**Decision support system for the differentiation of schizophrenia and mood disorders using multiple deep learning models on wearable devices data**

**Duc-Khanh Nguyen**
Department of Information Management, Yuan Ze University, Taoyuan, Taiwan

**Chien-Lung Chan**
Department of Information Management, Yuan Ze University, Taoyuan, Taiwan; Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan, Taiwan

**Ai-Hsien A Li**
Division of Cardiology, Far Eastern Memorial Hospital, Taipei, Taiwan; Graduate Program in Biomedical Informatics, Yuan Ze University, Taoyuan, Taiwan

**Dinh-Van Phan**
University of Economics, The University of Danang, Danang, Vietnam; Teaching and Research Team for Business Intelligence, University of Economics, The University of Danang, Danang, Vietnam

**Chung-Hsien Lan**
Department of Computer Science, Nanya Institute of Technology, Taoyuan, Taiwan

**Abstract**
In the modern world, with so much inherent stress, mental health disorders (MHDs) are becoming more common in every country around the globe, causing a significant burden on society and patients’ families. MHDs come in many forms with various severities of symptoms and differing periods of suffering, and as a result it is difficult to differentiate between them and simple to confuse...
them with each other. Therefore, we propose a support system that employs deep learning (DL) with wearable device data to provide physicians with an objective reference resource by which to make differential diagnoses and plan treatment. We conducted experiments on open datasets containing activity motion signal data from wearable devices to identify schizophrenia and mood disorders (bipolar and unipolar), the datasets being named Psykose and Depresjon. The results showed that, in both workflow approaches, the proposed framework performed well in comparison with the traditional machine learning (ML) and DL methods. We concluded that applying DL models using activity motion signal data from wearable devices represents a prospective objective support system for MHD differentiation with a good performance.

Keywords
mental health disorder detection, schizophrenia, mood disorders, wearable device, deep learning models

Introduction
With increasingly heavy stress in modern life, mental health disorders (MHDs) are becoming more and more common worldwide, the incidence rising in both developing and developed countries.1–4 MHDs occur in many forms with varying symptom severity and differing periods of suffering. These illnesses are difficult to detect and easily confused with each other. Underestimation of or incorrect treatment for these illnesses could result in difficult consequences. MHDs directly affect the mental and physical health not only of individual patients, but also those close to them, and even society as a whole.

In this study, we conducted experiments on two specific MHDs, schizophrenia and mood disorders (including bipolar and unipolar disorders). People with schizophrenia experience disturbed behavioral, cognitive, and thinking changes that can distort their perception of reality. Schizophrenia usually develops initially in people between the ages of 15 and 25, although it can appear later in life. The prevalence of schizophrenia in the general population is about one percent.5 Mood disorders are another mental disorder described by sleep disturbance, a feeling of emptiness or sadness, general loss of interest and initiative in activities, and anxiety.6 The severity of mood disorders is regulated by several factors, such as occupational function, their seriousness and duration, as well as the number of symptoms. Mood disorders are present in both unipolar and bipolar disorder (in major depressive episodes), which share common symptomatic and functional impairments;7,8 however, mania is not present in unipolar disorder, which is the main difference between unipolar and bipolar disorder. Therefore, we employed the term “Mood disorders” in this study to represent both unipolar and bipolar disorder. In general, patients with mood disorders develop a mood that is either prolonged or overly sad, depressed, and may even lead to suicide.9 These symptoms can severely disrupt a person’s life. Schizophrenia and mood disorders are both serious MHDs with extreme symptoms. Several studies have indicated that there are overlapping symptoms between these two illnesses, thus making it difficult for doctors to identify the disorder and provide the correct treatment.10,11 Therefore, early detection of MHDs, not only based on symptoms, is a very important and urgent issue to avert or decrease deleterious effects on patients in the future.

Alongside the rapid technological development and explosion of information, deep learning (DL) is rising in popularity as a powerful technique for data analysis and decision support, with a better and more reliable performance than ever.12,13 In particular, in the healthcare domain, there are numerous real-life deployed applications of DL, such as using DL with big biomedical data for translating; to predict cancer cells with high accuracy; for early disease diagnosis; to predict health
recovery trajectories, etc. Application of DL is becoming increasingly popular, and this technique is more widely-used than traditional data mining techniques.

In recent years, expenditure on wearable devices has been remarkable owing to their advantages and convenience. In addition to supporting users with daily reminders and notifications, these devices can track human activities such as sleep quality, count steps, calories burnt, etc. Thus, these devices provide a quick health condition overview. Furthermore, wearable devices are common for the remote management and monitoring of patients after leaving hospital. In addition, the devices could improve the quality of treatment and encourage patients to adhere to treatment, helping them to recover earlier as a result. For these reasons, wearable devices are becoming an innovative solution for medical observation, monitoring, and the provision of early medical interventions.

In fact, the application of wearable devices with DL in healthcare is an innovative solution, and many researchers are currently working in this field. This study employed a workflow that deploys DL models for MHD differentiation using activity motion signal data from wearable devices via two approaches. The aim was to build a support system using DL with wearable device data to provide physicians with an objective reference resource to assist in making differential diagnoses and planning treatment for patients. Instead of the long, traditional and costly process of diagnosis, such as interview, follow-up, and CT or MRI scan, by employing DL with wearable device data, physicians could obtain a pre-diagnosis – in other words, MHD early detection. We also generated three research questions. First, how can DL be applied with wearable device data to enable MHD differentiation? Second, what is the best approach for applying DL to detect and classify MHDs? And finally, how do the DL models perform in terms of prediction accuracy and sensitivity? The remainder of this study is presented as follows. In Literature review section, we review relevant studies of the differentiation of MHDs. We also run through several methodologies for applying DL with wearable devices in healthcare support systems. In Materials and methods section, we first introduce our target datasets with explanations, including several informatics charts and descriptive statistics, following which we present the DL concepts employed and the deployed DL model architectures. Experiments and results section shows our experimental process using two approaches and their results. Finally, Results section presents the discussion and conclusion.

Literature review

MHDs come in many different types (e.g. schizophrenia, bipolar disorder, unipolar disorder, dissociation, dissociative disorders, etc.) with serious symptoms and differing periods of suffering. In fact, schizophrenia and mood disorders (including bipolar and unipolar disorders) are both severe MHDs. These illnesses share common symptoms, such as delusions, depressed mood, sleep disturbance, social withdrawal, decreased concentration, and lack of motivation; this poses important challenges in terms of diagnosis and providing the right treatment for each individual patient. Classically, doctor and patient undergo an ask-and-answer conversation about mood, thought and behavior, and patients are required to complete questionnaires. This process is replicated again and again. However, this medical diagnosis process is also considered insufficient, and takes too much time. Furthermore, people believe that a diagnosis of MHD carries a stigma, and that when it is recorded it is not good for their career; thus, some patients can be afraid of seeking medical care and try to bury their MHD, which could be treated from the early stages with a low cost.

