

DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection

Bokai Cao¹, Lei Zheng¹, Chenwei Zhang¹, Philip S. Yu^{1,2}, Andrea Piscitello¹, John Zulueta³, Olu Ajilore³, Kelly Ryan⁴, and Alex D. Leow^{1,3,5}

¹Department of Computer Science, University of Illinois at Chicago

²Institute for Data Science, Tsinghua University

³Department of Psychiatry, University of Illinois at Chicago

⁴Department of Psychiatry, University of Michigan

⁵Department of Bioengineering, University of Illinois at Chicago

caobokai,lzheng21,czhang99,psyu,apisci2@uic.edu

jzulueta,oajilore,aleow@psych.uic.edu

karyan@med.umich.edu

ABSTRACT

The increasing use of electronic forms of communication presents new opportunities in the study of mental health, including the ability to investigate the manifestations of psychiatric diseases unobtrusively and in the setting of patients' daily lives. A pilot study to explore the possible connections between bipolar affective disorder and mobile phone usage was conducted. In this study, participants were provided a mobile phone to use as their primary phone. This phone was loaded with a custom keyboard that collected metadata consisting of keypress entry time and accelerometer movement. Individual character data with the exceptions of the backspace key and space bar were not collected due to privacy concerns. We propose an end-to-end deep architecture based on late fusion, named DeepMood, to model the multi-view metadata for the prediction of mood scores. Experimental results show that 90.31% prediction accuracy on the depression score can be achieved based on session-level mobile phone typing dynamics which is typically less than one minute. It demonstrates the feasibility of using mobile phone metadata to infer mood disturbance and severity.

KEYWORDS

typing dynamics; bipolar disorder; recurrent network; sequence prediction

1 INTRODUCTION

Mobile phones, in particular, "smartphones" have become near ubiquitous with 2 billion smartphone users worldwide. This presents new opportunities in the study and treatment of psychiatric illness including the ability to study the manifestations of psychiatric illness in the setting of patients' daily lives in an unobtrusive manner

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

KDD'17, August 13–17, 2017, Halifax, NS, Canada

© 2017 Association for Computing Machinery.

ACM ISBN 978-1-4503-4887-4/17/08...\$15.00

<https://doi.org/10.1145/3097983.3098086>



Figure 1: A sample of the collected data in time series.

and at a level of detail that was not previously possible. Continuous real-time monitoring in naturalistic settings and collection of automatically generated smartphone data that reflect illness activity could facilitate early intervention and have a potential use as objective outcome measures in efficacy trials [2, 6, 18].

While mobile phones are used for a variety of tasks the most widely and frequently used feature is text messaging. To the best of our knowledge, no previous studies [3, 19, 21, 37, 40, 46] have investigated the relationship between mobile phone typing dynamics and mood states. In this work, we aim to determine the feasibility of inferring mood disturbance and severity from such data. In particular we seek to investigate the relationship between the digital footprints and mood in bipolar affective disorder which has been deemed the most expensive behavioral health care diagnosis [35], costing more than twice as much as depression per affected individual [29]. For every dollar allocated to outpatient care for people with bipolar disorder, \$1.80 is spent on inpatient care, suggesting early intervention and improved prevention management could decrease the financial impact of this illness [35].

We study the mobile phone typing dynamics metadata on a session-level. A session is defined as beginning with a keypress which occurs after 5 or more seconds have elapsed since the last keypress and continuing until 5 or more seconds elapse between keypresses¹. The duration of a session is typically less than one minute. In this manner, each participant would contribute many samples, one per phone usage session, which could benefit data analysis and model training. Each session is composed of features that are represented in multiple views or modalities (e.g., alphanumeric characters, special characters, accelerometer values), each of which has different timestamps and densities, as shown in Figure 1. Modeling the multi-view time series data on such a fine-grained session-level brings up several formidable challenges:

¹5-second is an arbitrary threshold we set which can be changed and tuned easily.

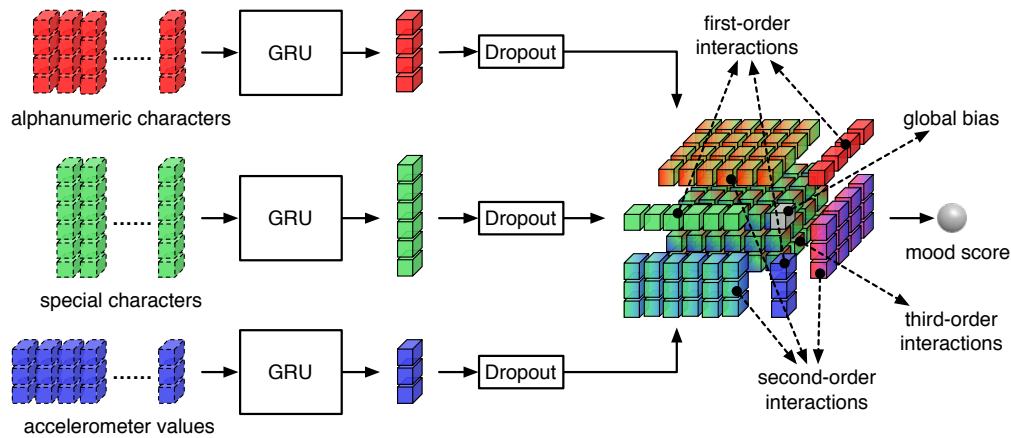


Figure 2: The architecture of DeepMood with a Multi-view Machine layer for data fusion.

- **Unaligned views:** An intuitive idea for fusing the multi-view time series is to align them with each unique timestamp. However, features defined in one view would be missing for data points collected in another view. For example, a data point in special characters has no *acceleration* in accelerometer values or *distance from last key* in alphanumeric characters².
- **Dominant views:** One may also attempt to do the fusion by concatenating the multi-view time series per session. However, the views usually have different densities in a session, because the metadata are collected from different sources or sensors. For example, character-related metadata collected following a person's typing behaviours are much sparser than accelerometer values collected in the background which have 16 times more data points in our dataset. Dense views could dominate a concatenated feature space and potentially override the effects of sparse but important views.
- **View interactions:** The multi-view time series from typing dynamics contains complementary information reflecting a person's mental health. The relationship between the digital footprints and mood states can be highly nonlinear. An effective fusion strategy is needed to explore feature interactions across different views.

