# Automatic Depression Detection among Higher Education Students Based on DeepFM

Ziling Ruan, Pengfei Yang, *Member, IEEE*, Jiayang Huang, Keyi Yang, Yidan Lv, and Zhi-Qiang Zhang, *Member, IEEE*

***Abstract*—Depressive disorder has become a common problem among higher education students, but it often gets undiagnosed and untreated due to unrecognized symptoms, poor access to medical resources, and fear of stigma. To improve the situation, automatic depression detection would be essential. In this paper, we explore the feasibility of depression detection in higher education students using their behavioral data automatically collected by the University system. First, a DeepFM network, which can not only take discrete-continuous mixed features as its input but also can learn linear and nonlinear relations between the input and the output, is presented for depression detection. A modified focal loss function (MFL) is then proposed to alleviate data imbalance impact caused by the fact that the proportion of healthy students outweighs those diagnosed with depression significantly. To verify the effectiveness of the proposed method, behavioral data from 3218 students were collected, of which 179 were diagnosed with depression by university psychologists using PHQ-9 scale scores. 5-fold cross-validations are performed, and the experiment results have illustrated that DeepFM obtains the highest average accuracy compared to Multilayer Perceptron (MLP), Factorisation Neural Network (FNN), and Product-based Neural Network (PNN), demonstrating the effectiveness of the proposed framework for depression detection among university students.**

***Index Terms*—Depression detection, data imbalance, deep learning, DeepFM, campus big data.**

## I. Introduction

Depression is a common illness worldwide, with an estimated 3.8% of the population affected, and it is a significant contributor to the overall global burden of disease [1]. Detection and diagnosis of depression early is a crucial step for proper treatment [2]. Recently, automatic objective assessment methods to assist mental health detection and diagnosis have been widely explored [3] [4] [5] [6].

Thus far, video analysis is regarded as the main approach for depression detection since it highly co-relates with head movements, facial expressions, gaze, etc. For instance, Uddin et al. [7] proposed a two-stream deep spatio-temporal framework for depression level prediction. They used a temporal median pooling approach and employed a multilayer bidirectional long short-term memory (Bi-LSTM) based model for depression analysis on video data. Niu et al. [8] developed Dual Attention (DA) and Element Recalibration (ER) blocks to construct the Dual Attention and Element Recalibration (DAER) network, which was used to extract the facial representations of individuals with different depression levels. Yang et al. [9] proposed a multi-modal framework for predicting the Patient Health Questionnaire depression scale (PHQ-8 score), and classifying an individual as depressed or not depressed, by hybridizing deep models and shallow methodologies. de Melo et al. [10] proposed a deep learning architecture called the Maximization and Differentiation Network to represent facial expression variations for depression assessment. Although these aforementioned studies have demonstrated the feasibility of depression analysis using video data, such data may implicate serious privacy concerns. Besides video data, EEG and mobile phone data have also been explored recently for depression analysis. For instance, Acharya et al. [11] and Seal et al. [12] both proposed convolutional neural networks (CNN) to automatically learn features to characterize depressed and normal EEG signals. Cai et al. [13] constructed a multimodal model to distinguish depressed patients from normal controls by fusing different EEG data sources, which were under neutral, negative, and positive audio stimulation. Shen et al. [14] presented an optimal channel selection method via Kernel-Target Alignment (KTA) and its application in depression detection. Although the recent advancements in EEG make it a powerful tool, it is relatively intrusive and unconformable during EEG data collection. Similar work has also been reported to use intrusive phone metadata [15] [16] [17].

Social media has proven to be an unintrusive and stable source of data, and some researchers recently also demonstrated the possibility of depression detection using social media data. For instance, Choing et al. [18] propose 90 unique features, through a combination of feature extraction using sentiment lexicons and content-based features from the social media messages themselves, to detect depression using machine learning classifiers. Zhihua Guo et. al [19] collected data from Sina Weibo and approached depression as a binary classification problem. The effectiveness of their approach was verified using classical machine learning methods. Jitimon Angskun et. al [20] also explored the effectiveness of data from Tweets in

Ziling Ruan, Pengfei Yang, Jiayang Huang, Keyi Yang, and Yidan Lv are with Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, the School of Computer Science and Technology, Xidian University, Xi'an, 710071, China (e-mail: 21031211775@stu.xidian.edu.cn; pfyang@xidian.edu.cn; jyhuang1@stu.xidian.edu.cn; 21031211507@stu.xidian.edu.cn; 19030500265@stu.xidian.edu.cn).

Zhi-Qiang Zhang is with the School of Electronic and Electrical Engineering, Institute of Robotics, Autonomous Systems and Sensing, University of Leeds, Leeds LS2 9JT, U.K(e-mail: z.zhang3@leeds.ac.uk).

detecting depression and found that machine learning models can capture depressive moods of depression sufferers. Tong et al. [21] proposed a classifier called Cost-sensitive Boosting Pruning Trees (CBPT) and tested its performance on two publicly accessible Twitter depression detection datasets. Shen et al. [22] proposed a cross-domain Deep Neural Network model with Feature Adaptive Transformation & Combination strategy (DNN-FATC) to enhance depression detection via social media with multi-source datasets. Although social media can provide a non-intrusive channel for collecting depressed text data, many people may not use social media at all.

Unlike the data collection challenges faced by depression detection for the general public, the emergence of digitization in education has made various types of student data readily available [23]. In this paper, we explore the feasibility of detecting depression in higher education students using the behavioral data automatically collected by the university system. With the behavioral data containing both discrete and continuous characteristics, the DeepFM model can automatically learn depression detection patterns among students. DeepFM is a deep learning structure for discrete-continuous mixed data as input, which has been widely used in multiple tasks, such as metal-organic properties prediction [24], passenger car sales prediction prediction [25], taxi pick-up area recommendation [26], and etc.

The main contributions of this study include 1) A DeepFM network, which can not only take input for discrete-continuous mixed features but also learn both linear and nonlinear relations between the input and the output, which is presented for depression detection. 2) A modified focal loss function (MFL)is then proposed to alleviate the data imbalance impact caused by the fact that the proportion of healthy students outweighs those diagnosed with depression significantly. The MFL can compensate for less training data for the depressed group by making it descend further in the stochastic gradient descent (SGD) process, thus reducing the bias of the model in the training phase. 3) To verify the effectiveness of the method, behavioral data from 3218 students were collected, of which 179 were diagnosed with depression. 5-fold cross-validations were performed, and the experiment results illustrate that DeepFM has obtained the highest average accuracy 96.4% compared to Multilayer Perceptron (MLP) 93.3%, Factorisation Neural Network (FNN) 91.8%, and Product-based Neural Network (PNN) 96.2%, which demonstrated the effectiveness of the proposed detection framework for enhancing depression detection performance.

The remaining paper is arranged as follows: Section II introduces the materials and methods. In section III, the experiment results are reported. Then, the discussion is presented in section IV. The conclusion is presented in the last section.

## II. Methodology and materials

In this section, we will introduce the proposed depression detection method: we will briefly describe the data preparation, and then we explain the proposed depression detection framework, which includes a DeepFM network architecture and a modified focal loss function. We will elaborate on each part below. For a better understanding of our proposed depression detection algorithm, Table I summarizes the symbols/notations used in this paper.