In addition, in MHD monitoring and follow-up, wearable devices are widely deployed and are being increasingly and progressively used, hence improving the quality of healthcare support and patient treatment. Numerous studies have been performed related to the use of wearable devices in MHDs. In particular, Naslund et al. proposed a framework using smartphone and Fitbit devices to
track seriously ill psychiatric patients. The participants were very pleased to take part in that experiment. They reported that they were encouraged to be more dynamic in order to reach their own activity goals, and in addition, they mentioned that the devices were useful for personal health status monitoring. Muaremi et al. examined a solution for stress experience evaluation by people using data from wearable chest belts and smartphones. In detail, they collect data derived from the heart rate during sleeping, physical activity, audio, and communication data during the workday to build multinomial logistic regression models. The results showed that by combining all features, the accuracy of predicting three stress levels (high perceived stress, moderate, and low) reached up to 61%. Byrne et al. used wearable devices to manage severe MHDs in the community. That study investigated the use of an inexpensive wearable device to detect physiological signs of stress from deviations in biometrics. The authors concluded that by employing a wearable device, it is possible to enhance treatment by improving early warning sign detection and improving communication between physician and patient. Another study by Cella et al. employed wearable devices in order to disclose the autonomic signature of the severity of schizophrenia. They conducted their experiment in groups of 30 schizophrenia patients and 25 control subjects. The participants were asked to wear a mHealth device that was able to measure autonomic activity and movement during the experimental period. These wearable devices proved acceptable and provided reliable behavior and autonomic activity measurement. The authors stated that the schizophrenia patients had lower levels of functioning and heart rate variability movement in comparison with the control participants. In addition, Hunkin et al. discussed the application of wearable devices and their perceived acceptability with regard to MHD treatment. In that study, they used a questionnaire that assessed perceptions of wearable and non-wearable treatments in current and former mental health help-seekers (N = 427). They reported that the wearable devices brought about strong interest in their use as an alternative to self-help options or as an adjunct to talking therapies. The aforementioned studies all applied wearable devices for patient monitoring, diagnosis and treatment. The physician obtains progressive information from their patient, thus providing a reliable resource for decision-making. However, several issues and challenges remain with regards to the real-life deployment of wearable devices in the healthcare setting due to factors such as stability, sensitivity, privacy, power source, continuous power supply limitation, user acceptability, safety, and even clinical knowledge.

On the other hand, the application of DL in healthcare, especially in MHDs, has attracted increasing attention of late. Kwak et al. presented an easy to understand review of more than 320 studies on applying DL in healthcare. They focused on the fields of medical image processing, genomics, electronic health, sensing, and online communication health from 2014 to 2019. In conclusion, the authors stated that with support from artificial intelligence, especially DL, healthcare informatics may ultimately change human life and open up a new paradigm for disease diagnosis, cancer detection, infectious diseases prediction, outpatient stroke prediction, etc. In terms of applying DL with wearable devices for healthcare support, Phan et al. proposed a DL approach to predict sleep quality using wearable device data. They conducted an experiment for 106 days in a row involving 30 participants (mean age = 20.79). Several DL models were deployed. The results showed that the DL models were able to predict sleep quality based on daytime physical activity with a highest accuracy of 62.2%. By examining a variety of deep ensemble learning approaches, Nguyen et al. ascertained that ensemble methods are advantageous in the healthcare support system to enhance prediction and diagnosis performance, and ensure high reliability from physicians. In terms of using DL for MHD prediction and prevention via wearable devices, Coutts et al. proposed a novel approach for mental and general health prediction using DL with wearable-based heart rate variability data. The experiment was conducted in 652 participants; subjective questionnaires were completed weekly or twice-weekly to evaluate the levels of general health, stress, anxiety, and depression. The experimental results showed that for mental health measures and classification, the proposed models achieved accuracies of 73% and 83% with two- and five-minute HRV data streams, respectively. Another study by Bashivan
et al. employed DL and other machine learning methods to recognize the mental state via a wearable electroencephalogram. In that work, electroencephalogram data were used, supported by machine learning, to differentiate between ‘emotional’ versus ‘logical’. The authors concluded that wearable EEG devices have significant potential in differentiating cognitive states among situations with different contexts.\(^{40}\)

Regarding major depressive disorder (MDD) and schizophrenia detection, Galvan et al. presented a feature selection and feature extraction with genetic algorithms methodology to detect depressive episodes in bipolar and unipolar patients.\(^{41}\) The experimental results revealed that the proposed model with a feature extraction approach reached a value of 0.734 for the area under the curve (AUC). The authors concluded that it is possible to differentiate between depressive states by using the activity signal from a smart-band, providing a real-time preliminary and automated tool for physicians to support the diagnosis of depression. Zancella et al. used feature extraction before a random forest to differentiate between healthy controls and MDD patients.\(^{42}\) The results showed that the proposed method achieved a sensitivity value of 0.867 and a specificity of 0.919. The authors concluded that the motor activity signal can be used to distinguish between healthy subjects and those with MDD. Jakobsen et al.\(^{43}\) presented a SMOTE class balancing technique with a deep neural network in motor activity time series of healthy controls as compared with unipolar and bipolar patients. Their proposed method performed a cut above the rest of the deployed machine learning techniques. Boeker et al.\(^{44}\) proposed hidden Markov model (HMM) parameters to classify healthy controls and schizophrenia patients. The work aimed to classify non-schizophrenic and schizophrenic participants based on the HMM, and the results showed that the features of the HMM were outperforming other models in terms of classifying non-schizophrenic and schizophrenic participants. Nguyen et al.\(^{35}\) presented a deep stacked generalization ensemble learning approach to classifying healthy controls and depressed patients in a study that shared a dataset with the current study. However, the method of processing the dataset likely led to underestimation of the true generalization error. In detail, by using overlapping windows by a step of 1 day, very similar samples were generated, and if a sample was in the training set and a very similar one was in the testing set, it would be almost a seen data sample, and therefore easily recognized.

In conclusion, subsequent to the development of wearable devices and the boom in DL in healthcare, the abovementioned studies demonstrated that the application of DL with monitoring wearable devices in healthcare is promising. Further, DL methods have been shown to outperform traditional machine learning techniques. In addition, the traditional diagnosis methods for MHDs have some limitations, as mentioned above. Thus, using DL with wearable device data to differentiate MHDs is a promising approach, providing greater reliability and more convenience than traditional diagnosis methods. As a result, this can form an objective support system for physicians as a reliable resource for diagnosis and treatment.

**Materials and methods**

**Materials**

Our experiments were conducted on two open datasets, named Psykose and Depresjon.\(^{46,47}\) The authors declared that the schizophrenic state of the Psykose dataset was assessed by Haukeland University Hospital’s medical experts. The Psykose dataset was collected from 32 healthy controls and 22 patients with schizophrenia. The Montgomery-Asberg Depression Rating Scale was used to label the severity of the patients’ depressive state in the Depresjon dataset according to the rating of medical experts. The Depresjon dataset was collected from 32 healthy controls and 23 unipolar and bipolar disorder patients. Both datasets were collected through a wrist-worn actigraphic device (Activwatch, fourth model, Cambridge Neurotechnology Ltd, United Kingdom), which participants were required to wear during the experiment for several days in a row. This device detects the peak amplitude of movement acceleration and later transfer it
into a transient voltage signal proportional to the rate of the acceleration. Activity counts are generated from
the raw digital voltage string being selected for each second.\textsuperscript{48} There was a common method of data
acquisition and data storage. The two mentioned datasets were first introduced in a study by Berle et al.,\textsuperscript{49}
focused on data of “schizophrenia” patients, “unipolar and bipolar disorder” patients, and “healthy control”
participants. The authors mentioned that, in the healthy control participants group, there were 23 hospital
employees, five students, and four patients without serious medical or psychiatric symptoms from primary
care. This group contained 12 men and 20 women, with an average age of 38.2 ± 13.0 years (mean ± std.
dev.), ranging from 21 to 66. They also indicated that there was no healthy control participant with a history
of psychotic or mood symptoms. However, the dataset was not published at that time. In two recent papers,
schizophrenia and unipolar and bipolar disorder were released as two separate datasets, including the
“healthy control” participants and corresponding patients. Therefore, the “healthy controls” across the
studies were the same, without serious medical or psychiatric symptoms. Hence, we separated and re-
ordered them into three classes corresponding to their characteristics, named Schizophrenia, Mood
disorders, and Healthy control.

