One-Dimensional Convolutional Neural Networks on Motor Activity Measurements in Detection of Depression

Joakim Ihle Frogner*
SimulaMet, Norway

Farzan Majeed Noori
Department of Informatics, University of Oslo, Norway

Pål Halvorsen†
SimulaMet, Norway

Steven Alexander Hicks†
SimulaMet, Norway

Enrique Garcia-Ceja
Software and Service Innovation, SINTEF Digital, Norway

Jim Torresen
Department of Informatics and RITMO, University of Oslo, Norway

Michael Alexander Riegler‡
SimulaMet, Norway

ABSTRACT

Nowadays, it has become possible to measure different human activities using wearable devices. Besides measuring the number of daily steps or calories burned, these datasets have much more potential since different activity levels are also collected. Such data would be helpful in the field of psychology because it can relate to various mental health issues such as changes in mood and stress. In this paper, we present a machine learning approach to detect depression using a dataset with motor activity recordings of one group of people with depression and one group without, i.e., the condition group includes 23 unipolar and bipolar persons, and the control group includes 32 persons without depression. We use convolutional neural networks to classify the depressed and non-depressed patients. Moreover, different levels of depression were classified. Finally, we trained a model that predicts Montgomery-Åsberg Depression Rating Scale scores. We achieved an average F1-score of 0.70 for detecting the control and condition groups. The mean squared error for score prediction was approximately 4.0.

CCS CONCEPTS

• **Human-centered computing → Mobile devices**;

KEYWORDS

Depression; Bipolar Disorder; Machine Learning

ACM Reference Format:
Joakim Ihle Frogner, Farzan Majeed Noori, Pål Halvorsen, Steven Alexander Hicks, Enrique Garcia-Ceja, Jim Torresen, and Michael Alexander Riegler. 2019. One-Dimensional Convolutional Neural Networks on Motor Activity Measurements in Detection of Depression, . In *4th International Workshop on Multimedia for Personal Health and Health Care (HealthMedia ’19), October 21, 2019, Nice, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3347444.3356238

1 INTRODUCTION

In recent years, the use of wearable devices to monitor mental and physical health has become quite normal. People are collecting data every day to improve their lives and to supervise their fitness levels. Besides measuring the quality of life, the data gathered by these devices may also be useful from a psychiatric perspective, where the data can be used to diagnose various mental health issues such as depression or changes in mood [15]. Currently, depression is one of the most frequent disorders and is expected to increase in upcoming years [19].

Mental illness is considered as disturbances in the brain which may cause changes in a person’s mood, thinking or behaviour [12]. As the relationship between mood and sensor data are not well understood, it is difficult to predict how changes in sensor data may be correlated with a person’s change in mood. In 2009, Scheffer et al. [18] discussed the phenomena of critical slowing down, which indicates the occurrence of early warning signals in critical transition periods preceding abrupt noticeable changes of state. Critical slowing down indicates that the system is unable to recover its original condition from small disturbances [1]. Unipolar depression and bipolar disorder are episodic mood disorders, where the pathological state and the healthy state might be understood as representing different stable states separated by abrupt changes between them [5].

Through the recording of motor activity, the biological system’s state is measurable. It has been noticed that reduced day-time activity and increased night-time activity indicates depressive state as compared to normal state [3].

In the field of mental health, activity and movement measurements have become an emerging topic. Several studies use sensors to diagnose or self-report patient movement over time [2, 16]. Different linear and non-linear statistical methods have been used to analyze the data. Reported findings include increased auto-correlations and variances as indicators of a critical slowing down [18], and

[*] Also affiliated with University of Oslo, Norway
[†] Also affiliated with Oslo Metropolitan University, Norway
[‡] Also affiliated with Kristiania University College, Norway

HealthMedia ’19, October 21, 2019, Nice, France

ACM ISBN 978-1-4503-6914-5/19/10…$15.00
https://doi.org/10.1145/3347444.3356238

increased skewness is also observed [10]. Such data also holds potential for machine learning applications which is used more in the context of psychiatry and psychology [11, 14].

In this paper, we use a dataset containing motor activity of depressed and non-depressed patients to perform depression classification using machine learning. One-Dimensional convolutional neural networks (1D-CNNs) are used on motor activity measurements to detect the depression. Afterwards, three levels of depression (i.e., no depression, mild, and severe depression) were detected based on the Montgomery-Åsberg Depression Rating Scale (MADRS) [13]. Our third model predicts the MADRS score of participants. The motivation behind using 1D-CNN is that feature extraction would be done automatically.

