Hindawi
Scientific Programming
Volume 2019, Article ID 2753152, 12 pages
https://doi.org/10.1155/2019/2753152

Research Article

Intelligent Behavior Data Analysis for Internet Addiction

Wei Peng,[1] Xinlei Zhang ,[2] and Xin Li [2]

[1]*Information Technology Support, East China Normal University, Shanghai 200062, China*
[2]*Shanghai Key Laboratory of Trustworthy Computing, MOE International Joint Lab of Trustworthy Software, East China Normal University, Shanghai 200062, China*

Correspondence should be addressed to Xinlei Zhang; xinleizhang1997@gmail.com

Received 26 August 2019; Revised 21 October 2019; Accepted 7 November 2019; Published 29 November 2019

Guest Editor: Aibo Song

Internet addiction refers to excessive internet use that interferes with daily life. Due to its negative impact on college students' study and life, discovering students' internet addiction tendencies and making correct guidance for them timely is necessary. However, at present, the research methods used in analyzing students' internet addiction are mainly questionnaires and statistical analysis, which relies on the domain experts heavily. Fortunately, with the development of the smart campus, students' behavior data such as consumption and trajectory information in the campus are stored. With this information, we can analyze students' internet addiction levels quantitatively. In this paper, we provide an approach to estimate college students' internet addiction levels using their behavior data in the campus. In detail, we consider students' addiction towards the internet is a hidden variable which affects students' daily time online together with other behavior. By predicting students' daily time online, we will find students' internet addiction levels. Along this line, we develop a linear internet addiction (LIA) model, a neural network internet addiction (NIA) model, and a clustering-based internet addiction (CIA) model to calculate students' internet addiction levels, respectively. These three models take the regularity of students' behavior and the similarity among students' behavior into consideration. Finally, extensive experiments are conducted on a real-world dataset. The experimental results show the effectiveness of our method, and it is also consistent with some psychological findings.

1. Introduction

Internet addiction disorder refers to excessive internet use that interferes with daily life [1]. Some research shows that the addiction towards the internet has a negative impact on college students, such as the backwardness of study, health, and social relationship [1–3]. Therefore, it is necessary to discover students' addiction tendencies towards the internet and make correct guidance for them.

At present, related works of internet addiction are concentrated on psychological fields. Such works focus on the causes, the influence of internet addiction, and internal mechanisms leading to internet addiction, together with methods to eliminate internet addiction. There are few works on calculating internet addiction levels quantitatively. Besides, the methods used for analyzing are mainly questionnaires and statistical analysis, which are cumbersome and relies on the domain experts heavily. Therefore, it is necessary to develop an approach to explore students' internet addiction level quantitatively and automatically.

Fortunately, with the development of the smart campus, students' behavior data are collected, such as the access data and consuming data. With these data, it is possible to analyze students' internet addiction levels quantitatively.

To this end, in this paper, we propose an approach to estimate students' internet addiction levels using their behavior data. Currently, there is no method to evaluate students' addiction level precisely, so we are unable to study it with supervised methods explicitly. Instead, we can calculate students' internet addiction level through another task. In detail, based on the definition of internet addiction, we consider that the student's internet addiction level is a hidden variable, which will affect students' daily time online. Besides, student's behavior data such as consuming data and the internet access gap reflect student's daily activities, which may also influence the time they spend online. Then, we can predict students' online time with their behavior data and internet addiction level. Through such a task, the internet addiction value can be inferred. Along this line, we propose a linear internet addiction (LIA) model, a neural network internet addiction (NIA) model, and a clustering-based internet addiction (CIA) model to capture the relationship between students' behavior data, internet addiction, and the time they spend online every day.

Furthermore, students have fixed disciplines every week, which leads to the regularity of time they spend online every week. LIA and NIA models take the regularity of students' behavior into consideration, and the CIA model mainly uses the relationship among students' behavior to learn their internet addiction level. Finally, we conduct extensive experiments on a real-world dataset from a Chinese college, including internet addiction calculation, internet addiction verification, and internet addiction analysis experiments. Particularly, to verify the internet addiction value we calculate is credible, we compare our results with the results evaluated from the psychological scale. The experimental results demonstrate the correctness and effectiveness of the model we propose. And the results are also consistent with some psychological findings.

2. Related Work

The main related work of this paper can be divided into two parts: internet addiction analysis and campus data mining.

2.1. Internet Addiction Analysis. Internet addiction analysis is a research direction in the psychological field. Some works focus on the causes of internet addiction. Researchers found that interpersonal difficulties, psychological factors, social skills, etc., are all reasons for internet addiction [1, 4, 5]. Other works aim at finding the influence of internet addiction. Upadhayay et al. claimed that excessive use of the internet would lead to the drawback of the study [2]. He et al. explored internet addiction's influence on the sensitivity towards punishment and award [6]. Their result shows that people with serious internet addiction are more sensitive to risk. There are also some works about the inner mechanism of forming internet addiction. Zhang et al. focused on the inner reason of family function's negative influence on internet addiction. They revealed that the stability and development of family might affect users' mental situations such as dignity and loneliness, and then such mental situations will have an influence on internet addiction [7]. Zhao et al. noticed that stressful life events make users feel depressed, which causes the user addicted to the internet [8].

