# Depression Diagnosis using Deep Convolutional Neural Networks

Mofassir ul Islam Arif, Mauricio Camargo, Jan Forkel, Guilherme Holdack, Rafael Rêgo Drumond, Nicolas Schilling, Tilman Hensch, Ulrich Hegerl and Lars Schmidt-Thieme

**Abstract** Depression is a prevalent psychiatric disorder that impacts the quality of life of 300 million people around the world. The complex nature of depression manifestations in patients and the lack of technological advances in the diagnosis process has left a lot of room for improvement in this particular domain. At present, the diagnosis is mainly made by physicians during a conversation comprising the exploration of the symptoms and the diagnostic criteria for depression. Recently, the electroencephalography (EEG) has regained interest as a promising approach to provide bio-markers which are of clinical value in the diagnostic process and for response prediction to therapy. In the present landscape, even the addition of EEG data has resulted in a semi-automated process, where the expert still has to heavily modify the raw data. This adds an

Mofassir ul Islam Arif · Mauricio Camargo · Jan Forkel · Guilherme Holdack · Rafael Rêgo Drumond · Nicolas Schilling · Lars Schmidt-Thieme
Universität Hildesheim, Universitätsplatz 1, 31141 Hildesheim
🖂 mofassir@ismll.uni-hildesheim.de
🖂 [camargo,holdack]@uni-hildesheim.de
🖂 jan.forkel134@gmail.com
🖂 [radrumond,schilling,schmidt-thieme]@ismll.de

Tilman Hensch · Ulrich Hegerl
Universitätsklinikum Leipzig, Liebigstraße 20, 04103 Leipzig
🖂 [Tilman.Hensch,Ulrich.Hegerl]@medizin.uni-leipzig.de

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/29
ISSN 2363-9881

inherent bias to the process based on the expert and incurs costs as well as time to the process of diagnosis. In this paper, we present a fast, effective and automated method that is able to quickly determine if the patient has depression while still maintaining a high accuracy of diagnosis. Our approach is built on using raw EEG-data, performing frequency domain preprocessing in order to split the data into its different frequency domains and to create EEG 'images'. These images are then treated by a convolutional neural network, which is a novel approach in this area. Experimental results have shown to provide outstanding results and to work without the need for feature engineering or any human interaction, which is a core strength of the model we are proposing.

# 1 Introduction

One of the worldwide leading causes of disability is depression. It is a frequently occurring mental disorder which has diverse symptoms and can lead to a decline in the quality of life. Depression impacts all aspects of the affected person's life whether it be school, work or even family life. In extreme cases, it leads to suicidal tendencies (Kupfer et al., 2012). Even though effective treatments exist today, it is seen that less than half of the affected people receives treatment. Reasons for this trend are not only the social stigma, lack of trained health-care providers and lack of resources, but also false diagnosis. There is a high rate of misdiagnosis for depression whereby people with non-psychiatric diseases could show similar symptoms and thus get diagnosed as affected . More than 300 million people of different ages and social background suffer from depression, thus in order to treat them best, a correct diagnosis is necessary (World Health Organization, 2017). Nevertheless, a major epidemiological study has shown, that primary care doctors identify only 55 % of patients with a clinically significant depression (Wittchen and Pittrow, 2002).

One method which might contribute to major depression diagnosis (MD diagnosis), apart from using questionnaires, is the analysis of the electroencephalogram (EEG) (Motomura et al., 2002), which measures the electrical activity of the brain cells over time and can thus record and track brain wave patterns. Usually, the EEG's electrodes are connected at multiple defined locations on the scalp at once, leading to a multi-channel EEG structure. Any irregularities in the recorded activity can be a sign of brain disorders and thus, can help with the medical diagnosis of patients. Depending on the device and need, different numbers of measurements per second (sampling rates) are used. A discriminant for depression, as it corresponds to the EEG behavior, is the inability of the human brain to go from an active state to a drowsy state. Therefore, the diagnosis of depression seeks to find this underlying trend in EEG readings taken during a session (Hegerl et al., 2012). Currently, the diagnosis of depression is handled primarily by doctors. The usage of EEG information along with the field expertise of doctors are the prime movers in the diagnosis, however this process is susceptible to errors due to the number of variables involved. Semi-automated techniques have been employed for the diagnostic process but they rely heavily on feature engineering and costly human operators. The aim of this paper is to provide a fully automated solution for analyzing raw EEG data while making use of machine learning tools.

