6-2020

Self-trained deep ordinal regression for end-to-end video anomaly detection

Guansong PANG
Singapore Management University, gspang@smu.edu.sg

Cheng YAN

Chunhua SHEN

Anton Van Den HENGEL

Xiao BAI

Deep Multi-task Learning for Depression Detection and Prediction in Longitudinal Data

Guansong Pang[1], **Ngoc Thien Anh Pham**[1], **Emma Baker**[1],
Rebecca Bentley[2], **Anton van den Hengel**[1]

[1] University of Adelaide, Adelaide SA 5005, Australia
[2] University of Melbourne, Parkville VIC 3010, Australia

Abstract

Depression is among the most prevalent mental disorders, affecting millions of people of all ages globally. Machine learning techniques have shown effective in enabling automated detection and prediction of depression for early intervention and treatment. However, they are challenged by the relative scarcity of instances of depression in the data. In this work we introduce a novel deep multi-task recurrent neural network to tackle this challenge, in which depression classification is jointly optimized with two auxiliary tasks, namely one-class metric learning and anomaly ranking. The auxiliary tasks introduce an inductive bias that improves the classification model's generalizability on small depression samples. Further, unlike existing studies that focus on learning depression signs from static data without considering temporal dynamics, we focus on longitudinal data because i) temporal changes in personal development and family environment can provide critical cues for psychiatric disorders and ii) it may enable us to predict depression before the illness actually occurs. Extensive experimental results on child depression data show that our model is able to i) achieve nearly perfect performance in depression detection and ii) accurately predict depression 2-4 years before the clinical diagnosis, substantially outperforming seven competing methods.

Introduction

Major Depressive Disorder (MDD), widely known as depression in the public, is a mental disorder characterized by a severe and persistent feeling of sadness, loss of interest in activities, or a sense of despair, causing significant impairment in daily life (Lamers et al. 2019). Globally over 300 million people of all ages are estimated to suffer from depression; depression is one major contributor to nearly 800 thousands suicide deaths per year (WHO 2017). Despite depression is among the most prevalent mental disorders, clinical diagnosis is difficult because i) clinical and self-assessment reports are highly dependent on specialist's expertise and the diagnosis methods used, involving a range of subjective ratings and potential biases, and ii) depression manifests itself in varying ways for different people and conditions. To prevent the huge health loss, it is highly desired to develop automated detection and prediction[1] approaches with objective assessment to complement clinical diagnosis.

Machine learning techniques have shown effective in enabling the automated detection and prediction of depression (Cohn et al. 2009; Yang, Fairbairn, and Cohn 2012; Nasir et al. 2016; Gong and Poellabauer 2017; Zhu et al. 2017; Zhou et al. 2018; Shen et al. 2018; Ay et al. 2019; Devlin et al. 2019; Uddin, Joolee, and Lee 2020; Lin et al. 2020). These approaches, especially deep learning approaches, normally require large labeled depression and non-depression samples to learn desired classification models, but in practice only small labeled depression samples are available as it is very difficult, if not impossible, to collect large depression samples. For instance, only 100-300 samples in total are available for modeling in most existing studies (Yang, Fairbairn, and Cohn 2012; Nasir et al. 2016; Gong and Poellabauer 2017; Zhou et al. 2018; Devlin et al. 2019; Uddin, Joolee, and Lee 2020; Chancellor and De Choudhury 2020). Training models with small samples can lead to overfitting, risking misdiagnosis of unseen depression cases.

In this work we introduce a novel deep multi-task recurrent neural network approach to tackle this challenge, in which depression classification is jointly optimized with two related tasks, one-class metric learning and anomaly ranking. Multi-task learning (Zhang and Yang 2017; Ruder 2017) improves generalization by introducing inductive biases contained in the related tasks to regularize the models; it is one of the most effective approaches to reduce overfitting. For depression classification, our two auxiliary tasks are devised to learn compact feature representations of normal samples and allow some variations in the representations of depression samples, enabling the detection of unseen depression cases that are clearly deviated from the normal samples in the representation space.

Further, existing studies (Yang, Fairbairn, and Cohn 2012; Nasir et al. 2016; Gong and Poellabauer 2017; Zhu et al. 2017; Shen et al. 2017; Zhou et al. 2018; Cai et al. 2018; Ay et al. 2019; Devlin et al. 2019; Chancellor and De Choudhury 2020; Uddin, Joolee, and Lee 2020) mainly focus on learning depression signs from static data without considering temporal dynamics. As a result, they can detect the depression cases often only when there are some clear depression symptoms presented. To enable very early intervention and prevention of the depression, we instead exploit lon-

[1] Here depression detection refers to the identification of subjects who already had depressive disorders in a set of samples, while prediction aims to identify subjects who currently do not have depression but are predicted to have in the near future.

gitudinal data, motivating by two key observations: i) temporal changes in personal development and family environment can provide critical cues for psychiatric disorders and ii) these cues may enable us to achieve more accurate depression detection and to predict depression well before the illness actually occurs.

In summary, this work makes two key contributions:

- We introduce a novel multi-task learning framework for depression classification. The two auxiliary tasks, one-class metric learning and anomaly ranking tasks, introduce an inductive bias that improves the classification model's generalizability to unseen depression cases.
- We further instantiate the framework into a multi-task recurrent neural network-based model, termed MTNet. MTNet effectively leverages longitudinal data to learn critical temporal-dependent depression cues. This is verified by extensive empirical results on child depression data. The results show that MTNet achieves nearly perfect performance in depression detection; it can accurately predict depression 2-4 years before the clinical diagnosis; and MTNet substantially outperforms seven competing non-temporal and temporal models in both tasks.

Related Work

Longitudinal Studies Related longitudinal studies focus on identifying factors or predictors that are associated with depression. They are based on structured questionnaire data from clinical interviews in varying waves of follow-up studies. The data may include a variety information of participated subjects, such as demographic information, socio-economic features, individual growth, family environment, etc. (Rushton, Forcier, and Schectman 2002; Zuckerbrot and Jensen 2006; Feng et al. 2009; Kessler et al. 2012; Henry et al. 2018; Korsten et al. 2019). Traditional statistic models like χ^2-based bivariate analysis or logistic regression-based multivariate analysis are often used for the association discovery. By contrast, our study is on learning detection and prediction models. Although the traditional methods can also be used for detection/prediction, they are ineffective to model the underlying complex temporal-dependent features. Further, the limited amount of labeled data presents another major challenge to these traditional methods.