In this paper, we propose a deep architecture based on late fusion, named DeepMood, to model mobile phone typing dynamics, as illustrated in Figure 2. The contributions of this work are threefold:

- **Data analysis (Section 2):** We obtain interesting insights related to the digital footprints on mobile phones by analyzing the correlation between patterns of typing dynamics metadata and mood in bipolar affective disorder.
- **A novel fusion strategy in a deep framework (Section 3):** Motivated by the aforementioned challenges that early fusion strategies (*i.e.*, aligning views with timestamps or concatenating views per session) would lead to the problems of unaligned or dominant views, we propose a two-stage late fusion approach for modeling the multi-view time series data. In the first stage, each

view of the time series is separately modeled by a Recurrent Neural Network (RNN) [34, 44]. The multi-view metadata are then fused in the second stage by exploring interactions across the output vectors from each view, where three alternative approaches are developed following the idea of Multi-view Machines [9], Factorization Machines [39], or in a fully connected fashion.

- **Empirical evaluations (Section 4):** We conduct experiments showing that 90.31% prediction accuracy on the depression score can be achieved based on session-level typing dynamics which reveals the potential of using mobile phone metadata to predict mood disturbance and severity. Our code is open-sourced at <https://www.cs.uic.edu/~bcao1/code/DeepMood.py>.

2 DATA

The data used in this work were collected from the BiAffect³ study which is the winner of the Mood Challenge for ResearchKit⁴. During a preliminary data collection phase, for a period of 8 weeks, 40 individuals were provided a Galaxy Note 4 mobile phone which they were instructed to use as their primary phone during the study. This phone was loaded with a custom keyboard that replaced the standard Android OS keyboard. The keyboard collected metadata consisting of keypress entry time and accelerometer movement and uploaded them to the study server. In order to protect participants' privacy, individual character data with the exceptions of the backspace key and space bar were not collected.

In this work, we study the collected metadata for participants including bipolar subjects and normal controls who had provided at least one week of metadata. There are 7 participants with *bipolar I* disorder that involves periods of severe mood episodes from mania to depression, 5 participants with *bipolar II* disorder which is a milder form of mood elevation, involving milder episodes of hypomania that alternate with periods of severe depression, and 8 participants with no diagnosis per DSM-IV TR criteria [28].

Participants were administered the Hamilton Depression Rating Scale (HDRS) [48] and Young Mania Rating Scale (YMRS) [53] once

²This is for privacy concerns, because malicious person may be able to unscramble and recover the texts using such information.

³<http://www.biaffect.com>

⁴<http://www.moodchallenge.com>

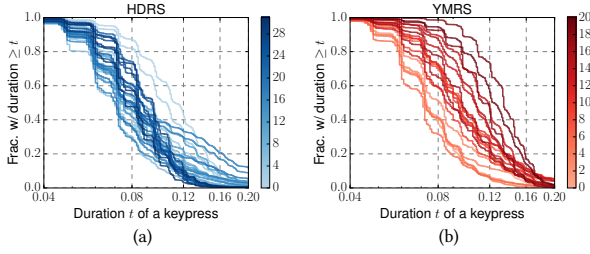


Figure 3: CCDFs of duration of a keypress.

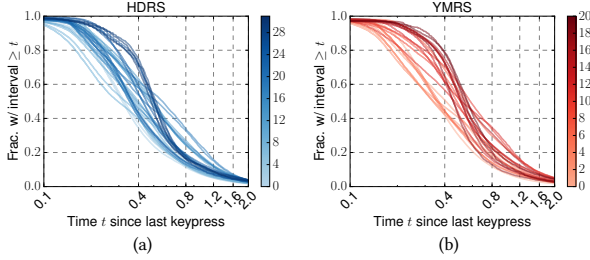


Figure 4: CCDFs of time since last keypress.

a week which are used as the golden standard to assess the level of depressive and manic symptoms in bipolar disorder. However, the use of these clinical rating scales requires a face-to-face patient-clinician encounter, and the level of affective symptoms is assessed during a clinical evaluation. Study findings may be unreliable when using rating scales as outcome measures due to methodological issues such as unblinding of raters and patients, differences in rater experiences and missing visits for outcome assessments [15, 18, 36]. Thus, it motivates us to explore more objective methods with real-time data for assessing affective symptoms.

2.1 Alphanumeric Characters

Due to privacy reasons, we only collected metadata for keypresses on alphanumeric characters, including duration of a keypress, time since last keypress, and distance from last key along two axes. Firstly, we aim to assess the correlation between duration of a keypress and mood states. The complementary cumulative distribution functions (CCDFs) of *duration of a keypress* are displayed in Figure 3. Data points with different scores are colored differently, and the range of mood scores corresponds to the colorbar. In general, the higher the score, the darker the color and the more severe the depressive or manic symptoms. According to the Kolmogorov-Smirnov test on two samples, for all the pairs of distributions, we can reject the null hypothesis that two samples are drawn from the same distribution with significance level $\alpha = 0.01$. As expected, we are dealing with a heavy-tailed distribution: (1) most keypresses are very fast with median 85ms, (2) but a non-negligible number have longer duration with 5% using more than 155ms. Interestingly, samples with mild depression tend to have shorter duration than

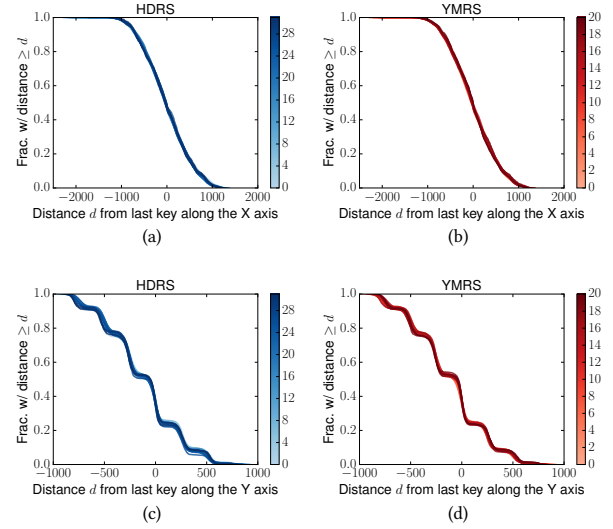


Figure 5: CCDFs of distance from last key along two axes. Note that lines are almost identical.

normal ones, while those with severe depression stand in the middle. Samples in manic symptoms seem to hold a key longer than normal ones.