## 1.1 Problem Formulation

In order to efficiently analyze EEG-data, the signal's artifacts must usually be removed by algorithms or experts, e.g., Sai et al. (2018) and Parvinnia et al. (2014). This step is essential for a correct interpretation of the EEG-signal (Tatum et al., 2011). Finding a method in which this step is not required, but still offers a good accuracy, can be useful for users. Additionally, false removals of artifacts by experts can be avoided by removing this step. Splitting up EEG-data into different frequency bands and analyzing the temporal order has shown differences between healthy and MD patients. This improves the classification of depression (Hegerl et al., 2012). Thus, it makes sense to split the EEG-signal of each channel and include all of the resulting time-series in the classification process.

Each EEG-signal is represented by a univariate time series $T = \{t_1, t_2, \ldots, t_n\}$, which is a sequence of data points measured equidistantly over time and $n$ samples. The EEG-device records the voltage (microvolts) at different locations at the same time. This can then be transformed into power of frequency bands as a multivariate time series M is generated. Each element $m_i$ represents a univariate time series, while any timestamp of $M$ now consists of $m_{\cdot t} = \{m_{1t}, m_{2t}, \ldots, m_{Ct}\}$, where by C we denote the number of channels the EEG-device is recording (Zheng et al., 2014). Given a specific number of EEG-channels $C$, length of the time series $L$ and class $y \in \{0, 1\}$ of the training data, the goal is to learn a model

$$\hat{y} : \mathbb{R}^{C \times L} \longrightarrow \{0, 1\} \tag{1}$$

that, given the EEG recording of one patient, predicts whether this patient suffers from depression or not. The classification model $\hat{y}$ is learned by minimizing a binary classification loss:

$$\min_{\theta} \ \mathcal{L}(y, \hat{y}) = \frac{1}{P} \sum_{i=1}^{P} -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \tag{2}$$

on a training data set of $P$ patients. Here the vector $\theta$ denotes the parameters learned during the optimization. These parameters are used by the model to predict whether or not the input signal corresponds to a depressed or healthy patient. By minimizing the above mentioned loss function, we maximize the accuracy – the number of correctly classified patients divided by the number of all patients – of our classifier.

There exist many different methods to approach this minimization problem. A variety of optimization algorithms such as gradient-based methods or even second order methods such as quasi Newton methods have been developed over time. More important is the choice of the prediction model, as the data is highly complex and structured. Lately, neural networks have gotten much attention due to their outstanding accuracies in fields such as image recognition, which also is a domain that uses complex and structured data.

Convolutional neural networks (CNNs) (Goodfellow et al., 2016) are a type of deep feed-forward artificial neural networks that have been applied to many areas in order to optimally classify objects. CNNs haven been especially useful in the classification of images, see e.g. Sharif Razavian et al. (2014) and Cireşan et al. (2012). A CNN takes a structured input, passes it through multiple convolutions and max-pooling layers, where it learns a rich feature representation of the data. Finally, fully connected layers are used on the downsampled input, to then compute a prediction. Convolutional layers change the dimension of the data by applying different kernels, which are automatically learned from the loss signal and correspond to the most important features. The max-pooling layers will save only the maximum value of a specific region and are used to downsample the input while keeping the most important aspects of the data. The fully connected layers have connections to all nodes in the previous layer and represent a high-level reasoning. The resulting network can classify patterns, independent of their location in the input image which is shown in (Robert, 2014). In the case of analyzing depression, the output layer must return a binary value, however, for problems with multiple classes, the extension is straightforward.

In order to optimally classify depressed patients, each EEG-signal will be split into its different frequency domains, then aligned and vertically added for each newly generated time series of every channel. The resulting matrix is then treated as an image and thus be used as input for the CNN. By trying multiple architectures, an optimally fitted architecture will be created. Using this approach, we build a model that functions without human interaction while still achieving a high accuracy.

## 2 Related Work

Manual time series classification, in order to classify major depression disorder, can be done by observing the vigilance states of the EEG signal. It is split up into different frequency bands, which are then analyzed regarding their trend (Hegerl et al., 2012). In order to classify depression via extracted features, common machine learning tools as logistic regression, k-nearest-neighbors and linear discriminant analysis can be used, if features of the EEG signal are extracted (Hosseinifard et al., 2013). There are many features that can be pulled off an EEG-signal in order to differentiate depressed patients from healthy controls. Examples are the power spectrum of different frequency bands, its standard deviation, mean and entropy. These can be computed and used for the classification via different machine learning algorithms (Katyal et al., 2014).