Detection and Prediction Another group of studies use machine learning to detect/predict depression using vocal/visual data taken during clinical interviews, or online social media data. The vocal features used include different prosodic, cepstral and spectral features (Cohn et al. 2009; Low et al. 2010a,b; Yang, Fairbairn, and Cohn 2012; Nasir et al. 2016; Gong and Poellabauer 2017). Visual features can be based on facial expression, gaze direction, position and orientation of the head, facial geometric features, eye movement (Suslow, Junghanns, and Arolt 2001; Cohn et al. 2009; Alghowinem et al. 2013; Nasir et al. 2016; Gong and Poellabauer 2017; Zhu et al. 2017; Zhou et al. 2018; Uddin, Joolee, and Lee 2020). The online features can be words in tweets or posts, language style, relevance to pre-defined topics, sentiments of tweets, engagement activities (Wang et al. 2013; Gong and Poellabauer 2017; Yang et al. 2017; Orabi et al. 2018; Devlin et al. 2019; Chancellor and De Choudhury 2020; Mann, Paes, and Matsushima 2020; Lin et al. 2020). A comprehensive way to harvest social media data is explored in (Shen et al. 2017, 2018) to fuse all these features together. In recent years electroencephalogram (EEG) data is also found to be useful for depression detection (Acharya et al. 2018; Cai et al. 2018; Ay et al. 2019).

In terms of depression classification model, support vector machines, logistic regression, and multi-layer perceptrons neural networks are commonly-used models (Cohn et al. 2009; Nasir et al. 2016; Gong and Poellabauer 2017; Cai et al. 2018). In recent years deep learning models, such as convolutional neural networks (CNN) (Yang et al. 2017; Rodrigues Makiuchi et al. 2019), long-short term memory (LSTM) networks (Orabi et al. 2018), or their combination CNN-LSTM (Ma et al. 2016), are explored for depression detection. These models are leveraged to automatically learn semantically rich feature representations from text, audio or visual data. However, these studies learn the models using static multimedia data, without considering the temporal changes of the subjects in different time periods.

Multi-task learning has led to many successes in a wide range of applications (Zhang and Yang 2017; Ruder 2017), including depression detection (Chao et al. 2015; Lu et al. 2018). Our work is fundamentally different from these two studies in: i) they focus on leverage multiple heterogeneous data sources, *e.g.*, audio and video data in (Chao et al. 2015), smartphone, wrist bands, and self-report data in (Lu et al. 2018), whereas we require a single data source only; and ii) they rely on additional correlated labeled data, *e.g.*, emotion data (Chao et al. 2015), self-evaluation depression results and depression severity level (Lu et al. 2018), whereas our model works well with only the binary depression labels. Our solution is thus more practical and easy-to-use in practice. Further, as far as we know, jointly optimizing classification with one-class metric learning and anomaly ranking is novel in the multi-task learning literature.

Multi-task Recurrent Neural Networks on Longitudinal Data

The Proposed Multi-task Learning: An Overview

Depression detection/prediction aims to learn a binary depression classification mapping function $\phi : \mathcal{X} \mapsto \mathcal{Y}$, where $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_N\}$ is a set of longitudinal data of N samples, and $\mathcal{Y} = \{0, 1\}$ is the output space, with '1' indicating the subject having depression and '0' otherwise. $\mathbf{X} \in \mathbb{R}^{w \times D}$, *i.e.*, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_w\}$, is a matrix input of an individual subject, where $\mathbf{x}_t \in \mathbb{R}^D$ is a feature vector derived from the t-th wave of questionnaire data.

Motivated by the fact that it is difficult to collect large depression samples, we introduce a novel multi-task learning framework to improve the generalizability of depression classification models on small depression samples. An overview of the approach is presented in Figure 1(a). Depression classification is our primary task and is jointly optimized with two auxiliary tasks, including anomaly ranking and one-class metric learning. The auxiliary tasks treat

depression samples as anomalies and enforce compact feature representations of normal samples and allow some variations in the representations of depression samples, serving as a regularizer of the classification model. This results in better generalized classification models than that in the single primary task. Formally, let $\tau : \mathcal{X} \mapsto \mathbb{R}$ be an anomaly ranking function that assigns an anomaly score to each subject; $\psi : \mathcal{X} \mapsto \mathcal{Q}$ be the one-class metric learning function, where $\mathcal{Q} \in \mathbb{R}^M$ with $M \ll D$ is a new feature space, then our overall objective function can be given as follows.

$$\arg\min_{\Theta_e,\Theta_a,\Theta_o} \sum_{i=1}^{N} \Big[\ell_e\big(\phi(\mathbf{X}_i;\Theta_e), y_i\big) + \alpha \ell_a\big(\tau(\mathbf{X}_i;\Theta_a), y_i\big) + \beta \ell_o\big(\psi(\mathbf{X}_i;\Theta_o), y_i\big) \Big], \quad (1)$$

where ℓ_e, ℓ_a and ℓ_o are respective loss functions for depression classification, anomaly ranking and one-class metric learning, y_i is the class label of $\mathbf{X}_i$, $\Theta = \{\Theta_e, \Theta_a, \Theta_o\}$ is the set of parameters to be learned, α and β are hyperparameters to control the importance of the two auxiliary tasks.

We instantiate the framework into a model called MTNet that leverages a shared LSTM neural network (Hochreiter and Schmidhuber 1997) to learn critical temporal changes in longitudinal data for depression classification. Supervised anomaly deviation and one-class support vector data description loss functions are defined to improve the model's generalization. A simple data augmentation method is also introduced to further enhance the generalizability. The details of each module of MTNet are presented as follows.

