression. However, features are extracted manually due to the use of ML techniques. Moreover, a patient's voice is sent to an external server for analysis, which does not protect data privacy. Therefore, we propose a deep learning model that allows us to automatically extract features and preserve data privacy.

Gideon et al. [3] discuss the impact of a recorded voice on the identification of depressive and manic episodes in BD. Only phone calls recorded during clinical trial are considered. Other phone calls are removed to protect data privacy. This means that this work does not analyze naturalistic and spontaneous voice, which might alter the detection results. However, our work is based on the analysis of spontaneous phone calls (i.e., recorded on a daily basis), thanks to our deep learning model that is locally running on a patient's smartphone.

Huang et al. [8] define two speech features based on speech landmark bigrams (i.e., bigram count and LDA bigram) for depression detection. Both features could be extracted from naturalistic phone calls including 6 elicitation tasks such as measuring the diadochokinetic rate. Landmarks are extracted using the SpeechMark® toolbox⁵, which is a Matlab tool that runs under Windows 7, Windows XP, Apple OSX (Lion or Mountain Lion), as mentioned on their web page. This means that the proposed work needs to send recorded phone calls to an external server for feature extraction. However, our work allows us to locally extract features on the smartphone.

Pan et al. [13] propose analysis approaches for detecting manic episodes in BD. The proposed approaches are based on the use of Support Vector Machines (SVM) and Generalized Markov Models (GMM). They record spontaneous phone calls and send them to an external server for analysis. This approach does not protect data privacy.

3 DL4DED

This section presents our novel approach, called DL4DED. It is a mobile voice analysis approach that:

1. monitors patients suffering from BD and MDD;
2. detects depressive episodes by locally analyzing their phone calls without storing them;
3. alerts patients, their family members, and their psychiatrists if a depressive episode is detected.

⁵ <https://speechmrk.com/speechmark-products-downloads/the-speechmark-matlab-toolbox/>.

DL4DED is based on the use of deep learning methods applied to spontaneous speech, recorded from phone calls, to identify depressed voice. The proposed deep learning model is locally running on a patient’s smartphone to preserve data privacy. Indeed, recorded phone calls are “temporarily” stored on the smartphone until the analysis is accomplished. Once the decision is sent to our external server, the recorded phone call is discarded. The architecture of DL4DED, the proposed deep learning model, and our optimizations are described below.

3.1 Architecture

Figure 1 shows the architecture of DL4DED. Our approach (1) records spontaneous phone calls, (2) stores them temporarily in a buffer belonging to the local memory of a patient’s phone, (3) processes them via our “mobile” deep learning model, (4) identifies the state of the recorded voice (i.e., depressed or not depressed), (5) discards the recorded voice (i.e., phone call), and finally (6) sends and stores the obtained decision to a database that is installed on a cloud server. The communication links between the smartphone and the cloud server are enabled via the HTTP protocol. A web dashboard is available for both patients and psychiatrists. It is used to display the patient’s mood on the basis of the voice status.

To identify the status of the voice, we use a novel deep learning model that is described in Sect. 3.2.

Fig. 1. Architecture of DL4DED

3.2 Deep Learning for Voice Status Identification

Figure 2 shows our deep learning model. It processes recorded phone calls and detects depressed voice. The proposed model takes as input a spectrogram. To obtain a spectrogram, a Short-Time Fourier Transform (STFT) is applied to the recorded phone calls. A spectrogram, also called voiceprint, represents the

Fig. 2. Deep learning model for voice status identification

Fig. 3. Energy bar: an audio segment

spectrum of frequencies of the recorded phone call, as a function of time. It is composed of a set of frequency bars. Each bar corresponds to a time-stamp t . It is a vector, having the dimension 513 corresponding to the quantities of energy, expressed in decibels (see Fig. 3).

In our case, the spectrogram is a matrix having the dimensions 513 (frequencies) and 120 (times). The temporal dimension has been experimentally chosen. However, the frequency dimension has been obtained by applying STFT on audio segments, recorded with a frequency of 16 kHz (standard frequency

for human voice recording). The intensity of colors represents the intensity of the energy of the recorded voice at the instant t , which might be useful for the identification of the patient's mood. Once the spectrogram is built, it is processed by a CNN network, having one convolution layer and one max-pooling layer. The convolution layer applies a $(1 * 3)$ filter/ $(1, 2)$, allowing us to keep all frequencies (i.e., all energy quantities describing the voice) and preserve the temporal continuity. The max-pooling layer applies a $(1 * 5)$ filter / $(1, 5)$, to extract medium-term features, while keeping all frequencies. The output of the max-pooling layer is processed by an LSTM that includes memory cells to save long-term information. A fully connected layer is applied afterwards, to transform the obtained matrix into a 128-dimensional vector. The latter is processed by a softmax classifier allowing a binary classification (i.e., depressed voice or not depressed voice). The parameters of our CNN model have been identified experimentally.

To run the proposed deep learning model on mobile devices, we considered two compression methods: quantization and pruning. These methods allow us to reduce the size of our deep learning model by removing weights or operations that are least useful for prediction.

