Audio-based Recognition of Bipolar Disorder Utilising Capsule Networks

Shahin Amiriparian[1], Arsany Awad[1], Maurice Gerczuk[1], Lukas Stappen[1], Alice Baird[1], Sandra Ottl[1], Björn Schuller[1,2]

***Abstract*—Bipolar disorder (BD) is an acute mood condition, in which states can drastically shift from one extreme to another, considerably impacting an individual's wellbeing. Automatic recognition of a BD diagnosis can help patients to obtain medical treatment at an earlier stage and therefore have a better overall prognosis. With this in mind, in this study, we utilise a Capsule Neural Network (CapsNet) for audio-based classification of patients who were suffering from BD after a mania episode into three classes of Remission, Hypomania, and Mania. The CapsNet attempts to address the limitations of Convolutional Neural Networks (CNNs) by considering vital spatial hierarchies between the extracted images from audio files. We develop a framework around the CapsNet in order to analyse and classify audio signals. First, we create a spectrogram from short segments of speech recordings from individuals with a bipolar diagnosis. We then train the CapsNet on the spectrograms with 32 low-level and three high-level capsules, each for one of the BD classes. These capsules attempt both to form a meaningful representation of the input data and to learn the correct BD class. The output of each capsule represents an activity vector. The length of this vector encodes the presence of the corresponding type of BD in the input, and its orientation represents the properties of this specific instance of BD. We show that using our CapsNet framework, it is possible to achieve competitive results for the aforementioned task by reaching a UAR of 46.2 % and 45.5 % on the development and test partitions, respectively. Furthermore, the efficacy of our approach is compared with a sequence to sequence autoencoder and a CNN-based neural network.**

***Index Terms*—capsule networks, spectrograms, bipolar disorder, deep learning, audio processing**

I. Introduction

Mental or neurological disorders affect millions of people yet a small minority of them receive treatment [1], [2]. As with any life altering condition, early diagnosis is advantageous as it quickens the chances of developing coping mechanisms for a diagnosed condition.

Bipolar disorder (BD) is a mental disorder that causes phases (known as episodes) of mania or depression leading to an abnormally elevated or deflated mood. These *episodes* could span for days or even months, and result in a significant variance in mood and motivation, affecting an individual's ability to carry out day-to-day tasks [3]–[7].

[1]Shahin Amiriparian, Arsany Awad, Maurice Gerczuk, Lukas Stappen, Alice Baird, and Björn Schuller are with the ZD.B. Chair of Embedded Intelligence for Health Care & Wellbeing, Univeristy of Augsburg, Germany. {shahin.amiriparian, arsany.awad}@tum.de {maurice.gerczuk, lukas.stappen, alice.baird, sandra.ottl, bjoern.schuller}@informatik.uni-augsburg.de

[2]Björn Schuller is also with GLAM – the Group on Language, Audio & Music, Imperial College London, UK.

The World Health Organisation has ranked BD as one of the top ten conditions prevalent in young adults with respect to the Disability-Adjusted Life Year (DALY). DALY is a measure for lifetime lost due to the effects of a severe disease [8]. One of the major challenges for BD is the resistance to treatment [9]–[12]. Previous studies failed to clearly identify the trigger for resistance to treatment, suggesting that patients either develop an aversion to the drugs because they are no longer potent or vice versa [9]–[12]. However, there is a common understanding that early detection of BD counteracts the risk of resistance to treatment at an early stage, for instance, by enabling early (pre-)treatment and also reducing frequent hospitalisation in later stages of the disease [9], [11], [13].

In order to increase the ability to automatically measure emotional states for better recognition of mental disorders, such as BD, various machine learning approaches have been developed in recent years. In the Audio/Visual Emotion Challenge 2018 (AVEC) [4], some of them, such as, Support Vector Machines (SVMs) and Deep Neural Networks have shown strong performance in classification of BD patients in the three severity classes Remission, Hypomania, and Mania based on audio-video recordings [4], [13]. In [4], the authors utilise supervised, semi-supervised, and unsupervised methods to extract and learn representations from BD data. For the supervised approach, which showed the best performance on the BD sub-challenge, the authors extracted low-level descriptors from continuous speech and video data. Moreover, Ringeval et al. applied a semi-supervised learning method using bag-of-words [14], and an unsupervised method using DEEP SPECTRUM features [15] for BD recognition with promising results. These results demonstrated that BD is most likely linked to affective states in speech, and using emotion-related feature representations were a key aspect for outperforming other approaches.

