

Physical activity based classification of serious mental illness group participants in the UK Biobank using ensemble dense neural networks

Tahmina Zebin, Niels Peek and Alexander J. Casson, *Senior Member, IEEE*

Abstract—Serious Mental Illnesses (SMIs) including schizophrenia and bipolar disorder are long term conditions which place major burdens on health and social care services. Locomotor activity is altered in many cases of SMI, and so in the long term wearable activity trackers could potentially aid in the early detection of SMI relapse, allowing early and targeted intervention. To move towards this goal, in this paper we use accelerometer activity tracking data collected from the UK Biobank to classify people as being either in a self-reported SMI group or an age and gender matched control group. Using an ensemble dense neural network algorithm we exploited hourly and average derived features from the wearable activity data and the created model obtained an accuracy of 91.3%.

I. INTRODUCTION

Serious Mental Illnesses (SMIs) including schizophrenia and bipolar disorder are long term conditions which place major burdens on health and social care services, often coupled with significant reduction in quality of life for the individual over a long term. Up to 40% of those discharged after hospital treatment for schizophrenia may relapse within 1 year [1].

To help reduce this rate there has long been an interest in creating smartphone based mHealth tools for helping with the management of SMIs. An example is *ClinTouch*, an Ecological Momentary Assessment (EMA) smartphone app [2], [3] which prompts the user to answer a structured set of questions multiple times per day about their thoughts and feelings. Response histories are available within the app and are also uploaded to a central server for remote monitoring by healthcare professionals, allowing potential changes in behaviour to be detected, and care planned and targeted as a result.

Today, non-invasive body-worn sensors such as accelerometers and gyroscopes can also help to measure a wide range of behavioral factors to inform this process. Indeed, impaired motor function is often found during episodes of SMI [4]. For example, people with bipolar disorder or schizophrenia can be significantly more sedentary than age- and gender-matched healthy controls [5]. Further examples of the use of such *passive sensing* for behavioural monitoring are

This work was supported by the UK Engineering and Physical Sciences Research Council grant number EP/P010148/1.

This work was conducted at the University of Manchester and Tahmina Zebin is currently working with the School of Computing Sciences, The University of East Anglia, Norwich, UK. Email: t.zebin@uea.ac.uk.

A. J. Casson is with the School of Electrical and Electronic Engineering, and N. Peek is with the Health e-Research Centre, The University of Manchester, Manchester, UK. Email: {alex.casson, niels.peek}@manchester.ac.uk.

presented in [4], [5] which indicate that passively collected behavioral data, using wearables, presents a potentially scalable and at present underutilized opportunity to help with the care of people with SMI.

However, long term wearable monitoring has only recently become available and prospective controlled trials of wearables with people with SMI can be very time consuming to perform, as participants go through different periods of relapse and remission, and the wanted state may not be present during the monitoring period. In this paper we investigate making use of activity tracking (accelerometer) data from the UK Biobank [6], giving a large cross-sectional population to investigate. Within this large dataset, 63 participants self-identified as having an SMI, and we investigate whether machine learning can be applied to the activity data from this population to classify the individuals separately from a matched control population also selected from the UK Biobank. Our classification scheme is based upon an ensemble of dense neural networks to obtain better performance compared to using a single dense neural network.

The rest of the paper is organized as follows. Section II introduces the UK Biobank data selection procedure for the classification scenario of this paper. Section III presents our proposed ensemble dense neural network learning architecture, with results presented and discussed in Section IV. Finally conclusions are drawn in Section V.

II. UK BIOBANK DATA SELECTION

A. UK Biobank overview

The UK Biobank is a very large scale open access database of a wide range genomic and bio-signal records, coming from more than 500,000 people for some data fields. The public showcase of the data available is online at <http://biobank.ctsu.ox.ac.uk/showcase/label.cgi>. In this paper we make use of the physical activity dataset (obtained through UK biobank application number 33693), which was collected from approximately 100,000 people wearing an Activity AX3 3-axis accelerometer device [8] for a week. During this time participants undertook their normal daily activities and the raw data is unlabelled. All data was sampled at 100 Hz, and both the raw data (in units of g) and a number of derived features (as described below) are available.