TABLE I
Table of Notations

| Notation | Description |
|---|---|
| $\boldsymbol{X}^s$ | Behavioral features of the $s$-th student |
| $y^s$ | Ground truth of the $s$-th student. |
| $\boldsymbol{\mathcal{X}}_{\mathrm{d}}^s$ | Discrete features after one-hot coding of the s-th student |
| $\boldsymbol{\mathcal{X}}_{\mathrm{c}}^s$ | Continuous features after Min-Max normalization of the $s$-th student |
| $\boldsymbol{\mathcal{X}}^s$ | Input of the FM component and the DNN component |
| $F_{\mathrm{FM}}$ | Output of FM component |
| $F_{\mathrm{DNN}}$ | Output of DNN component |
| $p$ | Probability of depression of $s$-th sample |
| $\hat{y}$ | Detection result given by the model |
| $\Theta$ | Parameters of the proposed model |
| d | Dimension of $\boldsymbol{\mathcal{X}}^s$ |
| $\boldsymbol{v}_i$ | The $i$-th row in $\boldsymbol{V}$ |
| $x_i$ | The $i$-th value of $\boldsymbol{\mathcal{X}}^s$ |
| L | The modified focal loss function |
| $S_1$,$S_0$ | Number of students with depression, Number of healthy students |
| $\lambda$ | The imbalance degree |
| $\Delta\ \theta$ | The gradient in backpropagation |
| $h_l$ | Number of neurons of the $l$-th layer of DNN |
| $\boldsymbol{a}^{(l)}$ | Output of $l$-th layer of DNN |
| $\sigma$ | The $sigmoid$ function |
| $\varphi$ | The $ReLU$ function |

### A. Data Preparation and Notations

Participants in the study were full-time undergraduates (aged from 19 to 21, enrolled between 2015 and 2017) from our University. Firstly, we distributed 12,000 questionnaires to all the students in these three cohorts and 5,000 students indicated they would like to participate in our study. After signing a non-disclosure agreement for data sharing with these student volunteers from the University, their behavioral data were downloaded from the University Data Center. Volunteers with more than 20% missing data and outliers (judged by Z-score [27]) were removed directly. For the remaining records, miceforest [28], a way of data interpolation, is performed to fill in missing values. Finally, the behavioral dataset containing 3218 undergraduates (2491 males and 727 females) was created. 179 of them have been diagnosed with depression by university psychologists. They used PHQ-9 scale [29] scores to determine if the student was suffering from depression. Students with a score greater than 4 are identified as depressed. Many researches have demonstrated that there is connection between depression and academic performance [30] [31] [32], daily behavior [33]–[37], movement/exercises [38] [39] [40] and de-

TABLE II
THE DETAILS OF THE BEHAVIORAL DATA

| | | |
|---|---|---|
| **Discrete Features (8)** | **Demographic Data (8)** | Gender, Age, Nationality, Native place, Family number, Family financial situation, Year of entry, Major |
| **Continuous Features (27)** | **Academic Performance Data (16)** | Sum of Courses scores, Mean of Courses scores, Std of courses scores, GPA, Number of scholarship awarded, Level of scholarship awarded, Class ranking, Number of courses absence, Number of courses late, Number of courses attendance, Number of courses leaving in advance, Number of retaking course, Number of resitting course, Sum of missing classes hours, Mean of missing classes hours, Std of missing classes hours |
| | **Physical Quality Data (7)** | PE scores, BMI, Total number weekly workouts, Average number of weekly workouts, Std of number of weekly workouts, Total hours of workouts, Average hours of workouts |
| | **Daily Data (4)** | Number of canteen consumption, Number of companions, Number of late getting back to dorms, Number of stay up late |

mography [41]–[44]. Therefore, demographic data, academic performance data, physical quality data, and daily life data were selected in this study. All these data were automatically collected by the University data center due to the recent advancement of university digitalization. The details of the data are listed in Table II. For any student $s(s = 1, 2, \cdots, 3218)$, his (or her) data record in the depression detection data set was depicted as $(\boldsymbol{X}^s, y^s)$, where $\boldsymbol{X}^s$ is a vector of dimension 35 that indicates the features of the behavioral data, and $y^s$ denotes the ground truth ($y^s = 1$ means $s$ is diagnosed with depression and $y^s = 0$ means $s$ is not depressed during the test). As shown in Table II, $\boldsymbol{X}^s$ consists of discrete features and continuous features. Since not all discrete features can be compared in their values, one-hot encoding is adopted to process them. After that, discrete features are converted to a high-dimensional sparse feature vector $\boldsymbol{\mathcal{X}}^s_{\mathrm{d}}$. For continuous features, Min-Max normalization was performed to get a feature vector $\boldsymbol{\mathcal{X}}^s_{\mathrm{c}}$.

### B. Depression Detection Framework Based on DeepFM

The main framework of the depression detection model based on DeepFM is depicted in Fig.1. Firstly, due to the high-dimensional and extremely sparse characteristic of discrete feature $\boldsymbol{\mathcal{X}}^s_{\mathrm{d}}$, an Embedding layer is used to reduce the dimension of the vector. Then, the dimension-reduced $\boldsymbol{\mathcal{X}}^s_{\mathrm{d}}$ is concatenated to the continuous feature $\boldsymbol{\mathcal{X}}^s_{\mathrm{c}}$, obtaining the feature $\boldsymbol{\mathcal{X}}^s$, as the input of the model. $\boldsymbol{\mathcal{X}}^s$ is taken as the input of the factorization-machine (FM) component and deep neural network(DNN) component, where FM learns linear relations and part nonlinear relations and DNN mainly learns nonlinear relations between the input and the output. Next,

TABLE III
DNN STRUCTURE. THE BATCHSIZE IS SET TO 64 IN OUR EXPERIMENTS. D IS THE DIMENSION OF THE INPUT VECTOR $\boldsymbol{\mathcal{X}}^s$. THE OUTPUT OF FC4 IS $F_{\mathrm{DNN}}$.

| Layer | Inputsize | Outputsize | Activation | Dropout |
|---|---|---|---|---|
| FC1 | batchsize × d | batchsize × 256 | ReLU | 0.1 |
| FC2 | batchsize × 256 | batchsize ×128 | ReLU | 0.1 |
| FC3 | batchsize × 128 | batchsize × 64 | ReLU | 0.1 |
| FC4 | batchsize × 64 | batchsize × 1 | - | - |

The output of FM ($F_{\mathrm{FM}}$) and DNN ($F_{\mathrm{DNN}}$) are summed up and the summation is used as an input of the activation function $sigmoid$.

### C. Architecture of DeepFM

Considering the discrete-continuous-mixed characteristic of our dataset, DeepFM is used as a classifier in this study. The DeepFM network consists of two components: the FM component and the DNN component.

First, the FM component parameters were estimated as $\boldsymbol{w} = [w_0, w_1, \ldots w_d]$, and a matrix $\boldsymbol{V} \in \mathbb{R}^{d\times k}$. A row $\boldsymbol{v_i}$ within $\boldsymbol{V}$ is a $k$-dimensional vector, which describes the $i$-th feature. The output of FM is defined as follows:

$$F_{\mathrm{FM}} = \boldsymbol{w}\begin{pmatrix} 1 \\ \boldsymbol{\mathcal{X}^s} \end{pmatrix} + \sum_{i=1}^{d-1}\sum_{j=i+1}^{d} \boldsymbol{v}_i\boldsymbol{v}_j^{\mathrm{T}} x_i x_j, \tag{1}$$

where $\boldsymbol{\mathcal{X}}^s = [x_1, x_2, \ldots, x_d]$. For each feature $x_i \in \boldsymbol{\mathcal{X}^s}$, a scalar $w_i \in \boldsymbol{w}$ is used to weigh its order-1 importance, and a latent vector is used to weigh its importance of interactions with other features.

Second, the DNN, a feed-forward neural network, is used to learn the nonlinear relations between the input and the output. The input of the DNN component is $\boldsymbol{\mathcal{X}}^s$ as well. As shown in Fig.1, The DNN component is mainly composed of 4 fully connected layers, each fully connected layer is followed by a ReLU layer and a Dropout layer. The detailed structure of DNN has been shown in Table III. After forward propagation, the output of the DNN component $F_{\mathrm{DNN}}$ is obtained.

Finally, after the action of the $sigmoid$ activation function, the output of the model, the probability of depression, is obtained. The details are as follows:

$$p = \sigma(F_{\mathrm{FM}} + F_{\mathrm{DNN}}), \tag{2}$$

where $\sigma$ is the $sigmoid$ activation function, $p$ is the estimated probability of depression, and the predicted label of a student is defined as:

$$\hat{y} = \begin{cases} 1 & \text{if } p > 0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where $\hat{y}$ is the detection result given by the model ($\hat{y} = 1$ means the student is classified as depressed. $\hat{y} = 0$ means the student is classified as healthy).

The model parameters $\Theta = \{\boldsymbol{w}, \boldsymbol{V}, \boldsymbol{U}\}$ are updated by stochastic gradient descent algorithm (SGD), where $\boldsymbol{w}$ and $\boldsymbol{V}$ are the parameters in the FM component and $\boldsymbol{U}$ represents the parameters in the DNN component.