Primary Task: Deep Depression Classification

Our classification model leverages a LSTM neural network layer to learn important temporal dependency in the longitudinal data. A LSTM layer consists of w LSTM cells, with each LSTM cell learning temporal-dependent representations of the input data at a specific wave (or time step). An internal structure of a LSTM cell is shown in Figure 1(b). A LSTM cell is composed by a memory cell and three gates, including forget, input, and output gates. The forget gate $\mathbf{f}_t$ controls how much information we throw away from the cell state $\mathbf{c}_{t-1}$ obtained at wave $t-1$, the input gate $\mathbf{i}_t$ decides the extent to which the new input at the current wave t flows into the cell, while the output gate $\mathbf{o}_t$ controls what information we want to output. The memory cell $\mathbf{c}_t$ keeps track of the dependencies among the multi-time-step inputs. The full operations within a basic LSTM cell are defined as follows.

$$\mathbf{f}_t = g_r(\mathbf{W}_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f), \quad (2)$$
$$\mathbf{i}_t = g_r(\mathbf{W}_i \mathbf{x}_t + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{b}_i), \quad (3)$$
$$\mathbf{o}_t = g_r(\mathbf{W}_o \mathbf{x}_t + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{b}_o), \quad (4)$$
$$\hat{\mathbf{c}}_t = \tanh(\mathbf{W}_c \mathbf{x}_t + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{b}_c), \quad (5)$$
$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \hat{\mathbf{c}}_t, \quad (6)$$
$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh(\mathbf{c}_t), \quad (7)$$

where $\mathbf{x}_t$ is the input feature vector at the t-th time step (*i.e.*, t-th wave questionnaire); $\mathbf{f}_t$, $\mathbf{i}_t$, and $\mathbf{o}_t$ are L-dimensional activation vectors that control the information flow within and between the LSTM cells; $\mathbf{h}_t \in \mathbb{R}^L$ is the output vector; $\hat{\mathbf{c}}_t$ and $\mathbf{c}_t$ are respectively the cell input activation vector and cell state vector; matrices $\mathbf{W}_* \in \mathbb{R}^{L \times D}$ and $\mathbf{U}_* \in \mathbb{R}^{L \times L}$ are the weight parameters of the input and recurrent network connections w.r.t. the three gates $\mathbf{f}_t$, $\mathbf{i}_t$, $\mathbf{o}_t$ and the memory cell $\mathbf{c}$; $\mathbf{b}_* \in \mathbb{R}^L$ are the parameters of the bias term; g_r is an activation function; '$\circ$' denotes the Hadamard product (element-wise product). $\mathbf{W}_*$, $\mathbf{U}_*$, and $\mathbf{b}_*$ need to be learned.

The full LSTM layer uses the recurrent LSTM cells to encode important temporal changes across all different questionnaire waves into the output vector $\mathbf{h} \in \mathbb{R}^L$ in the last LSTM cell. To learn more expressive representations, a FC layer is further used to project $\mathbf{h}$ onto a lower-dimensional feature representation space:

$$\mathbf{q} = g_s(\mathbf{W}_s \mathbf{h} + \mathbf{b}_s), \quad (8)$$

where $\mathbf{W}_s \in \mathbb{R}^{M \times L}$ and $\mathbf{b}_s \in \mathbb{R}^M$ are the learnable parameters, g_s is an activation function, and $\mathbf{q} \in \mathcal{Q}$ is the final feature representation of $\mathbf{X}$. Since Eqs. (2-8) are all differentiable, for brevity, they can be represented by a function ψ that is the composite function formed by Eqs. (2-8),

$$\mathbf{q} = \psi(\mathbf{X}; \Theta_r), \quad (9)$$

where Θ_r contains all the learnable parameters (*i.e.*, $\mathbf{W}_*$, $\mathbf{U}_*$, and $\mathbf{b}_*$) in the LSTM layer and the FC layer.

We then train a classifier on the $\mathcal{Q}$ representation space with a standard binary cross-entropy loss function:

$$\ell_e\big(\phi(\mathbf{X};\Theta_e), y\big) = -\big(y \log(p) + (1-y)\log(1-p)\big), \quad (10)$$

where y is the class label of $\mathbf{X}$ and

$$p = \phi(\mathbf{X};\Theta_e) = g_e(\mathbf{W}_e \psi(\mathbf{X};\Theta_r) + b_e). \quad (11)$$

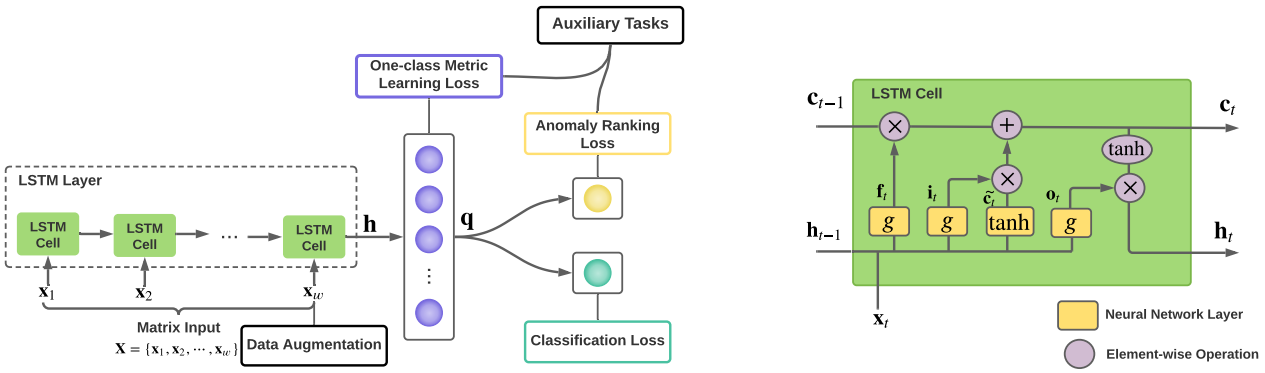
g_e is a sigmoid activation function; $\Theta_e = \{\Theta_r, \mathbf{W}_e, b_e\}$ are the parameters, with $\mathbf{W}_e \in \mathbb{R}^{1 \times M}$ and $b_e \in \mathbb{R}$. All parameters in Θ_e can be optimized in an end-to-end manner.