In addition to well-established machine learning approaches, Sabour et al. [16] have recently introduced a new neural network architecture for computer vision: The Capsule Neural Network (CapsNet). A major development behind the CapsNet is the replacement of pooling layers – an integral part of Convolutional Neural Networks (CNNs) – with routing-by-agreement [16]. The objective behind routing-by-agreement is to deduce relationships between low-level and higher-level features, thus learning the part-whole relationship during training, and to enable the learning of visual patterns in a new, more cohesive way. This property helps CapsNets to achieve consistent results irrespective of any audio-visual

transformations applied on the data. As a result, any dataset used, despite its size, would be utilised more efficiently during the training phase [17]. This is a crucial aspect in the medical field, given the scarcity of high-quality data.

Despite initial promising results, a CapsNet adaptation for audio-based or medical tasks is still in its infancy. For example, Afhar et al. [18] utilise a CapsNet for the classification of brain tumour types based on a Magnetic Resonance Imaging dataset. Another example is given in [19], which evaluates the ability of CapsNets for speech understanding, by first extracting high-level audio features using filter banks and bidirectional Gated Recurrent Units (GRUs), and then feeding them into what are known as *capsule layers* for classification. Lastly, Vesperini et al. [20] compared CapsNet to pure CNNs using spectrograms as input data to detect various audio events, and showing that CapsNets can additionally recognise some higher frequency spectral patterns.

In this study, we present a novel machine learning framework consisting of CapsNet components that learns the visual features from the extracted spectrograms to classify the severity of BD. This particular medical task seems to be heavily dependent on the extraction and learning of (new) audiovisual features, making it a promising starting point for the exploration of CapsNets with medical audio datasets. Furthermore, we compare our approach with two audio representation learning frameworks, AUDEEP and DEEP SPECTRUM, on this classification problem.

This paper is organised as follows. In the proceeding section, the BD dataset used in our experiments is presented. Then, the structure of our proposed framework as well as the baseline systems are introduced in Section III. Afterwards, the experimental results are discussed and analysed in Section IV. Finally, conclusions and future work plans are given in Section V.

II. BIPOLAR DISORDER DATASET

The dataset used for the aforementioned task is based on the BD dataset introduced by Çiftçi et al. [13]. This is one of the corpora used in the AVEC 2018 Challenge [4] and includes both audio and video recordings taken from structured interviews conducted with 100 Turkish natives diagnosed with a form of BD. Under the approval of the ethical committee, these patients were recruited from a mental health service hospital and have a prior diagnosis of BD following the DSM-5 inclusion criteria [21].

A group of patients were removed from the dataset on the grounds of some exclusion criteria, such as substance or alcohol abuse three month prior to an additional severe organic disease, signs of hallucinations, low mental capacity, or finally, disruptive behaviours during the session [13]. As well as this, some study participants refused to share their data publicly.

The final number of subjects is 46, consisting of 30 males and 16 females with a mean age of 36.5 and a standard deviation of 10.2 years. Sessions were recorded during a hospitalisation period as well as post their discharge on the third month. After each session, a rating on the Young Mania Rating Scale (YMRS) is given to each subject. These scores are then grouped into three classes: Remission, Hypomania, and Mania. As detailed in Table I, the *Mania* class includes patients with the highest YMRS scores, these exhibit severe BD symptoms. The second class is *Hypomania* and these subjects show milder symptoms. Finally, the *Remission* class with the lowest YMRS scores includes patients that show reduced symptoms.

After further processing of the final dataset, 88 audio samples are obtained for the Mania class, 82 for Hypomania, and 62 for Remission.

TABLE I
GROUPING OF THE YMRS VALUES INTO THREE LABELS.

Name	Label	YMRS
Remission	1	≤ 7
Hypomania	2	8–19
Mania	3	≥ 20

III. EXPERIMENTAL SETTINGS

As a means of showing the promise that CapsNets hold for this ternary classification task, the BD dataset has been used to conduct three deep learning experiments: i) the novel CapsNet framework (cf. Section III-A), ii) DEEP SPECTRUM, which utilises pre-trained image CNNs for feature extraction [15] (cf. Section III-B), and iii) AUDEEP, which is a recurrent sequence to sequence autoencoder (S2SAE) [22], [23] (cf. Section III-C). We apply AUDEEP and DEEP SPECTRUM in order to compare the efficacy of our CapsNet framework with alternative state-of-the-art deep learning methodologies.