We evaluated the performance of classification models using a *leave one participant out* cross-validation combined with majority voting. In depressed vs non-depressed classification, we achieved an average F1-score of 0.70 and 0.30 for detecting the different levels of depression. However, this model detects non-depressed participants with high performance. For the MADRS prediction model, we achieved an *mean squared error (MSE)* of 4.0.

The main contributions of this paper is to build 1D-CNN in order to: (i) Detect whether a participant is depressed or not; (ii) Detect whether a participant has no depression, mild depression, or severe depression; (iii) Predict participant's MADRS score which further can be used to distinguish between depressed and non-depressed patients.

The rest of this paper is organized as follows: Section 2 describes background and related work. Section 3 presents the dataset. Experiments and results are shown in Section 4. The paper is concluded in Section 5.

2 RELATED WORK

2.1 Mental Health Monitoring Systems

In the field of Mental Health Monitoring Systems (MHMS), research has already been done by many. In this section, we discuss some earlier research about depression and bipolar disorder, where they also applied machine learning to their study.

In a recent study conducted by Garcia et al. [8], the authors surveyed recent research works on the use of machine learning for MHMS. They classified different works by: study type (association/detection/forecasting), study duration (short-term or long term), and sensor type (wearable/external/software or social media). Association studies were conducted on those that help understand the relationships between variables. Methods used include linear regression, correlation analysis, t-tests and analysis of variance. Detection studies have a goal to detect/recognize the mental state, often using methods like classification models. Forecast studies aim to predict events about patients, for instance, epileptic seizures. The wearable sensor types include smart-watches and smart-phones. External sensors could, for example, be cameras or microphones installed in an institution where the participants are patients. Some studies used social media or software as a sensor type, where services like Instagram were used to collect data [8].

Grunerbl et al. [9] did a detection type study about bipolar disorder. The participants consisted of ten bipolar patients in Austria between 18 and 65 years old. In this study, the recorded data was phone calls and microphone data which achieved average recognition accuracy of 76% and precision and recall of over 97% of bipolar state detection. They also used the accelerometer and GPS as input data and achieved recognition accuracy of 70% (accelerometer) and 80% (GPS).

Faurholt-Jepsen et al. [4] presented an association type study about bipolar disorder. The participants were 29 bipolar patients, where the authors collected various actions from the patients smartphones such as the daily usage, the number of incoming calls, and the number of text messages sent and received. They found correlations between the mental state of the patients and the recorded information.

Andrew et al. [17] applied machine learning to photos posted on Instagram. They had 166 participants, who posted a total of 43,950 photos. By extracting statistical features using colors analysis, metadata and face detection, they achieved models that outperformed the average practitioner's success rate when diagnosing depression (70% of all depressed cases identified). Research on social media usage in the field of mental health is interesting because, for many, those are the platforms that they use to express their feelings.

Garcia-Ceja et al. [7] presented their work on motor activity based classification of depression in unipolar/bipolar patients. They applied machine learning for classifying depressed/non-depressed participants using Random Forest and a deep neural networks.

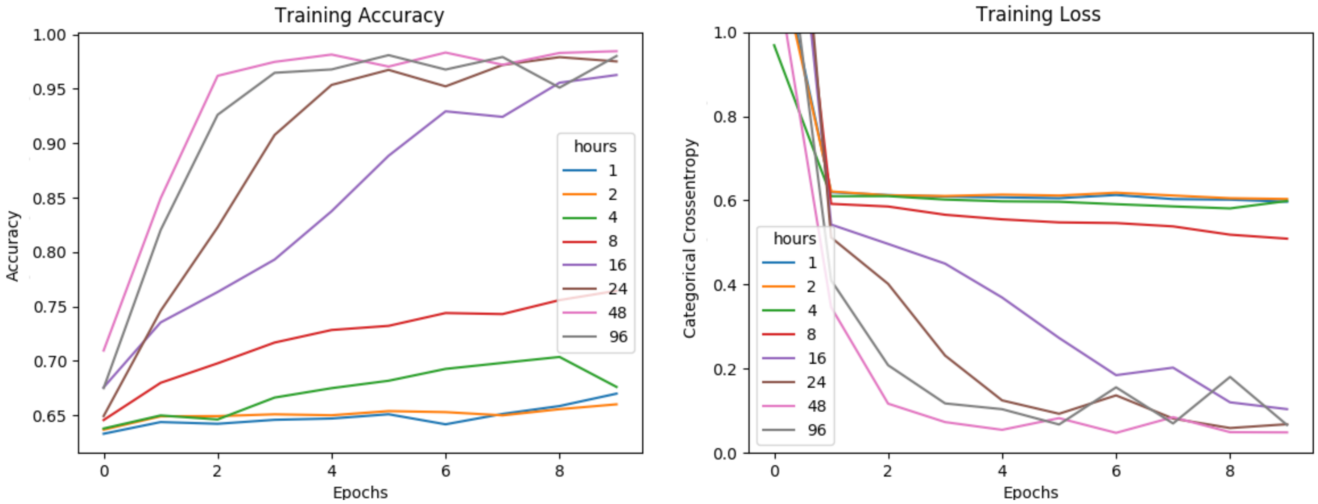
The main contribution between earlier research within MHMS and our work is that we apply CNNs to achieve our goal. Motor activity measurements can be related to mental health issues. However, the best methods for extracting this type of knowledge is not explored yet. With our experiments, we want to determine whether or not CNNs can do this job effectively.