An overview and descriptive statistics of these classes are presented in Tables 1 and 2. The
collection date and collection frequency differed among the participants. The maximum number of
collected days was 47, and the minimum collected days was 14. Moreover, the recording period
could be incoherent by reason of several subjective factors (e.g. the device needed to be charged, the
participant took off their device when taking a shower or sleeping, etc.). Figure 1 shows the boxplots
and a means counted activity line chart of activity counted during 24 h across the three different
Schizophrenia, Mood disorders, and Healthy control groups.

| Group by   | Dataset group | Min  | Max   | Mean   | Std   |
|------------|---------------|------|-------|--------|-------|
|            | Schizophrenia | 0.0  | 8000.0| 128.4  | 244.7 |
|            | Mood disorders| 0.0  | 8000.0| 163.1  | 320.8 |
|            | Healthy control| 0.0 | 8000.0| 188.5  | 378.8 |
| Hourly     | Schizophrenia | 0.0  | 92,295.0| 7699.0  | 9594.1 |
|            | Mood disorders| 0.0  | 130,377.0| 9765.1  | 13,063.4 |
|            | Healthy control| 0.0 | 240,442.0| 11,298.1  | 15,759.4 |
| Daily      | Schizophrenia | 0.0  | 601,370.0| 176,129.2 | 118,442.9 |
|            | Mood disorders| 0.0  | 841,370.0| 222,161.4 | 155,785.9 |
|            | Healthy control| 0.0 | 901,858.0| 260,146.7 | 217,800.4 |
| Weekly     | Schizophrenia | 22,844.0 | 3,358,302.0| 995,512.6 | 807,123.7 |
|            | Mood disorders| 561.0 | 4,203,291.0| 1,183,886.1 | 973,411.1 |
|            | Healthy control| 10.0 | 4,719,352.0| 1,537,987.3 | 1,405,367.4 |

Std.: Standard deviation.

Methods

In the proposed workflow, we deployed multiple deep learning models, namely, VGG16, Resnet50v2,
XceptionNet, and EfficientNetB1, all of which are classed as a Convolutional Neural Network (CNN), and
Long-short Term Memory (LSTM), Gated recurrent unit (GRU), Attention-based LSTM, and Attention-
based GRU, which all come under the concept of a Recurrent Neural Network (RNN). Recurrent Neural Network models are suitable for time series data extracted from wearable devices, while in contrast, CNN models are commonly used for image-based tasks. However, the results of several studies have shown that the CNN has a good performance in seizing important features of sequential data.50,51 Hence, we present a way in which to preprocess sequential data to fit into these models (see Section 4.1) to take advantage of the CNN. Two approaches were applied with these models. The first approach was to perform the MHD Prediction and Differentiating between Mood Disorders and Schizophrenia tasks separately, while the second approach was the Direct Differentiation of Mood Disorders, Schizophrenia, and Healthy Control. We then compared the performances of the two approaches and the models using the specific evaluation metrics detailed in Section 3.2.3.

Convolutional neural network. The Convolutional Neural Network (CNN) is a concept of neural network applied widely for image-based tasks. The main idea of this concept is a structure constructed by several convolution layers and pooling layers placed interchangeably with each other. A fully connected layer is set in the end to conduct the final prediction.52 The convolution layers extract the features from the right previous layers by using a use sliding windows. The pooling layers are deployed to reduce dimension from the right previous layers. The outputs of both the convolution layers and pooling layers are so-called feature maps. Right after the last convolutional-pooling layer block, the fully connected layer is deployed to flatten the output of this block into a simple neural network structure. At the end of this structure, the output layer yields the processed values that correspond to the prediction or differentiation result (Figure 2).

In this study, four models of the CNN concept were deployed, namely VGG16, Resnet, XceptionNet, and EfficientNet architectures. In detail, VGG16 is an upgraded version of a standard CNN following the arrangement of convolution and max pool layers consistently throughout the whole architecture. In the end, there are two fully connected layers with a Softmax activation function.53 Using a residual learning framework, Resnet is named after its architecture; this model can deal with the vanishing gradient problem seen in previous DL architectures.54 The modified version named Resnet50V2 was deployed in our experiment. The Xception model was designed based on Resnet, and is a linear stack of residual connections with depthwise separable convolution layers.55 EfficientNet is a simple architecture but gains a highly effective compound coefficient by uniformly scaling all dimensions of depth, width, and resolution in a gradually increasing way. Intuitively, whenever the input image is bigger or smaller, EfficientNet is always able to automatically adjust more or fewer layers and channels to increase the receptive field, capture more fine-grained patterns, and optimize the training time.56 A member of the EfficientNet family named EfficientNetB1 was deployed in this experiment.

| Dataset               | # Participants | Female | Age (mean ± std.) | Max of collected days | Min of collected days | Mean of collected days | Median of collected days | Std. of collected days |
|-----------------------|----------------|--------|-------------------|-----------------------|-----------------------|------------------------|-------------------------|-----------------------|
| Schizophrenia         | 22             | 20     | 46.7 ± 10.9       | 47                    | 16                    | 17.7                   | 16.0                    | 2.4                   |
| Mood disorders        | 23             | 13     | 42.8 ± 11.0       | 47                    | 14                    | 23.1                   | 16.0                    | 8.7                   |
| Healthy control       | 32             | 20     | 38.2 ± 13.0       | 47                    | 14                    | 17.6                   | 20.5                    | 4.0                   |

N/A: Not applicable. Std.: Standard deviation.
Recurrent neural network. As another concept of a neural network, a Recurrent Neural Network (RNN) is usually employed for time series data in order to make predictions relying on a sequence of previous information (Figure 3). It is a type of feedforward neural network, the way in which it delivers information along the cells in this network lending the term Recurrent. The RNN has the internal state on the neural cell to process the variable-length sequence of input, and the output of the current state will be the input of the next states of itself.