A. Capsule Neural Network

CapsNets are a state-of-the-art approach first utilised in the field of computer vision. The novelty of CapsNets lies in the introduction of *capsules* as a new deep learning component [16], [24]. A *capsule* is a group of several neurons that performs complex internal computations within themselves. A characteristic property of a capsule is the nature of its input and output: A capsule j receives as input a set of vectors $u_{1...n}$ and outputs a vector v_j. The activation of the neurons inside a capsule designates the presence of a number of properties that form an entity or a higher-level feature represented by that capsule [16], [24]. The output of a capsule is in the form of an activation vector; its length indicates the probability that the entity represented by the capsule is found in the input [16], [24]. This is a departure from regular neurons that would only output a scalar activation value.

The orientation of the activation vector of a capsule represents the "instantiation parameters" [16]. These parameters describe how deformed the input is compared to a canonical form of the feature implicitly learnt by the capsule. This property is essential for the CapsNet to achieve viewpoint invariance [16].

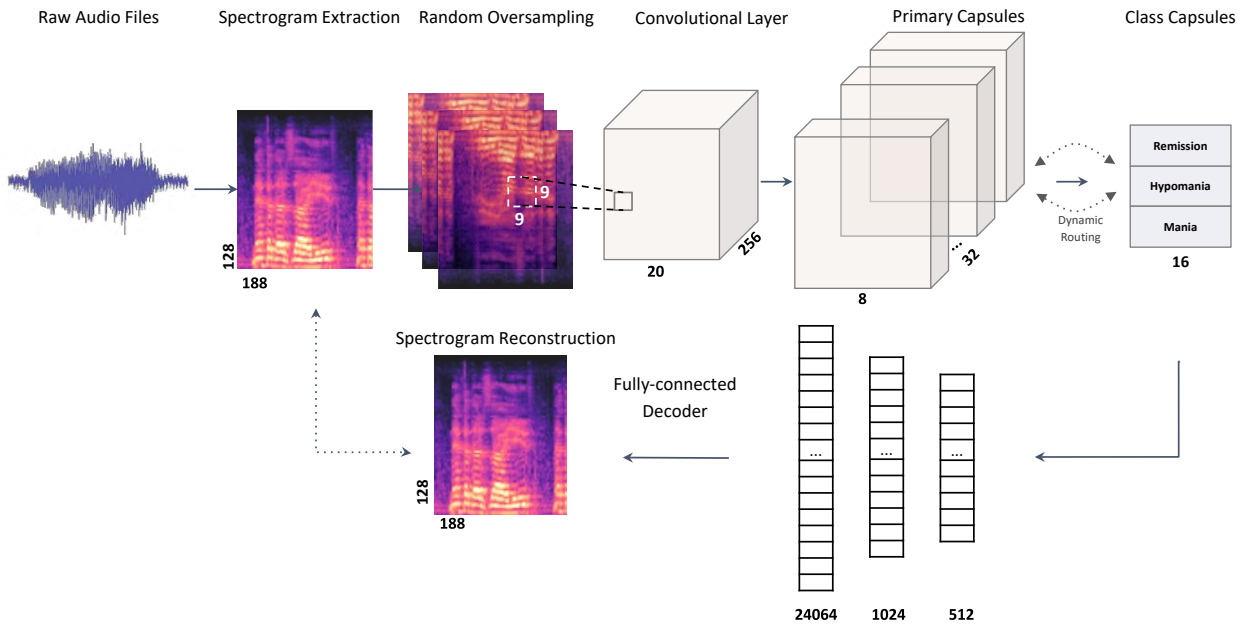

Fig. 1. The architecture of the proposed CapsNet framework for audio processing. First, spectrograms are generated from raw audio. Then, an optional step of data augmentation is executed, depending on the class imbalance. Finally, the generated spectrograms are fed into the CapsNets for training and classification. A detailed description of this procedure is given in Section III-A.