2.2 Depression Rating: MADRS

MADRS is a rating system for telling how depressed a patient is. It is more sensitive to changes than the Hamilton Rating Scale (HRS) for Depression for patients that go through antidepressant medication. The process for calculating a MADRS rating contains ten statements about the patient's behavior, where the topics are: apparent sadness, reported sadness, inner tension, reduced sleep, reduced appetite, concentration difficulties, lassitude, inability to feel, pessimistic thoughts, suicidal thoughts [13].

3 DATASET

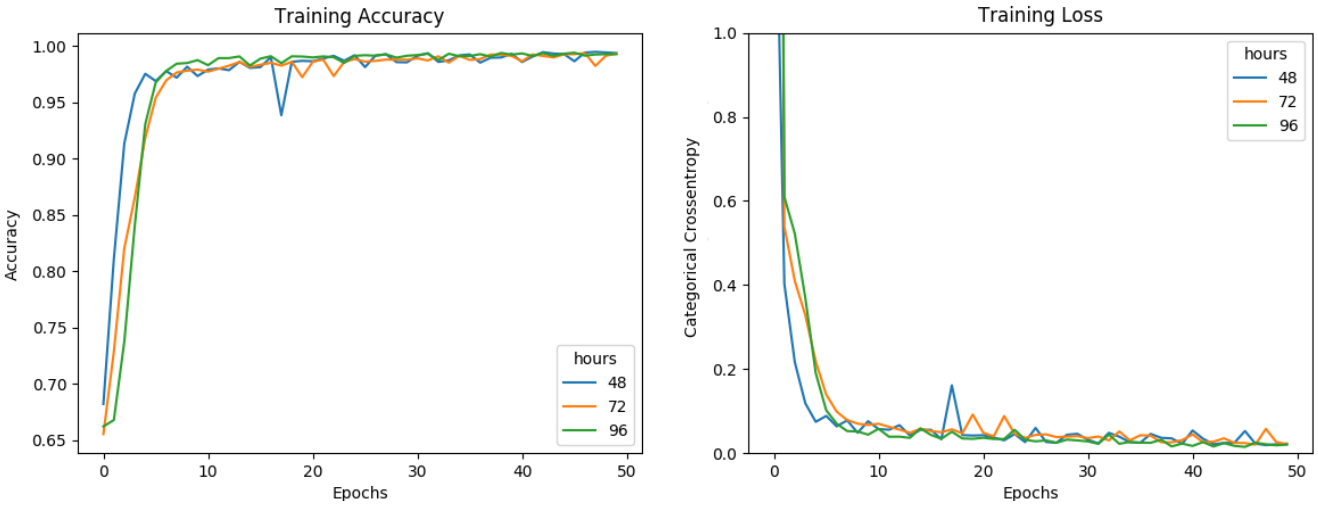
In our experiments, we used the *Depresjon* dataset [6], a publicly available dataset containing motor activity measurements from participants wearing an actigraph watch at their right wrist. The actigraph watch measures activity by using a piezoelectric accelerometer that is programmed to record the integration of intensity, amount and duration of movement in all directions. The sampling frequency was 32Hz. We did not perform any pre-processing of the data. The participants we focus on are 23 bipolar/unipolar patients and 32 non-depressed contributors. We label the bipolar/unipolar group as the *condition group*, and the non-depressed group as the *control group*.



(a) Training accuracy using 10 epochs

(b) Training loss using 10 epochs

Figure 1: Experiment 1. Model's performance using different segment lengths



(a) Training accuracy using 50 epochs

(b) Training loss using 50 epochs

Figure 2: Experiment 1. Model's performance using different segment lengths throughout 50 epochs

4 EXPERIMENTS AND RESULTS

In this section, we explain the performed experiments and discuss the results. Specifically, we conducted three experiments:

(1) Experiment 1: classify whether a participant belongs to the control group or the condition group.
(2) Experiment 2: classify how depressed the participants were, based on their MADRS score.
(3) Experiment 3: predict MADRS score of the participants.