Next we ask how the time since last keypress correlates with mood states. We show the CCDFs of *time since last keypress* in Figure 4. Based on the Kolmogorov-Smirnov test, for 98.06% in HDRS and 99.52% in YMRS of the distribution pairs, we can reject the null hypothesis that two samples are drawn from the same distribution with significance level $\alpha = 0.01$. Not surprisingly, this distribution is heavily skewed, with most time intervals being very short with median 380ms. However, there is a significant fraction of keypresses with much longer intervals where 5% have more than 1.422s. We can observe that the values of time since last keypress from the normal group (with light blue/red) approximate a uniform distribution on the log scale in the range from 0.1s to 2.0s. On the contrary, this metric from samples with mood disturbance (with dark blue/red) shows a more skewed distribution with a few values on the two tails and majority centered between 0.4s and 0.8s. In other words, healthy people show a good range of reactivity that gets lost in mood disturbance where the range is more restricted.

Figure 5 shows the CCDFs of *distance from last key* along two axes which can be considered as a sort of very rough proxy of the semantic content of people's typing. No distinction can be observed across different mood states, because there are no dramatic differences in the manner in which depressive or manic people type compared to controls.

2.2 Special Characters

In this view, we use one-hot-encoding for typing behaviors other than alphanumeric characters, including *auto-correct*, *backspace*, *space*, *suggestion*, *switching-keyboard* and *other*. They are usually sparser than alphanumeric characters. Figure 6 shows the scatter

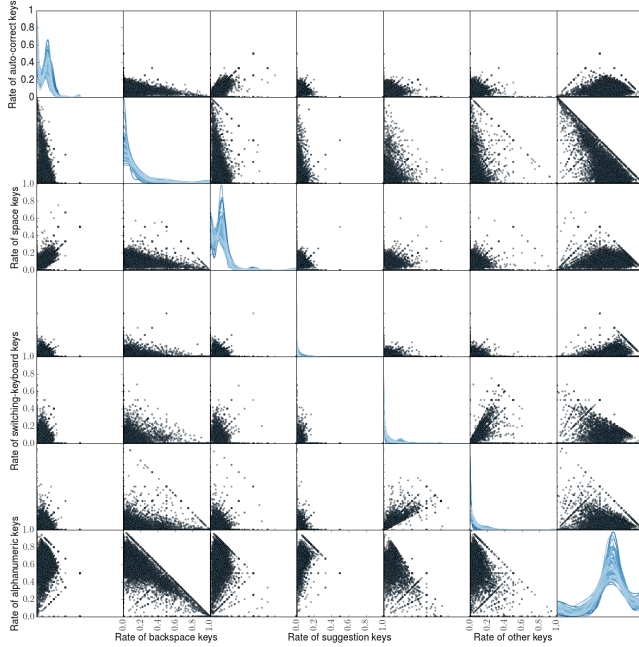


Figure 6: Scatter plot between rates of different keys.

plot between rates of these special characters as well as alphanumeric ones in a session where the color of a dot/line corresponds to the HDRS score. Although no obvious distinction can be found between mood states, we can observe some interesting patterns: the rate of alphanumeric keys is negatively correlated with the rate of *backspace* (from the subfigure at the 2nd row, 7th column), while the rate of *switching-keyboard* is positively correlated with the rate of *other keys* (from the subfigure at the 5th row, 6th column). On the diagonal there are kernel density estimations. It shows that the rate of alphanumeric characters is generally high in a session, followed by *auto-correct*, *space*, *backspace*, etc. Similar patterns can be found from the plot of YMRS which is omitted here.

2.3 Accelerometer Values

Accelerometer values are recorded every 60ms in the background during an active session regardless of a person's typing speed, thereby making them much denser than alphanumeric characters. The CCDFs of absolute accelerometer values along three axes are displayed in Figure 7. Data points with different mood scores are colored differently, and the higher the score, the more severe the depressive or manic symptoms. According to the Kolmogorov-Smirnov test on two samples, for all the pairs of distributions, we can reject the null hypothesis that two samples are drawn from the same distribution with significance level $\alpha = 0.01$. Note that the vertical axis of the non-zoomed plots is on a log scale. We observe a heavy-tailed distribution for all three axes and for both HDRS and YMRS, with more than 99% of data points being less than 7.45, 9.97 and 10.56 along X, Y and Z axis, respectively. By zooming into data points at the "head" of the distribution on a regular scale, we can see different patterns on the absolute acceleration along different axes. There is a nearly uniform distribution of absolute acceleration along

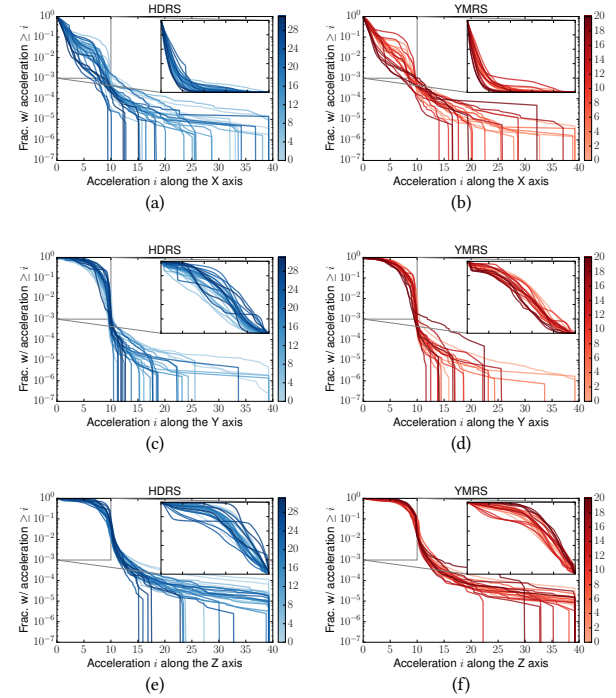


Figure 7: CCDFs of absolute acceleration along three axes.

Table 1: Statistics of the dataset.

Statistics	ALPH.	SPEC.	ACCEL.
# data points	836,027	538,520	14,237,503
# sessions	34,993	33,385	37,647
mean length	24	16	378
median length	14	9	259
maximum length	538	437	90,193

the Y axis in the range from 0 to 10, while the majority along the X axis lie between 0 and 2, and the majority along the Z axis lie between 6 and 10. An interesting observation is that compared with normal ones, samples with mood disturbance tend to have larger accelerations along the Z axis, and smaller accelerations along the Y axis. Hence, we suspect that people in a normal mood state prefer to hold their phone towards to themselves, while people in depressive or manic symptoms are more likely to lay their phone with an angle towards to the horizon, given that data were collected only when the phone was in a portrait position.

See Table 1 for more information about the statistics of the dataset. Note that the length of a sequence is measured in terms of the number of data points in a sample rather than the duration in time.

