Applying Machine Learning to Bipolar Disorder Categorization Using the Motor Activity Dataset

Rabab Azeem
Queen's University
azeem.rabab@gmail.com

Julien Pierre Chanel
Queen's University
19jpc@queensu.ca

Jingjing Mao
Queen's University
18jm69@queensu.ca

Aria Maz
Queen's University
aria.maz@queensu.ca

Naser AI-Obeidat
Queen's University
21nao3@queensu.ca

***Abstract*—Bipolar disorder, especially bipolar II disorder, is known to have a high suicide and self-harm rate, and a high misdiagnosis rate. This project is an attempt to identify patients with bipolar II disorder among a group of healthy controls and patients with major depressive disorder. We use the motor activity data collected by the motion-sensitive sensor at Haukeland University Hospital, Bergen, Norway. We discuss the preprocessing methods performed and a potential method of feature extraction. We applied a convolutional neural network and a long short-term memory network and the accuracy was 63.48% and 75.77% respectively. Finally, we discussed future directions and suggested methods of improving the accuracy.**

I. Introduction

Beyond the surface of one of the most prevalent mental illnesses lies a harsh reality. Over the past 2 decades, bipolar disorder has resulted in self-harm among 30-40% of patients and led to the suicide of 6% [7]. There are 3 main categories of bipolar disorder, although sometimes their types can blend [5]:

- Bipolar I disorder: is defined by manic episodes that last for at least 7 days (nearly every day for most of the day) or by manic symptoms that are so severe that the person needs immediate medical care [5].
- Bipolar II disorder: is defined by a pattern of depressive episodes and hypomanic episodes. The hypomanic episodes are less severe than the manic episodes in bipolar I disorder [5].
- Cyclothymia: (also called cyclothymia) is defined by recurring hypomanic and depressive symptoms that are not intense enough or do not last long enough to qualify as hypomanic or depressive episodes [5].

Despite its significant consequences and risk, bipolar disorder is commonly known as being hard to be diagnosed. Specifically, Bipolar type II, which has a less severe manic episode, can be mistaken as unipolar depression. Compared to depression and anxiety, the literature about applying machine learning to bipolar disorder diagnosis is rather limited. A literature review paper [4] of bipolar disorder diagnosis and machine learning suggests that bipolar disorder detection is a rapidly improving area of research, even though before 2021 there were only 2 publications from Canada that talked about machine learning and bipolar disorder diagnosis. Among all machine learning techniques applied to bipolar disorder research, the most used were classification models and regression models. Due to the effect of bipolar disorder on daily activities, previous work has been done on applying machine learning algorithms to the motor activity dataset collected from around 50 patients over several days [1]. Further improvements in models and machine learning algorithms were achieved and machine learning has been proven promising in bipolar disorder diagnosis [6] [3]. This project aimed to address the misdiagnosis of Bipolar type II and major depressive disorder. Therefore, we focused on the classification of healthy controls, bipolar II disorder and major depressive disorder.

II. Methodology

A. Introduce the dataset

The proposed data set is the Depression dataset [1] which consists of motor activity recordings of 23 unipolar and bipolar depressed patients and 32 healthy controls at Haukeland University Hospital, Bergen, Norway [1]. The activity level was recorded with an actigraph watch worn that can detect movements over 0.05g at the patient's wrist. The sampling frequency of the actigraph watch is 32Hz, and the output activity level is saved every minute.

The paper associated with the Depression dataset used machine learning to classify patients into depressed and non-depressed. Consisting of two folders, the dataset contained data for the control and condition groups. Each patient had a designated csv file containing timestamped data (in minutes) of physical activity, date of measurement, and activity measurements collected from their actigraph watch. Additionally, vital patient-specific data were included such as afftype (bipolar II, unipolar depressive, bipolar I), melanch (melancholia, no melancholia), gender, age groups, education levels, marriage status, and MADRS scores at the start and end of the measurement period. MADRS scores are a rating scale widely used and validated tool for measuring the severity of depressive symptoms in individuals [2].

B. Data preprocessing

There are mainly three steps to clean the data. The only demographic features available for healthy controls are gender and age, meaning if we want to include healthy controls in the training dataset, we need to drop all the demographic columns except for age and gender. Secondly, we transformed the raw dataset. For the activity recordings of each patient, we extract the recording of each day, and attach the demographic information of the corresponding patient after the daily recording. Therefore, each row of the output dataset is the 24hr recording of a patient with the patient's demographic info attached. During our preprocessing, we also realized that some healthy controls forgot to wear their actigraph watch after several days of recording. Therefore, we dropped days with too many identical data points.

Our analysis aimed to establish the correlation between physical activity and mental health using both statistical learning and deep learning techniques. To achieve this, we leveraged the vast collection of motor activity signals that were timestamped every minute for a large number of patients. However, due to the sheer volume and complexity of the data, accurate model building necessitated feature selection. Therefore, we plan to apply a feature selection process to identify the most relevant features in the motor activity recordings, which could assist in generating accurate MADRS scores and diagnoses. To simplify the data for machine learning algorithms, we transformed the daily numerical motion data into categorical labels consisting of 0, 1, 2, or 3. The labels were determined based on the average motion within 10-minute intervals. A label of 0 indicates that the averaged motion was 1 standard deviation below the daily average, and the non-zero values of the original motion data within the 10-minute interval were less than 40%. A label of 1 indicates that the averaged motion was 1 standard deviation below the daily average, but the non-zero values of the original motion data within the 10-minute interval were more than 40%. A label of 3 indicates that the averaged motion was 1 standard deviation above the daily average. A label of 2 is used for all other cases. By categorizing the data in this way, we hope to make it easier for machine learning algorithms to interpret the data and identify patterns. Due to time constrain, we didn't use the proposed data transformation method in our training dataset.