1) Dynamic Routing:: Another significant aspect to CapsNets is the routing of the input from one layer to the next. In order to ensure that the output of a capsule is sent to the suitable parent in the above layer, dynamic routing is applied [16]. In doing so, a part-whole relationship is learnt between low-level and high-level features. At the start of the training, the output of any given capsule is routed to all possible capsules in the above layer by means of coupling coefficients [16]. All coupling coefficients between a given capsule and all capsules of the above layer should sum to 1. During the training phase, these coefficients are refined to reflect the relationship between learnt features.

2) Our Framework for Audio Processing: To use CapsNets on the audio files of the BD dataset, we develop a process which starts with a set of raw audio clips and ends with training and classification. This is depicted in Figure 1. In the first step (depicted left of Figure 1), spectrograms are generated from 2 s short segments of speech recordings with an overlap of 1 s from the audio recordings of individuals with a bipolar diagnosis. From these spectrograms, we then compute 128 mel-frequency bands. Subsequently, we use random oversampling to compensate the class imbalance problem. The training partition contains 4 511, 7 888, and 12 687 spectrograms of the Remission, Hypomania, and Mania classes, respectively. Afterwards, as illustrated in Figure 1, the CapsNet is trained on the extracted spectrograms with three high-level capsules, one for each of the YMRS levels. These capsules attempt both to learn the correct BD class (YMRS level), and to build a robust representation of the input data. The output of each capsule represents an activity vector. The length of this vector encodes the presence of the corresponding type of BD in the input, and its orientation represents the properties of this specific instance of BD. In the final step, the predictions are obtained to assess the performance of the network. For the evaluation metric we use Unweighted Average Recall (UAR), which is the average of the recall from each of the BD classes. The chance-level for our three-class classification problem is 33.3 % UAR.

The CapsNet in our framework is composed of the following layers (cf. Figure 1):

- A convolutional layer having 256 kernels with a size of 9, a stride of 1, and a Rectified Linear Unit (ReLU) activation [25].
- 32 primary capsules: Each capsule is 8 dimensional, containing 8 convolutional units having a 9×9 kernel and a stride of 2.
- 3 top-level class capsules, one for each BD class (Remission, Hypomania, and Mania).
- A decoder layer which takes a 16×3 vector as input. This is then fed into three fully connected layers containing 512, 1 024, and 24 064 neurons, respectively.

3) Hyperparameters: We train the described CapsNet architecture on the training partition of the BD dataset [13], and use the independent development partition to observe the training process of the network. Most hyperparameter values are adopted by the CapsNet-Keras implementation[1]. We apply an initial learning rate of 0.001 which is decayed with a factor of 0.9 after every epoch (cf. Table II). The decoder reconstruction loss is given a weight of 0.392, and we feed

[1]The implementation alongside reasoning for the chosen hyperparameter values can be found at https://github.com/XifengGuo/CapsNet-Keras

batches of 10 spectrograms each to the network. We further balance the training partition by randomly oversampling the minority classes (4 205, 11 260 and 13 144 spectrograms of Remission, Hypomania, and Mania classes, respectively) to avoid the CapsNet getting halted in a local minimum early in the training process. This means that for the Remission and Hypomania classes additional, duplicated samples are picked with replacement at random from the respective classes. Specifically, we use the RandomOversampler from the imblearn[2] python library. The development and testing partitions are left untouched.

B. DEEP SPECTRUM

Our second deep learning approach is DEEP SPECTRUM[3] [15], which is an open-source Python toolkit with process parallelisation for rapid GPU-based deep feature extraction from audio data by applying pre-trained CNNs, such as AlexNet [26], GoogLeNet [27], or VGG networks [28]. DEEP SPECTRUM features have shown strong performance for audio-based recognition tasks, including classification of autistic child vocalisation [29], sentiment analysis [30], classification of various speech and vocalisation types [31], [32], and emotion recognition [33].

To extract the DEEP SPECTRUM features, we first generate mel-spectrograms from the audio recordings in the BD dataset using a Hanning window of width 2 s and an overlap of 1 s. From these, we then compute 128 mel-frequency bands. Afterwards, the generated mel-spectrograms are forwarded through AlexNet [26], a pre-trained image CNN, and the activations of the second-last fully connected layer (fc7) with 4 096 neurons are extracted (cf. Table II). This results in a 4 096 dimensional DEEP SPECTRUM feature set, which can be considered as a high-level representation of the input mel-spectrograms [15].