Some methods use only one channel for the task of classifying depression. In order to do so, the spectral asymmetry index SASI and detrended fluctuation analysis (DFA) get computed, combined and then their accuracy determined (Bachmann et al., 2017). Other methods use multiple channels to classify depression. They tend to compute features for every channel and use all of them for classification. The adaptive weighted distance nearest neighbor algorithm can be used to classify multi-channel EEG signals. It assigns weights to the training samples in order of their importance by using nearest neighbor classifier with leave-one-out cross validation. These weights can then be used to optimize the nearest neighbor search of new input queries. In this setting, coefficients of an autoregressive model (AR), the Higuchi fractal dimension and the power of the 4 frequency bands alpha, beta, delta and theta, of each channel, are used as features (Parvinnia et al., 2014). Another approach is the determination of wavelets in the EEG-pattern. These can be used to compute relative wavelet energy and the coefficients of a discrete wavelet transformation. These features then can be used as input for standard machine learning algorithms as SVM, multilayer perceptrons, Naive Bayes and k-nearest neighbors (Amin et al., 2017). The coefficients extracted from the wavelets can also be used as input of feed forward neural networks (Hazarika et al., 1997).

Neural networks also play their part in the classification of EEG-data. One method to classify multi-channel EEG-data, is the Multi-Channel Deep Convolution Neural Network (MC-DCNN). It takes a time series as input and computes features separately for every channel via CNN and feeds them into a multilayer perceptron to perform classification (Zheng et al., 2014). Convolutional neural networks also have proven to be a good tool for image classification (Sharif Razavian et al., 2014) (Robert, 2014). They offer many possibilities and can be applied to different fields. CNN and other neural networks have also already been used in order to classify schizophrenia (Chu et al., 2017) and other disorders using the patients EEG-data.

# 3 Data Foundation

As the basis for this work, the dataset of Hegerl et al. (2012) is used which is provided by the Department of Psychiatry and Psychotherapy of the University Leipzig, Germany. This dataset is completely labeled, and was collected making use of the same EEG equipment on a 15 minutes timeframe per session from 60 patients. It will be referred to as the *Leipzig Dataset*.

## 3.1 Proposed Method: Frequency Domain Exploration

By knowing that the decrease on vigilance stages is said to be an important characteristic of depressed patients (Hegerl et al., 2012), it is important to first comprehend what a decline on vigilance stages means. The brain operates in 4 different vigilance stages: Stages 0, A, B and C. Stage 0 is characterized by a total awake state, while stage C corresponds to a sleep onset. Stages A and B are also subdivided into more specific states. In that sense, a decline on vigilance stages represents a state of higher relaxation, but in order to detect a decline, first it is necessary to properly identify each stage.

A decrease in vigilance states is not necessarily associated with a decrease in a channel energy level along time. Instead, it takes into account the shape, the amplitude, the presence of eye movement and also the predominant range of frequencies which composes the signal (Hegerl et al., 2017). The four predominant ranges are referred to as delta (2–4 Hz), theta (4–8 Hz), alpha (8–12 Hz) and beta (12–25 Hz). Therefore, a classification method based on a frequency dimension analysis was proposed. The benefits are twofold: The analysis converts the data into the frequency domain allowing a different analysis and it provides a more compact data representation. The intuition of the method can be seen in Figure 1.