3 DEEPMOOD ARCHITECTURE BASED ON LATE FUSION

In this paper, we propose an end-to-end deep architecture, named DeepMood, to model mobile phone typing dynamics. Specifically,

DeepMood provides a late fusion framework. It first models each view of the time series data separately using Gated Recurrent Unit (GRU) [12], a simplified version of Long Short-Term Memory (LSTM) [24]. It then fuses the output of the GRU from each view. As the GRU extracts a latent feature representation out of each time series, where the notions of sequence length and sampling time points are removed from the latent space, this avoids the problem of dealing directly with the heterogeneity of the time series from each view. Following the idea of Multi-view Machines [9], Factorization Machines [39], or in a conventional fully connected fashion, three alternative fusion layers are designed to integrate the complementary information in the multi-view time series to produce a prediction on the mood score. The architecture is illustrated in Figure 2.

3.1 Modeling One View

Each view in the metadata is essentially a time series whose length can vary a lot across sessions that largely depends on the duration of a session. In order to model the dynamic sequential correlations in each time series, we adopt the RNN architecture [34, 44] which keeps hidden states over a sequence of elements and update the hidden state \mathbf{h}_k by the current input \mathbf{x}_k as well as the previous hidden state \mathbf{h}_{k-1} where $k > 1$ with a recurrent function:

$$\mathbf{h}_k = f(\mathbf{x}_k, \mathbf{h}_{k-1}) \quad (1)$$

The simplest form of an RNN is as follows:

$$\mathbf{h}_k = \sigma(\mathbf{W}\mathbf{x}_k + \mathbf{U}\mathbf{h}_{k-1}) \quad (2)$$

where $\mathbf{W} \in \mathbb{R}^{d_h \times d_x}$, $\mathbf{U} \in \mathbb{R}^{d_h \times d_h}$ are model parameters that need to be learned, d_x and d_h are the input dimension and the number of recurrent units, respectively. $\sigma(\cdot)$ is a nonlinear transformation function such as tanh, sigmoid, and rectified linear unit (ReLU). Since RNNs in such a form would fail to learn long term dependencies due to the exploding and the vanishing gradient problem [5, 23], they are not suitable to learn dependencies from a long input sequence in practice.

To make the learning procedure more effective over long sequences, the GRU [12] is proposed as a variation of the LSTM unit [24]. The GRU has been attracting great attentions since it overcomes the vanishing gradient problem in traditional RNNs and is more efficient than the LSTM in some tasks [14]. The GRU is designed to learn from previous timestamps with long time lags of unknown size between important timestamps via memory units that enable the network to learn to both update and forget hidden states based on new inputs.

A typical GRU is formulated as:

$$\begin{aligned} \mathbf{r}_k &= \text{sigmoid}(\mathbf{W}_r \mathbf{x}_k + \mathbf{U}_r \mathbf{h}_{k-1}) \\ \mathbf{z}_k &= \text{sigmoid}(\mathbf{W}_z \mathbf{x}_k + \mathbf{U}_z \mathbf{h}_{k-1}) \\ \tilde{\mathbf{h}}_k &= \tanh(\mathbf{W}_h \mathbf{x}_k + \mathbf{U}_h (\mathbf{r}_k \odot \mathbf{h}_{k-1})) \\ \mathbf{h}_k &= \mathbf{z}_k \odot \mathbf{h}_{k-1} + (1 - \mathbf{z}_k) \odot \tilde{\mathbf{h}}_k \end{aligned} \quad (3)$$

where \odot is the element-wise multiplication operator, a reset gate \mathbf{r}_k allows the GRU to forget the previously computed state \mathbf{h}_{k-1} , and an update gate \mathbf{z}_k balances between the previous state \mathbf{h}_{k-1} and the candidate state $\tilde{\mathbf{h}}_k$. The hidden state \mathbf{h}_k can be considered as a compact representation of the input sequence from \mathbf{x}_1 to \mathbf{x}_k .

3.2 Late Fusion on Multiple Views

Here we pursue a late fusion strategy to integrate the output vectors of the GRU units on these time series data from different views. This avoids the issues of alignment and diverse frequencies among the time series under different views when performing early fusion directly on the input data.

In the following we study alternative methods for performing late fusion. These include not only the straightforward approach based on adding a fully connected layer to concatenate the features from different views, but also novel approaches to capture interactions among the features across multiple views by exploring the concept of Factorization Machines [39] to capture the second-order interactions as well as the concept of Multi-view Machines [9] to capture higher order interactions as shown in Figure 8.

We denote the output vectors at the end of a sequence from the p -th view as $\mathbf{h}^{(p)}$. We can consider $\{\mathbf{h}^{(p)} \in \mathbb{R}^{d_h}\}_{p=1}^m$ as multi-view data where m is the number of views.

Fully connected layer. In order to generate a prediction on the mood score, a straightforward idea is to first concatenate features from multiple views together, i.e., $\mathbf{h} = [\mathbf{h}^{(1)}; \mathbf{h}^{(2)}; \dots; \mathbf{h}^{(m)}] \in \mathbb{R}^d$, where d is the total number of multi-view features, and typically $d = md_h$ for one-directional RNNs and $d = 2md_h$ for bidirectional RNNs. We then feed forward \mathbf{h} into one or several fully connected neural network layers with a nonlinear function $\sigma(\cdot)$ in between.

$$\begin{aligned} \mathbf{q} &= \text{relu}(\mathbf{W}^{(1)}[\mathbf{h}; 1]) \\ \hat{\mathbf{y}} &= \mathbf{W}^{(2)}\mathbf{q} \end{aligned} \quad (4)$$

where $\mathbf{W}^{(1)} \in \mathbb{R}^{k' \times (d+1)}$, $\mathbf{W}^{(2)} \in \mathbb{R}^{c \times k'}$, k' is the number of hidden units, c is the number of classes, and the constant signal “1” is to model the global bias. Note that here we consider only one hidden layer between the input layer and the final output layer as shown in Figure 8(a).