As shown in Figure 1(a), the LSTM and FC layers are shared by all three tasks. Thus, the feature representation function ψ is jointly optimized with the two auxiliary tasks.

Auxiliary Tasks

The cause of different depression cases can vary significantly from each other, leading to highly different feature expressions in the longitudinal data. Further, the depression cases available for training are often limited; consequently, they do not span the entire set of all possible depression cases. A single classification task can therefore fail to recognize novel depression cases that are not covered by the limited depression data. Motivated by these observations, two auxiliary tasks, anomaly ranking and one-class metric learning, are incorporated to introduce an inductive bias that prefers compact feature representations of normal samples and allows some variations in the representations of depression samples. They effectively regularize the classification model, preventing overfitting of small depression samples.

Anomaly Ranking A partially-supervised anomaly ranking task is introduced to enforce the model to assign significantly larger anomaly scores for depression samples than that of non-depression samples. Motivated by (Pang, Shen,



(a) The Proposed Multi-task Neural Network Approach

(b) An Overview of a LSTM Cell in the LSTM Layer

Figure 1: An Overview of the Proposed Multi-task Learning Framework. As shown in (a), our approach jointly optimizes depression classification with one-class metric learning and anomaly ranking tasks to learn well generalized models with small depression samples. Each subject sample is a matrix input $\mathbf{X} \in \mathbb{R}^{w \times D}$, *i.e.*, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_w\}$, where $\mathbf{x}_t \in \mathbb{R}^D$ is a feature vector derived from the t-th wave of questionnaire data. MTNet leverages a LSTM layer with w basic LSTM cells (see (b) for the internal of a LSTM cell) to capture temporal dependency in $\mathbf{X}$ across different waves, followed by a fully-connected (FC) layer and an output layer with two independent heads for depression classification and anomaly ranking respectively. The one-class metric learning is applied to the FC layer. A simple yet effective data augmentation is applied to the last wave of data.

and van den Hengel 2019), a prior-driven anomaly ranking loss function, called deviation loss, is used to fulfill this goal. Particularly, a Gaussian prior $\mathcal{N}(\mu, \sigma^2)$ is imposed on the anomaly scores of all samples, which posits that the anomaly scores of non-depression samples are centered around a Gaussian mean value μ while the anomaly scores of depression samples have at least $a\sigma$ deviations from μ. Note that this prior is imposed on the anomaly scores rather than the feature representations, enabling better variability in optimizing the classification model.

Formally, we add another network output head with one linear unit to learn an anomaly score for each sample:

$$\tau(\mathbf{X}; \Theta_a) = \mathbf{W}_a \psi(\mathbf{X}; \Theta_r) + b_a, \tag{12}$$

where $\Theta_a = \{\Theta_r, \mathbf{W}_a, b_a\}$ are the parameters to be learned. We then define the deviation using the well-known Z-Score:

$$dev(\mathbf{X}) = \frac{\tau(\mathbf{X}; \Theta_a) - \mu}{\sigma}, \tag{13}$$

The deviation function is then plugged into the widely-used contrastive loss (Hadsell, Chopra, and LeCun 2006) to define our anomaly ranking loss function:

$$\ell_a\big(\tau(\mathbf{X}; \Theta_a), y\big) = (1-y)|dev(\mathbf{X})| + y \max\big(0, a - dev(\mathbf{X})\big). \tag{14}$$

By minimizing ℓ_a, our model pushes the anomaly scores of non-depression samples as close as possible to μ while enforcing at least $a\sigma$ between μ and the anomaly scores of depression samples in the upper tail of the Gaussian distribution. Following (Pang, Shen, and van den Hengel 2019), the prior $\mathcal{N}(0, 1)$ is used with $a = 5$ to guarantee significant deviations of depression samples from normal samples.

One-class Metric Learning Unlike the anomaly ranking task that introduces the inductive bias using an output layer independent from the classification output, the one-class metric learning task exerts itself directly on the feature layer. Both auxiliary tasks thus enable complementary constraints at the output level and the feature representation level.

One-class learning aims at learning a compact one-class data description of given data samples without any class labels. We adapt it to our application by using depression samples as anomalies to learn more compact one-class representations of non-depression samples. Specifically, we enhance the popular one-class model, support vector data description (SVDD) (Tax and Duin 2004), by adding a large margin between the one-class samples and the presumed anomalies. The loss function is defined as

$$\ell_o\big(\psi(\mathbf{X}; \Theta_o), y\big) = (1-y)||\psi(\mathbf{X}; \Theta_r) - \mathbf{n}||_2 + y \max\big(0, m - ||\psi(\mathbf{X}; \Theta_r) - \mathbf{n}||_2\big), \tag{15}$$

where $\Theta_o = \{\Theta_r\}$, $\mathbf{n} \in \mathbb{R}^M$ is the one-class center vector of normal samples and m is a hyperparameter to control the margin. This loss enforces a large margin between non-depression and depression samples in the ψ-induced representation space while minimizing the $\mathbf{n}$-centered hypersphere's volume. We found empirically that MTNet can perform well with varying settings of $\mathbf{n}$, *e.g.*, $\mathbf{n} \sim \mathcal{N}(0, 1)$ or $\mathbf{n} \sim \mathcal{U}(0, 1)$. We use $\mathbf{n} \sim \mathcal{N}(0, 1)$ by default, *i.e.*, generating $\mathbf{n}$ by randomly drawing a vector from a standard Gaussian distribution. $m = 1$ is used to enforce a sufficiently large distance margin in the feature representation space.

Data Augmentation

A simple data augmentation method is introduced to augment depression samples and further enhance the model's generalizability. Specifically, a pair of depression samples are randomly selected, and then a small percentage of randomly selected values in the last wave data of one sample are

replaced with the corresponding values in another sample to create a new depression sample. The augmented sample can well retain the original depression-relevant information while at the same time enriching the depression samples. By using this method, we increase the number of depression samples in the training data by a factor of 10. In our experiment, we randomly replaced 5% feature values by default.