Also available is a web based questionnaire for these participants, giving information regarding the presence and absence of any mental health condition, including present and past depression, bipolar affective disorder, generalized anxiety disorder, harm behaviour, and subjective wellbeing.

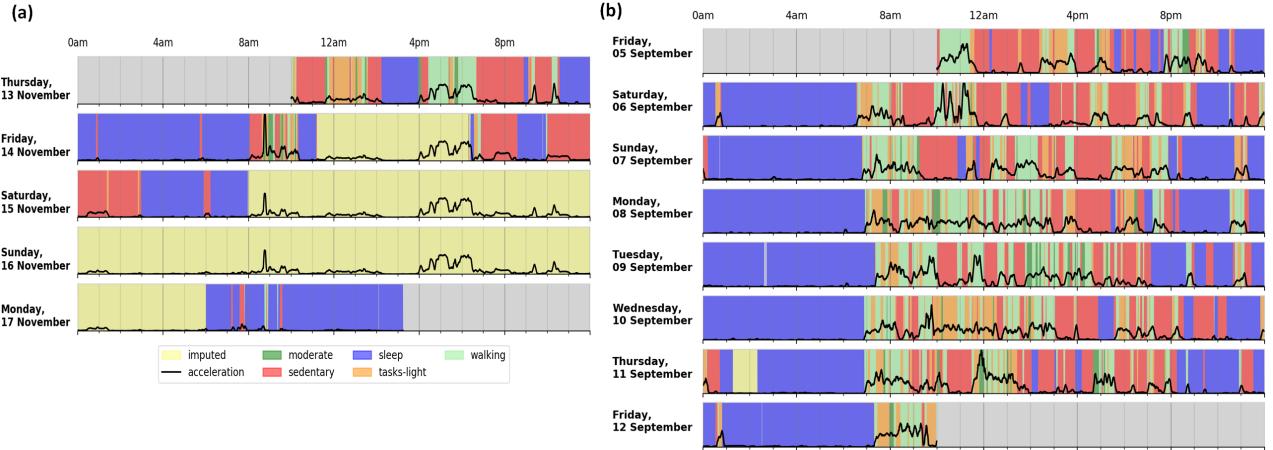


Fig. 1. Illustration of daily activity pattern for one participant. Images generated with toolbox provided by [7]. The colour code identifies moderate activities, sleep, walking and sedentary patterns. (a) Participant from SMI group. (b) Participant from control group.

Features (Total =43)		
Gender and Age(Categorically encoded to five groups) Data fields: 31-0.0; 34-0.0	Weekly Average Acceleration Week Days: Monday, Tuesday, Wednesday, Thursday, Friday Weekend: Saturday, Sunday UK Biobank Data Fields: 90019-0.0 to 90025-0.0	No-wear time bias adjusted average acceleration, standard deviation, median, minimum UK Biobank Data Fields: 900870.0 to 90091-0.0
Overall acceleration Average(milli-gravity). UK Biobank Data Field: 90012-0.0	Daily Average Acceleration 24 hourly Values UK Biobank Data Fields: 90027-0.0 to 90050-0.0	
Standard deviation of acceleration. UK Biobank Data Field: 90013-0.0		

Fig. 2. List of the accelerometer features used in this study and the associated UK Biobank data fields.

The outcomes of these questions are encoded in the Biobank (data field ID 21054) in sixteen fields and allows us to use the raw unlabelled activity dataset in a supervised classification experiment to group participants between SMI and control groups.

B. Accelerometry features

An example of the accelerometry signals available in the dataset, processed using the open source toolbox from [7] to threshold the amount of activity into sedentary, moderate, and similar levels of activity, is shown in Fig. 1. In addition to the raw data a number of derived features based upon averaging are available including: daily acceleration averages (7 measurements); per hour average acceleration (24 measurements); no-wear time (5 different metrics); categorically encoded age and gender (5 different metrics); overall weekly acceleration average and standard deviation (2 features). The list of all of the features we used, and the associated UK Biobank data fields, is given in Fig. 2. As different features have different value ranges we have used a *StandardScaler* normalization applied to the raw data to avoid any potential model bias due to larger values in any individual feature.