Factorization Machine layer. Rather than capturing nonlinearity through the transformation function, we consider explicitly modeling feature interactions between input units as shown in Figure 8(b).

$$\begin{aligned} \mathbf{q}_a &= \mathbf{U}_a \mathbf{h} \\ b_a &= \mathbf{w}_a^T [\mathbf{h}; 1] \\ \hat{y}_a &= \text{sum}([\mathbf{q}_a \odot \mathbf{q}_a; b_a]) \end{aligned} \quad (5)$$

where $\mathbf{U}_a \in \mathbb{R}^{k \times d}$, $\mathbf{w}_a \in \mathbb{R}^{d+1}$, k is the number of factor units, and a denotes the a -th class. By denoting $\bar{\mathbf{h}} = [\mathbf{h}; 1]$, we can rewrite the decision function of \hat{y}_a in Eq. (5) as follows:

$$\begin{aligned} \hat{y}_a &= \sum_{f=1}^k \left(\sum_{i=1}^d \mathbf{U}_a(f, i) \mathbf{h}(i) \right)^2 + \sum_{i=1}^{d+1} \mathbf{w}_a(i) \bar{\mathbf{h}}(i) \\ &= \sum_{f=1}^k \left(\sum_{i=1}^d \mathbf{U}_a(f, i) \mathbf{h}(i) \right) \left(\sum_{j=1}^d \mathbf{U}_a(f, j) \mathbf{h}(j) \right) + \sum_{i=1}^{d+1} \mathbf{w}_a(i) \bar{\mathbf{h}}(i) \\ &= \sum_{f=1}^k \sum_{i=1}^d \sum_{j=1}^d \mathbf{U}_a(f, i) \mathbf{U}_a(f, j) \mathbf{h}(i) \mathbf{h}(j) + \sum_{i=1}^{d+1} \mathbf{w}_a(i) \bar{\mathbf{h}}(i) \\ &= \sum_{i=1}^d \sum_{j=1}^d \langle \mathbf{U}_a(\cdot, i), \mathbf{U}_a(\cdot, j) \rangle \mathbf{h}(i) \mathbf{h}(j) + \sum_{i=1}^d \mathbf{w}_a(i) \mathbf{h}(i) + \mathbf{w}_a(d+1) \end{aligned} \quad (6)$$

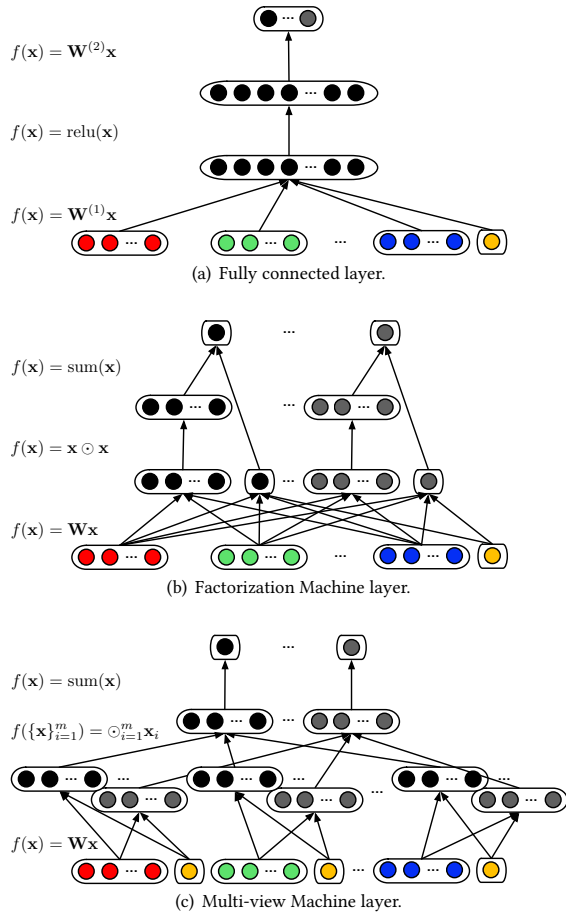


Figure 8: A comparison of different strategies for fusing multi-view data from the perspective of computational graph. Red, green and blue represent features coming from different views, and yellow represents a constant signal “1” which models the bias.

One can easily see that this is similar to the two-way Factorization Machines [39] except that the subscript j ranges from $i + 1$ to d in the original form.

Multi-view Machine layer. In contrast to modeling up to the second-order feature interactions between all input units as in the Factorization Machine layer, we could further explore all feature interactions up to the m -th order between inputs from m views as shown in Figure 8(c).

$$\begin{aligned} \mathbf{q}_a^{(p)} &= \mathbf{U}_a^{(p)} [\mathbf{h}^{(p)}; 1] \\ \hat{\mathbf{y}}_a &= \text{sum}([\mathbf{q}_a^{(1)} \odot \dots \odot \mathbf{q}_a^{(m)}]) \end{aligned} \quad (7)$$

where $\mathbf{U}_a^{(p)} \in \mathbb{R}^{k \times (d_h+1)}$ is the factor matrix of the p -th view for the a -th class. By denoting $\bar{\mathbf{h}}^{(p)} = [\mathbf{h}^{(p)}; 1]$, $p = 1, \dots, m$, we can

verify that Eq. (7) is equivalent to Multi-view Machines [9].

$$\begin{aligned} \hat{\mathbf{y}}_a &= \sum_{f=1}^k \prod_{p=1}^m \left(\sum_{i_p=1}^{d_h+1} \mathbf{U}_a^{(p)}(f, i_p) \bar{\mathbf{h}}^{(p)}(i_p) \right) \\ &= \sum_{f=1}^k \sum_{i_1=1}^{d_h+1} \dots \sum_{i_m=1}^{d_h+1} \left(\prod_{p=1}^m \mathbf{U}_a^{(p)}(f, i_p) \bar{\mathbf{h}}^{(p)}(i_p) \right) \\ &= \sum_{i_1=1}^{d_h+1} \dots \sum_{i_m=1}^{d_h+1} \left(\sum_{f=1}^k \prod_{p=1}^m \mathbf{U}_a^{(p)}(f, i_p) \right) \left(\prod_{p=1}^m \bar{\mathbf{h}}^{(p)}(i_p) \right) \end{aligned} \quad (8)$$

As shown in Figure 2, the full-order feature interactions across multiple views are modeled in a tensor, and they are factorized in a collective manner.

Note that a dropout layer [22] is applied before feeding the output from GRU to the fusion layer which is a regularization method designed to prevent co-adaptation of feature detectors in deep neural networks. The dropout method randomly sets each unit as zero with a certain probability. The dropout units contribute to neither the feed-forward process nor the back-propagation process.

Following the computational graph, it is straightforward to compute gradients for model parameters in both the Factorization Machine layer and the Multi-view Machine layer, as we do for the conventional fully connected layer. Therefore, the error messages generated from the loss function on the final mood score can be back-propagated through these fusion layers all the way to the very beginning, *i.e.*, \mathbf{W}_r , \mathbf{U}_r , \mathbf{W}_z , \mathbf{U}_z , \mathbf{W} , \mathbf{U} in GRU for each input view. In this manner, we can say that DeepMood is an end-to-end learning framework for mood detection.