4 Implementation Issues

Figure 4 shows the technical architecture of DL4DED. The smartphone is in charge of recording and analyzing phone calls, while running our deep learning model that has been implemented using the Keras⁶ library. Generated decisions are then sent to a Flask RESTful API 1.0.2, a web service platform that stores received decisions in a RethinkDB 2.3.6 0xenial (GCC 5.3.1) database. A nodeJS v4.2.6 server is installed to build a real-time web application. It is composed of a set of dashboards, displaying analysis results and an estimation of the patient's mood/status.

To predict depressed voice, recorded phone calls are pre-processed first. Afterwards, the deep learning model is loaded into a mobile application to trigger predictions. Implementation details are presented below.

4.1 Data Processing

To process a phone call by a CNN model, a spectrogram is built. For this purpose, the recorded voice (i.e., voice signal) is processed by a pre-emphasis filter. A pre-emphasis filter allows us to (1) improve the Signal-to-Noise Ratio (SNR); (2) avoid numerical problems that might appear during the Fourier Transform operation and (3) balance the frequency spectrum. Actually, high frequencies usually have smaller magnitudes compared to lower frequencies.

⁶ <https://keras.io>.

Fig. 4. DL4DED implementation

After applying the pre-emphasis filter, the signal is decomposed into short-time and overlapping frames. This step allows us to avoid applying the STFT across the entire signal and consequently losing the frequency contours of the signal over time. In this case, STFT will be applied on short-term frames allowing us to obtain a good approximation of the frequency contours of the signal by concatenating adjacent frames.

A window function (e.g., Hamming window) and a STFT are applied to each frame. This allows us to compute the power spectrum that is used to extract the frequency bands by applying triangular filters on a mel scale. The mel scale aims to mimic the non-linear human ear perception of sound, by being more discriminative at lower frequencies and less discriminative at higher frequencies.

After applying the filter bank to the power spectrum (i.e., periodogram) of the signal, we obtain the spectrogram that is processed by our CNN.

All steps described above were implemented in Android studio to create our mobile application (see first step of Fig. 5).

4.2 Implementation on Mobile Devices

To implement our deep learning model on mobile devices, we followed three steps (see Fig. 5). First, our Keras-based deep learning model is converted to Tensorflow. Actually, Tensorflow allows an easier integration of deep learning models on mobile devices (i.e., Android or IOS). For this purpose, we use a method called “Keras_to_Tensorflow” provided by the Tensorflow library. This method converts a Keras model file into a Tensorflow file which contains both the network architecture and its associated weights. Second, we optimize the generated model by removing weights and operations that are least useful for predictions by applying pruning and quantization methods. Third, we load our model (i.e.,

*.pb file including weights and architecture) into our mobile application to allow real-time detection of depressed voice.

Fig. 5. Implementation of our deep learning model on mobile devices

5 Experimental Results

Our experiments were conducted on (1) an Ubuntu Server 16.04.5 LTS, with 2 GPUs Nvidia GeForce GTX1080 Ti Turbo 11 GB GDDR5X-RAM PCIe x16 HDMI and (2) a OnePlus A6003 Smartphone, running under an Android OS 8.1.0, with 8 GB of RAM and 128 GB of storage. We use the DAIC-WOZ data set⁷ to evaluate DL4DED. The DAIC-WOZ dataset was compiled by the USC Institute of Creative Technologies and published as part of the Audiovisual Emotional Challenge 2016 (AVEC 2016). The DAIC WOZ data set includes 189 sessions, with an average duration of 16 min, between a participant and a virtual interviewer, controlled by a human interviewer in another room via a “Wizard of Oz” approach. Prior to the interview, each participant completed a psychiatric questionnaire (PHQ-8), from which a binary classification (depressed, non-depressed) was derived [4]. To evaluate DL4DED, we conducted two groups of experiments. The first group is used to evaluate the performance of our approach in terms of accuracy. The second group aims to assess the power consumption of DL4DED on mobile devices.

5.1 Performance

The main objective of this group of experiments is to compare our “optimized deep learning model” (i.e., with model compression, running on a smartphone)

⁷ <http://dcapswoz.ict.usc.edu>.

and the “original deep learning model” (i.e., the originally created deep learning model without compression and optimization), in terms of accuracy that has been calculated using Tensorflow libraries. This experiment does not depend on the used data since it assesses the accuracy loss related to the use of DL4DED. Therefore, we evaluate both models on the same database (i.e., the DAIC-WOZ dataset). The obtained results demonstrate that the accuracy of DL4DED (0.5) is slightly lower than the accuracy of the original deep learning model (0.52). This means that the applied compression techniques do not significantly alter the analysis results.

5.2 Power Consumption

To assess the power consumption of DL4DED, we used the battery monitoring functionality provided by the Android OS that allows us to measure the power usage of each application in mAh (see Fig. 6).

Fig. 6. Power monitoring functionality of Android OS

mAh stands for milli-Ampere-hours and expresses the number of milliamperere (i.e., electric charge quantity), the mobile application has used per hour. Actually, the battery capacity is usually expressed in mAh.