The Algorithm of MTNet

The algorithmic procedure of our model MTNet is presented in Algorithm 1. After random initialization of the network parameters in Step 1, stochastic gradient descent is used to optimize the model in Steps 2-8. In Step 4, as the number of depression samples is typically far smaller than that of non-depression samples, we generate sample batches with balanced class distribution to achieve more effective optimization. This shares the same spirit as oversampling in imbalanced learning (He and Garcia 2009). Step 5 calculates the batch-wise loss for the three tasks, and Step 6 performs gradient descent steps to learn all the parameters in Θ. Note that Θ_r are shared parameters in $\{\Theta_e, \Theta_a, \Theta_o\}$, and thus, the feature representations in MTNet are jointly optimized by all three tasks. At the inference stage, only the classification function ϕ is used to produce the class label.

Algorithm 1 MTNet

Input: $\mathcal{X} \in \mathbb{R}^{w \times D}$ - training samples, and binary class labels $\mathbf{y}$
Output: $\phi : \mathcal{X} \mapsto \mathcal{Y}$ - a depression classification network

1: Randomly initialize $\Theta = \{\Theta_e, \Theta_a, \Theta_o\}$
2: **for** $j = 1$ to $\#epochs$ **do**
3: **for** $k = 1$ to $\#batches$ **do**
4: $\mathcal{B} \leftarrow$ Randomly sample the same number of depression and non-depression samples
5: Calculate the loss using $\frac{1}{|\mathcal{B}|} \sum_{\mathbf{X}_i \in \mathcal{B}} \Big[\ell_e\big(\phi(\mathbf{X}_i; \Theta_e), y_i\big) + \alpha \ell_a\big(\tau(\mathbf{X}_i; \Theta_a), y_i\big) + \beta \ell_o\big(\psi(\mathbf{X}_i; \Theta_o), y_i\big) \Big]$
6: Perform a gradient descent step w.r.t. the parameters in Θ
7: **end for**
8: **end for**
9: **return** ϕ

Experiments

Datasets

Our model is evaluated on large child depression data based on the Longitudinal Study of Australian Children (LSAC) data (Gray, Sanson et al. 2005). LSAC consists of multiple waves of questionnaire-based interview data of 10,090 children across Australia. Initially, children aged from infant to 5 years and their families are interviewed between August 2003 and February 2004. This routine is repeated every two years afterwards. At the time of writing seven waves of data are available. LSAC provides a dataset of 4,983 children aged 4 to 5 years when taking the first wave of questionnaire. These children are all healthy until 287 children are confirmed to have depression at the 6/7-th wave of interview in the years 2013 to 2015. After a simple feature screening to remove uninformative features (*e.g.*, features with very large percentage of missing values), 210 social-demographic features related to individual growth and development (*e.g.*, age, gender, living location, schooling performance), and family environment (*e.g.*, social, educational, economic, employment, household income, housing conditions) are used. In the selected features, missing values are filled with the mean/mode value in each feature; categorical features are then converted into numeric features by using one-hot encoding. The resulting dataset contains 762 features in each wave of data. Thus, the dataset used has 4,983 samples, with each sample represented by a 7×762 matrix. We further perform a stratified random split of the dataset into three subsets, including 60% data as a training set, and respective 20% data for validation and testing sets.

Competing Methods

MTNet is compared with seven popular and/or state-of-the-art temporal and non-temporal methods.

- **Non-temporal Methods**. Three popular classification methods, including logistic regression (LR), support vector machines (SVM) and multi-layer perceptron (MLP) neural networks, are used as competing methods that are not designed to capture temporal dependence. They are used as baselines to verify whether temporal dependence is important to depression classification. They are two main ways to apply these methods to the longitudinal data. One way is to build the classification model using the most recent single wave data only, performing classification based on the present information. The second way is to use the data from all the waves, in which for each subject we concatenate the feature vectors derived from all the waves into one lengthy unified feature vector; the classifiers are then built upon this concatenated data. This way helps capture some temporal-dependent changes. All three methods are evaluated in both ways, with LR/SVM/MLP-s denoting the classifier using the single wave data and LR/SVM/MLP-m denoting the classifier using the concatenated multi-wave data.
- **Temporal Methods**. One of the most advanced deep temporal methods, long short-term memory (LSTM) neural networks, is used as competing temporal methods. This method is set to be equivalent to a simplified MTNet using only the classification loss function, *i.e.*, ℓ_e in Eq. (10).

Implementation Details

MTNet is implemented with one LSTM layer with 200 units, followed by a fully-connected (FC) layer with 20 units and a classification output layer. The sigmoid function is used in g_r in the LSTM layer by default; the widely-used ReLU activation function is used in g_s in the FC layer. A dropout layer with a dropout rate of 0.5 is applied to the LSTM and FC layers. The competing method LSTM uses exactly the same network architecture as MTNet but does not have the metric learning and anomaly ranking modules. In addition to the dropout, a l_2-norm regularizer is also applied to regularize MTNet and LSTM models. MLP uses a similar network structure with two hidden layers of respectively 200

and 20 units, with each layer having a dropout rate of 0.5. MTNet, LSTM, and MLP are implemented using Keras[2] and optimized using RMSprop (Tieleman and Hinton 2012) with a batch size of 256 and 20 batches per epoch. They are trained with 30 epochs as their performance can converge early. $\alpha = 0.5$ and $\beta = 2.0$ are used in MTNet by default.

LR and SVM are taken from the open-source scikit-learn package[3]. Due to a large percentage of irrelevant features presented in the data, our extensive results showed that applying the l_1-norm regularizer to MLP, LR and SVM obtains significantly better performance than the l_2-norm regularizer. Thus, the l_1-norm regularizer is applied to these three classifiers to bring sparsity to the model. The regularization hyperparameter is probed with $\{0.001, 0.01, 0.1, 1\}$, with the best performance reported. The same oversampling method as MTNet is used in LR, SVM, MLP and LSTM to alleviate the class-imbalanced problem in our dataset.

Performance Evaluation Measures

Three widely-used evaluation measures are used, including the Area Under Receiver Operating Characteristic Curve (AUC-ROC), Area Under Precision-Recall Curve (AUC-PR), and F_1-score (F-score for brevity). AUC-ROC summarizes the ROC curve of true positives against false positives, while AUC-PR summarizes the curve of precision against recall. AUC-ROC is popular due to its good interpretability. AUC-PR is more indicative than AUC-ROC in evaluating performance on imbalanced data. F-score is the harmonic mean of precision and recall. We also report the precision and recall results to gain more insights into the performance. The reported results are averaged over five independent runs. For all measures, larger values indicate better performance.