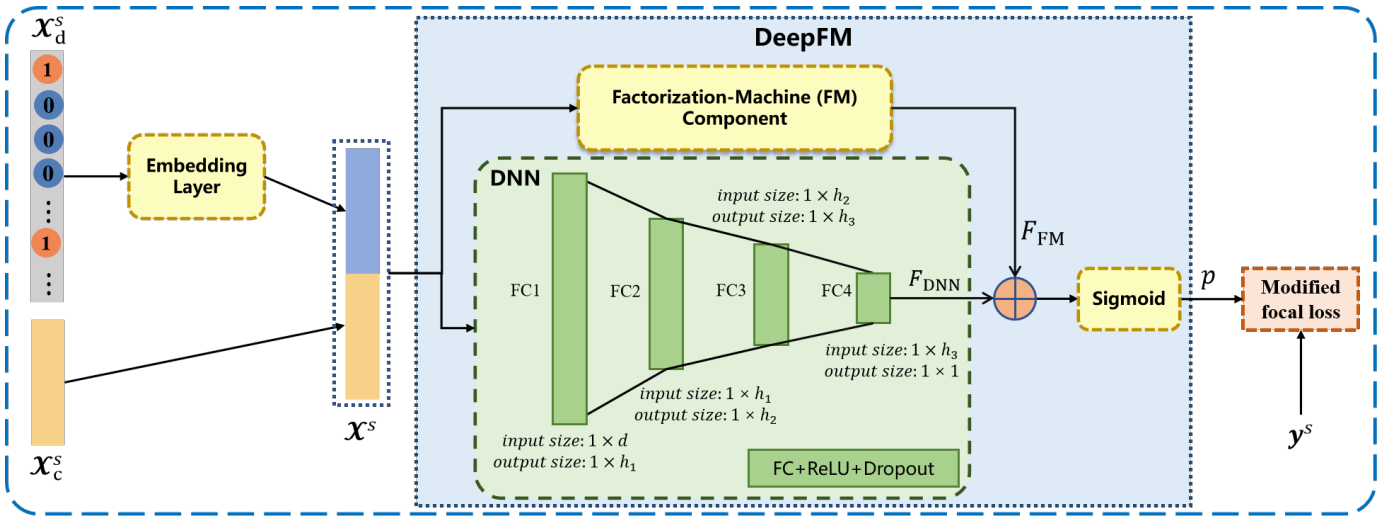



Fig. 1. Main framework of depression detection model. The framework mainly includes 2 components, FM and DNN, and the proposed loss function, the modified focal loss function. The input of the framework is a vector $\boldsymbol{\mathcal{X}}^s$ composed of depressive features. The output, $p$, is the probability of depression.

### D. The Modified Focal Loss

To reduce the influence of the imbalance characteristics of the dataset, we proposed a class imbalance loss function named Modified Focal Loss (MFL) to reduce the bias of the model. With the probability $p$ obtained in Eq. 2 and the ground truth $y^s$ defined in Section II-A, MFL is designed as:

$$\begin{aligned} L = &- y^s((1-\sin(\frac{\pi}{2}\hat{p}))\cos(\frac{\pi}{2}\hat{p})f(\lambda,S_1,S_0))(1-\hat{p})^\gamma \log\hat{p} \\ &- (1-y^s)((1-\sin(\frac{\pi}{2}\hat{p}))\cos(\frac{\pi}{2}\hat{p}))(1-\hat{p})^\gamma \log\hat{p} \end{aligned} \tag{4}$$

and

$$\hat{p} = \begin{cases} p & \text{if } y^s = 1 \\ 1-p & \text{otherwise,} \end{cases} \tag{5}$$

where $\sin()$ and $\cos()$ respectively denote sine function and cosine function, and $\gamma$ is an adjustable parameter. The function $f()$ introduces the distribution of the training dataset into the loss function. In this way, the gradients of the minority class will descend further to make up for the disadvantage of less training data. $f(\lambda, S_1, S_0)$ is defined as follows:

$$f(\lambda,S_1,S_0) = \begin{cases} (S_0/S_1)^{\frac{1}{2}} & \text{if } \lambda \le 1 \\ (S_0/S_1)^{\frac{1}{8}} & \text{otherwise,} \end{cases} \tag{6}$$

where $\lambda$ is the imbalance degree and is defined as:

$$\lambda = -\frac{1}{2}\log\frac{S_1}{S_0}. \tag{7}$$

Here, $S_0$ denotes the number of healthy students and $S_1$ denotes the number of students with depression.

During the model training, any parameter $\theta \in \Theta$ in the model would be updated as follows:

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta \times \Delta\theta \tag{8}$$

$$\Delta\theta = \frac{\partial L}{\partial \theta} \tag{9}$$

where $\theta_{\text{old}}$ denotes the value of parameter $\theta$ of the previous iteration, and $\theta_{\text{new}}$ denotes the current value updated by SGD. $\eta$ is the learning rate. $\Delta\theta$ denotes the current gradient. According to the chain rule:

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial \hat{p}}\frac{\partial \hat{p}}{\partial p}\frac{\partial p}{\partial \theta} \tag{10}$$

$$\frac{\partial p}{\partial \theta} = p(1-p)\frac{\partial(F_{\text{FM}}+F_{\text{DNN}})}{\partial \theta} \tag{11}$$

$$\frac{\partial(F_{\text{FM}}+F_{\text{DNN}})}{\partial \theta} = \begin{cases} 1 & \theta = w_0 \\ x_i & \theta = w_i \\ x_i\sum_{e=1}^{k} v_{j,e}x_j - v_{i,e}x_i^2 & \theta = v_{i,e} \\ \left(\varphi^{'}(\sum_{r=1}^{h_{l-1}} u_{r,g}^{(l-1)}a_r^{(l-1)} + b_r^{(l-1)})a_r^{(l-1)}\right) & \theta = u_{r,g}^{(l-1)}, \end{cases} \tag{12}$$

where $\varphi$ is the $ReLU$ function, $u_{r,g}^{(l-1)} \in \boldsymbol{U}^{(l-1)}$, $\boldsymbol{U}^{(l-1)} \in \mathbb{R}^{h_l \times h_{l-1}}$ is the in the weight matrix $l-1$-th layer of DNN component, $h_l$ denotes the number of neurons of the $l$-layer, so $0 < r \le h_{l-1}, 0 < g \le h_l$, $a_r^{(l-1)} \in \boldsymbol{a}^{(l-1)}$ represents the output of $r$-th neuron of $(l-1)$-th layer and $b_r^{(l-1)} \in \boldsymbol{b}^{(l-1)}$ represents the bias of $r$-th neuron of $(l-1)$-th layer. And $a_r^{(l-1)} = \varphi(\sum_{z=1}^{h_{(l-2)}} u_{z,r}^{(l-2)}a_z^{(l-2)} + b_z^{(l-2)})$.

$$\frac{\partial \hat{p}}{\partial p} = \begin{cases} 1 & \text{if } y^s = 1 \\ -1 & \text{otherwise,} \end{cases} \tag{13}$$

$$\frac{\partial L}{\partial \hat{p}}=\begin{cases}\begin{pmatrix}-f(\lambda,S_1,S_0)(1-\hat{p})^{\gamma-1}((1-\hat{p})\log\hat{p}\\(\sin(\frac{\pi}{2}\hat{p})+\cos(\pi\hat{p}))(-\frac{\pi}{2})+(\frac{(1-\hat{p})}{\hat{p}\ln 10}\\-\gamma\log\hat{p})(1-\sin(\frac{\pi}{2}\hat{p}))\cos(\frac{\pi}{2}\hat{p}))\end{pmatrix} & \text{if } y^s=1,\\ \begin{pmatrix}-(1-\hat{p})^{\gamma-1}((1-\hat{p})\log\hat{p}\\(\sin(\frac{\pi}{2}\hat{p})+\cos(\pi\hat{p}))(-\frac{\pi}{2})+(\frac{(1-\hat{p})}{\hat{p}\ln 10}\\-\gamma\log\hat{p})(1-\sin(\frac{\pi}{2}\hat{p}))\cos(\frac{\pi}{2}\hat{p}))\end{pmatrix} & \text{otherwise.}\end{cases} \tag{14}$$

From Eq 8 to Eq 14, we can conclude that the length of gradient of positive class($y^s=1$ ) is $f(\lambda,S_1,S_0)(f(\lambda,S_1,S_0)>1)$ times as long as the negative class($y^s=0$), which indicates that the minority class will descend further during model training. Besides, the value of function $f()$ is not as large as possible. This is because an enormous gradient will result in gradient vanishing, loss not being declined, and unstable output.