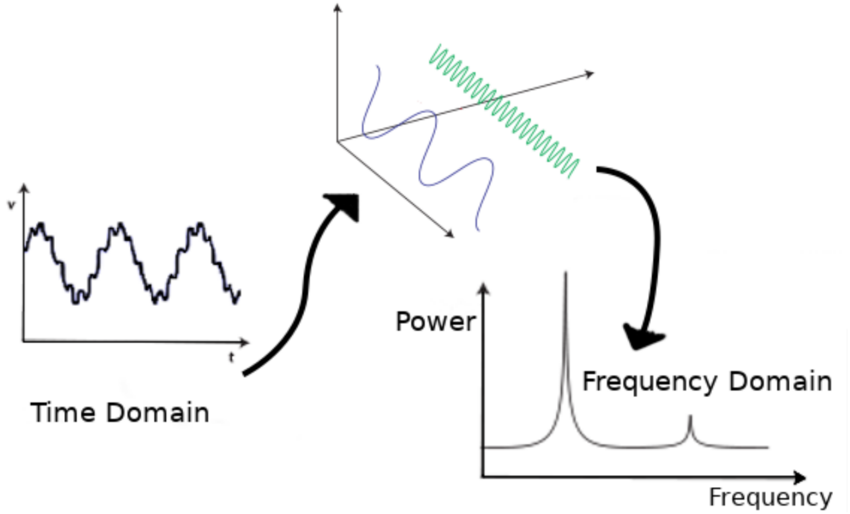

**Figure 1:** Frequency Domain transformation applied to EEG data. A Fast Fourier Transform (FFT) was applied, followed by the extraction of the Power Spectrum.

## 3.2 Preprocessing

The preprocessing phase followed the work of Hegerl et al. (Hegerl et al., 2012). The dataset was initially filtered (high-pass 0.5 Hz, low-pass 70 Hz, and notch filter 50 Hz), and then divided into 1 second segments. With respect to the proposed method (*frequency domain exploration*), the power spectrum technique presented in Hjorth (1970) was applied independently to each of the 1 second segments. Due to the Nyquist limit, only frequencies up to half of the sampling rate can be obtained. In case of the Leipzig dataset, the reading was available with a 500 Hz sampling rate, thus only frequencies up to 250 Hz can be obtained. The initial approach was to concatenate the results obtained for the consecutive segments, creating a new series as depicted in Figure 2 (each color represents a different channel and for each second the full range of frequencies from 0 to the selected threshold is used).

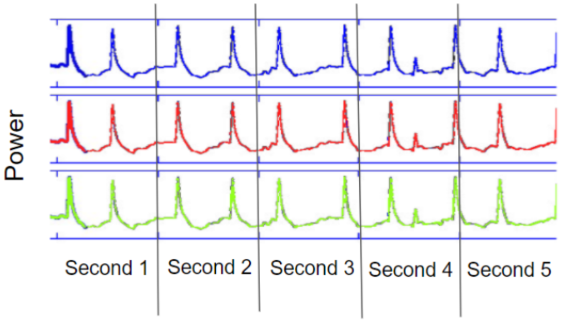

**Figure 2:** Schematic illustration of power spectrums applied to channels – in this example, three channels. Each second of each power spectrum was determined by using the method described in Figure 1.

Initial experiments on the loaded data identified the need of reducing its original physical size due to memory and time constraints. In the dataset used, a one-second segment corresponds to 500 readings for each channel (34 channels), collected for 15 minutes (900 seconds). This represents almost 19 million sampled values – per patient. Having initially 60 patients (Hegerl et al., 2012), a complete representation of the data 918 million sampled energy levels. Table 1 represents the initial dimensions without dimensionality reduction.

**Table 1:** Sample values from dataset.

| Amount of Seconds | Readings per Second | # of Channels | # of Patients | Total Number of Sample Values |
|---|---|---|---|---|
| 900 | 500 | 34 | 60 | **918 millions** |

In order to reduce the data size, *time skipping* has been applied to collect one out of every 4 seconds of reading, a *frequency threshold* has been set to only use data within the ranges of the vigilant states and, finally, a *signal resample technique* was performed to compress the characteristics of the data. The basic idea is simply to obtain a more compact representation which still preserves the overall trends along time.

## 4 Model Architecture

A Deep Neural Network has been proposed to learn the discriminant features for correctly classifiying depression after the frequency domain treatment of the raw data. After the preprocessing we are left with a reduced sized instance which is then treated as an image input to the model. The channel inputs serve the height $H$ and the frequency range (considered after the power spectrum) acts as the width $W$ of the image, having $N$ seconds concatenated. Secondly we downsample the preprocessed values by using 4 different regimes i.e. 1000, 2000, 3000, and 4000 samples. These EEG frequency domain images of size $H \times W'$, where $W'$ is the downsampled width, is then used as an input for the model.

The basis of the architecture was derived using the ImageNet (Krizhevsky et al., 2012) architecture that has proven to have excellent performance in the image recognition domain. The treatment of the preprocessed data as images with one color channel allows us to leverage the pattern recognition power of the deep CNNs to extract the discriminant patterns of depression. Given that the input coming in from the EEG recording device is multi-channel in nature, this enables us to treat each of these channels as a component of the height of the image. Thereby, we stack them one on top of the other to create an EEG “image”. The height of these images are the EEG channels and the width is the EEG readings inside these channels over time. The multi-channel nature of the data also enables us to make the image by stacking multiple channels.