In this study, we deployed two architectures based on the RNN concept, named Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU), and the other two models were the Attention-based LSTM and Attention-based GRU. LSTM is able to store long-term dependency information owing to four different gates with different activation functions – namely, forget gate, input gate, and output gate - to keep and store or let the information pass to the next state. Specifically, the forget gate decides whether the information, which is combined from the previous state output and current state input, should be thrown away or kept. The input gate is used to add relevant information from the current state. The output gate’s role is determining the
next hidden state. The GRU network is an updated version of LSTM. Instead of three gates as in LSTM, the GRU has only two gates, a reset gate and an update gate. In detail, the reset gate is used to determine how much information from the previous state need to be removed, and the update gate has similar functions to the forget and input gates of LSTM, deciding whether to let the information go through or not and what information needs to be added. Attention is a mechanism that is usually used in RNN architectures for natural language processing tasks that refer to sequence data. Attention’s idea comes from making an encoder-decoder architecture able to have a variable length of internal representation. In order to do that, this mechanism keeps the intermediate outputs from the encoder layer, and after that, the model is trained on attention-selected information. In short, the output information is conditional on the attention-selected item from the input information.\(^\text{59}\)

**Performance metrics.** Commonly, a confusion matrix and corresponding derived metrics are employed to evaluate supervised learning models for classification tasks. The basic components of this matrix are so-called True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP). Specifically, TN describes truly negative samples that are predicted as negative, TP describes truly positive samples that are predicted as positive, FN describes truly positive samples that are predicted as negative, and FP describes truly negative samples that are predicted as positive (Figure 4).

In this study, we evaluated our experimental models in terms of accuracy, precision, sensitivity, F1 score, and the Matthews correlation coefficient (MCC), which are the derived metrics of the confusion matrix (equations (1)–(5)). A description of these metrics is presented in Table 3. The F1 score and MCC were outstanding as compared with the others in terms of evaluating the performance on the imbalanced dataset. The F1 score considers the number of wrong predictions and the types of error that the model makes. The MCC was even better than the F1 score, as proved by Chicco et al. in a study of six synthetic use cases and in a real genomics scenario,\(^\text{60}\) its value ranges from \(-1\) to 1, whereby \(-1\) represents a perfect misclassification and one represents a perfect classification of the model.

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{1}
\]

---

*Nguyen et al.*

![Figure 3. A basic Recurrent Neural Network structure and its unfold state. (h<sub>i</sub> is the neural network cell or layers, x<sub>i</sub> is the input at time step \(i\), y<sub>i</sub> is the output at time step \(i\), \(V_i\) is the output for the next step, \(U\) and \(W\) are the weights of the hidden layers.)*
The confusion matrix and corresponding derived metrics are usually employed for binary-class classification tasks. However, in our experiments, there was a 3-class classification task named “Direct differentiation of Mood disorders, Schizophrenia and Healthy control groups”. For the multi-class classification task, TP described the samples that were predicted as class $\mathcal{C}_i$ and truly belonged to class $\mathcal{C}_i$. TN described the samples that were predicted as non-class $\mathcal{C}_i$ and truly belonged to non-class $\mathcal{C}_i$. FP described the samples that were predicted as class $\mathcal{C}_i$ but actually belonged to non-class $\mathcal{C}_i$. FN described the samples that were predicted as non-class $\mathcal{C}_i$ but actually belonged to class $\mathcal{C}_i$. Furthermore, the distribution among the groups of Healthy control, Schizophrenia, and Mood disorders was approximately 2:1:1. Moreover, schizophrenia and mood disorders detection were more important than healthy control detection. Under these circumstances, we applied the weighted probability on the output layer of every single prediction of the DL models. In detail, the probability of schizophrenia and mood disorders was duplicated corresponding to the distribution ratio among classes. In the decision of prediction, the highest probability would be chosen as the class that the sample belonged to. The corresponding added weight of each class is presented in Table 4. The confusion matrix and its derived metrics were applied as mentioned above.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{2}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}
\]

\[
\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{4}
\]

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}
\]
Experiments and results

Proposed framework

We conducted the experiments in three phases, namely Data preprocessing, Building and training models, and Evaluation models (Figure 5). The details of these phases are presented below.

Phase 1 began with reading and cleaning of the raw data, which was the set of number of accelerometer data points collected in every single minute from the individual groups, Schizophrenia, Mood disorders, and Healthy control. Next, we used 5-fold cross-validation by participant ID; in other words, for each fold in the cross-validation process, 20% of the participants in each group were kept for testing, and the remaining 80% of participants were used for training. We then transformed the data to fit the deployed models. In detail, the experiments were conducted under two different concepts of DL, of which the RNN concept is widely-used for sequential data, which only has one dimension. These datasets were originally time-series data, which are apparently suitable for RNN concept models. However, the CNN concept is widely-used for image-based datasets, which have width, height, and depth dimensions. Hence, we converted these datasets into a 1440xDx1 pseudo image for each sample, where 1440 was the height, D the width and one the depth of the pseudo generated image corresponding to the total minutes in a day, the total number of recorded dates, and the number of activity types. Subsequently, for every participant from each dataset, we applied a 7-days moving forward window, with a 1-day moving step that covered 7 days for each sample from the beginning until the date on which the moving window reached the end of the collection term of the individual participant. In detail, we generated \((D_i - 7)\) samples for each participant, where \(D_i\) is the total days collected from participant \(i\). The moving forward window was designed in a different way depending on the deployed model (Figure 6).

In phase 2, we have a different combination of these three individual datasets for each approach, which is described in section 4.2. Multiple DL models with different approaches were trained, namely Resnet50v2, VGG16, XceptionNet, and EfficientNetB1, which belong to the CNN concept (Tables A1-A4), and LSTM, GRU, Attention-based LSTM, and Attention-based GRU, which belong to the RNN concept (Tables A5-A8).

In phase 3, we evaluated and compared the approaches and models based on the proposed performance metrics. Finally, the well-trained models were saved for further applications. We implemented our experiments on Anaconda v1.10.0, Python 3.7, with free machine learning software and a DL library named Scikit-learn version 0.24.1 and Tensorflow version 2.1.0. We

| Metric | Description |
|--------|-------------|
| Accuracy | The ratio of correct predictions to total predictions made |
| Precision | Metric that quantifies the number of correct positive predictions made |
| Sensitivity | Metric that quantifies the number of correct positive predictions made out of all actual positive samples in the dataset. |
| F1 score | The harmonic mean of sensitivity and precision |
| MCC | The MCC presents the Pearson product-moment correlation coefficient between the actual value of samples and the predicted samples. It is a reliable model performance evaluation tool, even if the classes are imbalanced. |

MCC: Matthews correlation coefficient.
Table 4. Added weight value for each class.

| Class              | Added weight value |
|--------------------|--------------------|
| Schizophrenia      | 2                  |
| Mood disorders     | 2                  |
| Healthy control    | 1                  |

Figure 5. Proposed framework. (Phase 1: Data preprocessing; Phase 2: Building and training models; Phase 3: Evaluation models) MHD: Mental health disorder.

Figure 6. 7-Day moving forward window. A: The 7-Day moving forward window for sequence data with a length of 10,080, where 10,080 is the total minutes in 7 days. B: The 7-Day moving forward window for image-based data, which had a size of 7x1440; seven is the days that this window covers and 1440 is the total minutes for each day.
conducted the whole process on a computer with the following specifications: Intel(R) Core(TM) i9-10900K CPU, GeForce GTX 1080 Ti 11 GB GPU memory, and 32 GB DDR4 RAM.

Experiments

To prove the advantages of DL in differentiating MHDs using wearable device data, we conducted our experiments using two approaches (Figure 7). A description of the data in each approach is presented in Table 5.