Furthermore, for every model, the optimal segment length was calculated based on different segment lengths and cross-validation was performed. The segment length is what we thought was going to impact the result. Even though we reduced the chance of over-fitting by keeping aside randomized training and testing data, there is a high chance that the training data contains samples from all participants. Garcia-Ceja et al. [7] did *leave one participant out validation*, which means that for each participant in the dataset, keep them outside the training data, train on the rest of the participants, then make predictions on the participant that was left out. Each prediction for the left out participant can be different, so to determine the final label they used majority voting (using the most common predicted label).

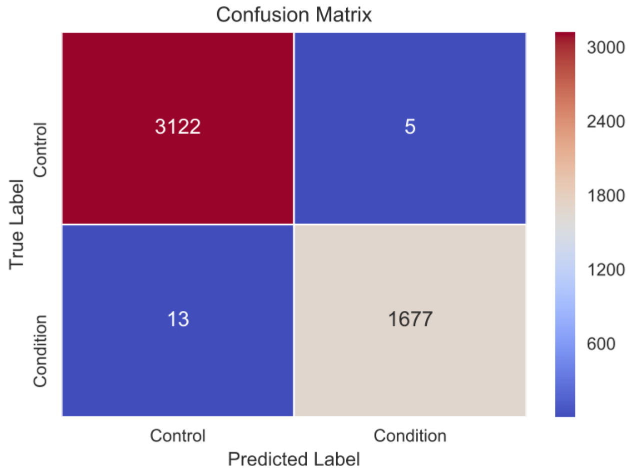


Figure 3: Experiment 1. Confusion matrix for testing the classifier on unseen data

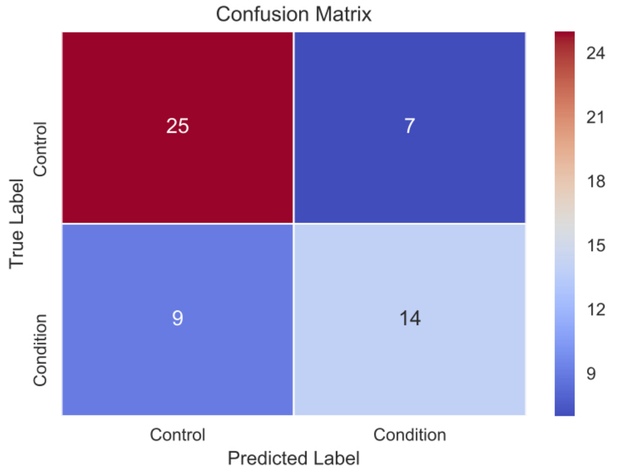


Figure 4: Experiment 1. Confusion matrix containing detected classes after leave one participant out experiment

4.1 Experiment 1: Control vs Condition groups

To find the optimal segment length, we trained the model with segment lengths of 1, 2, 4, 8, 16, 24, 48 and 96 hours. The input data was split into 80% data for training and the rest for testing. 40% of training data was used as validation data. Initially, the training was done using 10 epochs for each of the eight different input sets. We used a batch size of 16 and the Adam optimizer with a learning rate of 0.001 throughout this experiment.

The primary goal here was to find the best segment length to use, and not to train the models to be perfect, so these hyper-parameters were fine for this purpose. As shown in Figure 1, the best results are achieved with segments of 48 hours. Afterwards the results were almost similar. To find the best segment length, we needed to experiment with more epochs. We reran the same experiment for 50 epochs, with 48, 72 and 96 hour long segments. Herein, we see that nothing more was achieved with segments longer than 48 hours, as shown in Figure 2. Figure 3 shows the confusion matrix of control and condition patients.

Table 1: Performance metrics for leave one participant out experiment for control vs condition

Label	Accuracy	Precision	Recall	Specificity	F1
Control	0.71	0.74	0.78	0.61	0.76
Condition	0.71	0.67	0.61	0.78	0.64
Mean	**0.71**	**0.71**	**0.70**	**0.70**	**0.70**