Depression Detection

Experimental Settings We investigate the capability of our model in identifying small depression cases from highly-imbalanced test data. In this experiment, it is assumed that the positive samples have developed the depression illness. Thus, MTNet, LSTM, and LR/SVM/MLP-m are trained and tested with *all seven waves of data*. LR/SVM/MLP-s are trained and tested with the 7-th wave data only.

Results The detection results of MTNet and its seven competing methods are shown in Table 1. MTNet achieves very promising results in this context, having perfect performance in both AUC-ROC and AUC-PR. It is able to achieve a recall of one and a precision of over 0.97, resulting in a F-score of nearly 0.99. This substantially outperforms all seven competing methods by 4%-28%. Additionally, LSTM, SVM-m, LR-m and MLP-m consistently outperform SVM-s, LR-s and MLP-s by large margins in AUC-ROC, AUC-PR, and F-score, indicating the significance of incorporating features in different time periods into the modeling. *Note that the nearly perfect performance here is mainly due to the fact that the positive samples have exhibited clear depression signs at the 6/7-th wave data, and thus, they can be well distinguished given a good model fitting*.

[2]https://keras.io/

[3]https://scikit-learn.org/

Table 1: Performance Results (mean±std) of Depression Detection. The best result per measure is boldfaced.

	AUC-ROC	AUC-PR	F-score	Precision	Recall
LR-s	0.846±0.004	0.628±0.072	0.773±0.042	0.877±0.113	0.698±0.013
LR-m	0.982±0.011	0.873±0.059	0.932±0.033	0.896±0.049	0.972±0.021
SVM-s	0.840±0.003	0.693±0.013	0.807±0.007	**0.990**±0.020	0.681±0.007
SVM-m	0.977±0.007	0.860±0.054	0.925±0.030	0.892±0.053	0.961±0.013
MLP-s	0.886±0.008	0.715±0.019	0.769±0.018	0.968±0.009	0.639±0.026
MLP-m	0.999±0.001	0.984±0.010	0.939±0.021	0.921±0.033	0.919±0.042
LSTM	0.999±0.001	0.993±0.004	0.953±0.011	0.975±0.021	0.933±0.023
MTNet	**1.000**±0.000	**1.000**±0.000	**0.988**±0.004	0.976±0.008	**1.000**±0.000

Depression Prediction

Experimental Settings This section evaluates the capability of our model in leveraging longitudinal data to predict the possible occurrence of having depression in the near future. To achieve this goal, the models, including MTNet, LSTM, and LR/SVM/MLP-m, are trained and tested using the first five waves of data. LR/SVM/MLP-s are trained with the fifth-wave data only. The task is to predict whether a child will have depression at the upcoming wave 6 or 7.

Results The results of depression prediction are shown in Table 2. Our model MTNet is the best performer in AUC-ROC, AUC-PR and F-score. MTNet substantially outperforms all of its competing methods by 2%-38% in AUC-PR and 3%-20% in F-score. Impressively, MTNet obtains a recall of 0.8, achieving at least 7.8% improvement over its contenders. Note that given a sufficiently high precision of 0.7, the high recall rate in MTNet would enable accurate intervention and treatment of most (*i.e.*, 80%) forthcoming depression cases at a very early stage (2-4 years before the depression actually occurs), effectively preventing and reducing the depression cases. Although LSTM gains a precision of over 0.75, its recall is rather low, indicating the intervention of at most 59% forthcoming depression cases only. Clearly MTNet can provide significantly better solutions than LSTM in early prediction.

Table 2: Performance Results (mean±std) of Depression Prediction. The best result per measure is boldfaced.

	AUC-ROC	AUC-PR	F-score	Precision	Recall
LR-s	0.648±0.020	0.595±0.018	0.632±0.021	0.659±0.032	0.611±0.037
LR-m	0.648±0.019	0.596±0.016	0.620±0.023	0.670±0.027	0.579±0.038
SVM-s	0.666±0.012	0.610±0.012	0.646±0.005	0.682±0.023	0.614±0.011
SVM-m	0.684±0.013	0.631±0.010	0.646±0.026	0.729±0.013	0.582±0.045
MLP-s	0.771±0.007	0.780±0.010	0.706±0.023	0.679±0.032	0.742±0.071
MLP-m	0.814±0.009	0.808±0.019	0.718±0.022	0.730±0.049	0.718±0.082
LSTM	0.779±0.012	0.775±0.010	0.662±0.034	**0.751**±0.031	0.593±0.039
MTNet	**0.818**±0.009	**0.823**±0.008	**0.743**±0.037	0.697±0.006	**0.800**±0.080

Sample Efficiency

Experimental Settings This section examines the model's generalizability from the sample efficiency aspect, *i.e.*, how is the performance if less labeled training data is available? Specifically, in both depression detection and prediction tasks, each model is trained on a new training set that is a random subset of the original training data, and then it is evaluated using the same test data as that used in Tables 1 and 2. We report the results of LR/SVM/MLP-m and omit

LR/SVM/MLP-s since the former methods consistently outperforms the latter ones as shown in Tables 1 and 2.

Results The F-score performance w.r.t. the amount of training data used is shown in Figure 2. Remarkably, MTNet is substantially more sample-efficient than its competing methods; it performs much better than, or comparably well to, the best performance of its competing methods even when it uses 50% less training data. This superiority of MTNet benefits from the integrated metric learning and anomaly ranking tasks in its multi-task objective function, which enable better generalization to unseen depression cases. This is manifested by the large performance gap between MTNet and LSTM, since the only difference between them is the two auxiliary tasks integrated into MTNet. Nevertheless, as expected, the performance of all models degrades with decreasing training data; it should also be noted that complex temporal models such as MTNet and LSTM may not as effective as their simpler alternatives like MLP when only very limited (*e.g.*, 10%) training data is available, since MTNet and LSTM may be under-fitting in such cases.