C. SMI group and control group identification

As reported in [9], there are 137 participants who self-indicated any form of mental health condition. Here we make use of the responses to questionnaire A3 and A4 (UK biobank data field 20544-0.1 and 20544-0.2) for which 63 participants (21 female, 42 male) were identified as having an SMI, and use this cohort as our SMI group.

For the control group, potentially a very large number of participants could be used (all of the remaining participants) but this is computationally impractical for the current work, and it is likely that some participants with undisclosed or un-diagnosed SMI would be present in such a control group. Moreover, it would give an extremely imbalanced data set for the machine learning analysis. Here we overcome this by using a one-sided Selection with synthetic Minority Over-sampling Technique (SMOTE) technique [10] to resample the data present. We randomly selected an age and gender matched group of 200 participants from the Biobank to form the control group (86 female, 114 male), giving a dataset of a good size, which is also of a practical size to work with for initial machine learning development.

III. ACTIVITY BASED SMI CLASSIFICATION

In this paper, we utilised an ensemble dense neural network architecture to process the hourly accelerometer features available and classify participants as being in either the SMI or control group. We opted for a Deep Neural Network (DNN) architecture due to its capability to process raw time-series data, large error handling capability, and potential for high model accuracy [11].

Our model is implemented using the *Keras* open source library in Python [12], and we have utilized the *sequential* model and the *dense*, *dropout*, *concatenate*, and *batch normalization* layers. At the very first layer, we fed the 43 features from Fig. 2: the thirty-eight derived accelerometer features, one-hot encoded gender (2 channels), and age features (3 channels). These features create a non-linear,

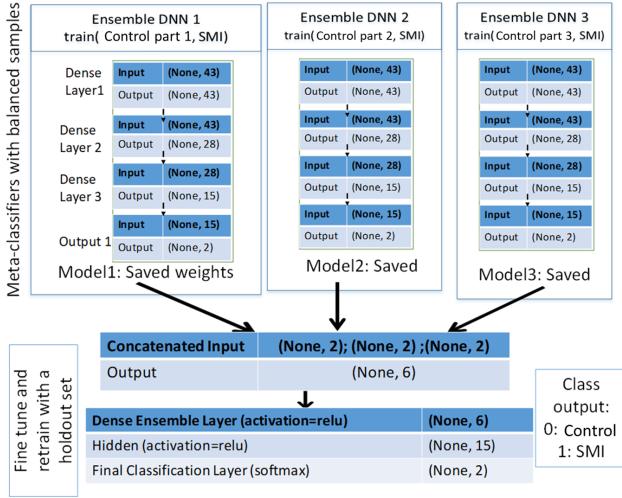


Fig. 3. Visualization of the proposed ensemble Dense Neural Network (DNN) model. The model handles a class imbalance using a synthetic minority oversampling technique.

distributed representation of the input. A visualisation of the proposed network is given in Fig. 3.

Ensemble methods are commonly used for achieving better performance compared to single models. As can be seen from Fig. 3, we started with three baseline models to obtain initial feature processing and classification. In each of these networks there were four dense layers to learn from the features according to their priority. These baseline meta-classifiers were identical in nature for feature exploration purposes, but fed with separate balanced training sets. As the majority (control) group instances (sample size = 200) were almost three times than the minority SMI group (sample size = 63) we partitioned the the majority group into three equal parts to avoid the majority bias on the ensemble layers. We then applied the SMOTE technique on the minority group to further equalize the class imbalance. A balanced test set of 80 participants containing both the groups was kept separately for the purpose of model evaluation.

After the three baseline classifiers a three layer dense neural network is employed (seen at the bottom of Fig. 3), trained by using the concatenated output of the three ensemble DNNs as inputs. The output layer uses a softmax activation, and all layers are fine-tuned using a holdout dataset. The final model was trained with an Adam optimizer with a learning rate of 0.002 and a binary cross-entropy. The training of the model has been done offline on a PC with 2.4 GHz Intel core i7 processor and 32 GB memory.