In the first approach, we conducted the MHD prediction task and the Differentiating between mood disorders and schizophrenia task separately. In detail, the aim of the first task was to detect patients who have MHDs (no matter what type of disorder) from the participants. To do that, we first combined the Schizophrenia group and Mood disorders group into one, named the Mental health disorder group. The samples belonging to this made-up group were notated as class 1. On the other hand, the samples belonging to the Healthy control group were notated as class 0. The total samples in each group were 525 and 547, respectively. The objective of the second task was to classify the MHD among the patients. In this task, we only conducted the experiment on the Schizophrenia group and Mood disorders group. Specifically, the samples belonging to the Schizophrenia group were notated as class 1, and the samples belonging to the Mood disorders group were notated as class 0. The total samples in each group were 258 and 267, respectively. Hence, both of these two tasks are of binary classification, but with different aims. In fact, a physician could deploy these two tasks separately according to their specific requirements.

On the other hand, in the second approach, we conducted our experiment on all three datasets. Samples belonging to the Healthy control group were notated as class 0, samples belonging to the Mood disorders group were notated as class 1, and samples belonging to the Schizophrenia group were notated as class 2. The total numbers of samples of each class were 547, 267, and 258, respectively. Consequently, the second approach was a multiclass classification task with an imbalanced distribution.

By deploying 24 trained-by-task DL models, much effort was put into tuning and choosing the hyperparameters of the DL models. Essentially, we used the training data for training and tuning the hyperparameters and the test data for evaluating and comparing the optimal learned models. We considered two hyperparameters, namely, the number of epochs and the learning rate. A grid search over the hyperparameters based on the MCC was performed for each architecture. Table A10 presents the performance of each model in terms of deployed learning rate and epoch. The optimization algorithm for all models was fixed with the Adam optimizer. The selected hyper-parameter sets for each model are presented in Table A11. Some tricky techniques were used to improve the models’
learning performance. In detail, the LearningRateScheduler function was applied to scale down the models’ learning rate by 10% after every 40 epochs during the process of training, thereby enhancing the learning performance of the DL models. In addition, the ModelCheckpoint function was applied to save the best weight of the model once validation loss reached the minimum value.

**Results**

In this study, we conducted experiments using eight well-known DL models belonging to the CNN and RNN concepts for every single task in each approach. In total, 24 trained-by-task DL models were deployed in our experiments; we conducted 16 trained-by-task DL models using the first approach and eight using the second approach. The average evaluation metrics on mean were calculated for each concept in order to make a comparison between them for each task.

In approach 1, in the MHD prediction task, the CNN models showed better performances than the RNN models when comparing the average of each architecture on mean values. In detail, the XceptionNet model outperformed the rest, with an accuracy, sensitivity, F-score, and MCC of 0.86, 0.92, 0.87, and 0.73, respectively. The Attention-based GRU was the best candidate among the RNN models, at 0.79, 0.79, 0.78, and 0.58, respectively. The Random forest and ANN models showed inferior performances. On the Differentiating between Mood disorders and Schizophrenia task, the CNN and RNN models performed equally on average. Resnet50v2 was the best model under the CNN concept, with an accuracy, sensitivity, F-score, and MCC of 0.88, 0.93, 0.88, and 0.77, respectively. The LSTM model was the best model based on the RNN concept, with values of 0.88, 0.93, 0.88, and 0.77, respectively. However, LSTM presented as a more stable model as compared with Resnet50v2 on 5-fold cross-validation based on the lower standard deviation (Table 6).

Approach 2 aimed to directly differentiate between the Mood disorders, Schizophrenia, and Healthy control groups with added weight. The models based on the CNN concept showed better performances than the models under the RNN concept, with values almost 4% higher in all
evaluation average metrics. In particular, VGG16 outperformed all the other models, with an accuracy, sensitivity, F-score, and MCC of 0.75, 0.72, 0.71, and 0.60, respectively. Among the RNN models, the Attention-based GRU showed the best performance over other models, with values of 0.72, 0.71, 0.68, and 0.57, respectively (Table 7).

Overall, the models belonging to the CNN concept showed a better performance than the RNN-concept models under both approaches in corresponding tasks. In detail, in the first approach, XceptionNet dominated the others models in the MHD prediction task. Resnet50v2 and LSTM were in first place for the Differentiating between Mood disorders and Schizophrenia task. In the second approach, Resnet50v2 and XceptionNet showed good performances, staying in second place and closely following the best model - VGG16. VGG16 performed well in approach 2, dealing with the 3-class differentiation task. However, in both the MHD prediction and Differentiating between Mood disorders and Schizophrenia tasks in approach 1, VGG16 was overshadowed by other models. The Attention-based LSTM and Attention-based GRU models were the best-performing models based on the RNN concept, but nonetheless, they still could not surpass the CNN models.

**Discussion and conclusion**

In this study, we proposed a workflow and two approaches to apply DL with wearable device data to conduct a MHD prediction task, a Differentiating between Mood disorders and Schizophrenia task, and a Direct Differentiation between Mood disorders, Schizophrenia and Healthy control task. Overall, both proposed approaches using DL models were found to be suitable for the differentiation of MHDs in subjects sporting wearable devices. In approach 1, the first task was to detect patients among the participants, and the second differentiation task was to differentiate between two kinds of MHD among the patients. Both of these are binary-class classification tasks. According to the specific requirements, a physician could deploy these two tasks separately, or in sequence from MHD detection to Schizophrenia and Mood disorders differentiation. On the other hand, in approach 2, Healthy control, Schizophrenia, and Mood disorders differentiation were performed concurrently; hence, this is suitable for the classification of MHDs directly. In addition, these datasets were collected from different participants with distinct characteristics in terms of physical condition and habits. Furthermore, they contained missing data owing to subjective factors (e.g. the device needed to be charged, the participant removed their device when taking a shower or sleeping, etc.). Nevertheless, our proposed workflow and models still performed well. These datasets are time-series data, which basically work well for RNN-concept models; however, in this study, the CNN models showed a better performance than the RNN models in all approaches and their corresponding tasks. By treating the sequence data as images, our results showed that CNN-concept models work well not only on images, but also on sequence datasets.

In this study, bipolar disorder and unipolar disorder were combined into a single mood disorder category. However, these two disorders do have some differences, such as unipolar disorder being less episodic than bipolar disorder, and the treatment of unipolar disorder differs from that of bipolar disorder due to the risk of mania in the latter. Therefore, there is room for improvement in MHD differentiation using DL models to discriminate unipolar and bipolar disorders separately. There are several challenges in deploying wearable devices and using DL to analyze data in the healthcare setting. In particular, the collected data may experience noise and distortion due to the hairs on the skin and the constant body motion, leading to a decrease in adhesion between the wearable device and the skin. Furthermore, battery life is one of the major challenges that need to be considered in maintaining continual monitoring, as well as the general design of the wearable device. In addition, consideration of wearer data is important with regards to healthcare ethics, and the development of a
Table 6. Comparison of DL models in approach 1. Each model is repeated 5 times, corresponding to 5-fold validation. The average results are presented as the mean ± standard deviation across the five experiments, of the best hyper-parameter sets.