### *E. Hyper-parameter Setting of Our Proposed Framework*

As shown in Table III, the sizes of the outputs from FC1 to FC3 are 256, 128, and 64 respectively. The output size of the last layer of DNN is 1. In the training phase, the batchsize was set to 64, and the learning rate was set to 0.01. We used stochastic gradient descent (SGD) for training, and the dropout rate in each dropout layer was 0.1. The length of $\boldsymbol{v}_i$, $k$ (proposed in section C) was set to 8. Additionally, the parameter $\gamma$ in the modified focal loss was set as 3. 5-fold cross-validations with each fold of 10 random runs (using different random seeds) were performed. Therefore, in each experiment, we use 80% of the data (i.e., 2574 students, diagnosed vs. undiagnosed: 143 vs. 2431) for training and 20% (i.e., 644 students, diagnosed vs. undiagnosed: 36 vs. 608 ) for testing.

### *F. Evaluations Metrics*

In this study, in addition to classification accuracy, sensitivity, specificity, G-mean, F1score, and AUC (area under receiver operating characteristic curve) were calculated to evaluate the performance of the proposed method. Sensitivity and specificity represent the proportion of all positive samples predicted correctly to all actual positive samples and the proportion of all negative samples predicted correctly to all actual negative samples, respectively. In this paper, positive samples are diagnosed students and negative samples are healthy students. F1score, G-mean, and AUC are the balance of two evaluation indexes, which are commonly used in imbalance classification studies [45]. They are defined as:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

$$\text{Sensitivity}=\frac{TP}{TP+FN} \tag{16}$$

$$\text{Specificity}=\frac{TN}{TN+FP} \tag{17}$$

$$\text{G-mean}=\sqrt{\text{Sensitivity}\times\text{Specificity}} \tag{18}$$

$$\text{F1score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{19}$$

where precision $=\frac{TP}{TP+FP}$, recall $=\frac{TP}{TP+FN}$. The terms $TP$, $TN$, $FP$, and $FN$ represent the number of true positives, true negatives, false positives, and false negatives respectively.

## III. Experimental Results

In this section, we first introduce the baseline methods. Then, we evaluate the performance of the proposed method on the student behavioral dataset. Extensive comparisons of classification performance are implemented between the proposed method and other neural networks. Finally, the influence of the proposed loss function such as the convergence of the loss function, the influence of the parameters in the modified loss function, and comparisons with other loss functions were also reported.

### *A. Baseline Methods*

To verify the advantages of DeepFM in processing discrete-continuous mixed features, we compared the performances with Factorisation Machine supported neural network (FNN) [46], Product-based neural network (PNN) [47] and MLP [48]. The structure of MLP is the same as that of the DNN component. So the performance of MLP expresses the role of high-order feature interactions. FNN (Factorisation Machine Supported Neural Network) is a deep learning model for predicting user behavior. To minimize the influence of other factors, the structure of the fully connected layers in FNN is set to be the same as that of MLP. PNN (Product-based Neural Network) was a deep learning model based on multiplication to represent feature crossing. The PNN network structure adds the Product layer to the traditional deep neural network, so as to realize the crossover of features. Similarly, the fully connected network portion of the PNN in this document was structured in the same way as the MLP setup. In the training phase, all of the models updated their parameters according to the loss given by the modified focal loss function.

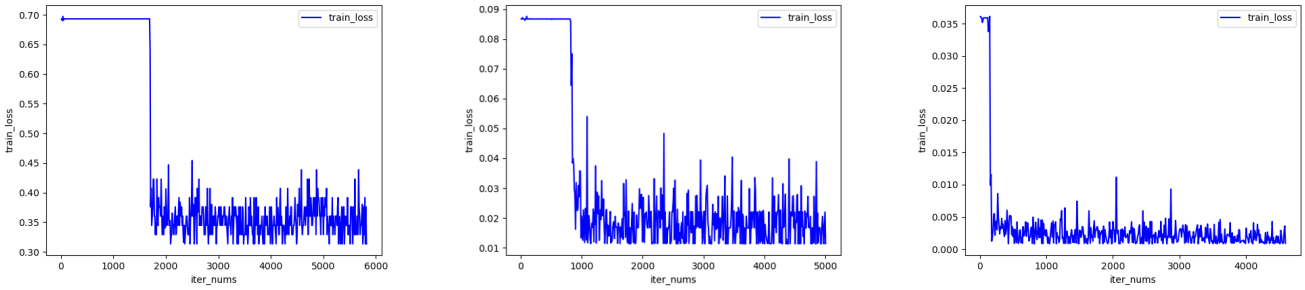
### *B. Depression Detection Performance*

The depression detection performance comparison was conducted between the proposed method DeepFm with the modified focal loss (DeepFM-MFL), and Multilayer Perceptron with the modified focal loss (MLP-MFL), Factorisation Machine supported neural network with modified focal loss (FNN-MFL), and Product-based neural network with the modified focal loss (PNN-MFL). All of them were implemented with Pytorch[1]on our Dell PC with the Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz, 16 GB RAM, and 64-bit Windows 10 OS using Pycharm 2020.1.2 x64. The average classification accuracy, sensitivity, F1score, AUC, G-mean, and specificity were calculated over 5-fold cross-validation with 10 runs (each run with a different random seed) to evaluate the detection performance of these methods. Table IV illustrates

[1]https://pytorch.org/

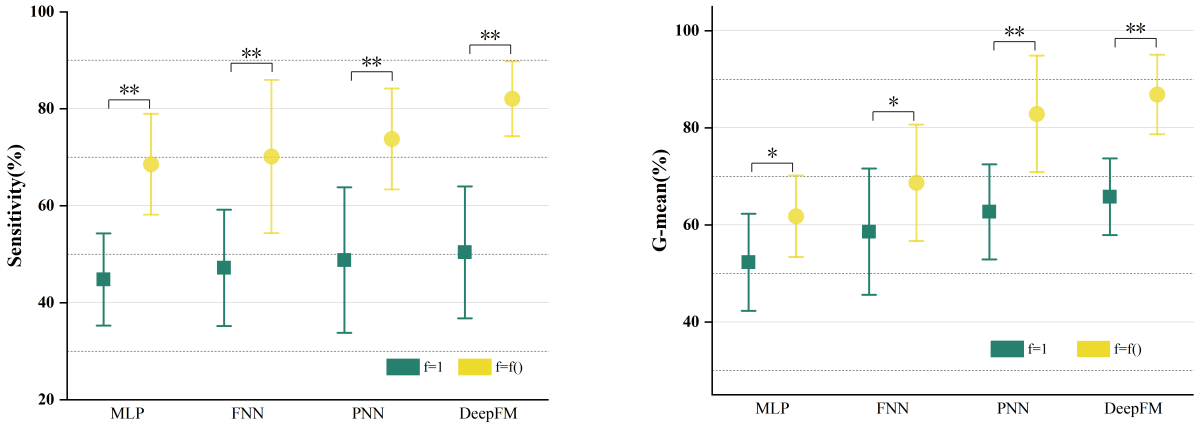
TABLE IV
COMPARISON OF CLASSIFICATION RESULTS FOR DIFFERENT MODELS (MEAN±STD UNIT:%). "A-B" REPRESENTED MODEL A WITH LOSS FUNCTION B DURING THE TRAINING PHASE. "MFL": THE MODIFIED FOCAL LOSS FUNCTION.

| Models | Sensitivity | Specificity | F1score | AUC | GM | Accuracy |
|---|---|---|---|---|---|---|
| **FNN-MFL** | 65.4±15.8 | 94.4±1.3 | 49.1±7.9 | 84.1±8.6 | 59.8±8.4 | 91.8±2.7 |
| **MLP-MFL** | 68.6±10.3 | 95.6±0.9 | 52.6±16.2 | 86.4±5.3 | 61.8±8.4 | 93.3±1.3 |
| **PNN-MFL** | 69.4±10.3 | 97.3±1.8 | 74.3±8.7 | 84.3±7.4 | 80.2±7.3 | 96.2±2.4 |
| **DeepFM-MFL** | **82.1±7.7** | **99.0±0.1** | **81.2±5.6** | **91.9±4.5** | **86.9±8.2** | **96.4±0.8** |




(a) Cross entropy loss (b) Focal loss (c) **Modified focal loss (proposed)**

Fig. 2. Different loss function curves during the DeepFM model training. The horizontal axis represents the number of iterations "iter_nums" and the vertical axis represents the current loss value.