C. Recurrent Sequence to Sequence Autoencoders

Our third deep learning system is AUDEEP[4], a highly effective deep architecture for unsupervised representation learning from audio data utilising recurrent S2SAEs [22], [23]. We apply AUDEEP, since commonly used deep representation learning methods, such as CNNs, Restricted Boltzmann Machines [25], [34], or stacked autoencoders [35] do not explicitly model the sequential nature of acoustic data, and generally need inputs of a fixed dimensionality [36].

To learn deep representations using AUDEEP, we first chunk the BD speech recordings into 2 s segments, from which we create power spectra. A S2SAE is then trained on these spectra, and the learnt representations are extracted for use as feature sets for the classification.

The spectrograms are generated with Hanning windows of width 0.08 s and an overlap of 0.04 s, from which 128 log-scaled mel-frequency bands are computed (cf. Table II). The spectrograms are normalised between [-1; 1], as the outputs of the S2SAE are constrained to this range [22], [23].

For our S2SAE, we apply two recurrent layers, each with 256 GRUs, a unidirectional encoder and a bidirectional decoder. The S2SAE is trained on the generated spectrograms using Adam optimiser with a fixed learning rate of 0.001 [37] for 10 epochs in batches of 64 samples. In order to minimise the problem of overfitting, a dropout of 20 % is applied to the output of each recurrent layer [38]. We also apply amplitude clipping with a threshold of -60 dB to filter some background noise. The list of all selected hyperparameters is given in Table II).

IV. RESULTS

A summary of the results from all three deep learning approaches and the applied hyperparameters is given in Table II. In the following sections, the training process of all deep learning approaches are described, and the results are analysed.

We train **CapsNet** for a maximum of 20 epochs and use the UAR achieved on the development partition to choose the best model checkpoint for evaluation on the test partition. After the fifth epoch of training the network achieves a UAR of 46.2 % and 45.5 % on the development and test partitions, respectively. A confusion matrix from the labels of the test set is shown in Figure 2, from which it can be seen that there is a minimal confusion between the Remission, and Hypomania classes, and a higher confusion between Mania and the other two classes.

In order to evaluate the extracted **DEEP SPECTRUM** features, the LibLINEAR library with the L2-regularised L2-loss dual solver [39] is used via the open-source linear SVM implementation provided in the scikit-learn machine learning library [40]. Furthermore, feature standardisation is applied, and the SVM complexity parameter $C \in [10^{-9}, 10^{0}]$ is optimised on the development partition with a factor of 10. Finally, implementing the best configuration on the development set ($C = 10^{-3}$) the DEEP SPECTRUM features are evaluated on the test set. Using this configuration and the parameters shown in Table II, a UAR of 45.0 % and 45.6 % can be achieved on the development and test partitions, respectively. It can be observed that both CNN-based approaches, DEEP SPECTRUM and CapsNet, show highly similar performance on the BD dataset. However, it should be noted that due to the extremely high computational cost and long training periods needed for the CapsNet implementation, we were not able to optimise the hyperparameters of this framework. Therefore, we cannot fully evaluate the extent to which the CapsNet framework may have achieved on this audio-based recognition task.

For the evaluation of the **AUDEEP** features, a Multilayer Perceptron (MLP) with two hidden fully connected layers with ReLU activation, and a softmax output layer is used. Each hidden layer contains 150 units, and the output layer has one unit for each class (i. e. three neurons for three classes). A dropout of 40 % is applied to all layers except the output layer [38], and the network has been trained for 400 epochs with a fixed learning rate of 0.001 [37] (cf. Table II). Using

[2]https://github.com/scikit-learn-contrib/imbalanced-learn
[3]https://github.com/DeepSpectrum/DeepSpectrum
[4]https://github.com/auDeep/auDeep

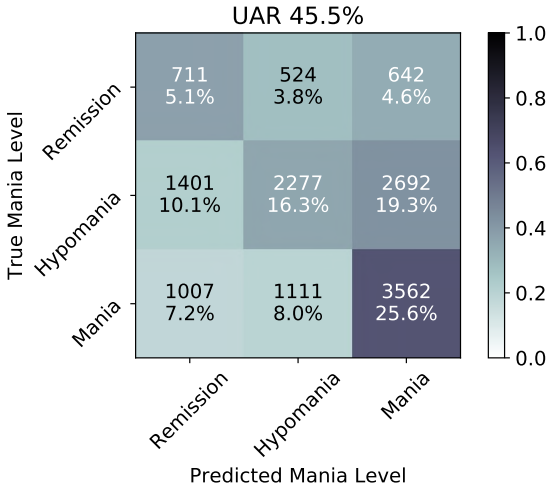

Fig. 2. Confusion matrix of the predictions by the CapsNet from the labels of the test partition.

this configuration, it was possible to reach a UAR of 47.2 % and 49.8 % on the development and test sets, respectively.