| Task                                         | Model      | Accuracy   | Precision | Sensitivity | F-score    | MCC        |
|----------------------------------------------|------------|------------|-----------|-------------|------------|------------|
| MHD prediction                               | VGG16      | 0.81 ± 0.04| 0.87 ± 0.05| 0.77 ± 0.08 | 0.81 ± 0.05| 0.62 ± 0.08|
|                                              | Resnet50v2 | 0.84 ± 0.03| 0.84 ± 0.08| 0.84 ± 0.09 | 0.84 ± 0.06| 0.68 ± 0.07|
|                                              | XceptionNet| 0.86 ± 0.04| 0.84 ± 0.03| 0.92 ± 0.06 | 0.87 ± 0.04| 0.73 ± 0.08|
|                                              | EfficientNetB1 | 0.81 ± 0.04| 0.83 ± 0.07| 0.82 ± 0.04 | 0.80 ± 0.06| 0.63 ± 0.07|
| Average of CNN models on mean                |            | 0.83 ± 0.02| 0.85 ± 0.02| 0.84 ± 0.06 | 0.83 ± 0.03| 0.67 ± 0.05|
| LSTM                                         |            | 0.77 ± 0.06| 0.78 ± 0.10| 0.82 ± 0.18 | 0.78 ± 0.08| 0.54 ± 0.09|
| GRU                                          |            | 0.75 ± 0.10| 0.78 ± 0.17| 0.82 ± 0.23 | 0.76 ± 0.13| 0.55 ± 0.14|
| Attention-based LSTM                         |            | 0.77 ± 0.05| 0.78 ± 0.05| 0.77 ± 0.06 | 0.76 ± 0.05| 0.55 ± 0.10|
| Attention-based GRU                          |            | 0.79 ± 0.06| 0.79 ± 0.05| 0.79 ± 0.06 | 0.78 ± 0.06| 0.58 ± 0.10|
| Average of RNN models on mean                |            | 0.77 ± 0.02| 0.78 ± 0.01| 0.80 ± 0.02 | 0.77 ± 0.01| 0.56 ± 0.02|
| ANN                                          |            | 0.76 ± 0.03| 0.78 ± 0.13| 0.80 ± 0.07 | 0.78 ± 0.04| 0.55 ± 0.05|
| Random forest                                |            | 0.72 ± 0.06| 0.70 ± 0.11| 0.82 ± 0.10 | 0.74 ± 0.04| 0.47 ± 0.08|
| Random guessing                              |            | 0.50 ± 0.00| 0.52 ± 0.02| 0.50 ± 0.00 | 0.50 ± 0.00| 0.00 ± 0.00|
| Differentiating between mood disorders and schizophrenia | VGG16  | 0.79 ± 0.16| 0.83 ± 0.14| 0.80 ± 0.20 | 0.80 ± 0.15| 0.60 ± 0.30|
|                                              | Resnet50v2 | 0.88 ± 0.07| 0.85 ± 0.10| 0.92 ± 0.11 | 0.88 ± 0.07| 0.77 ± 0.15|
|                                              | XceptionNet| 0.86 ± 0.10| 0.82 ± 0.14| 0.91 ± 0.12 | 0.86 ± 0.12| 0.72 ± 0.20|
|                                              | EfficientNetB1 | 0.86 ± 0.09| 0.85 ± 0.10| 0.86 ± 0.12 | 0.85 ± 0.09| 0.71 ± 0.18|
| Average of CNN models on mean                |            | 0.85 ± 0.04| 0.84 ± 0.02| 0.87 ± 0.06 | 0.85 ± 0.03| 0.70 ± 0.07|

(continued)
Table 6. (continued)

| Task                  | Model              | Accuracy | Precision | Sensitivity | F-score | MCC |
|----------------------|--------------------|----------|-----------|-------------|---------|-----|
| LSTM                 | 0.88 ± 0.06        | 0.84 ± 0.06 | 0.93 ± 0.05 | 0.88 ± 0.06 | 0.77 ± 0.11 |
| GRU                  | 0.86 ± 0.04        | 0.81 ± 0.09 | 0.93 ± 0.12 | 0.79 ± 0.10 | 0.73 ± 0.07 |
| Attention-based LSTM | 0.83 ± 0.07        | 0.83 ± 0.07 | 0.84 ± 0.07 | 0.82 ± 0.07 | 0.66 ± 0.13 |
| Attention-based GRU  | 0.82 ± 0.10        | 0.83 ± 0.08 | 0.83 ± 0.08 | 0.82 ± 0.09 | 0.66 ± 0.17 |
| Average of CNN models on mean | 0.85 ± 0.03 | 0.83 ± 0.01 | 0.88 ± 0.06 | 0.83 ± 0.04 | 0.71 ± 0.05 |
| ANN                  | 0.78 ± 0.08        | 0.83 ± 0.05 | 0.73 ± 0.18 | 0.77 ± 0.11 | 0.57 ± 0.16 |
| Random forest        | 0.83 ± 0.09        | 0.84 ± 0.13 | 0.86 ± 0.15 | 0.84 ± 0.09 | 0.69 ± 0.18 |
| Random guessing      | 0.49 ± 0.01        | 0.51 ± 0.02 | 0.49 ± 0.01 | 0.50 ± 0.01 | 0.00 ± 0.01 |

CNN: Convolutional Neural Network; RNN: Recurrent Neural Network; MHD: Mental Health Disorder; MCC: Matthews correlation coefficient; LSTM: Long-short Term Memory; GRU: Gated Recurrent Unit.

*The best-performing metrics on mean value are in bold.
Table 7. Comparison of DL models in approach 2 using weight-added value. Each model is repeated 5 times, corresponding to 5-fold validation. The average results are presented as the mean ± standard deviation across the five experiments, of the best hyper-parameter sets.

| Task                                                                 | Model        | Accuracy  | Precision | Sensitivity | F-score | MCC    |
|----------------------------------------------------------------------|--------------|-----------|-----------|-------------|---------|--------|
| Direct differentiation between mood disorders, schizophrenia, and healthy control groups | VGG16      | 0.75 ± 0.07* | 0.72 ± 0.08 | 0.72 ± 0.08 | 0.71 ± 0.08 | 0.60 ± 0.10 |
|                                                                      | Resnet50v2   | 0.74 ± 0.07 | 0.68 ± 0.10 | 0.69 ± 0.08 | 0.68 ± 0.09 | 0.56 ± 0.09 |
|                                                                      | XceptionNet  | 0.74 ± 0.05 | 0.71 ± 0.04 | 0.69 ± 0.02 | 0.68 ± 0.02 | 0.58 ± 0.06 |
|                                                                      | EfficientNetB1 | 0.71 ± 0.09 | 0.68 ± 0.07 | 0.69 ± 0.09 | 0.67 ± 0.09 | 0.58 ± 0.12 |
|                                                                      | Average of CNN models on mean | 0.74 ± 0.02 | 0.70 ± 0.02 | 0.70 ± 0.02 | 0.69 ± 0.02 | 0.57 ± 0.03 |
|                                                                      | LSTM         | 0.67 ± 0.11 | 0.66 ± 0.08 | 0.69 ± 0.09 | 0.65 ± 0.09 | 0.49 ± 0.14 |
|                                                                      | GRU          | 0.62 ± 0.05 | 0.60 ± 0.07 | 0.62 ± 0.08 | 0.58 ± 0.06 | 0.42 ± 0.10 |
|                                                                      | Attention-based LSTM | 0.71 ± 0.06 | 0.68 ± 0.06 | 0.69 ± 0.08 | 0.67 ± 0.07 | 0.54 ± 0.09 |
|                                                                      | Attention-based GRU | 0.72 ± 0.09 | 0.69 ± 0.09 | 0.71 ± 0.11 | 0.68 ± 0.10 | 0.57 ± 0.14 |
|                                                                      | Average of RNN models on mean | 0.68 ± 0.05 | 0.66 ± 0.04 | 0.68 ± 0.04 | 0.65 ± 0.05 | 0.51 ± 0.07 |
|                                                                      | ANN          | 0.61 ± 0.07 | 0.57 ± 0.09 | 0.57 ± 0.08 | 0.55 ± 0.08 | 0.36 ± 0.13 |
|                                                                      | Random forest | 0.62 ± 0.09 | 0.65 ± 0.06 | 0.66 ± 0.06 | 0.60 ± 0.07 | 0.46 ± 0.08 |
|                                                                      | Random guessing | 0.55 ± 0.03 | 0.67 ± 0.03 | 0.53 ± 0.01 | 0.53 ± 0.02 | 0.31 ± 0.02 |