Table 2: Loss and Accuracy of 3-Fold cross-validation for control vs condition

Fold	Loss	Accuracy
1	0.06	0.98
2	0.07	0.98
3	0.06	0.98
Mean	**0.063**	**0.98**

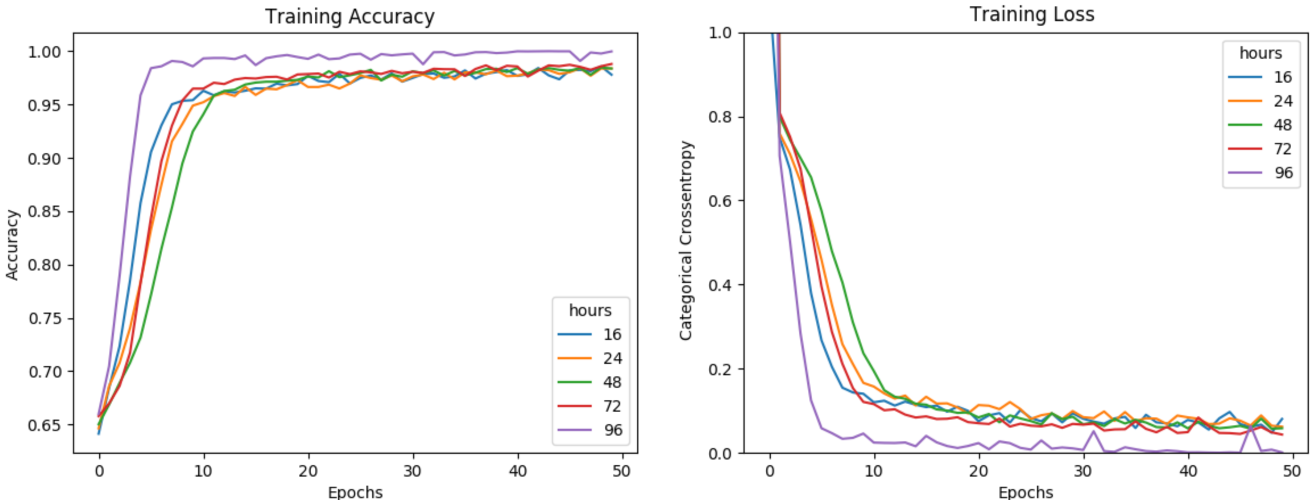
To ensure the model is not over-fitting, 3-fold cross-validation was performed. First, we split the dataset into a train and test set. Then, we generated three folds containing training and validation parts, where for each fold a model was trained to fit the inputs. Each epoch the model was validated against the validation split. After training a model for a fold, we evaluated them by looking at the mean accuracy/loss against the global test split. If the accuracy was still high and the loss was still low, the model would have a good chance of doing correct classifications on unseen data.

To make this process time efficient, each fold was trained for only ten epochs. The goal was to prove consistency in the model and not achieve high performance. As one can see in the cross-validation results (Table 2), we have a mean loss of 0.06 and a mean accuracy of 0.98, which means that the model is consistently correct in most classifications.

In leave one participant out experiment, for each participant, we generated input data that did not contain any activity data from the participant. Earlier results were promising, so we expected the results of this experiment to be better, which we can see in Figure 4. The model was able to detect true negatives (where the correct label and the predicted label is *control*), but the number of false positives and false negatives were a bit too high. Overall, we calculated a mean F1-score of 0.70 for this model (see Table 1). We could train the models for more than ten epochs and hope for better results, but we assume it would not be much difference in training accuracy and loss (as shown in Figure 2).

4.2 Experiment 2: Depression Levels

The second experiment was based on classification of depression levels. We labeled participants with MADRS score 0 to 10 as not depressed, between 11 and 19 as mildly depressed, and above 20 as moderately depressed. The steps were similar like before. Initially, we find the optimal segment length, then use the best segment length in cross-validation and make sure that the performance is consistent. We skipped the shortest segments of 1, 2, 4 and 8 hours,



(a) Training accuracy using 50 epochs

(b) Training loss using 50 epochs

Figure 5: Experiment 2. Model's performance using different segment lengths throughout 50 epochs

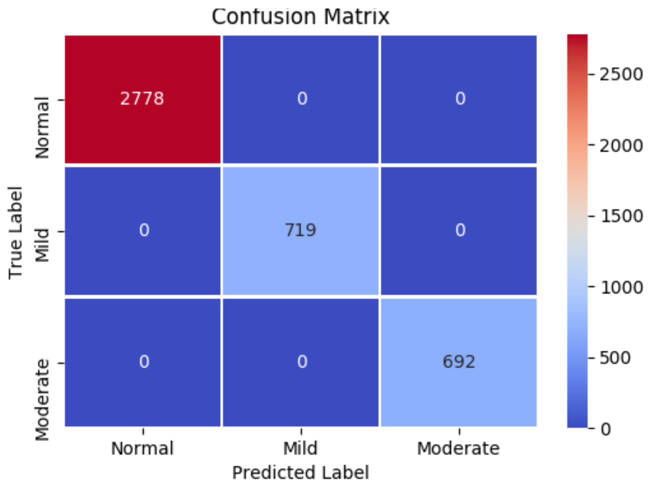


Figure 6: Experiment 2. Trained model on 96 hour long segments for classifying the degree of depression