4 EXPERIMENTS

We investigate a session-level prediction problem. That is to say, we use features of alphanumeric characters, special characters and accelerometer values in a session to predict the mood score of the associated participant.

4.1 Experimental Setup

The implementation is completed using Keras [13] with Tensorflow [17] as the backend. The code has been made available at the author’s homepage⁵. Specifically, a bidirectional GRU is applied on each view of the metadata. RMSProp [45] is used as the optimizer. We truncate sessions that contain more than 10 keypresses, and we remove sessions if any of their views contain less than 10 keypresses. It leaves us with 14,613 total samples which are then split by time for training and validation. Each user contributes first 80% of her sessions for training and the rest for validation. We empirically set other parameters, including the number of epochs, batch size, learning rate and dropout fraction. The number of recurrent units and factor units are selected on the validation set. Detailed configurations of the hyper-parameters are summarized in Table 2.

Experiments on the depression score HDRS are conducted as a binary classification task where $c = 2$. We consider sessions with the HDRS score between 0 and 7 (inclusive) as negative samples (normal) and those with HDRS greater than or equal to 8 as positive samples (from mild depression to severe depression). On the other hand, the mania score YMRS is more complicated without a widely

⁵<https://www.cs.uic.edu/~bcao1/code/DeepMood.py>

Table 2: Parameter configuration.

Parameter	Value
# recurrent units (d_h)	4, 8, 16
# factor units (k)	4, 8, 16
# epochs	500
batch size	256
learning rate	0.001
dropout fraction	0.1
maximum sequence length	100
minimum sequence length	10

adopted threshold. Therefore, YMRS is directly used as the label for a regression task where $c = 1$. Accuracy and F-score are used to evaluate the classification task, and root-mean-square error (RMSE) is used for the regression task.

4.2 Compared Methods

The compared methods are summarized as follows:

- **DMVM**: The proposed DeepMood architecture with a Multi-view Machine layer for data fusion.
- **DFM**: The proposed DeepMood architecture with a Factorization Machine layer for data fusion.
- **DNN**: The proposed DeepMood architecture with a conventional fully connected layer for data fusion.
- **XGB**: The implementation of a tree boosting system from XGBoost⁶ [10] is used. We concatenate the sequence data with the maximum length 100 (padding 0 for short ones) of each feature as the input.
- **SVM** and **LR**: These are two linear models. With the same input setting as XGB, the implementations of Linear Support Vector Classification/Regression and Logistic/Ridge Regression from `scikit-learn`⁷ are used for Classification/Regression tasks.

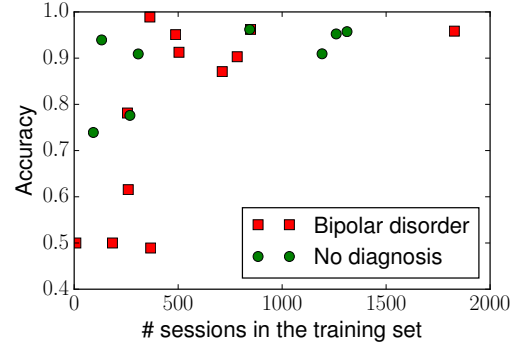
In general, DMVM, DFM and DNN can be categorized as late fusion approaches, while XGB, SVM and LR are early fusion strategies for the sequence prediction problem on multi-view time series. Note that the number of model parameters for fusing multi-view data in DMVM and DFM is $ck(d + m)$ and $ckd + c(d + 1)$, respectively, thereby leading to approximately the same model complexity $O(ckd)$ due to $m \ll d$. For DNN, the number of model parameters for fusion is $ck' + k'(d + 1)$. For a fair comparison, we need to control the model complexity of the compared methods at the same level. Therefore, in all experiments, we always set $k' = ck$.

4.3 Prediction Performance

Experimental results are shown in Table 3. We can see that the late fusion based DeepMood methods are the best on the prediction for the dichotomized HDRS scores, especially DMVM and DFM with 90.31% and 90.21%, respectively. It demonstrates the feasibility of using passive typing dynamics from mobile phone metadata to predict the disturbance and severity of mood states. In addition, it is found that SVM and LR are not a good fit to this task, or sequence prediction in general. XGB performs reasonably well as an ensemble

Table 3: Prediction performance of compared methods.

Task	Classification		Regression
	Accuracy	F-score	RMSE
DMVM	0.9031	0.9070	3.5664
DFM	0.9021	0.9029	3.6767
DNN	0.8868	0.8929	3.7874
XGB	0.8555	0.8562	3.9634
SVM	0.7323	0.7237	4.1257
LR	0.7293	0.7172	4.1822

**Figure 9: Prediction performance of DMVM on each of the 20 participants.**

method, but DMVM still outperforms it by a significant margin 5.56%, 5.93% and 10.02% in terms of accuracy, F-score and RMSE, respectively. Among the DeepMood variations, the improvement of DMVM and DFM over DNN reveals the potential of replacing a conventional fully connected layer with a Multi-view Machine layer or Factorization Machine layer for data fusion in a deep framework. This is because DMVM and DFM can explicitly capture higher order interactions among features, while DNN does not capture any feature interaction.

In practice, it is important to understand how the model works for each individual when monitoring her mood states. Therefore, we investigate the prediction performance of DMVM on each of the 20 participants in our dataset. Results are shown in Figure 9 where each dot represents a participant with the number of her contributed sessions in the training set and the corresponding prediction accuracy. We can see that the proposed model can steadily produce accurate predictions ($\geq 87\%$) of a participant's mood states when she provides more than 400 valid typing sessions in the training phase. Note that the prediction we make in this work is per session which is typically less than one minute. We can expect more accurate results on the daily level by ensembling sessions occurring during a day.

4.4 Convergence Efficiency

In this section, we show more details about the learning procedure of the proposed DeepMood architecture with different fusion layers and that of XGB. Figure 10 illustrates how the accuracy on the validation set changes over epochs. We observe that different fusion layers have different convergence performance in the first 300

⁶<https://github.com/dmlc/xgboost>

⁷<http://scikit-learn.org>

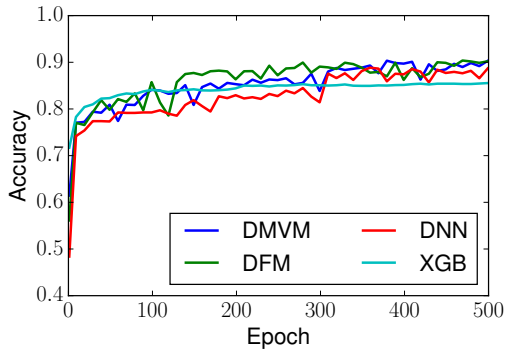


Figure 10: Learning procedure.