To act as comparison cases, in Section IV we report the performance of both the ensemble DNN, and a single (non-ensemble) DNN which is identical in set up to the above and trained on the unbalanced training data set. We also report the performance of a standard random forest classifier applied to the data, implemented from the python *sklearn* library.

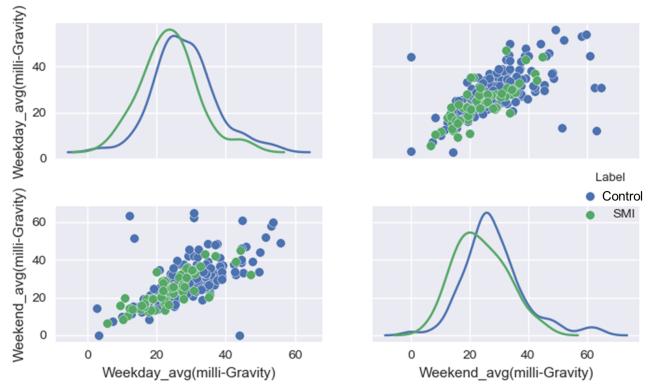


Fig. 4. Comparative pair-plot for weekday and weekend average acceleration for the control (blue) and SMI (green) participants.

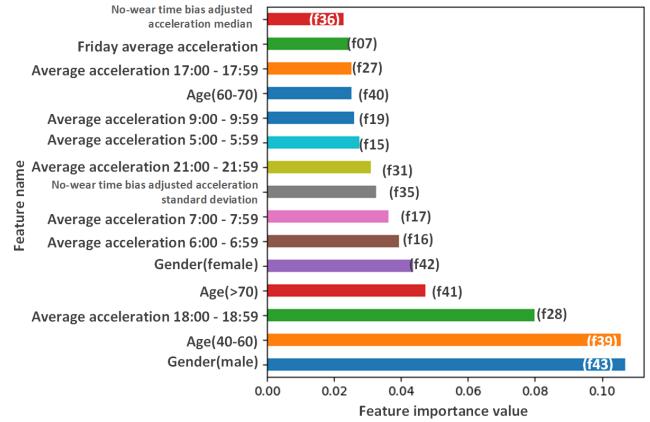


Fig. 5. Top fifteen discriminatory features for this classification task (presented in an ascending order). #f indicates the feature number according to our internal numbering scheme.

IV. RESULTS AND DISCUSSION

To illustrate the two groupings, we present a comparative pair-plot for weekday and weekend average acceleration for the control and SMI group participants in Fig. 4. The control group is seen to have greater average daily acceleration than the SMI group. We have also provided an estimate of feature importance for the classification task in hand in Fig. 5. This has identified age, gender, and some of the hourly average acceleration features during the morning, evening and night hours to be highly important in separating the two groups.

Confusion matrices for the single DNN model and ensemble DNN model are given in Fig. 6. It can be seen that the ensemble DNN substantially outperforms the single DNN with an overall accuracy of 91.3% compared to 83.8%. The number of records correctly classified into the control state is in fact the same between the two DNN set ups, with the ensemble DNN getting better performance due to more participants being correctly classified into the SMI class.

A comparison to the performance of a random forest classifier is given in Table I. The random forest model outperformed the single DNN in overall performance, but not the ensemble DNN. In addition, the random forest had much

TABLE I
QUANTITATIVE COMPARISON OF ENSEMBLE DNN WITH TRADITIONAL SINGLE DNN AND A RANDOM FOREST CLASSIFIER.