(a) (b)

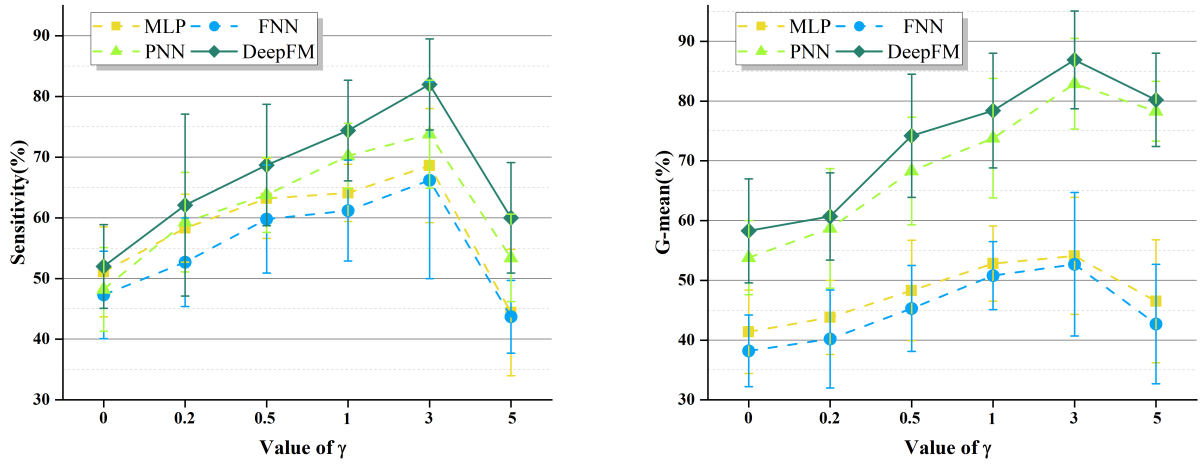
Fig. 3. Performance comparisons under different values of $f$. when $f = 1$, according to Eq. (4), the modified focal loss would ignore the distribution of training data. When $f = f()$ according to Eq. (6), the modified focal loss would consider the distribution of training data. The asterisks indicate significant differences between the seven methods obtained by paired t-tests (*p <0.05, and ** p <0.01).

the classification results across all students in the test set. As we can see from the table, the proposed DeepFM-MFL method can achieve the best detection performance compared to the others with a sensitivity of 82.1±7.7%, F1-score of 81.2±5.6%, AUC of 91.9±4.5%, GM of 92.3±7.9% and accuracy of 96.4±0.8%, specificity of 99.0±0.1%.

### C. The Influence of the Modified Focal Loss

In order to further evaluate the performance of the proposed modified focal loss, in this part, firstly, we study the convergence of the loss functions, the role of $f()$ function, and the influence of the $\gamma$ parameter in the loss function. Then, the classification performances of cross entropy loss (CE loss), focal loss (FL) [49] and the proposed loss (MFL) were compared. In this paper, the parameter $\gamma$ of $(1-\hat{p})^{\gamma}$ in focal loss was set to 3 (equals to $\gamma$ of MFL).

*1) Convergence of the proposed framework:* To demonstrate the convergence of the proposed framework, we illustrate the convergence process of the modified focal loss during the DeepFM model training phase in Fig. 2(c). As shown in Fig. 2(c), the proposed loss begins to converge with large fluctuations after 200 iterations. After 3000 iterations, the fluctuation decreases and the convergence becomes stable. By contrast, Fig. 2(a) shows the convergence process of cross entropy loss under the same conditions as the modified focal loss. As shown in Fig. 2(a), it is not until 1600 iterations that the CE loss begins to converge and the loss curve fluctuates greatly, which indicates that the output of the model is very




(a) (b)

Fig. 4. Performance comparisons under different values of the parameter $\gamma$ in Eq (4).

unstable. And Fig. 2(b) shows the focal loss curve. In short, comparing the 3 figures in Fig. 2, the CE loss curve and focal loss curve have more vibration than the proposed loss curve. Besides, the proposed loss function converges faster than others. We conjecture that the main reason is that our loss function takes into account the distribution of the training data so that the minority class will get more attention during the model training.

*2) Influence of the Parameters in Modified Focal Loss:* The main idea of the modified focal loss is that it introduces the distribution of the training dataset into the loss function by defining the function $f()$. Therefore, in order to investigate the role of the function $f()$, we respectively set $f$ as 1 and $f()$(whose value is calculated by Eq. (6). The results of 4 models (mentioned in part A) are shown in Fig. 3. This figure shows that sensitivity and g-mean are improved when $f = f()$. Next, we analyze the impact of the value of $\gamma$ in Eq. (4). From Eq. (4), we can see that with the increase of the value of $\gamma$ the value of loss will decrease. In this way, well-classified samples will get a smaller loss so that the model will pay less attention to these samples but more attention to others during the model training. Fig. 4 shows how the values of sensitivity and g-mean change with the value of $\gamma$. As can be seen from the figure, both sensitivity and g-mean reach their maximum when $\gamma$ is equal to 3.

*3) Classification performance comparisons of different loss functions:* In this part, we compare the classification performances of CE loss, focal loss, and the proposed loss on MLP, FNN, PNN, and DeepFM models. Fig. 5 shows the sensitivity values of 4 models with different loss functions. The figure demonstrates that the proposed loss function is superior to CE loss and focal loss under the depression detection task, which significantly improves the models' ability to identify positive samples(which refer to students who have been diagnosed with depression in this paper). Fig. 6 gives information about the distribution of estimated probabilities of depression of samples in a test set in a randomized experiment on the DeepFM model with 3 different loss functions. We can see immediately from the figure that more points are distributed above 0.5 when MFL is used in classification models which indicates that more positive samples are found. That is the reason why the sensitivity is improved.

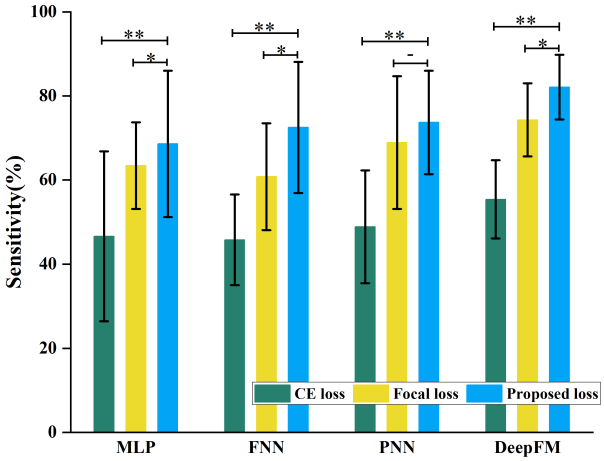



Fig. 5. Sensitivity comparisons of different loss functions. The asterisks indicate significant differences between the seven methods obtained by paired t-tests (*p <0.05, and ** p <0.01).

## IV. Discussion

### A. The Effect of Different Types of Data on Depression Detection

We further investigated the importance of 4 types of data (as shown in Table II ) in determining student depression via an ablation study. We trained the model using three of the four types of data mentioned in Table II to investigate the importance of the remaining type of data. In this way, the larger the decrease in detection performance compared to the model trained using all types of data, the more important that type of data is in determining whether the student is depressed. Fig. 7 illustrates the averaged sensitivity across 5-fold cross-validation with 10 runs (each run with a different

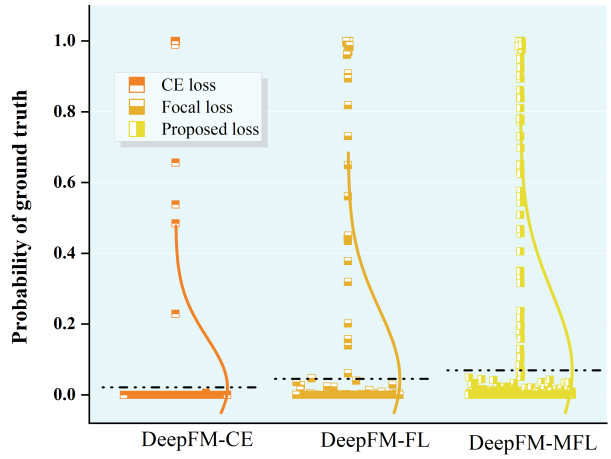



Fig. 6. Scatter diagram and normal distribution curve of estimated probability of ground truth of test set for DeepFM model with different loss functions; "A-B": A denotes the chosen classification model, and B denotes the chosen loss function. "CE": Cross entropy loss function; "FL": Focal loss function; "MFL": Modified focal loss function.