## 4.1 Architecture Details

The base architecture is a 5 layer CNN interleaved with max-pooling layers and using the ReLu (Rectified Linear Unit) activation function as described by $y = \max\{0, x\}$. Two fully connected layers were used before the output layer and the final output was softmax to provide “Depressed” or “Not-Depressed”. An overview over this architecture can be seen in Figure 3. In order to arrive at the best model for this particular setup, 4 variants of the base architecture were created which went from shallow to deep in terms of the filters they incorporated. The details of the architecture of the convolutional layers can be seen in Table 2. The kernel size has been kept constant for all the variants in order to have a practical learning time. The aim of the architecture is to capture the general trends in the time series using the first three convolutional layers and capture the higher level trends using the last two layers.

**Table 2:** Architectures breakdown with respect to the number of filters applied in each convolutional layer. Each convolutional layer is followed by a max pooling layer with a stride of 2.

| Name | Filters / Complexity |
|---|---|
| Arch_1 | 3, 6, 8, 12, 18 |
| Arch_2 | 3, 6, 12, 24, 28 |
| Arch_3 | 3, 6, 12, 20, 30 |
| Arch_4 | 3, 6, 18, 36, 72 |

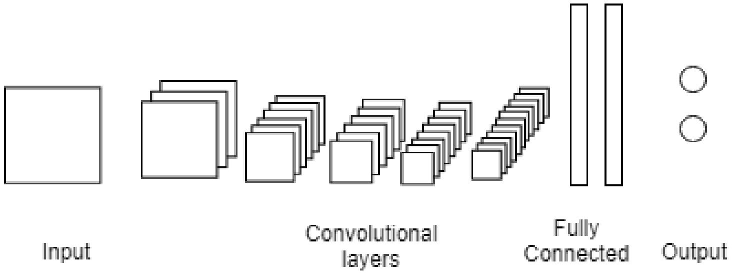

**Figure 3:** The above figure shows the base architecture components. Each convolutional layer is followed by a ReLU activation layer and a maxpooling layer. The different architectures mentioned in Table 2 correspond to the filters used in each layer respectively. Maxpooling stride has been chosen to decrease the input size by half. This technique is used to enhance the feature space while reducing the spatial size of the images ( that have been created after the preprocessing).

Along with the variants of the architecture, we also investigated the degree to which the data could be down-sampled in order to achieve comparable performance while simultaneously reducing the time needed to learn the model. The downsampling technique was inspired by Esling and Agon (2012), where they showed that time series could be effectively downsampled while still maintaining the integrity of the underlying trends of interest to us in the depression classification as stated in Hegerl et al. (2017).

The downsampling of the preprocessed data enabled us to extract the relevant trends within the sampling rate while bringing down the volume of the data at the same time. This was necessary, as mentioned in Section 3.2 and also as shown by empirical evidence.

# 5 Experiments

## 5.1 Baselines

According to Hosseinifard et al. (2013) good accuracies for the classification of depressed patients can be achieved by analyzing non-linear features of EEG-data. In their research, they compute the Higuchi-fractal, detrended fluctuation analysis (DFA), correlation dimension and Lyapunov exponent for every channel of the EEG-data and use those as input for classifiers as linear discriminant analysis (LDA), logistic regression (LR) and k-nearest-neighbors (KNN). For further improvement of the results, the features were combined and also applied to the classifiers. The data was split into two thirds training data and the remaining third for testing purposes. A genetic algorithm[1] on the training data was used for feature selection. In order to make most of the data, leave-one-out cross validation was used in this step.

In our research, the data only consists of 60 instead of 90 patients. Also, artifacts of the EEG-data are not removed. This resulted in a high volatility of the feature values. This problem is known and usually the reason, why artifacts get discarded (Tatum et al., 2011). The Lyapunov exponent did not return good results and was thus not used. Apart from that, the steps of Tatum et al. (2011) were followed. The best accuracy was computed when the data was downsampled

[1] https://github.com/manuel-calzolari/sklearn-genetic

to 2000 samples. Average accuracies for different classifiers and features over multiple runs can be seen in Table 3. Performing feature selection via a genetic algorithm with leave-one-out cross-validation slightly increased the accuracy for the combined features using an LDA-classifyer to an average of **68 %**. LR and KNN did not benefit from feature selection. This was verified by also performing reccursive feature elimination.