Table 4: Prediction performance using different views of the metadata.

Task	Classification		Regression
	Accuracy	F-score	RMSE
DMVM w/o ALPH.	0.8125	0.8164	3.9833
DMVM w/o SPEC.	0.9008	0.9034	3.8166
DMVM w/o ACCEL.	0.8318	0.8253	3.9499
DMVM w/ all	0.9031	0.9070	3.5664
DFM w/ ALPH.	0.8322	0.8224	3.9515
DFM w/ SPEC.	0.6260	0.5676	4.1040
DFM w/ ACCEL.	0.8015	0.8089	3.9722
DFM w/ all	0.9021	0.9011	3.6767

epochs, and afterwards they steadily outperform XGB. Among the DeepMood methods, it is found that DMVM and DFM converge more efficiently than DNN in the first 300 epochs, and they reach a better local minima of the loss function at the end. This again shows the importance of the fusion layer in a deep framework. It is also interesting to see the convergence process of XGB considering its popularity and success on many tasks in practice. We found that the generalizability of XGB on the sequence prediction task is limited, although its training error could perfectly converge to 0 at an early stage.

4.5 Importance of Different Views

To better understand the role that different views play in the buildup of mood detection by DeepMood, we examine separate models trained with or without each view. Since DMVM is designed for heterogeneous data fusion, *i.e.*, data with at least two views, we train DMVM on every pairwise views. Moreover, we train DFM on every single view. Experimental results are shown in Table 4. First, we observe that SPEC. are poor predictors of mood states. ALPH. and ACCEL. have significantly better predictive performance. ALPH. are the best individual predictors of mood states. It validates a high correlation between the mood disturbance and typing patterns including duration of a keypress, time interval since the last keypress, as well as accelerometer values.

5 RELATED WORK

This work is studied in the context of supervised sequence prediction. Xing et al. provide a brief survey on the sequence prediction problem where sequence data are categorized into five subtypes: simple symbolic sequences, complex symbolic sequences, simple time series, multivariate time series, and complex event sequences [50]. Sequence classification methods are grouped into three subtypes: feature based methods, sequence distance based methods, and model based methods. Feature based methods first transform a sequence into a feature vector and then apply conventional classification models [1, 25, 30, 31, 52]. Distance based methods include K nearest neighbor classifier [16, 26, 27, 38, 47, 49] and SVM with local alignment kernel [32, 41, 42] by measuring the similarity between a pair of sequences. Model based methods assume that sequences in a class are generated by an underlying probability distribution, including Naive Bayes [11], Markov Model [51] and Hidden Markov Model [43].

However, most of the works focus on simple symbolic sequences and simple time series, with a few on complex symbolic sequences and multivariate time series. The problem of classifying complex event sequence data (a combination of multiple numerical measurements and categorical fields) still needs further investigation which motivates this work. Furthermore, most of the methods are devoted to shallow models with feature engineering. Inspired by the great success of deep RNNs in the applications of other sequence tasks, including speech recognition [20] and natural language processing [4, 34], in this work, we propose a deep architecture to model complex event sequences of mobile phone typing dynamics.

On multi-view learning, Cao et al. propose to fuse multi-view data through the operation of tensor product and assume that the effects of feature interactions across views have a low rank [7, 9]. Lu et al. extend it to multi-task learning [33]. Zhang et al. use Factorization Machines to initialize the bias terms and embedding vectors for multi-field categorical data at the bottom layer of a deep architecture [55]. There are also some work incorporating multiple views into the process of subgraph mining [8] and deep learning [54] to help identify meaning patterns from data.

6 CONCLUSION

It appears that mobile phone metadata could be used to predict the presence of mood disorders. The proposed DeepMood architecture is able to achieve 90.31% prediction accuracy, where late fusion is indeed more effective than early fusion and more sophisticated fusion layer also helps. The ability to passively collect data that can be used to infer the presence and severity of mood disturbances may enable providers to provide interventions to more patients earlier in their mood episodes. Models such as the one presented here may also lead to deeper understanding of the effects of mood disturbances in the daily activities of people with mood disorders.

7 ACKNOWLEDGEMENTS

This work is supported in part by NSF through grants IIS-1526499, and CNS-1626432, and NSFC 61672313.