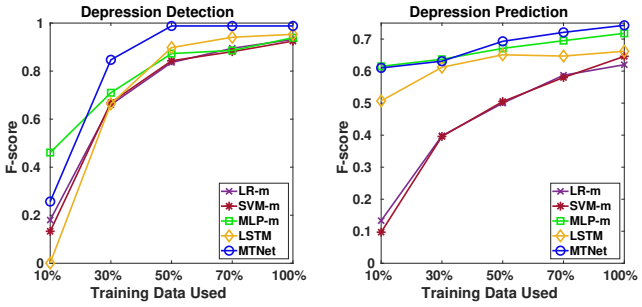


Figure 2: F-score Results w.r.t. the Amount of Training Data.

Ablation Study

Experimental Settings This section evaluates the contribution of each module in MTNet to its overall performance. LSTM is used as a baseline to evaluate the effect of incorporating one or more of the following three modules, including the anomaly ranking loss ℓ_a, the one-class metric learning loss ℓ_o, and the data augmentation (DA).

Results The ablation study results are reported in Table 3. It is clear that, in both of the depression detection and prediction tasks, all three modules in MTNet make important contribution to its superior overall performance; the anomaly ranking loss ℓ_a and the one-class metric learning loss ℓ_o are complementary to each other. These observations are more evident in the task of depression prediction because this task is much more complex and requires well designed methods to achieve desired performance.

Parameter Sensitivity Tests

Experimental Settings This section evaluates the performance sensitivity of MTNet w.r.t. its two key hyperparameters, α and β, which respectively adjust the importance of the anomaly ranking loss and one-class metric learning loss.

Table 3: Ablation Study Results. DA is short for data augmentation. The best result per measure is boldfaced.

Depression Detection					
Method	AUC-ROC	AUC-PR	F-score	Precision	Recall
LSTM	0.999	0.993	0.953	0.975	0.933
LSTM+ℓ_a	**1.000**	0.999	0.986	**0.976**	0.996
LSTM+ℓ_o	**1.000**	0.997	0.983	0.973	0.993
LSTM+ℓ_a+ℓ_o	**1.000**	0.999	0.986	0.973	**1.000**
LSTM+ℓ_a+ℓ_o+DA	**1.000**	**1.000**	**0.988**	**0.976**	**1.000**
Depression Prediction					
Method	AUC-ROC	AUC-PR	F-score	Precision	Recall
LSTM	0.779	0.775	0.662	**0.751**	0.593
LSTM+ℓ_a	0.785	0.786	0.688	0.704	0.684
LSTM+ℓ_o	0.804	0.821	0.702	0.673	0.737
LSTM+ℓ_a+ℓ_o	0.817	**0.834**	0.721	0.709	0.737
LSTM+ℓ_a+ℓ_o+DA	**0.818**	0.823	**0.743**	0.697	**0.800**

Results The sensitivity test results are presented in Figure 3. The results show that MTNet generally performs stably with both α and β in a wide range of setting choices. Relatively small α and large β are needed for MTNet to achieve the best performance. This indicates that MTNet is dependent more on the one-class metric learning than the anomaly ranking. This can also be observed in the results of the prediction task in Table 3.

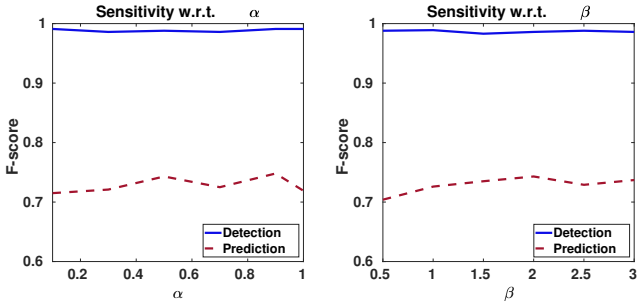


Figure 3: Sensitivity Tests w.r.t. α and β.

Conclusions and Future Work

In this work we propose a novel multi-task learning framework and its instantiation MTNet for depression classification. MTNet leverages the two auxiliary tasks, one-class metric learning and anomaly ranking tasks, to improve the depression classification model's generalization to unseen depression cases. The resulting model substantially enhances the depression classification in both of the detection and prediction tasks. Remarkably, our empirical results show that MTNet is able to i) detect almost all depression cases when the subjects already have depression and ii) accurately predict 80% depression cases 2-4 years before the depression actually occurs. The improved generalizability of MTNet is also supported by the sample efficiency experiment, in which MTNet requires significantly less labeled depression samples to perform comparably well to, or substantially better than, the competing methods. In future work, to improve the accountability of the model, we plan to incorporate an interpretation module into our model to provide some insightful explanation for each of its classification result.

References

Acharya, U. R.; Oh, S. L.; Hagiwara, Y.; Tan, J. H.; Adeli, H.; and Subha, D. P. 2018. Automated EEG-based screening of depression using deep convolutional neural network. *Computer Methods and Programs in Biomedicine* 161: 103–113.

Alghowinem, S.; Goecke, R.; Wagner, M.; Parker, G.; and Breakspear, M. 2013. Eye movement analysis for depression detection. In *Proceedings of the IEEE International Conference on Image Processing*, 4220–4224. IEEE.

Ay, B.; Yildirim, O.; Talo, M.; Baloglu, U. B.; Aydin, G.; Puthankattil, S. D.; and Acharya, U. R. 2019. Automated depression detection using deep representation and sequence learning with EEG signals. *Journal of Medical Systems* 43(7): 205.

Cai, H.; Han, J.; Chen, Y.; Sha, X.; Wang, Z.; Hu, B.; Yang, J.; Feng, L.; Ding, Z.; Chen, Y.; et al. 2018. A pervasive approach to EEG-based depression detection. *Complexity* 2018.

Chancellor, S.; and De Choudhury, M. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digital Medicine* 3(1): 1–11.

Chao, L.; Tao, J.; Yang, M.; and Li, Y. 2015. Multi task sequence learning for depression scale prediction from video. In *International Conference on Affective Computing and Intelligent Interaction*, 526–531. IEEE.