Classifier	Overall accuracy	Recall (SMI class)	Recall (Control class)	Precision (SMI class)	Precision (Control class)
Random forest	85.2%	67.1%	87.0%	64.2%	83.0%
Single DNN	83.8%	85.3%	82.4%	84.4%	78.8%
Ensemble DNN	91.3%	89.7%	92.7%	92.1%	90.5%

		True class		
		Control	SMI	Precision
Predicted class	Control	38 47.5%	5 6.3%	88.4% 11.6%
	SMI	8 10.0%	29 36.3%	78.4% 21.6%
	Recall	82.6% 17.4%	85.3% 14.7%	83.8% 16.2%

		True class		
		Control	SMI	Precision
Predicted class	Control	38 47.5%	4 5.0%	90.5% 9.5%
	SMI	3 3.8%	35 43.8%	92.1% 7.9%
	Recall	92.7% 7.3%	89.7% 10.3%	91.3% 8.8%

Fig. 6. Class-wise confusion matrix of the test dataset for the (a) non-ensemble and (b) ensemble DNN.

lower recall in the SMI class than either of the DNN based approaches. In terms of model execution time, the ensemble-DNN model took longer (10.9 seconds) in the training phase than the single DNN (3.41 seconds) and random forest (9.66 seconds).

These high classification performances show that it is potentially possible to differentiate between the two classes based upon a week of activity data. This acts as a starting point to show that there is separability between the two classes, although we note that our control group was only matched for age and gender. Differences in socio-economic status and disease co-morbidity may be present which could be the underlying source of the separability reported here and should be taken as a limitation of the current work. Future investigations will investigate whether the accelerometry data sampled at different points in time could be used to give an indicator of potential relapse or remission events, with the ultimate aim of using wearables to help with earlier interventions.

V. CONCLUSIONS

In this work we applied a data-driven ensemble dense neural network technique aimed to separately classify a serious mental illness participant group from a control group using their physical activity data available from the UK Biobank. The results show that the two classes were potentially separable based upon accelerometer activity recorded over a week. Using an ensemble classifier approach substantially improved the performance compared to using a single dense neural network.

REFERENCES

- [1] I. Barnett, J. Torous, P. Staples, *et al.*, “Relapse prediction in schizophrenia through digital phenotyping: A pilot study,” *Neuropsychopharmacology*, vol. 43, no. 8, pp. 1660–1666, 2018.
- [2] S. Shiffman, A. A. Stone, and M. R. Hufford, “Ecological momentary assessment,” *Annu. Rev. Clin. Psychol.*, vol. 4, no. 1, pp. 1–32, 2008.
- [3] P. Whelan, M. Machin, S. Lewis, *et al.*, “Mobile early detection and connected intervention to coproduce better care in severe mental illness,” *Stud. Health Technol. Inform.*, vol. 216, no. 1, pp. 123–126, 2015.
- [4] E. Reinertsen and G. D. Clifford, “A review of physiological and behavioral monitoring with digital sensors for neuropsychiatric illnesses,” *Physiol. Meas.*, vol. 39, no. 5, 05TR01, 2018.
- [5] D. Vancampfort, J. Firth, F. B. Schuch, *et al.*, “Sedentary behavior and physical activity levels in people with schizophrenia, bipolar disorder and major depressive disorder: A global systematic review and meta-analysis,” *World Psychiatry*, vol. 16, no. 3, pp. 308–315, 2017.
- [6] C. Sudlow, J. Gallacher, N. Allen, *et al.*, “UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age,” *PLoS Med.*, vol. 12, no. 3, e1001779, 2015.
- [7] A. Doherty, D. Jackson, N. Hammerla, *et al.*, “Large scale population assessment of physical activity using wrist worn accelerometers: The UK biobank study,” *PLoS One*, vol. 12, no. 2, e0169649, 2017.
- [8] Axivity. (2019). AX3 3-axis logging accelerometer, [Online]. Available: <https://axivity.com/product/ax3/>.
- [9] K. A. S. Davis, J. R. I. Coleman, M. Adams, *et al.*, “Mental health in UK biobank: Development, implementation and results from an online questionnaire completed by 157 366 participants,” *BJPsych Open*, vol. 4, no. 3, pp. 83–90, 2018.
- [10] I. H. Witten, E. Frank, M. A. Hall, *et al.*, *Data Mining: Practical machine learning tools and techniques*. Burlington: Morgan Kaufmann, 2016.
- [11] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, p. 436, 2015.
- [12] F. Chollet. (2013). Keras: The Python Deep Learning library, [Online]. Available: <https://keras.io/>.