The overall results indicate the promise for CapsNets, as even without any hyperparameter optimisation we see that this framework can achieve comparable results to DEEP SPECTRUM and AUDEEP. We observe that AUDEEP achieves the highest performance (4.2 and 4.3 percentage points higher than the DEEP SPECTRUM and CapsNet results, respectively), confirming findings in previous works regarding the suitability of Recurrent Neural Networks for sequential data, such as natural speech [41]–[44]. In the context of BD sub-challenge [4], it is shown that an approach built around CapsNets yields competitive results.

V. CONCLUSIONS AND FUTURE WORK

CapsNets were originally introduced to compensate the generalisation problem of CNNs towards observing novel viewpoints in images [16], [24]. CNNs using spectrograms as input often encounter difficulties when recognising data from classes with fine-grained similarities [31], [45], [46]. With this in mind, we developed our framework around CapsNets to evaluate whether a better generalisation for spectrogram classification can be achieved. We showed that using our approach, it is possible to achieve state-of-the-art results for the highly complex task of bipolar speech classification. We assume that by optimising the parameters of the CapsNet higher performance can be achieved.

The presented CapsNet result motivates exploring this framework in more depth. For such a deeper analysis the quality of the results should be shown numerically with the inclusion of a significance measure. A suggested metric is the two-tailed t-test, where the null hypothesis would be the baseline framework (AVEC 2018 challenge [4], AUDEEP, or DEEP SPECTRUM). To achieve this, development work should be made, so that the CapsNets code outputs the predictions of each set.

TABLE II
COMPARISON OF THE CLASSIFICATION RESULTS ACHIEVED WITH CAPSNET, DEEP SPECTRUM, AND AUDEEP. FOR EACH NETWORK THE LIST OF HYPERPARAMETERS AND PARAMETERS FOR THE SPECTROGRAM CREATION ARE GIVEN. THE CHANCE-LEVEL IS 33.3 % UAR.

System		development	test
CapsNet			
chunk size [s]	2		
chunk overlap [s]	1		
primary capsules	32		
class capsules	3		
initial learning rate	0.001		
learning rate decay	0.9		
reconstruction weight	0.392		
batchsize	10		
epochs	5		
UAR [%]	–	46.2	45.5
Baseline 1: DEEP SPECTRUM (pre-trained CNNs)			
colour map	viridis		
chunk size [s]	2		
chunk overlap [s]	1		
number of mel-filters	128		
CNN descriptor	AlexNet fc7		
feature dimension	4 096		
SVM complexity	10^{-3}		
UAR [%]	–	45.0	45.6
Baseline 2: AUDEEP (S2SAE)			
chunk size [s]	1		
window width [s]	0.08		
window overlap [s]	0.04		
number of mel-filters	128		
amplitude clipping [dB]	-60		
S2SAE number of layers	2		
S2SAE units (GRU cell)	256		
S2SAE epochs	10		
S2SAE batchsize	64		
S2SAE dropout	0.2		
S2SAE learning rate	0.001		
MLP number of layers	2		
MLP units	150		
MLP epochs	400		
MLP dropout	0.4		
MLP learning rate	0.001		
UAR [%]	–	47.2	49.8

The performance of CapsNet came at a high computational cost. This is due to the complexity of the computations inside each capsule and the nature of their input and output being vectors as opposed to the scalar values in conventional neural networks [24], [47]. Another reason for such computational costs is the high number of parameters that need to be trained. For our BD classification task, the number of trainable parameters was as high as 89 724 672. As a result, training the CapsNet for one epoch using the BD dataset (after random oversampling) of size 38 061 spectrograms typically takes 2.5 hours on an Nvidia GTX TITAN X with 12 GB of VRAM. This means, that for 20 epochs of training, $20 \times 2.5 = 50$ hours of training were needed; that is 2 days and 2 hours. In this way, the authors propose that it would be valuable to research ways to optimise the computations inside a CapsNet capsule, as well as trying to utilise the recent advancement in GPU hardware to more easily allow for their inherent ability to perform more

complex computations.