Table 3: Performance metrics for the leave one participant out experiment for different depression levels

Label	Accuracy	Precision	Recall	Specificity	F1
Normal	0.69	0.66	0.97	0.30	0.79
Mild	0.70	0.14	0.09	0.86	0.11
Moderate	0.76	0.0	0.0	0.98	0.00
Mean	**0.72**	**0.30**	**0.35**	**0.71**	**0.30**

as we were positive these segments would not be any good. We proceeded to train segment lengths of 16, 24, 48, 72 and 96 hours. Training accuracy and loss curves in Figure 5 show that the results for 96-hour segments were promising. We achieved an accuracy of 100% on the testing set. Figure 6 presents the confusion matrix with no error in classification.

Table 4: Loss and Accuracy of 3-Fold cross-validation for different depression levels

Fold	Loss	Accuracy
1	0.04	0.989
2	0.01	0.998
3	0.05	0.985
Mean	**0.033**	**0.991**

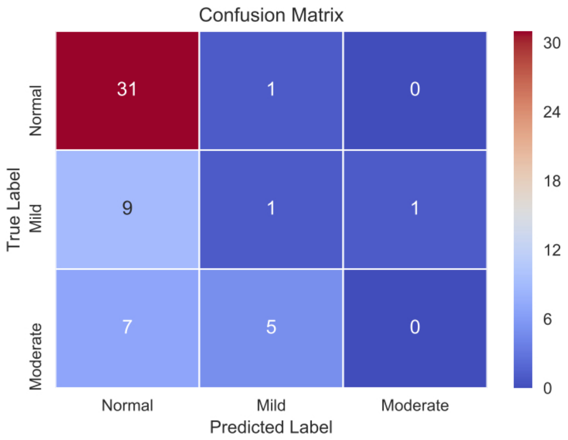


Figure 7: Experiment 2. Confusion matrix of leave one participant out for classifying the degree of depression

For cross-validation, we trained the three models for 15 epochs each. Table 4 lists the accuracies of all 3 folds. For leave one participant out the model was trained on all participants except one.

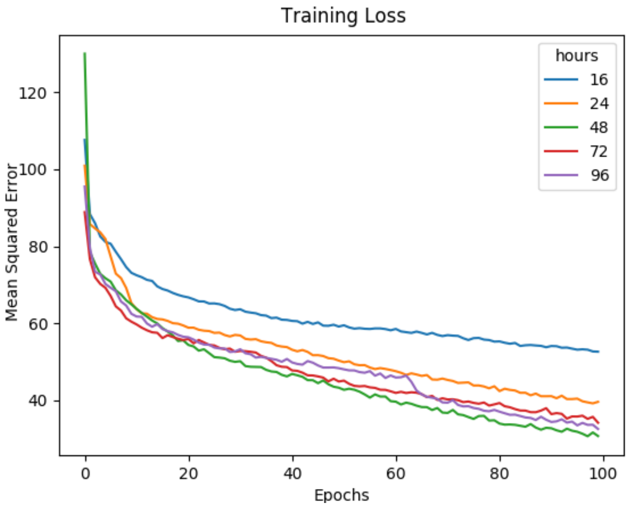


Figure 8: Experiment 3. Training the MADRS prediction model for 100 epochs with different segment lengths. The model trained on 48-hour segments performed best.

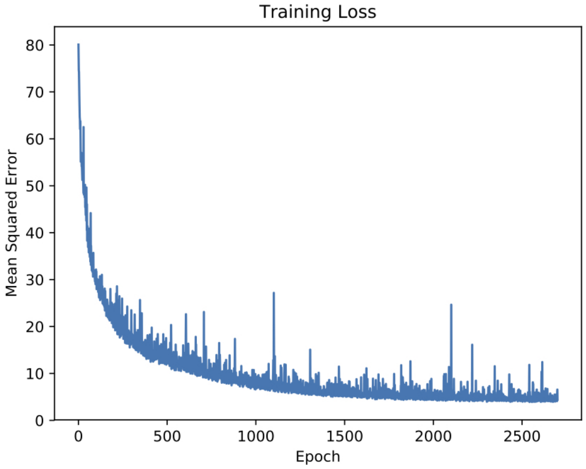


Figure 9: Experiment 3. MADRS Prediction: Training history throughout 2700 epochs. The MSE is approximately 4.0 after 2000 epochs.

Figure 7 illustrates well at detecting non-depressed participants (F1-score of 0.79), and terrible at everything else (F1-score of 0.11 for mild depression and the model did not detect any participant with moderate depression). Overall, we calculated a mean F1-score of 0.30 for this model (see Table 3).