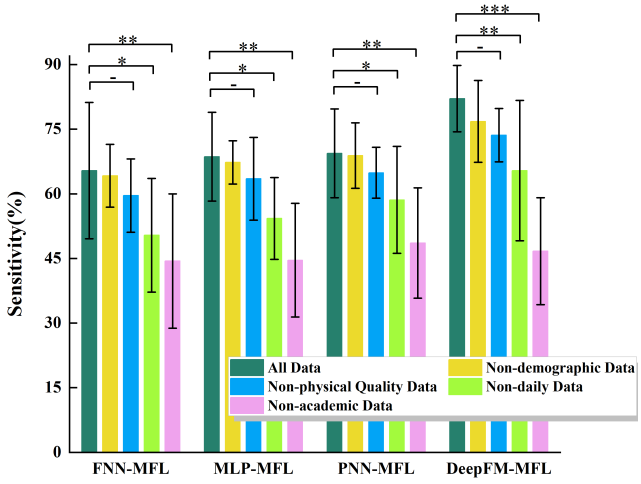



Fig. 7. Performance comparisons under different data. "All Data" means all of the data shown in Table II were used during model training. "Non-type Data" means apart from this type of data, the remaining data were used for model training. The asterisks indicate significant differences between the five conditions obtained by paired t-tests (* p <0.05, ** p <0.01, and *** p <0.001).

random seed). As can ben seen from the figure, all 4 types of data we collected are crucial for the detection of depression among students. Because removing any one of them would negatively affect the results of the experiment (sensitivities are reduced). The four types of data in descending order of importance are academic data, daily data, physical quality data, and demographic data. The results of the significance analysis show that academic data and daily data have a more significant effect on depression detection. This illustrates that depression among higher education students is mainly manifested through academic performance and some daily behaviors.

### *B. FM vs DNN vs DeepFM*

To investigate the role of FM and DNN, we trained them separately and tested their performance. As we can see from

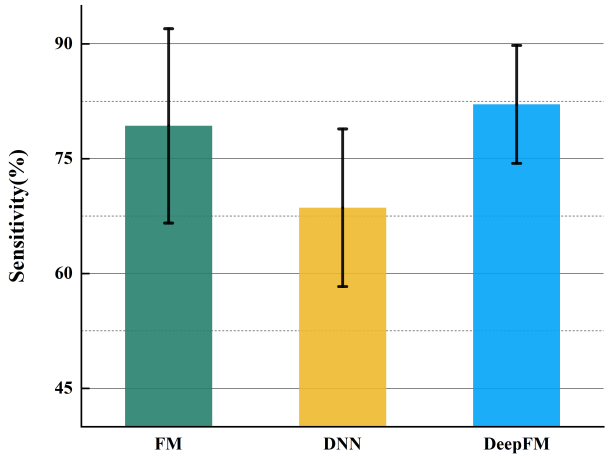



Fig. 8. The sensitivity comparisons of the FM component, the DNN component, and DeepFM.

Fig. 8, DeepFM has achieved the highest sensitivity compared with FM and DNN. Additionally, FM outperforms DNN, indicating that the linear correlation between behavioral data and students' depression is more pronounced compared to nonlinearity. Besides, in order to estimate the complexity of FM, DNN, and DeepFM, we compared their training time and test time. The average training times for FM, DNN, and DeepFM are 11.851s, 28.048s, and 54.486s, respectively. Their average test times are 0.034 ms, 0.067 ms, and 0.062 ms, respectively. All of these experiments were performed on a Dell PC with the Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz, 16 GB RAM, and 64-bit Windows 10 OS using Pycharm 2020.1.2 x64. We would like to point out that although DeepFM has a more complicated training process than FM and DNN, and the training time is a bit longer, it won't affect the computational speed of the target detection during the test phase.

### *C. Comparisons with Similar Studies*

To the best of our knowledge, we are the first to collect data for student depression detection in a campus big data scenario. And there are no similar studies using university data for depression detection, so our dataset type is unique at this time. In this case, it is not fair to compare our approach with other studies because we use different data. Although that said, there are several methods, i.e., Random Forest (RF), Support Vector Machine (SVM) [19] [20], Random Forest with Random Oversampling (RF-ROS), Random Forest with Tomek link (RF-TL) and Random Forest with Random Oversampling and Tomek link (RF-ROSTL) [17] that might apply to our data. We thus implemented them and tested them on our dataset. Fig. 9 illustrates the averaged classification accuracy across 5-fold cross-validation with 10 runs (each run with a different random seed). As we can see from the figure, the proposed DeepFM-MFL method can in general achieve the highest accuracy with regards to others.

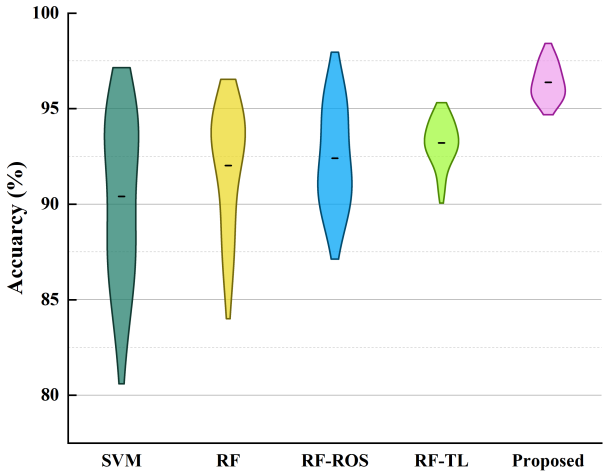



Fig. 9. Violin plots represent the distributions of classification accuracy achieved by 5 methods. The thick black line in each violin indicates the mean value.

TABLE V
THE DISTRIBUTIONS OF GENDER AND YEAR OF ENTRY

| | Gender | | Year of entry | | |
|---|---|---|---|---|---|
| Lable | Male | Female | 2015 | 2016 | 2017 |
| **Depressed** | 120 | 59 | 56 | 63 | 60 |
| **Normal** | 2371 | 668 | 1030 | 996 | 1013 |

### D. Comparisons among diverse student populations

To explore whether there exists a significant difference between diverse student populations, we conducted some ablation studies on the feature of "Gender" and "Year of entry", respectively. The distribution of "Gender" and "Year of entry" have been shown in Table V. We find that the depression rate of females is higher than that of males. The male base is larger than the female base, which may be due to the fact that the ratio of males to females in this university is 7:3. There is no significant difference in the number of students from different "Year of entry". Fig.10 and Fig.11 show the depression detection performance between different "Gender" and "Year of Entry" on the proposed model, which can be seen from the 2 figures that there indeed don't exist significant differences across "Gender" and "Year of entry".

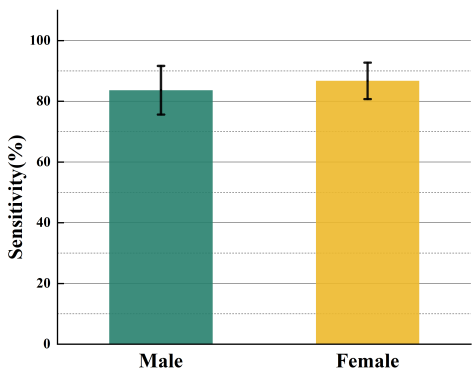



Fig. 10. The sensitivity comparisons between different "Gender" on DeepFM-MFL.