**Table 3:** Average accuracies for baseline-model using different features and classifiers.

| Classifier / Feature | Correlation | Higuchi | DFA | Combination |
|---|---|---|---|---|
| LDA | 55.6 % | 52.5 % | 52.5 % | **66.2 %** |
| LR | 54.5 % | 57.5 % | 52.6 % | 60.1 % |
| KNN | 61.7 % | 57.2 % | 53.7 % | 62.01 % |

## 5.2 Our Model

In order to test our model, the preprocessing was done to create two versions from the Leipzig dataset. One subset included frequencies up to 40 Hz while the other contained frequencies up to 100 Hz. These datasets were then brought together into one combined dataset. This artificial augmentation combined with the subsampling over the individual instances enabled us to double the amount of data that was available for the training of the model. This augmentation also enabled us to sample from the varying frequency ranges and still capture the underlying data trend which is the discriminant feature of depression. The combined dataset was then treated with the hold-out technique with 25 % of the data being used for testing the model. All models were trained using an Nvidia Tesla K80 GPU. The criteria for convergence was classification error based on the cross-entropy loss (Robert, 2014), the training was stopped when there was no noticeable reduction in the error anymore.

Each architecture as shown in Table 2 was tested with the down-sampling rate from 1000 to 4000 in steps of 1000 samples. This enabled us to find the right trade-off between information used vs training time. The results are presented in Table 4. It also shows the best average accuracy of the baseline model in Section 5.1. It is easy to see, that our model outperforms the baseline considerably.

**Table 4:** Results: Mean Accuracy for the different architectures (± standard deviation) for 10 runs.

| Name / Sampling | 1000 | 2000 | 3000 | 4000 |
|---|---|---|---|---|
| Arch_1 | 85.56 % ± 4.10 | 84.78 % ± 2.76 | 85.75 % ± 5.34 | 83.58 % ± 6.60 |
| Arch_2 | 84.78 % ± 2.63 | 84.06 % ± 2.86 | 86.94 % ± 4.95 | 84.30 % ± 1.90 |
| Arch_3 | **88.14** % ±**1.31** | 84.05 % ± 6.31 | 88.62 % ± 2.63 | 89.18 % ±5.82 |
| Arch_4 | 87.42 % ± 3.46 | 85.99 % ± 3.25 | 82.14 % ± 1.56 | 87.54 % ±2.31 |
| Baseline | | 68 % | | |

The mean accuracy listed in Table 4 shows that comparable performance can be achieved with an intermediate architecture. The need for a more complex architecture like Arch_4 does not enhance the accuracy significantly. Therefore, the most complex architecture, namely Arch-4, is not selected. This results in a reduction of the training time of the model due to a decrease in the model's complexity. This is further corroborated by the standard deviation which can also be seen in the results. Figure 4 shows the comparison of accuracy of the 4 architectures for the 4 sampling rates. The standard deviation (SD) for the runs is shown by the black bar atop the columns. The length of the bar corresponds to the SD (shorter the bar, smaller the SD).

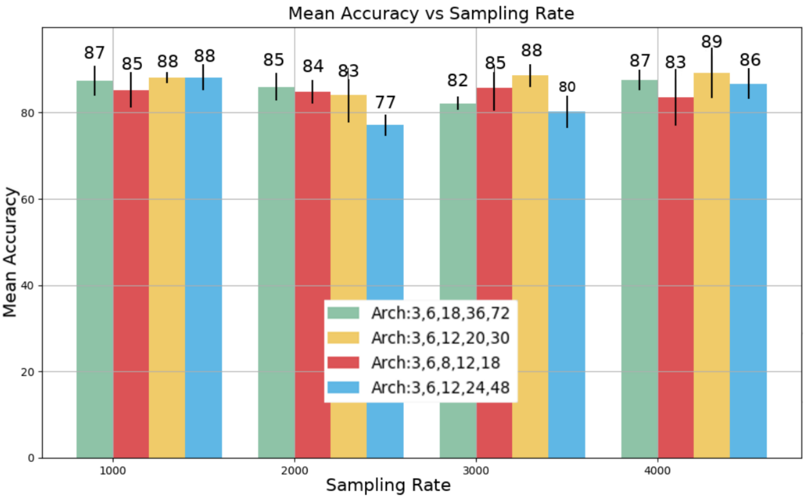

**Figure 4:** Comparison of architectures.

It is evident from Figure 4 that it is possible to get a very good performance by using an intermediate architecture such as Arch_3. We can also see that a more rigid downsampling regime leads to a more consistent performance in terms of the variance of the results. This can be due to the models overfitting on the higher samples of the data. Therefore a more rigid downsampling adds an intrinsic regularization to the model and leads to a more consistent performance.