2.2. Campus Data Mining. Data are produced everywhere in our daily life activities, for example, the consumption records, chatting records, web browsing records, and so on. Using such data, we are able to make some interesting applications, such as tag recommendation, which suggests a list of tags when a user wants to annotate an item. Wang et al. proposed the TAPITF model to combine both time awareness and personalization aspects into tag recommendation task [9]. Campus data mining refers to solving problems on campus with data mining methods. Some works mainly analyze students' daily behavior in life. Guan et al. predicted students' financial hardship through their smart card usage, internet usage, and students' trajectories on campus (Dis-HARD model) so that the school can offer those students with stipend portfolios [10]. Based on this work, Ye et al. proposed a CS^{3}G model [11], which predicted stipend portfolios with multimodal data. Their work has higher accuracy compared to the Dis-HARD model and protects students' privacy. The Bayesian method is widely used in many fields. Wang et al. proposed a Bayesian probabilistic multitopic matrix factorization model for rating prediction [12]. And similarly Zhu et al. proposed an unsupervised method under the framework of empirical Bayes to calculate students' procrastination value with their borrow info in the library [13]. Peng et al. proposed a deep topical correlation analysis approach to track students' thoughts and serve the development of smart campus using multimodal data [14]. There are also some works aiming at analyzing students' studying process and improving their performance in class, which is called educational data mining (EDM). For example, Burlak et al. identified if a student is cheating in an exam by analyzing their interactive data with online course systems such as start time, end time, IP address, and access frequency [15]. Abdi et al. predicted students' grades based on their answers to usual work and duration of stay on a question [16].

Above all, to the best of our knowledge, there is no work on analyzing internet addiction using students' daily behavior. And we are the first to analyze internet addiction based on their behavior data with data mining methods.

3. Preliminaries

Internet addiction is an abstract concept in the psychological field, so it is hard to give a measurable definition of internet addiction. To solve this problem, we first make a reasonable assumption about internet addiction. Then, based on this assumption, we calculate the internet addiction value using students' behavior data.

3.1. Internet Addiction Assumption. Psychological research shows that most college students are addicted to the internet [17]. And we mentioned that internet addiction refers to excessive use of internet interfering with daily life. Therefore, students with different internet addiction levels are very likely to spend different time online. Besides, different behaviors show the different activities in school, which in turn also leads to different online time. And students of different genders or departments will also have some differences in the internet use.

Based on such fact, we assume that internet addiction is a hidden factor, which may influence students' daily time online together with their behavior and profile information. Therefore, we will learn such factors by modelling how students' internet addiction and behavior influence daily online time. To simplify the problem, we also assume students' internet addiction level will not change in a semester.

3.2. Problem Formulation. Since we do not have any label about internet addiction level, we cannot use supervised methods to study students' internet addiction value. Thus, we need to estimate it through some known data. Based on our assumption that the internet addiction value is a hidden variable, which may affect the time students spend online, the value can be learned by predicting students' daily online time.

Formally, we define a_u as the internet addiction level of student u. Daily time online sequence of student u during a period T is represented as $\{T_u(t)\}$. And the daily behavior sequence of u during the same period is represented as $\{B_u(t)\}$. We also define the personal profile information of student u as $\{p_u\}$. Our task is to model the relationship $\{a_u, p_u, B_u(t)\} \longrightarrow \{T_u(t)\}$, which is how students' behavior and internet addiction influence their daily time online. Then the internet addiction level a_u can be calculated from this model. Note that t above is in the set T.

4. Internet Addiction Calculation Model

To calculate students' internet addiction level, we propose three internet addiction calculation models: the linear internet addiction (LIA) model, the neural network internet addiction (NIA) model, and the clustering-based internet addiction (CIA) model. For the LIA model, we mainly consider the linear relation between students' behavior, internet addiction level, and their daily online time. Furthermore, since the neural network is powerful to capture the higher order relation among features, we explore the NIA model to find that nonlinear relation between students' behavior, internet addiction level, and their daily online time.

As for the CIA model, instead of directly studying the relation between students' behavior, internet addiction level, and their daily online time, we think that students who spend more time online than the normal online time are more likely to be addicted to the internet. So we devise a clustering-based method to find the normal online time and then regard the difference between students' actual online time and the normal online time as their internet addiction level.

In this chapter, we first describe these three models in detail, and then we will discuss the advantages and disadvantages of each model.