These performance shortcomings are partly addressed by Hinton et al. in their recently published paper [47]. The authors replace routing-by-agreement with an Estimation Maximisation algorithm, and apply new regularisation methods instead of a fully connected decoder [47]. We plan to investigate these computational improvements in more detail in future work. We also want to verify whether it is possible to extract meaningful representations by using the activations of the neurons in the decoder layers of the CapsNet. As well as this, it is also interesting to evaluate the efficacy of the CapsNet framework on a wider range of speech and audio recognition tasks, such as autism severity classification [29], speech emotion recognition [48], [49], or acoustic scene classification [50].

Acknowledgements

This work is funded by the EUs Horizon 2020 Programme under grant agreement No. 688835 (RIA DE-ENIGMA), the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B) and the BMW Group research.

References

[1] R. C. Kessler, M. Angermeyer, J. C. Anthony, R. De Graaf, K. Demyttenaere, I. Gasquet, G. De Girolamo, S. Gluzman, O. Gureje, J. M. Haro *et al.*, "Lifetime prevalence and age-of-onset distributions of mental disorders in the world health organization's world mental health survey initiative," *World psychiatry*, vol. 6, no. 3, p. 168, 2007.

[2] WHO, "Mental disorders affect one in four people," 2001. [Online]. Available: http://www.who.int/whr/2001/media_centre/press/release/en/

[3] P. E. Keck Jr, S. L. McElroy, S. M. Strakowski, S. A. West, K. W. Sax, J. M. Hawkins, M. L. Bourne, and P. Haggard, "12-month outcome of patients with bipolar disorder following hospitalization for a manic or mixed episode," *American Journal of Psychiatry*, vol. 155, no. 5, pp. 646–652, 1998.

[4] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. Ciftci, H. Gülec, A. A. Salah, and M. Pantic, "AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition," in *Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC'18, co-located with the 26th ACM International Conference on Multimedia, MM 2018*, F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, Eds., ACM. Seoul, South Korea: ACM, October 2018, 10 pages, to appear.

[5] J. Angst and R. Sellaro, "Historical perspectives and natural history of bipolar disorder," *Biological psychiatry*, vol. 48, no. 6, pp. 445–457, 2000.

[6] G. L. Dion, M. Tohen, W. A. Anthony, and C. S. Waternaux, "Symptoms and functioning of patients with bipolar disorder six months after hospitalization," *Psychiatric Services*, vol. 39, no. 6, pp. 652–657, 1988.

[7] M. J. Gitlin, J. Swendsen, T. L. Heller, and C. Hammen, "Relapse and impairment in bipolar disorder," *The American journal of psychiatry*, vol. 152, no. 11, p. 1635, 1995.

[8] A. Table, "3: burden of disease in dalys by cause, sex and mortality stratum in who regions, estimates for 2002," *The world health report*, 2004.

[9] M. Gitlin, "Treatment-resistant bipolar disorder," *Focus*, vol. 11, no. 1, pp. 227–63, 2007.

[10] I. E. Bauer, J. C. Soares, S. Selek, and T. D. Meyer, "The link between refractoriness and neuroprogression in treatment-resistant bipolar disorder," in *Neuroprogression in Psychiatric Disorders*. Karger Publishers, 2017, vol. 31, pp. 10–26.

[11] M. Berk, P. Conus, N. Lucas, K. Hallam, G. S. Malhi, S. Dodd, L. N. Yatham, A. Yung, and P. McGorry, "Setting the stage: from prodrome to treatment resistance in bipolar disorder," *Bipolar disorders*, vol. 9, no. 7, pp. 671–678, 2007.

[12] C.-T. Li, Y.-M. Bai, Y.-L. Huang, Y.-S. Chen, T.-J. Chen, J.-Y. Cheng, and T.-P. Su, "Association between antidepressant resistance in unipolar depression and subsequent bipolar disorder: cohort study," *The British Journal of Psychiatry*, vol. 200, no. 1, pp. 45–51, 2012.

[13] E. Çiftçi, H. Kaya, H. Güleç, and A. A. Salah, "The turkish audio-visual bipolar disorder corpus," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 2018, pp. 1–6.