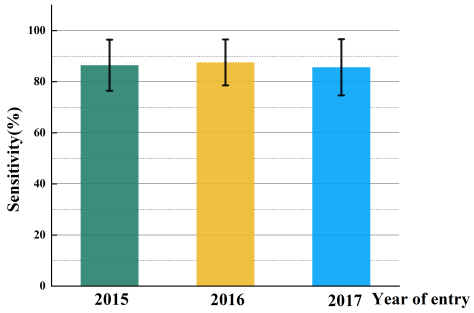



Fig. 11. The sensitivity comparisons between different "Year of entry" on DeepFM-MFL.

### E. Future Work

In this study, we propose a new method to detect depression among higher education students. Real-world student behavioral data is used to detect student depression. And we have achieved a high average accuracy of 96.4%. However, there is still room for improvement, and optimization can be done in both method and application.

Firstly, for the task of imbalanced classification, apart from the class imbalance loss function proposed in this paper, an algorithm-level method, the imbalance classification problem can be also addressed at the data level. For example, generative models [50] [51] [52] can synthesize samples from the minority class. Besides, oversampling methods [53] [54] [55] [56]can also be used to achieve data augmentation. In the future, data augmentation can be used at the data level to further address the imbalance problem. Secondly, identifying the levels of students' depression is definitely important for depression detection and is beneficial. In situations of scarce healthcare resources, major depressive disorder deserves primary attention as it is a significant contributor to poor academic performance, substance abuse, and attempted and completed suicide [57] [58] [59]. However, we haven't achieved this in this preliminary study drawback as its main purpose is to validate the feasibility of depression detection using the data automatically collected by the university system. Finally, in the next phase, we will also further process the data to extract the features that indicate students' behavioral changes, verify the role of students' behavioral changes in depression detection, and design a model to capture the characteristics of such changes.

## V. CONCLUSION

This study presented a preliminary study of depression detection using student behavioral data automatically collected by the University system. DeepFM-MFL was proposed to enhance depression detection performance via learning linear and nonlinear relations between student behavioral data and depression, and MFL was proposed to alleviate the data imbalance caused by the fact that the proportion of healthy students outweighed those diagnosed with depression significantly. The experiment results have illustrated that the proposed DeepFM-MFL obtained the best performance compared with MLP, FNN and PNN.

## ACKNOWLEDGMENT

All authors would like to thank the volunteers for data acquisition in this study.

## REFERENCES

[1] "Depression," 2021, https://www.who.int/news-room/fact-sheets/detail/depression/.

[2] C. G. Davey and P. D. McGorry, "Early Intervention for Depression in Young People: A Blind Spot in Mental Health Care," *Lancet Psychiatry*, vol. 6, no. 3, pp. 267–272, 2019.

[3] R. M. Mehmood, H.-J. Yang, and S.-H. Kim, "Children Emotion Regulation: Development of Neural Marker by Investigating Human Brain Signals," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[4] G. Sharma and A. M. Joshi, "SzHNN: A Novel and Scalable Deep Convolution Hybrid Neural Network Framework for Schizophrenia Detection Using Multichannel EEG," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.

[5] X. Sun, C. Ma, P. Chen, M. Li, H. Wang, W. Dang, C. Mu, and Z. Gao, "A Novel Complex Network-Based Graph Convolutional Network in Major Depressive Disorder Detection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–8, 2022.

[6] D. Yang, B. Gao, W. Woo, H. Wen, Y. Zhao, and Z. Gao, "Wearable Structured Mental-Sensing-Graph Measurement," *IEEE Trans. Instrum. Meas.*, pp. 1–1, 2022.

[7] M. A. Uddin, J. B. Joolee, and Y.-K. Lee, "Depression Level Prediction using Deep Spatiotemporal Features and Multilayer Bi-ltsm," *IEEE Trans. Affect. Comput.*, 2020.

[8] M. Niu, Z. Zhao, J. Tao, Y. Li, and B. W. Schuller, "Dual Attention and Element Recalibration Networks for Automatic Depression Level Prediction," *IEEE Trans. Affect. Comput.*, 2022.

[9] L. Yang, D. Jiang, and H. Sahli, "Integrating Deep and Shallow Models for Multi-modal Depression Analysis—Hybrid Architectures," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 239–253, 2018.

[10] W. C. de Melo, E. Granger, and M. B. Lopez, "MDN: A Deep Maximization-differentiation Network for Spatio-temporal Depression Detection," *IEEE Trans. Affect. Comput.*, 2021.

[11] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated EEG-based Screening of Depression using Deep Convolutional Neural Network," *Comput. Methods Programs Biomed.*, vol. 161, pp. 103–113, 2018.

[12] A. Seal, R. Bajpai, J. Agnihotri, A. Yazidi, E. Herrera-Viedma, and O. Krejcar, "DeprNet: A Deep Convolution Neural Network Framework for Detecting Depression Using EEG," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[13] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, and B. Hu, "Feature-level Fusion Approaches Based on Multimodal EEG Data for Depression Recognition," *Inf. Fusion*, vol. 59, pp. 127–138, 2020.

[14] J. Shen, X. Zhang, X. Huang, M. Wu, J. Gao, D. Lu, Z. Ding, and B. Hu, "An Optimal Channel Selection for EEG-based Depression Detection via Kernel-target Alignment," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2545–2556, 2020.

[15] R. Razavi, A. Gharipour, and M. Gharipour, "Depression Screening Using Mobile Phone Usage Metadata: A Machine Learning Approach," *J Am. Med. Inform. Assoc.*, vol. 27, no. 4, pp. 522–530, 2020.

[16] X. Xu, P. Chikersal, J. M. Dutcher, Y. S. Sefidgar, W. Seo, M. J. Tumminia, D. K. Villalba, S. Cohen, K. G. Creswell, J. D. Creswell *et al.*, "Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 1, pp. 1–27, 2021.

[17] S. Sawangarreerak and P. Thanathamathee, "Random Forest with Sampling Techniques for Handling Imbalanced Prediction of University Student Depression," *Information*, vol. 11, no. 11, p. 519, 2020.

[18] R. Chiong, G. S. Budhi, and S. Dhakal, "Combining Sentiment Lexicons and Content-based Features for Depression Detection," *IEEE Intell. Syst.*, vol. 36, no. 6, pp. 99–105, 2021.

[19] Z. Guo, N. Ding, M. Zhai, Z. Zhang, and Z. Li, "Leveraging Domain Knowledge to Improve Depression Detection on Chinese Social Media," *IEEE Trans. Comput. Soc. Syst.*, 2023.

[20] J. Angskun, S. Tipprasert, and T. Angskun, "Big Data Analytics on Social Networks for Real-time Depression Detection," *J. Big Data*, vol. 9, no. 1, p. 69, 2022.

[21] L. Tong, Z. Liu, Z. Jiang, F. Zhou, L. Chen, J. Lyu, X. Zhang, Q. Zhang, A. Sadka, Y. Wang *et al.*, "Cost-sensitive Boosting Pruning Trees for Depression Detection on Twitter," *IEEE Trans. Affect. Comput.*, 2022.

[22] T. Shen, J. Jia, G. Shen, F. Feng, X. He, H. Luan, J. Tang, T. Tiropanis, T. S. Chua, and W. Hall, "Cross-domain Depression Detection via Harvesting Social Media," in *IJCAI Int. Joint Conf. Artif. Intell.*, 2018.

[23] C. Fischer, Z. A. Pardos, R. S. Baker, J. J. Williams, P. Smyth, R. Yu, S. Slater, R. Baker, and M. Warschauer, "Mining Big Data in Education: Affordances and Challenges," *Rev. Res. Educ.*, vol. 44, no. 1, pp. 130–160, 2020.

[24] M. Feng, M. Cheng, X. Ji, L. Zhou, Y. Dang, K. Bi, Z. Dai, and Y. Dai, "Finding the optimal CO2 adsorption material: Prediction of multi-properties of metal-organic frameworks (MOFs) based on DeepFM," *Sep. Purif. Technol.*, vol. 302, p. 122111, 2022.

[25] W. Liu, Z. Zhang, Y. Bai, Y. Liu, A. Yang, and J. Li, "A DeepFM-Based Non-Parametric Model Enabled Big Data Platform for Predicting Passenger Car Sales in Sustainable Way," *IEEE trans. Intell. Transp. Syst.*, 2023.

[26] Y. Liu, R. Qing, Y. Zhao, X. Wang, Z. Liao, Q. Li, and B. Cao, "Fusing Spatio-Temporal Contexts into DeepFM for Taxi Pick-Up Area Recommendation." *Comput. Syst. Sci. Eng.*, vol. 46, no. 1, 2023.