*The best performing metrics on mean value are in bold.
CNN: Convolutional Neural Network; RNN: Recurrent Neural Network; MCC: Matthews correlation coefficient; LSTM: Long-short Term Memory; GRU: Gated Recurrent Unit.
platform for integration and data security is needed. On the other hand, DL has been demonstrated to perform well in diagnosis and prediction, as mentioned above. However, there is difficulty in tracing a prediction back to the importance of features, because DL is considered a black box. As a result, it becomes less preferable for use in healthcare, where the physician needs to clearly comprehend the hidden factors that cause a disease, and in situations where decisions become a matter of life and death.\textsuperscript{65} Furthermore, DL techniques require a much greater hardware resource and a longer processing duration than other traditional ML techniques, in addition to more domain knowledge in order to fine-tune and deploy suitable models.

In fact, more kinds of data could be collected using these wearable devices, such as sleep quality, step count, heart rate, etc. Therefore, we still have room to improve the performance of DL models for MHD identification by including these kinds of data in the training and prediction steps. Furthermore, we conducted our experiment on a 7-days moving forward window; in other words, 7 days of gravitational acceleration signals were collected for each sample. We may change the number of this moving forward window to observe the differences in results and enhance the accuracy of prediction. On the other hand, to overcome the unintelligible DL ‘black box’, an explainable AI system could be deployed to encourage the use of DL in healthcare.\textsuperscript{66} Furthermore, a federated learning system could be developed that is able to enhance privacy and enable the production of scalable smart healthcare networks and applications.

In conclusion, by analyzing data from wearable devices collected in 1 week using DL models, the experimental results showed that the proposed approaches provided a good performance. As a result, this may become an objective reliable resource that can assist physicians in making diagnostic decisions and ascertaining the best treatment plans for their patients, besides the usual diagnostic method of interview, questionnaire, follow-up, and examination. Applying DL to wearable device data is a prospective approach for MHD differentiation with high accuracy.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Ministry of Science and Technology, Taiwan (MOST 111-2221-E-155-024).

**ORCID iDs**

Duc-Khanh Nguyen https://orcid.org/0000-0001-8388-7425
Ai-Hsien A Li https://orcid.org/0000-0001-7118-7819

**Supplemental Material**

Supplemental material for this article is available online.

**References**

1. Lépine J-P and Briley M. The increasing burden of depression. *Neuropsychiatr Disease Treatment* 2011; 7(Suppl 1): 3–7.
2. Wang PS, Aguilar-Gaxiola S, Alonso J, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *Lancet* 2007; 370(9590): 841–850.

3. Das J, Do QT, Friedman J, et al. Mental health and poverty in developing countries: revisiting the relationship. *Soc Science Med* 2007; 65(3): 467–480.

4. Merten EC, Cwik JC, Margraf J, et al. Overdiagnosis of mental disorders in children and adolescents (in developed countries). *Child Adolesc Psychiatry Ment Health* 2017; 11(1): 1–11.

5. Insel TR. Rethinking schizophrenia. *Nature* 2010; 468(7321): 187–193.

6. Pilling S, Anderson I, Goldberg D, et al. Depression in adults, including those with a chronic physical health problem: summary of NICE guidance. *Bmj* 2009; 339: 339.

7. Serafini G, Pompili M, Borgwardt S, et al. Brain changes in early-onset bipolar and unipolar depressive disorders: a systematic review in children and adolescents. *Eur Child Adolesc Psychiatry* 2014; 23(11): 1023–1041.

8. Cuellar AK, Johnson SL and Winters R. Distinctions between bipolar and unipolar depression. *Clin Psychology Review* 2005; 25(3): 307–339.

9. Twenge JM, Joiner TE, Rogers ML, et al. Increases in depressive symptoms, suicide-related outcomes, and suicide rates among US adolescents after 2010 and links to increased media screen time. *Clin Psychol Sci* 2018; 6(1): 3–17.

10. Laursen TM and Munk-Olsen T. A comparison of selected risk factors for unipolar depressive disorder, bipolar affective disorder, schizoaffective disorder, and schizophrenia from a Danish population-based cohort. *J Clin Psychiatry* 2007; 68(11): 1673–1681.

11. Kirli S and Çaliskan M. A comparative study of sertraline versus imipramine in postpsychotic depressive disorder of schizophrenia. *Schizophrenia Res* 1998; 33(1–2): 103–111.

12. Oztemel E and Gursev S. Literature review of Industry 4.0 and related technologies. *J Intell Manufact* 2020; 31(1): 127–182.

13. Chen X-W and Lin X. Big data deep learning: challenges and perspectives. *IEEE Access* 2014; 2: 514–525.

14. Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 2017; 19(6): 1236–1246.

15. Xia M, Zhang G, Mu C, et al. Cervical cancer cell detection based on deep convolutional neural network. In 2020 39th Chinese Control Conference (CCC). Shenyang, China, 27–29 July 2020.

16. Pham T, Tran T, Phung D, et al. Predicting healthcare trajectories from medical records: a deep learning approach. *J Biomed Inform* 2017; 69: 218–229.

17. Shrestha A and Mahmood A. Review of deep learning algorithms and architectures. *IEEE Access* 2019; 7: 53040–53065.

18. Lee C-N, Fong VH, Chu Y-T, et al. A Wearable Device Of Gait Tracking For Parkinson’S Disease Patients. In 2018 International Conference on Machine Learning and Cybernetics (ICMLC), Chengdu, China, 15–18 July 2018.

19. Appelboom G, Camacho E, Abraham ME, et al. Smart wearable body sensors for patient self-assessment and monitoring. *Arch Public Health* 2014; 72(1): 1–9.

20. Wu M and Luo J. Wearable technology applications in healthcare: a literature review. *Online J Nurs Inf* 2019; 23: ■■■.

21. Musliner KL, Munk-Olsen T, Mors O, et al. Progression from unipolar depression to schizophrenia. *Acta Psychiatr Scand* 2017; 135(1): 42–50.