REFERENCES

- [1] Charu C Aggarwal. 2002. On effective classification of strings with wavelets. In *KDD*. ACM, 163–172.
- [2] David Ankers and Steven H Jones. 2009. Objective assessment of circadian activity and sleep patterns in individuals at behavioural risk of hypomania. *Journal of clinical psychology* 65, 10 (2009), 1071–1086.
- [3] American Psychiatric Association and others. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.
- [4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [5] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* 5, 2 (1994), 157–166.
- [6] Jedediah M Bopp, David J Miklowitz, Guy M Goodwin, Will Stevens, Jennifer M Rendell, and John R Geddes. 2010. The longitudinal course of bipolar disorder as revealed through weekly text messaging: a feasibility study. *Bipolar disorders* 12, 3 (2010), 327–334.
- [7] Bokai Cao, Lifang He, Xiangnan Kong, Philip S Yu, Zhifeng Hao, and Ann B Ragin. 2014. Tensor-based Multi-view Feature Selection with Applications to Brain Diseases. In *ICDM*.
- [8] Bokai Cao, Xiangnan Kong, Jingyuan Zhang, Philip S Yu, and Ann B Ragin. 2015. Mining Brain Networks using Multiple Side Views for Neurological Disorder Identification. In *ICDM*.
- [9] Bokai Cao, Hucheng Zhou, Guoqiang Li, and Philip S Yu. 2016. Multi-view Machines. In *WSDM*. ACM, 427–436.
- [10] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *KDD*. ACM.
- [11] Betty Yee Man Cheng, Jaime G Carbonell, and Judith Klein-Seetharaman. 2005. Protein classification based on text document classification techniques. *Proteins: Structure, Function, and Bioinformatics* 58, 4 (2005), 955–970.
- [12] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
- [13] François Chollet. 2015. Keras. <https://github.com/fchollet/keras>. (2015).
- [14] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
- [15] Mark A Demitrack, Doug Faries, John M Herrera, David J DeBrot, and William Z Potter. 1998. The problem of measurement error in multisite clinical trials. *Psychopharmacology bulletin* 34, 1 (1998), 19.
- [16] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. *VLDB* 1, 2 (2008), 1542–1552.
- [17] Martin Abadi et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015). <http://tensorflow.org/> Software available from tensorflow.org.
- [18] Maria Faurholt-Jepsen, Maj Vinberg, Mads Frost, Sune Debel, Ellen Margrethe Christensen, Jakob E Bardram, and Lars Vedel Kessing. 2016. Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. *International journal of methods in psychiatric research* 25, 4 (2016), 309–323.
- [19] Mads Frost, Afsaneh Doryab, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E Bardram. 2013. Supporting disease insight through data analysis: refinements of the monarca self-assessment system. In *UBICOMP*. ACM, 133–142.
- [20] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *ICASSP*. IEEE, 6645–6649.
- [21] Agnes Gruenerbl, Venet Osmani, Gernot Bahle, Jose C Carrasco, Stefan Oehler, Oscar Mayora, Christian Haring, and Paul Lukowicz. 2014. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In *AH*. ACM, 38.
- [22] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).
- [23] Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6, 02 (1998), 107–116.
- [24] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [25] Xiaonan Ji, James Bailey, and Guozhu Dong. 2007. Mining minimal distinguishing subsequence patterns with gap constraints. *Knowledge and Information Systems* 11, 3 (2007), 259–286.
- [26] Eamonn Keogh and Shrutti Kasetty. 2003. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery* 7, 4 (2003), 349–371.
- [27] Eamonn J Keogh and Michael J Pazzani. 2000. Scaling up dynamic time warping for datamining applications. In *KDD*. ACM, 285–289.
- [28] Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Kathleen R Merikangas, and Ellen E Walters. 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry* 62, 6 (2005), 593–602.
- [29] Kiran E Laxman, Kate S Lovibond, and Mariam K Hassan. 2008. Impact of bipolar disorder in employed populations. *The American journal of managed care* 14, 11 (2008), 757–764.
- [30] Neal Lesh, Mohammed J Zaki, and Mitsunori Ogiwara. 1999. Mining features for sequence classification. In *KDD*. ACM, 342–346.
- [31] Christina Leslie and Rui Kuang. 2004. Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research* 5, Nov (2004), 1435–1455.
- [32] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research* 2, Feb (2002), 419–444.
- [33] Chun-Ta Lu, Lifang He, Weixiang Shao, Bokai Cao, and Philip S Yu. 2017. Multi-linear Factorization Machines for Multi-Task Multi-View Learning. In *WSDM*.
- [34] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, Vol. 2. 3.
- [35] Pamela B Peele, Ying Xu, and David J Kupfer. 2003. Insurance expenditures on bipolar disorder: clinical and parity implications. *American Journal of Psychiatry* 160, 7 (2003), 1286–1290.
- [36] Bruce M Psaty and Ross L Prentice. 2010. Minimizing bias in randomized trials: the importance of blinding. *Jama* 304, 7 (2010), 793–794.
- [37] Alessandro Puiatti, Steven Mudda, Silvia Giordano, and Oscar Mayora. 2011. Smartphone-centred wearable sensors network for monitoring patients with bipolar disorder. In *EMBC*. IEEE, 3644–3647.
- [38] Chotirat Ann Ratanamahatana and Eamonn Keogh. 2004. Making Time-series Classification More Accurate Using Learned Constraints. In *SDM*. SIAM, 11.
- [39] Steffen Rendle. 2012. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology* 3, 3 (2012), 57.
- [40] O Schlessing, Ph Renevey, M Bertschi, J-M Koller, R Paradiso, and others. 2011. Monitoring physiological and behavioral signals to detect mood changes of bipolar patients. In *ISMICT*. IEEE, 130–134.
- [41] Rong She, Fei Chen, Ke Wang, Martin Ester, Jennifer L Gardy, and Fiona SL Brinkman. 2003. Frequent-subsequence-based prediction of outer membrane proteins. In *KDD*. ACM, 436–445.
- [42] Sören Sonnenburg, Gunnar Rätsch, and Bernhard Schölkopf. 2005. Large scale genomic sequence SVM classifiers. In *ICML*. ACM, 848–855.
- [43] Prashant K Srivastava, Dhvani K Desai, Soumyadeep Nandi, and Andrew M Lynn. 2007. HMM-ModE-Improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. *BMC bioinformatics* 8, 1 (2007), 1.
- [44] Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *ICML*. 1017–1024.
- [45] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning* 4, 2 (2012).
- [46] Gaetano Valenza, Mimma Nardelli, Antonio Lanata, Claudio Gentili, Gilles Bertschy, Rita Paradiso, and Enzo Pasquale Scilingo. 2014. Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis. *IEEE Journal of Biomedical and Health Informatics* 18, 5 (2014), 1625–1635.
- [47] Li Wei and Eamonn Keogh. 2006. Semi-supervised time series classification. In *KDD*. ACM, 748–753.
- [48] Janet BW Williams. 1988. A structured interview guide for the Hamilton Depression Rating Scale. *Archives of general psychiatry* 45, 8 (1988), 742–747.
- [49] Xiaopeng Xi, Eamonn Keogh, Christian Shelton, Li Wei, and Chotirat Ann Ratanamahatana. 2006. Fast time series classification using numerosity reduction. In *ICML*. ACM, 1033–1040.
- [50] Zhengzheng Xing, Jian Pei, and Eamonn Keogh. 2010. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter* 12, 1 (2010), 40–48.
- [51] Oksana Yakhnenko, Adrian Silvescu, and Vasant Honavar. 2005. Discriminatively trained markov model for sequence classification. In *ICDM*. IEEE, 8–pp.
- [52] Lixiang Ye and Eamonn Keogh. 2009. Time series shapelets: a new primitive for data mining. In *KDD*. ACM, 947–956.
- [53] RC Young, JT Biggs, VE Ziegler, and DA Meyer. 1978. A rating scale for mania: reliability, validity and sensitivity. *The British Journal of Psychiatry* 133, 5 (1978), 429–435.
- [54] Jingyuan Zhang, Bokai Cao, Sihong Xie, Chun-Ta Lu, Philip S Yu, and Ann B Ragin. 2016. Identifying Connectivity Patterns for Brain Diseases via Multi-side-view Guided Deep Architectures. In *SDM*.
- [55] Weinan Zhang, Tianming Du, and Jun Wang. 2016. Deep Learning over Multi-field Categorical Data. In *ICIR*. Springer, 45–57.