We measured the power consumption of (1) DL4DED and (2) a baseline mobile application, while varying the duration of phone calls (from 1 min to 16 min). The baseline mobile application only records phone calls and sends them to an external server that loads and runs our “original deep learning model”

(i.e., without compression and optimization) to detect depressed voice. Due to the throughput of our communication links, the baseline approach does not send the whole duration of the phone call. However, DL4DED processes the whole duration of the recorded phone call to allow better classification. The obtained results show that the average power consumption of DL4DED (5 mAh) is higher than the power consumption of the baseline mobile application (1 mAh), for a phone call duration of 4 min. This is quite plausible. Actually, DL4DED records phone calls, saves it to a temporary buffer and loads a deep learning model on the smartphone, to allow real-time prediction of the depressed voice. Loading an optimized deep learning model on a smartphone should logically increase power consumption. Our experiments show that the average difference of power usage (4 mAh) for a phone call of 4 min is reasonably low and therefore acceptable. As shown in Fig. 7, it is clear that the power consumption of DL4DED increases when the duration of the phone call increases. This is related to the fact that we process the whole duration of the phone call in contrast to the baseline approach. This issue will be solved in future work by processing only pertinent parts of the recorded phone call.

Fig. 7. Power consumption: DL4DED vs. baseline approach

6 Conclusion

We presented a novel mobile deep learning approach for depressive episode detection, called DL4DED. It is a combination of CNN and LSTM networks. The proposed deep learning model was optimized using compression techniques to be loaded onto smartphones. DL4DED records spontaneous phone calls on a daily basis, stores them temporarily locally on a smartphone, and loads our optimized deep learning model to allow real-time detection of depressed voice. DL4DED preserves data privacy since recorded phone calls are not sent to external servers. DL4DED was evaluated on the DAIC-WOZ database. The results demonstrated the efficiency of DL4DED in terms of accuracy and power consumption.

There are several directions for future research. First, we aim to consider the number of recorded phone calls in our study to improve analysis results.

Actually, the absence of phone calls could be seen as a severe sign of depression. Second, we plan to extend our deep learning approach to detect manic episodes in BD. Third, we aim to build a realistic and balanced database to improve the performance of DL4DED. Finally, further optimization methods should be investigated to improve our implementation on mobile devices.

Acknowledgment. This work is supported by the German Academic Exchange Service (DAAD) (Transformation Partnership: Theralytics Project).

References

1. Cornet, V.P., Holden, R.J.: Systematic review of smartphone-based passive sensing for health and wellbeing. *J. Biomed. Inform.* **77**, 120–132 (2018)
2. Eralp, A., Orhan, Z., Kabil, M.: Psychodroid: a mobile psychological disorder detection application by dynamic question generation and content analysis. *Proc. - Soc. Behav. Sci.* **159**, 691–696 (2014)
3. Gideon, J., Provost, E.M., McInnis, M.G.: Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2359–2363 (2016)
4. Gratch, J., et al.: The distress analysis interview corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, pp. 3123–3128. LREC, Reykjavik (2014)
5. Grünerbl, A., et al.: Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE J. Biomed. Health Inform.* **19**(1), 140–148 (2015)
6. Haque, A., Guo, M., Miner, A.S., Fei-Fei, L.: Measuring depression symptom severity from spoken language and 3D facial expressions, pp. 1–7. *CoRR abs/1811.08592* (2018)
7. Huang, K.Y., Wu, C.H., Su, M.H.: Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses. *Pattern Recogn.* **88**, 668–678 (2019)
8. Huang, Z., Epps, J., Joachim, D.: Speech landmark bigrams for depression detection from naturalistic smartphone speech. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ICASSP 2019, pp. 5856–5860 (2019)
9. Jiménez-Serrano, S., Tortajada, S., García-Gómez, J.M.: A mobile health application to predict postpartum depression based on machine learning. *Telemed. J. E-Health: Off. J. Am. Telemed. Assoc.* **21**(7), 567–574 (2015)
10. Khorram, S., Gideon, J., McInnis, M.G., Provost, E.M.: Recognition of depression in bipolar disorder: leveraging cohort and person-specific knowledge. In: INTER-SPEECH, pp. 1215–1219 (2016)
11. Maxhuni, A., Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O., Morales, E.F.: Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive Mob. Comput.* **31**, 50–66 (2016)
12. Mohammadi, M., Al-Fuqaha, A., Sorour, S., Guizani, M.: Deep learning for IoT big data and streaming analytics: a survey. *IEEE Commun. Surv. Tutor.* **20**(4), 2923–2960 (2018)

13. Pan, Z., Gui, C., Zhang, J., Zhu, J., Cui, D.: Detecting manic state of bipolar disorder based on support vector machine and gaussian mixture model using spontaneous speech. *Psychiatry Invest.* **15**(7), 695–700 (2018)
14. Su, M., Wu, C., Huang, K., Yang, T.: Cell-coupled long short-term memory with L-skip fusion mechanism for mood disorder detection through elicited audiovisual features. *IEEE Trans. Neural Netw. Learn. Syst.* 1–12 (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.