It is also possible to inspect the confusion matrix in Table 5, which describes the best model performance both for depressed and non-depressed patients. It was obtained by running twice a 5-Fold Cross-Validation (with different random splits) on the Leipzig dataset and using the described data augmentation technique (different frequency threshold).

**Table 5:** Confusion Matrix (n = 240).

| **Predicted Value** \ **True Value** | **Depressed** | **Non-depressed** |
|---|---|---|
| **Depressed** | 106 (44.2 %) | 20 (8.3 %) |
| **Non-depressed** | 14 (5.8 %) | 100 (41.7 %) |

## 6 Conclusion

In this paper, we have shown that the advances made in the field of machine learning can contribute to improvement of MD diagnosis. We have presented a novel architecture which could be used with the frequency domain approach of the time series data coming in from conventional EEG equipment. The proposed methodology has been shown to perform better than the baseline methods tested here as well as to provide an end to end mechanism with which raw EEG data could be used for the diagnosis of depression. We have investigated several downsampling regimes with roughly comparable accuracy. Lastly, we have analyzed several architectures in terms of their complexity, to arrive at one that provides the best trade-off for training time versus accuracy. In contrast to the other methods, there is also no need to set the specific frequency ranges which will be used to set each of the vigilance states, since the model learns that by itself. Thus, future work on this model could possibly lead to a broader generalization with respect to other mental health-related applications. The small dataset size did not allow us to keep a holdout set. Also, the accuracy might get further increased by using a higher or lower sampling size, that we did not try.

**Acknowledgements** This work would not have been possible without the supervision of the University Hospital of Leipzig. We are grateful to Prof. Dr. med. Ulrich Hegerl and Dr. Tilman Hensch and their team for their work regarding the acquisition and maintenance of the dataset that they have provided for this study. We are also indebted to the University of Hildesheim and the team under Prof. Dr. Dr. Lars Schmidt-Thieme for their guidance and support which made this project possible.

## References

Amin HU, Mumtaz W, Subhani AR, Saad MNM, Malik AS (2017) Classification of EEG Signals Based on Pattern Recognition Approach. Frontiers in Computational Neuroscience 11:103. DOI: 10.3389/fncom.2017.00103.

Bachmann M, Lass J, Hinrikus H (2017) Single Channel EEG Analysis for Detection of Depression. Biomedical Signal Processing and Control 31:391–397. DOI: 10.1016/j.bspc.2016.09.010.

Chu L, Qiu R, Liu H, Ling Z, Shi X (2017) Individual Recognition in Schizophrenia Using Deep Learning Methods with Random Forest and Voting Classifiers: Insights from Resting State EEG Streams. arXiv preprint arXiv:1707.03467. URL: https://arxiv.org/abs/1707.03467.

Cireşan D, Meier U, Schmidhuber J (2012) Multi-column Deep Neural Networks for Image Classification. arXiv preprint arXiv:1202.2745. URL: https://arxiv.org/abs/1202.2745.

Esling P, Agon C (2012) Time-series Data Mining. ACM Computing Surveys (CSUR) 45(1):12, Association of Computing Machinery (ACM). DOI: 10.1145/2379776.2379788.

Goodfellow I, Bengio Y, Courville A, Bengio Y (2016) Deep Learning. MIT Press, Cambridge (USA).

Hazarika N, Chen JZ, Tsoi AC, Sergejew A (1997) Classification of EEG Signals Using the Wavelet Transform. In: Fakotakis N, Likothanassis S, Mourtzopoulos S, Philips W, Psarakis E, Rue V (eds.), Proceedings of 13th International Conference on Digital Signal Processing (DSP'97), Institute of Electrical and Electronics Engineers (IEEE), New York (USA), Vol. 1, pp. 89–92. DOI: 10.1109/ICDSP.1997.627975.