[27] "Z-score," 2024, https://pypi.org/project/miceforest/.

[28] "Miceforest," 2024, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.zscore.html.

[29] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9: Validity of A Brief Depression Severity Measure," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, 2001.

[30] A. Hysenbegasi, S. L. Hass, and C. R. Rowland, "The Impact of Depression on The Academic Productivity of University Students," *J. Ment. Health Policy Econ.*, vol. 8, no. 3, p. 145, 2005.

[31] Y. Deng, J. Cherian, N. U. N. Khan, K. Kumari, M. S. Sial, U. Comite, B. Gavurova, and J. Popp, "Family and Academic Stress and Their Impact on Students' Depression Level and Academic Performance," *Front. Psychiatry*, vol. 13, p. 869337, 2022.

[32] S. Awadalla, E. B. Davies, and C. Glazebrook, "A Longitudinal Cohort Study to Explore the Relationship Between Depression, Anxiety and Academic Performance among Emirati University Students," *BMC Psychiatry*, vol. 20, pp. 1–10, 2020.

[33] R. Katikalapudi, S. Chellappan, F. Montgomery, D. Wunsch, and K. Lutzen, "Associating Internet Usage with Depressive Behavior among College Students."

[34] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones," in *Proc. 2014 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 3–14.

[35] K. Opoku Asare, Y. Terhorst, J. Vega, E. Peltonen, E. Lagerspetz, and D. Ferreira, "Predicting Depression from Smartphone Behavioral Markers Using Machine Learning Methods, Hyperparameter Optimization, and Feature Importance Analysis: Exploratory Study," *JMIR mHealth uHealth*, vol. 9, no. 7, p. e26540, 2021.

[36] J. Sander, M. Moessner, and S. Bauer, "Depression, Anxiety and Eating Disorder-related Impairment: Moderators in Female Adolescents and Young Adults," *Int. J. Environ. Res. Public Health*, vol. 18, no. 5, p. 2779, 2021.

[37] S. C. Garcia, M. E. Mikhail, P. K. Keel, S. A. Burt, M. C. Neale, S. Boker, and K. L. Klump, "Increased Rates of Eating Disorders and Their Symptoms in Women with Major Depressive Disorder and Anxiety Disorders," *Int. J. Eat. Disord.*, vol. 53, no. 11, pp. 1844–1854, 2020.

[38] B. G. G. da Costa, J.-P. Chaput, M. V. V. Lopes, L. E. A. Malheiros, and K. S. Silva, "Movement Behaviors and Their Association with Depressive Symptoms in Brazilian Adolescents: A Cross-Sectional Study," *J. Sport Health Sci.*, vol. 11, no. 2, pp. 252–259, 2022.

[39] X. Wang, Z.-d. Cai, W.-t. Jiang, Y.-y. Fang, W.-x. Sun, and X. Wang, "Systematic Review and Meta-Analysis of the Effects of Exercise on Depression in Adolescents," *Child Adolesc. Psychiatry Ment. Health*, vol. 16, no. 1, p. 16, 2022.

[40] F. B. Schuch and B. Stubbs, "The Role of Exercise in Preventing and Treating Depression," *Curr. Sports Med. Rep.*, vol. 18, no. 8, pp. 299–304, 2019.

[41] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic Nonverbal Behavior Indicators of Depression and Ptsd: The Effect of Gender," *J. Multimodal User Interfaces*, vol. 9, pp. 17–29, 2015.

[42] T. L. Osborn, K. E. Venturo-Conerly, A. R. Wasil, J. L. Schleider, and J. R. Weisz, "Depression and Anxiety Symptoms, Social Support, and Demographic Factors among Kenyan High School Students," *J. Child Fam. Stud.*, vol. 29, pp. 1432–1443, 2020.

[43] K. L. Taylor, E. J. Hadgkiss, G. A. Jelinek, T. J. Weiland, N. G. Pereira, C. H. Marck, and D. M. van der Meer, "Lifestyle Factors, Demographics and Medications Associated with Depression Risk in An International Sample of People with Multiple Sclerosis," *BMC Psychiatry*, vol. 14, no. 1, pp. 1–12, 2014.

[44] T. for Adolescents with Depression Study (TADS) Team *et al.*, "The Treatment for Adolescents with Depression Study (TADS): Demographic and Clinical Characteristics," *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 44, no. 1, pp. 28–40, 2005.

[45] B. S. Raghuwanshi and S. Shukla, "Class Imbalance Learning Using Under Bagging Based Kernelized Extreme Learning Machine," *Neurocomputing*, 2019.

[46] W. Zhang, T. Du, and J. Wang, "Deep Learning over Multi-field Categorical Data: –A Case Study on User Response Prediction," in *Lect. Notes Comput. Sci.* Springer, 2016, pp. 45–57.

[47] Y. Qu, B. Fang, W. Zhang, R. Tang, M. Niu, H. Guo, Y. Yu, and X. He, "Product-Based Neural Networks for User Response Prediction over Multi-Field Categorical Data," *ACM Trans. Inf. Syst*, vol. 37, no. 1, pp. 1–35, 2018.

[48] M. W. Gardner and S. Dorling, "Artificial Neural Networks (the Multilayer Perceptron)—A Review of Applications in the Atmospheric Sciences," *Atmospheric Environ.*, vol. 32, no. 14-15, pp. 2627–2636, 1998.

[49] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2980–2988.

[50] K. Huang and X. Wang, "ADA-INCVAE: Improved Data Generation Using Variational Autoencoder for Imbalanced Classification," *Appl. Intell.*, vol. 52, no. 3, pp. 2838–2853, 2022.

[51] B. Zhu, X. Pan, S. vanden Broucke, and J. Xiao, "A GAN-based Hybrid Sampling Method for Imbalanced Customer Classification," *Inf. Sci.*, vol. 609, pp. 1397–1411, 2022.

[52] H. Ding, L. Chen, L. Dong, Z. Fu, and X. Cui, "Imbalanced Data Classification: A KNN and Generative Adversarial Networks-Based Hybrid Approach for Intrusion Detection," *Future Gener. Comput. Syst.*, vol. 131, pp. 240–254, 2022.

[53] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, "FW-SMOTE: A Feature-weighted Oversampling Approach for Imbalanced Classification," *Pattern Recognit.*, vol. 124, p. 108511, 2022.

[54] K. Schultz, S. Bej, W. Hahn, M. Wolfien, P. Srivastava, and O. Wolkenhauer, "Convex Space Learning Improves Deep-Generative Oversampling for Tabular Imbalanced Classification on Smaller Datasets," *arXiv*, 2022.

[55] R. Liu, "A Novel Synthetic Minority Oversampling Technique Based on Relative and Absolute Densities for Imbalanced Classification," *Appl. Intell.*, pp. 1–18, 2022.

[56] Y. Li, X. Lai, M. Wang, and X. Zhang, "C-SASO: A Clustering-Based Size-Adaptive Safer Oversampling Technique for Imbalanced SAR Ship Classification," *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–12, 2022.

[57] E. Stockings, L. Degenhardt, Y. Y. Lee, C. Mihalopoulos, A. Liu, M. Hobbs, and G. Patton, "Symptom Screening Scales for Detecting Major Depressive Disorder in Children and Adolescents: A Systematic Review and Meta-Analysis of Reliability, Validity and Diagnostic Utility," *J. Affect. Disord.*, vol. 174, pp. 447–463, 2015.

[58] S. Byun, A. Y. Kim, E. H. Jang, S. Kim, K. W. Choi, H. Y. Yu, and H. J. Jeon, "Detection of Major Depressive Disorder from Linear and Nonlinear Heart Rate Variability Features during Mental Task Protocol," *Comput. Biol. Med.*, vol. 112, p. 103381, 2019.

[59] B. Levis, Y. Sun, C. He, Y. Wu, A. Krishnan, P. M. Bhandari, D. Neupane, M. Imran, E. Brehaut, Z. Negeri *et al.*, "Accuracy of the PHQ-2 Alone and In Combination With the PHQ-9 for Screening to Detect Major Depression: Systematic Review and Meta-Analysis," *Jama*, vol. 323, no. 22, pp. 2290–2300, 2020.