C. Models

A Multi-Layer Perceptron (MLP) is a artificial neural network which consists of different layers of nodes that process data, and each node is connected to all nodes in the previous layer. The nodes in the input layer of the MLP are representative of the attributes of the data, and the nodes in the output layer are representative of the predicted output. In between these two layers is the hidden layer, which is responsible for executing activation functions such as the sigmoid, tanh (hyperbolic tangent), and the ReLu (Rectified Linear Unit). Activation functions allow neural networks to establish nonlinear relationships in data, and this is crucial as the vast majority of real world problems cannot be modeled by linear functions. During a process called backpropagation, the weights in the MLP are adjusted using an optimization algorithm called gradient descent to minimize the error found in the predicted output. CNNs use specific layers such as pooling and convolutional layers to extract features from images. Consequently, the number of parameters necessary for training is reduced, leading to a more efficient pattern recognition model. Knowing the power of pattern recognition of CNN, we deduce that CNN may be able to find the difference between depression and bipolar disorder since the activity level of a patient with bipolar disorder tends to have a large fluctuation within a day.

Long short-term memory (LSTM) networks were developed to solve the problem of processing long-term dependencies in sequential data, such as speech or text. Contrary to MLPs, LSTMs use memory gates and cells that enable them to retain or ignore data from earlier time steps in a sequence. We think LSTM is a reasonable choice because the time series nature of the dataset and the time of activity level reflects the patient's circadian rhythm. People with mental health disorders tend to have sleep difficulties and different circadian rhythms from healthy controls. Here are the model architectures of each model:

1) *Convolutional neural network:*

- 1 dimensional convolutional layer
- 1 dimensional convolutional layer
- dropout
- 1 dimensional Pooling using maximum
- flattening
- dense layer
- dense layer

2) *Long short-term memory network:*

- LSTM layer
- dropout
- dense layer
- dense layer

D. Results

The accuracy of CNN was 63.48%, and the accuracy of LSTM was 75.77%. The accuracy of LSTM is higher, meaning the circadian rhythm of the patients contributes to the diagnosis of bipolar disorder, and future modelling should take the time series nature of the dataset into account.

III. Conclusion

In conclusion, the performance of CNN and LSTM without feature extraction and further preprocessing was limited. The lack of reliable datasets is the biggest difficulty in applying machine learning algorithms to mental health disorder detection. The number of data entries of the dataset we used is less than 1k after cleaning, which may limit the model from being exposed to a more diverse dataset and recognizing the more general pattern. The lack of data points can also lead to bias against underrepresented groups in the chosen dataset. Moreover, extracting daily activity has a risk of erasing the long-term pattern of the dataset and reducing the amount of information the model can interpret since mental health disorders can only be diagnosed after days of symptom monitoring. Hence, we suggest future studies consider the long-term effect of mental health disorders. Also, after extracting the daily activity level of each patient, the training and testing data points may contain the activity level data from the same patient, which may cause overfitting. We also suggest using data with heart rate monitoring or other physiological information to assist in the interpretation of activity levels.

In real life, the portion of the population with mental health disorders is relatively small, meaning the technique of dealing with an imbalanced dataset needs to be used, for instance, SMOTE. It is also worth trying to apply the proposed data transformation method to the cleaned dataset. One can use the transformed data with some statistical methods such as logistic regression, classification tree, and random forest. If the transformed dataset erased too much information, one can also combine the transformed dataset with the original dataset. A more in-depth validation of the result is needed. For instance, the model may be a lot less sensitive to a certain type of disorder. When the number of data points is small, we suggest a Leave-One-User-Out validation strategy.

Overall, deep learning algorithms are promising in bipolar disorder diagnosis, but we face the challenge of the lack of datasets and the long-term effect of mental health disorders.

REFERENCES

[1] Enrique Garcia-Ceja, Michael Riegler, Petter Jakobsen, Jim Tørresen, Tine Nordgreen, Ketil J. Oedegaard, and Ole Bernt Fasmer. Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients. In *Proceedings of the 9th ACM Multimedia Systems Conference*, MMSys '18, page 472–477, New York, NY, USA, 2018. Association for Computing Machinery.

[2] Wolfgang Hiller, Gabriele Dichtl, Heidemarie Hecht, Wolfgang Hundt, Werner Mombour, and Detlev von Zerssen. Evaluating the new icd-10 categories of depressive episode and recurrent depressive disorder. *Journal of Affective Disorders*, 31(1):49–60, 1994.

[3] Petter Jakobsen, Enrique Garcia-Ceja, Michael Riegler, Lena Antonsen Stabell, Tine Nordgreen, Jim Torresen, Ole Bernt Fasmer, and Ketil Joachim Oedegaard. Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls. *PLOS ONE*, 15(8), 2020.

[4] Zainab Jan, Noor AI-Ansari, Osama Mousa, Alaa Abd-alrazaq, Arfan Ahmed, Tanvir Alam, and Mowafa Househ. The role of machine learning in diagnosing bipolar disorder: Scoping review. *Journal of Medical Internet Research*, 23(11), 2021.

[5] National Institute of Mental Health. Bipolar disorder. https://www.nimh.nih.gov/health/topics/bipolar-disorder. Accessed: 2023-03-01.

[6] Praveen Manoj Singh and P. S. Sathidevi. Design and implementation of a machine learning-based technique to detect unipolar and bipolar depression using motor activity data. *Lecture Notes in Networks and Systems*, page 99–107, 2021.

[7] wikipedia. Bipolar disorder. https://en.wikipedia.org/wiki/Bipolar_disorder. Accessed: 2023-03-01.