Hegerl U, Wilk K, Olbrich S, Schoenknecht P, Sander C (2012) Hyperstable Regulation of Vigilance in Patients with Major Depressive Disorder. The World Journal of Biological Psychiatry 13(6):436–446. DOI: 10.3109/15622975.2011.579164.

Hegerl U, Sander C, Ulke C, et al. (2017) Vigilance Algorithm Leipzig (VIGALL) Version 2.1 Manual. Leipzig (Germany). URL: https://research.uni-leipzig.de/vigall/.

Hjorth B (1970) EEG Analysis Based on Time Domain Properties. Electroencephalography and Clinical Neurophysiology 29(3):306–310. DOI: 10.1016/0013-4694(70)90143-4.

Hosseinifard B, Moradi MH, Rostami R (2013) Classifying Depression Patients and Normal Subjects Using Machine Learning Techniques and Nonlinear Features from EEG Signal. Computer Methods and Programs in Biomedicine 109(3):339–345. DOI: 10.1016/j.cmpb.2012.10.008.

Katyal Y, Alur SV, Dwivedi S, Menaka R (2014) EEG Signal and Video Analysis Based Depression Indication. In: IEEE International Conference on Advanced Communications, Control and Computing Technologies (ICACCCT), Institute of Electrical and Electronics Engineers (IEEE), New York (USA), pp. 1353–1360. DOI: 10.1109/ICACCCT.2014.7019320.

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges C, Bottou L, Weinberger K (eds.), Advances in Neural Information Processing Systems (NIPS'12), Curran Associates, Inc., Red Hook (USA), Vol. 25, pp. 1097–1105. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kupfer DJ, Frank E, Phillips ML (2012) Major Depressive Disorder: New Clinical, Neurobiological, and Treatment Perspectives. The Lancet 379(9820):1045–1055. DOI: 10.1016/S0140-6736(11)60602-8.

Motomura E, Inui K, Nakase S, Hamanaka K, Okazaki Y (2002) Late-onset Depression: Can EEG Abnormalities Help in Clinical Sub-typing? Journal of Affective Disorders 68(1):73–79. DOI: 10.1016/s0165-0327(00)00330-x.

Parvinnia E, Sabeti M, Jahromi MZ, Boostani R (2014) Classification of EEG Signals Using Adaptive Weighted Distance Nearest Neighbor Algorithm. Journal of King Saud University – Computer and Information Sciences 26(1):1–6. DOI: 10.1016/j.jksuci.2013.01.001.

Robert C (2014) Machine Learning: A Probabilistic Perspective. Taylor & Francis, Milton Park (United Kingdom). DOI: 10.1080/09332480.2014.914768.

Sai CY, Mokhtar N, Arof H, Cumming P, Iwahashi M (2018) Automated Classification and Removal of EEG Artifacts With SVM and Wavelet-ICA. IEEE Journal of Biomedical and Health Informatics 22(3):664–670. DOI: 10.1109/JBHI.2017.2723420.

Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S (2014) CNN Features Off-the-shelf: An Astounding Baseline for Recognition. In: OConner L (ed.), IEEE Conference on Computer Vision and Pattern Recognition Workshops, Institute of Electrical and Electronics Engineers (IEEE), New York (USA), pp. 806–813. DOI: 10.1109/CVPRW.2014.131.

Tatum WO, Dworetzky BA, Schomer DL (2011) Artifact and Recording Concepts in EEG. Journal of Clinical Neurophysiology 28(3):252–263. DOI: 10.1097/WNP.0b013e31821c3c93.

Wittchen HU, Pittrow D (2002) Prevalence, Recognition and Management of Depression in Primary Care in Germany: The Depression 2000 Study. Human Psychopharmacology: Clinical and Experimental 17:S1–S11. DOI: 10.1002/hup.398.

World Health Organization (2017) Depression and Other Common Mental Disorders: Global Health Estimates. Tech. Rep., World Health Organization. URL: `https://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/`.

Zheng Y, Liu Q, Chen E, Ge Y, Zhao JL (2014) Time Series Classification Using Multi-channels Deep Convolutional Neural Networks. In: Li F, Li G, Hwang S, Yao B, Zhang Z (eds.), International Conference on Web-Age Information Management, Springer, Cham (Switzerland), Lecture Notes in Computer Science, Vol. 8485, pp. 298–310. DOI: 10.1007/978-3-319-08010-9_33.