[14] M. Schmitt and B. Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3370–3374, 2017.

[15] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. Stockholm, Sweden: ISCA, August 2017, pp. 3512–3516.

[16] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *CoRR*, vol. abs/1710.09829, 2017. [Online]. Available: http://arxiv.org/abs/1710.09829

[17] A. Géron, "Introducing capsule networks," https://www.oreilly.com/ideas/introducing-capsule-networks, accessed: 12.01.2019.

[18] P. Afshar, A. Mohammadi, and K. N. Plataniotis, "Brain tumor type classification via capsule networks," *arXiv preprint arXiv:1802.10200*, 2018.

[19] V. Renkens *et al.*, "Capsule networks for low resource spoken language understanding," *arXiv preprint arXiv:1805.02922*, 2018.

[20] F. Vesperini, L. Gabrielli, E. Principi, and S. Squartini, "Polyphonic sound event detection by using capsule neural network," *arXiv preprint arXiv:1810.06325*, 2018.

[21] A. P. Association, *Diagnostic Criteria and Codes*. American Psychiatric Association, 2013, ch. 5. [Online]. Available: https://dsm.psychiatryonline.org/doi/abs/10.5555/appi.books.9780890425596.Section2

[22] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proceedings of the 2nd Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE 2017)*. Munich, Germany: IEEE, November 2017, pp. 17–21.

[23] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *Journal of Machine Learning Research*, vol. 18, no. 173, pp. 1–5, 2018.

[24] G. E. Hinton, A. Krizhevsky, and S. Wang, "Transforming auto-encoders," in *International Conference on Artificial Neural Networks*, vol. 6791. Springer, 06 2011, pp. 44–51.

[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*. Haifa, IS: ACM, 2010, pp. 807–814.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA, USA: IEEE, 2015, pp. 1–9.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[29] A. Baird, S. Amiriparian, N. Cummins, A. M. Alcorn, A. Batliner, S. Pugachevskiy, M. Freitag, M. Gerczuk, and B. Schuller, "Automatic classification of autistic child vocalisations: A novel database and results," in *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. Stockholm, Sweden: ISCA, August 2017, pp. 849–853.

[30] S. Amiriparian, N. Cummins, S. Ottl, M. Gerczuk, and B. Schuller, "Sentiment analysis using image-based deep spectrum features," in *Proceedings of the 7th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, 2017, pp. 26–29.

[31] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, S. Pugachevskiy, and B. Schuller, "Bag-of-deep-features: Noise-robust deep feature representations for audio analysis," in *Proceedings of the 31st International Joint Conference on Neural Networks (IJCNN)*. Rio de Janeiro, Brazil: IEEE, July 2018, pp. 2419–2425.

[32] S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. Schuller, "CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," in *Proceedings of the 7th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, 2017, pp. 340–345.

[33] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 25th ACM International Conference on Multimedia, MM 2017*. Mountain View, CA: ACM, October 2017, pp. 478–484.

[34] G. E. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 599–619.

[35] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of machine learning research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[36] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[37] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, Banff, CA, 2014.

[38] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[39] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[41] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 ieee international conference on Acoustics, speech and signal processing (icassp)*. IEEE, 2013, pp. 6645–6649.

[42] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[43] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth annual conference of the international speech communication association*, 2014.

[44] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *Proceedings of INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association*. Hyderabad, India: ISCA, September 2018, pp. 122–126.

[45] S. Amiriparian, A. Baird, S. Julka, A. Alcorn, S. Ottl, S. Petrović, E. Ainger, N. Cummins, and B. Schuller, "Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks," in *Proceedings of INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association*. Hyderabad, India: ISCA, September 2018, pp. 2334–2338.

[46] S. Amiriparian, N. Cummins, M. Gerczuk, S. Pugachevskiy, S. Ottl, and B. Schuller, ""are you playing a shooter again?!" deep representation learning for audio-based video game genre recognition," *IEEE Transactions on Games*, 2018, 11 pages, to appear.

[47] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with em routing," 2018.

[48] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[49] A. Kumar and B. Raj, "Audio event and scene recognition: A unified approach using strongly and weakly labeled data," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 3475–3482.

[50] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference (EUSIPCO), 2016*. IEEE, 2016, pp. 1128–1132.