22. Tasic L, Larcadera AL, Pontes JG, et al. Peripheral biomarkers allow differential diagnosis between schizophrenia and bipolar disorder. *J Psychiatr Res* 2019; 119: 67–75.
23. Regier DA, Kaelber CT, Rae DS, et al. Limitations of diagnostic criteria and assessment instruments for mental disorders: implications for research and policy. *Arch General Psychiatry* 1998; 55(2): 109–115.

24. Stolzenburg S, Freitag S, Evans-Lacko S, et al. The stigma of mental illness as a barrier to self labeling as having a mental illness. *J Nervous Mental Disease* 2017; 205(12): 903–909.

25. Bharadwaj P, Pai MM and Suziedelyte A. Mental health stigma. *Econ Let* 2017; 159: 57–60.

26. Glaros C and Fotiadis DI. Wearable devices in healthcare. In: *Intelligent Paradigms for Healthcare Enterprises*. Berlin: Springer, 2005, pp. 237–264.

27. Guk K, Han G, Lim J, et al. Evolution of wearable devices with real-time disease monitoring for personalized healthcare. *Nanomaterials* 2019; 9(6): 813.

28. Marakhimov A and Joo J. Consumer adaptation and infusion of wearable devices for healthcare. *Comput Hum Behav* 2017; 76: 135–148.

29. Naslund JA, Aschbrenner KA and Bartels SJ. Wearable devices and smartphones for activity tracking among people with serious mental illness. *Ment Health Phy Activity* 2016; 10: 10–17.

30. Muaremi A, Arnrich B and Tröster G. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience* 2013; 3(2): 172–183.

31. Byrne S, Kotze B, Ramos F, et al. Using a mobile health device to manage severe mental illness in the community: what is the potential and what are the challenges? *Aust New Zealand J Psychiatry* 2020; 54(10): 964–969.

32. Cella M, Okruszek L, Lawrence M, et al. Using wearable technology to detect the autonomic signature of illness severity in schizophrenia. *Schizophrenia Res* 2018; 195: 537–542.

33. Hunkin H, King DL and Zajac IT. Perceived acceptability of wearable devices for the treatment of mental health problems. *J Clin Psychol* 2020; 76(6): 987–1003.

34. Iqbal SM, Mahgoub I, Du E, et al. Advances in healthcare wearable devices. *Npj Flexible Electro* 2021; 5(1): 1–14.

35. Lu L, Zhang J, Xie Y, et al Wearable health devices in health care: Narrative systematic review. *JMIR Mhealth and Uhealth* 2020; 8(11): e18907.

36. Kwak GH-J and Hui P. *Deephealth: Deep Learning for Health Informatics*. ACM Transactions on Computing for Healthcare, Kamakura City, Japan, 14-16 August 2019.

37. Phan D-V, Chan C-L and Nguyen D-K. Applying deep learning for prediction sleep quality from wearable data. In Proceedings of the 4th International Conference on Medical and Health Informatics. 2020.

38. Nguyen D-K, Lan C-H and Chan C-L. Deep ensemble learning approaches in healthcare to enhance the prediction and diagnosing performance: the workflows, deployments, and surveys on the statistical, image-based, and sequential datasets. *Int J Environ Res Public Health* 2021; 18(20): 10811.

39. Coutts LV, Plans D, Brown AW, et al. Deep learning with wearable based heart rate variability for prediction of mental and general health. *J Biomed Inform* 2020; 112: 103610.

40. Bashivan P, Rish I and Heisig S, Mental state recognition via wearable eeg. *arXiv* preprint arXiv: 1602.00985, 2016.

41. Galván-Tejada CE, Zanella-Calzada LA, Gamboa-Rosales H, et al. Depression episodes detection in unipolar and bipolar patients: a methodology with feature extraction and feature selection with genetic algorithms using activity motion signal as information source. *Mobile Inf Syst* 2019; 2019: 1–12.

42. Zanella-Calzada LA, Galvan-Tejada C, Chavez-Lamas N, et al. Feature extraction in motor activity signal: towards a depression episodes detection in unipolar and bipolar patients. *Diagnostics* 2019; 9(1): 8.
43. Jakobsen P, Garcia-Ceja E, Riegler M, et al. Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls. Plos One 2020; 15(8): e0231995.
44. Boeker M, Riegler MA, Hammer HL, et al. Diagnosing schizophrenia from activity records using hidden markov model parameters. In: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), Aveiro, Portugal, 07–09 June 2021.
45. Nguyen D-K, Chan C-L, Adams Li A-H, et al. Deep stacked generalization ensemble learning models in early diagnosis of depression illness from wearable devices data. In: 2021 5th International Conference on Medical and Health Informatics, Kyoto, Japan, 14-16 May 2021.
46. Garcia-Ceja E, Rielger MA, Jakobsen P, et al. Depresjon: a motor activity database of depression episodes in unipolar and bipolar patients. in Proceedings of the 9th ACM multimedia systems conference, Amsterdam, Netherlands, 12-15 June 2018.
47. Jakobsen P, Garcia-Ceja E, Stabell LA, et al. PSYKOSE: a motor activity database of patients with schizophrenia. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 2020.
48. Routen AC, Upton D, Edwards M, et al. Intra-and inter-instrument reliability of the actiwatch 4 accelerometer in a mechanical laboratory setting. J Hum Kinetics 2012; 31: 17–24.
49. Berle JO, Hauge ER, Oedegaard KJ, et al. Actigraphic registration of motor activity reveals a more structured behavioural pattern in schizophrenia than in major depression. BMC Research Notes 2010; 3(1): 1–7.
50. Chen W and Shi K. A deep learning framework for time series classification using Relative Position Matrix and Convolutional Neural Network. Neurocomputing 2019; 359: 384–394.
51. Sezer OB and Ozbayoglu AM. Algorithmic financial trading with deep convolutional neural networks: time series to image conversion approach. Appl Soft Comput 2018; 70: 525–538.
52. LeCun Y and Bengio Y. Convolutional networks for images, speech, and time series. Handbook Brain Theory Neural Networks 1995; 3361(10): 1995.
53. Simonyan K and Zisserman A, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
54. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016. 2016.
55. Chollet F, Xception: deep learning with depthwise separable convolutions. CoRR abs/1610.02357 (2016). arXiv preprint arXiv:1610.02357, 2016.
56. Tan M and Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, Long Beach, 9–15 June 2019.
57. Hochreiter S and Schmidhuber J. Long short-term memory. Neural Computation 1997; 9(8): 1735–1780.
58. Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.
59. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017.
60. Chicco D and Jurman G. The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020; 21(1): 1–13.
61. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta (BBA)-Protein Struct 1975; 405(2): 442–451.
62. Van RG and Drake F. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace, 2009.
63. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Machine Learn Res 2011; 12: 2825–2830.
64. Abadi M, Agarwal A, Barham P, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015.

65. Adadi A and Berrada M. Explainable AI for healthcare: from black box to interpretable models. In: *Embedded Systems and Artificial Intelligence*. Singapore: Springer, 2020, pp. 327–337.

66. Xu F, Uszkoreit H, Du Y, et al. Explainable AI: a brief survey on history, research areas, approaches and challenges. In CCF international conference on natural language processing and Chinese computing, Dunhuang, China, 9-14 October 2019.