Cohn, J. F.; Kruez, T. S.; Matthews, I.; Yang, Y.; Nguyen, M. H.; Padilla, M. T.; Zhou, F.; and De la Torre, F. 2009. Detecting depression from facial actions and vocal prosody. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 1–7. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 4171–4186.

Feng, X.; Keenan, K.; Hipwell, A. E.; Henneberger, A. K.; Rischall, M. S.; Butch, J.; Coyne, C.; Boeldt, D.; Hinze, A. K.; and Babinski, D. E. 2009. Longitudinal associations between emotion regulation and depression in preadolescent girls: Moderation by the caregiving environment. *Developmental Psychology* 45(3): 798.

Gong, Y.; and Poellabauer, C. 2017. Topic modeling based multimodal depression detection. In *Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge*, 69–76.

Gray, M.; Sanson, A.; et al. 2005. Growing up in Australia: The longitudinal study of Australian children. *Family Matters* (72): 4.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 1735–1742.

He, H.; and Garcia, E. A. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9): 1263–1284.

Henry, M.; Rosberger, Z.; Ianovski, L. E.; Hier, M.; Zeitouni, A.; Kost, K.; Mlynarek, A.; Black, M.; MacDonald, C.; Richardson, K.; et al. 2018. A screening algorithm for early detection of major depressive disorder in head and neck cancer patients post-treatment: Longitudinal study. *Psycho-oncology* 27(6): 1622–1628.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8): 1735–1780.

Kessler, R. C.; Avenevoli, S.; Costello, E. J.; Georgiades, K.; Green, J. G.; Gruber, M. J.; He, J.-p.; Koretz, D.; McLaughlin, K. A.; Petukhova, M.; et al. 2012. Prevalence, persistence, and sociodemographic correlates of DSM-IV disorders in the National Comorbidity Survey Replication Adolescent Supplement. *Archives of General Psychiatry* 69(4): 372–380.

Korsten, L. H.; Jansen, F.; de Haan, B. J.; Sent, D.; Cuijpers, P.; Leemans, C. R.; and Verdonck-de Leeuw, I. M. 2019. Factors associated with depression over time in head and neck cancer patients: A systematic review. *Psycho-oncology* 28(6): 1159–1183.

Lamers, F.; Milaneschi, Y.; Smit, J. H.; Schoevers, R. A.; Wittenberg, G.; and Penninx, B. W. 2019. Longitudinal association between depression and inflammatory markers: results from the Netherlands study of depression and anxiety. *Biological Psychiatry* 85(10): 829–837.

Lin, C.; Hu, P.; Su, H.; Li, S.; Mei, J.; Zhou, J.; and Leung, H. 2020. SenseMood: Depression Detection on Social Media. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 407–411.

Low, L.-S. A.; Maddage, N. C.; Lech, M.; Sheeber, L.; and Allen, N. 2010a. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5154–5157. IEEE.

Low, L.-S. A.; Maddage, N. C.; Lech, M.; Sheeber, L. B.; and Allen, N. B. 2010b. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering* 58(3): 574–586.

Lu, J.; Shang, C.; Yue, C.; Morillo, R.; Ware, S.; Kamath, J.; Bamis, A.; Russell, A.; Wang, B.; and Bi, J. 2018. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2(1): 1–21.

Ma, X.; Yang, H.; Chen, Q.; Huang, D.; and Wang, Y. 2016. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 35–42.

Mann, P.; Paes, A.; and Matsushima, E. H. 2020. See and Read: Detecting Depression Symptoms in Higher Education Students Using Multimodal Social Media Data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 440–451.

Nasir, M.; Jati, A.; Shivakumar, P. G.; Nallan Chakravarthula, S.; and Georgiou, P. 2016. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 43–50.

Orabi, A. H.; Buddhitha, P.; Orabi, M. H.; and Inkpen, D. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 88–97.

Pang, G.; Shen, C.; and van den Hengel, A. 2019. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 353–362.

Rodrigues Makiuchi, M.; Warnita, T.; Uto, K.; and Shinoda, K. 2019. Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 55–63.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* .

Rushton, J. L.; Forcier, M.; and Schectman, R. M. 2002. Epidemiology of depressive symptoms in the National Longitudinal Study of Adolescent Health. *Journal of the American Academy of Child & Adolescent Psychiatry* 41(2): 199–205.

Shen, G.; Jia, J.; Nie, L.; Feng, F.; Zhang, C.; Hu, T.; Chua, T.-S.; and Zhu, W. 2017. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3838–3844.

Shen, T.; Jia, J.; Shen, G.; Feng, F.; He, X.; Luan, H.; Tang, J.; Tiropanis, T.; Chua, T.-S.; and Hall, W. 2018. Cross-domain depression detection via harvesting social media. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 1611–1617.

Suslow, T.; Junghanns, K.; and Arolt, V. 2001. Detection of facial expressions of emotions in depression. *Perceptual and Motor Skills* 92(3): 857–868.

Tax, D. M.; and Duin, R. P. 2004. Support vector data description. *Machine Learning* 54(1): 45–66.

Tieleman, T.; and Hinton, G. 2012. Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. URL http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Uddin, M. A.; Joolee, J. B.; and Lee, Y.-K. 2020. Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm. *IEEE Transactions on Affective Computing* .

Wang, X.; Zhang, C.; Ji, Y.; Sun, L.; Wu, L.; and Bao, Z. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 201–213. Springer.

WHO. 2017. Depression and other common mental disorders: global health estimates. Technical report, World Health Organization.

Yang, L.; Jiang, D.; Xia, X.; Pei, E.; Oveneke, M. C.; and Sahli, H. 2017. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge*, 53–59.

Yang, Y.; Fairbairn, C.; and Cohn, J. F. 2012. Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing* 4(2): 142–150.

Zhang, Y.; and Yang, Q. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114* .

Zhou, X.; Jin, K.; Shang, Y.; and Guo, G. 2018. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing* .

Zhu, Y.; Shang, Y.; Shao, Z.; and Guo, G. 2017. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. *IEEE Transactions on Affective Computing* 9(4): 578–584.

Zuckerbrot, R. A.; and Jensen, P. S. 2006. Improving recognition of adolescent depression in primary care. *Archives of Pediatrics & Adolescent Medicine* 160(7): 694–704.