4.3 Experiment 3: MADRS Score Prediction

For MADRS score prediction, the optimal segment length was 48-hours (see Figure 8). Before training the model, we did 3-fold cross-validation to check its consistency. For each fold, the model was trained for 100 epochs to fit the corresponding training data and validated on the corresponding validation data. Then, after a model had completed its training, it was evaluated against the global test data (same procedure as the two previous experiments). Finally, the Mean Squared Error for each fold was saved and compared with other folds to see how they averaged (see Table 5). Afterwards, the model was trained over night (2700 epochs). To summarize, this model was trained to fit segments of length 48 hours (2880 minutes). Hyper-parameters were similar as for previous models. We split the dataset into 60% training data and 40% testing data. Based on the time it took to train each of the 100-epoch experiments, we calculated that around 2700 epochs of training would be a realistic amount.

The training resulted in a mean squared error approximately at 4.0 (on validation data). The training graph (Figure 9) shows that further training would not necessarily give any better results. It can be consider as a baseline for comparison. The baseline would be the average score from the training set without taking any input features for consideration. Predictions on the test data (Figure 10) looked very promising. The graph shows correct MADRS scores in the x-axis and the predicted MADRS scores in the y-axis. Each blue dot is a prediction, and the dotted black line is a linear guideline for the perfect predictions (where the predicted and correct scores are the same).

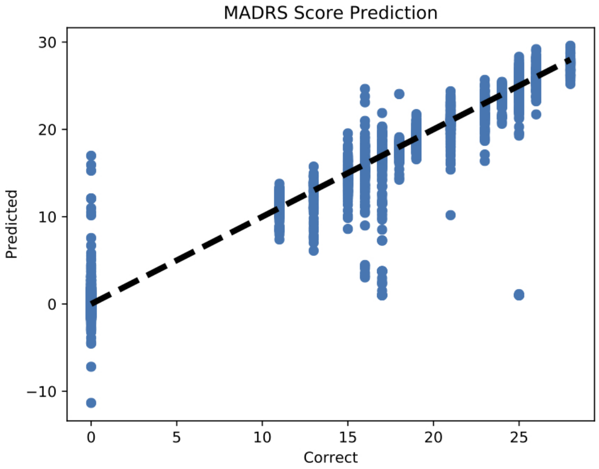


Figure 10: Running the MADRS prediction model on unseen segments. The predictions are not perfect, but they somewhat follow the line where predictions and correct MADRS scores are the same.

Table 5: 3-Fold cross-validation for the prediction model. Only small variation between the folds tells us that the model is consistent enough.

Fold	Mean Squared Error
1	30.18
2	33.61
3	30.40
Mean	**31.40**

4.4 Discussion

Overall, we achieved an F1-score of 0.70 for classifying the control and condition groups, which is slightly better than the F1-scores from the research of Garcia-Ceja et al. [7] without oversampling. They achieved 0.66 for the deep neural network and 0.67 for the random forest. When using Synthetic Minority Over-sampling Technique (SMOTE) as a technique for generating more data, they increased their random forest F1-score to 0.73. An extension of this work could be to attempt to use the same sampling strategies as Garcia-Ceja et al. did on the data passed into our CNN.

Garcia-Ceja et al. also suggested future research to explore classification based on the MADRS scale, which we executed. The results were similar to our other findings; overall performance is not exceptionally high, but we are able to classify most non-depressed participants correctly from the predicted scores.

5 CONCLUSION

In this paper, we have presented a 1D-CNN to detect depression using activity measurements. The used dataset contains motor activity measurements for each minute in the measured period for each participant. Three machine learning models are trained to fit time-sliced segments of these measurements. The first model classifies participants into a condition group (depressed) and a control group (non-depressed). We trained another model to classify the depression level of participants (*normal*, *mild* or *moderate*). Finally, we trained a model that predicts MADRS scores. We evaluate the performance of the classification models using leave-one-participant-out validation as a technique, in which we achieved an average F1-score of 0.70 for detecting the control and condition groups, and 0.30 for detecting the depression levels. The MADRS score prediction resulted in a mean squared error of approximately 4.0.

The current results indicate that depression detection using very easy to obtain activity data is definitely possible within a clinical setting. To make the automatic depression detection possible in a real world scenario, we would need to collect more data also from end user devices, for example, Apple or Fitbit smart watches. For future work, we plan to first collect such a dataset, and secondly, we will develop models that can be applied to the general population outside of the clinical setting which will have a much greater impact and benefit for society as a whole.

6 ACKNOWLEDGMENT

This publication is partially supported by The Research Council of Norway as a part of the INTROducing Mental health through Adaptive Technology (INTROMAT) project under grant agreement 259293.

REFERENCES

[1] Atiyeh Bayani, Fatemeh Hadaeghi, Sajad Jafari, and Greg Murray. 2017. Critical slowing down as an early warning of transitions in episodes of bipolar disorder: A simulation study based on a computational model of circadian activity rhythms. *Chronobiology international* (2017), 235–245.

[2] Jan O Berle, Erik R Hauge, Ketil J Oedegaard, Fred Holsten, and Ole B Fasmer. 2010. Actigraphic registration of motor activity reveals a more structured behavioural pattern in schizophrenia than in major depression. *BMC research notes* (2010), 149–156.

[3] Christopher Burton, Brian McKinstry, Aurora Szentagotai Tătar, Antoni S Blanco, Claudia Pagliari, and Maria Wolters. 2013. Activity monitoring in patients with depression: a systematic review. *Journal of affective disorders* (2013), 21–28.

[4] Maria F Jepsen, Maj Vinberg, Mads Frost, Sune Debel, Ellen M Christensen, Jakob E Bardram, and Lars Vedel Kessing. 2016. Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. *International journal of methods in psychiatric research* (2016), 309–323.

[5] Ole B Fasmer, Hagop S Akiskal, John R Kelsoe, and Ketil J Oedegaard. 2009. Clinical and pathophysiological relations between migraine and mood disorders. *Current Psychiatry Reviews* (2009), 93–109.

[6] Enrique G Ceja, Michael Riegler, Petter Jakobsen, Jim Tørresen, Tine Nordgreen, Ketil J Oedegaard, and Ole Bernt Fasmer. 2018. Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients. In *Proceedings of the 9th ACM Multimedia Systems Conference*. ACM, 472–477.

[7] Enrique G Ceja, Michael Riegler, Petter Jakobsen, Jim Torresen, Tine Nordgreen, Ketil J Oedegaard, and Ole Bernt Fasmer. 2018. Motor activity based classification of depression in unipolar and bipolar patients. In *Proceedings of the 31st International Symposium on Computer-Based Medical Systems*. IEEE, 316–321.

[8] Enrique G Ceja, Michael Riegler, Tine Nordgreen, Petter Jakobsen, Ketil J Oedegaard, and Jim Tørresen. 2018. Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing* (2018), 1–26.

[9] Agnes Grünerbl, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Oehler, Gerhard Tröster, Oscar Mayora, Christian Haring, and Paul Lukowicz. 2014. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE Journal of Biomedical and Health Informatics* (2014), 140–148.

[10] Vishwesha Guttal and Ciriyam Jayaprakash. 2008. Changing skewness: an early warning signal of regime shifts in ecosystems. *Ecology letters* (2008), 450–460.

[11] Lee B Leng, Lee B Giin, and Wan-Young Chung. 2015. Wearable driver drowsiness detection system based on biomedical and motion sensors. In *IEEE SENSORS*. IEEE, 1–4.

[12] Ronald W Manderscheid, Carol D Ryff, Elsie J Freeman, Lela M Eily, Satvinder Dhingra, and Tara W Strine. 2009. Evolving definitions of mental illness and wellness. *Preventing chronic disease* (2009), 19–19.

[13] Stuart A Montgomery and MARIE Åsberg. 1979. A new depression scale designed to be sensitive to change. *The British journal of psychiatry* (1979), 382–389.

[14] Oscar M Mozos, Virginia Sandulescu, Sally Andrews, David Ellis, Nicola Bellotto, Radu Dobrescu, and Jose Manuel Ferrandez. 2017. Stress detection using wearable physiological and sociometric sensors. *International journal of neural systems* (2017), 1–17.

[15] Frank J Penedo and Jason R Dahn. 2005. Exercise and well-being: a review of mental and physical health benefits associated with physical activity. *Current opinion in psychiatry* (2005), 189–193.

[16] Nadja Razavi, Helge Horn, Philipp Koschorke, Simone Hügli, Oliver Höfle, Thomas Müller, Werner Strik, and Sebastian Walther. 2011. Measuring motor activity in major depression: the association between the Hamilton Depression Rating Scale and actigraphy. *Psychiatry research* (2011), 212–216.

[17] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* (2017), 6–15.

[18] Marten Scheffer, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H Van Nes, Max Rietkerk, and George Sugihara. 2009. Early-warning signals for critical transitions. *Nature* (2009), 53–59.

[19] Jean M Twenge. 2015. Time period and birth cohort differences in depressive symptoms in the U.S., 1982–2013. *Social Indicators Research* (2015), 437–454.