4.1. Linear Internet Addiction (LIA) Model. In this section, we first introduce how we use a linear model to reveal the relationship of $\{a_u, p_u, B_u(t)\} \longrightarrow \{T_u(t)\}$. Then to strengthen the model, we take the regularity of students' behaviors into consideration.

4.1.1. Naive LIA. Based on the internet addiction assumption, the behavior is a factor which will influence students' online time. However, different kinds of behavior may have a different effect. Therefore, a weight vector is necessary to represent the different effects of each kind of behavior. The impact of behavior on online time is not different in individuals, so every student shares this weight vector. We deal with different kinds of personal attributes in the same way. Besides, even two students have the same behavior and personal attributes, and they may still spend different time online because of the difference in their addiction level towards the internet. We suppose that different internet addiction level is the only reason which causes different time online with the same behavior and personal attributes. Here comes our naive linear internet addiction model:

$$y_u(t) = wx_u(t) + a_u, \tag{1}$$

where $y_u(t)$ represents the duration student u spend online at time t. $x_u(t)$ refers to the combination of behavior vector and personal attributes of student u at time t, and w is the weight vector of that combined vector. a_u here is the internet addiction level of student u. Our task is to find the value of a_u and w that minimize the loss function, that is,

$$\arg\min_{w,a_u} \sum_{u\in U}\sum_{t\in T}\left(y_u(t) - w^T x_u(t) - a_u\right)^2 + \lambda\|w\|^2 + \mu\sum_{u\in U} a_u^2. \tag{2}$$

The item $\lambda\|w\|^2$ is used to prevent the model from overfitting. $\mu\sum_{u\in U} a_u^2$ can be used to adjust the weight between the behavior and internet addiction.

4.1.2. LIA with Regular Behavior. College students usually have a fixed curriculum. Therefore, their behavior has some regularity every week, which will also lead to the regularity of the time they spend online. Take student u as an example; courses on Monday are kind of boring, so he spends a lot of time surfing the internet. However, courses on Tuesday are hard, which means he must pay attention to the class, so he may not surf the internet in class. Based on such facts, it is necessary to take the regular online time into consideration.

So, we modify our linear internet addiction model by adding an item $d_u(\pi(t))$ to represent the regular online time of student u at time t. Due to the characteristics of the college study, they perform similar online habits every week. So here $\pi(t)$ means which day of time t is of the week it belongs to, and $d_u(x)$ means the regular online time of the day x of the week. Here comes our new model:

$$y_u(t) = wx_u(t) + a_u + d_u(\pi(t)). \tag{3}$$

For the convenience of calculation, we define $x_{2u}(t)$ as an 8-dimensional vector with the first item one standing for the internet addiction and others being a one-hot representation of the week. The formula above is equal to

$$y_u(t) = w^T x_u(t) + w_u^T x_{2u}(t), \tag{4}$$

with x_{2u} being equal to

$$\left(1, \pi_1(t), \pi_2(t), \pi_3(t), \pi_4(t), \pi_5(t), \pi_6(t), \pi_7(t)\right), \tag{5}$$

$$\pi_i(t) = \begin{cases} 1, & \text{if } \pi(t) = i; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Our task is to find a suitable w and w_u that will minimize the loss function, the first item of w_u is the internet addiction level of student u:

$$\arg\min_{w,w_u} \sum_{u\in U}\sum_{t\in T}\left(y_u(t) - w^T x_u(t) - w_u^T x_{2u}(t)\right)^2 + \lambda\|w\|^2 + \mu\sum_{u\in U}\left\|w_u\right\|^2. \tag{7}$$

Similarly, we add $\lambda\|w\|^2$ to prevent the formula from overfitting, and we use the formula $\mu\|w_u\|^2$ to adjust the weights between behavior, personal attributes, internet addiction level, and regular habits.

4.2. Neural Network Internet Addiction (NIA) Model. The neural network is able to model the high-level relationship among features. It is powerful in a variety of application scenarios [18–20]. For example, in the tag recommendation task, Yuan et al. utilized the multilayer perceptron to model the nonlinearities of interactions among users, items, and tags [21]. In this section, we develop a neural network internet addiction (NIA) model to represent the nonlinear influence of students' behaviors, personal attributes, internet addiction, and their regular behavior on their daily online time.

4.2.1. Network Structure. The neural network consists of two parts: the public part and the private part. We use the public part to represent that the effect of the behavior and personal attributes on daily online time is not different in individuals, which means the input of the public part is the combination of the behavior vector of student u on time t and his personal attributes vector $x_u(t)$. The weight matrix V_c and the threshold vector λ_c of this part will update every iteration.

Because the internet addiction level and regular behavior are different in individuals, we use a private part to depict such characteristics. Every student has his own weight matrix V_u and threshold vector λ_u, and the parameters will only be updated when the corresponding student's data are used as the input. The private input x_2 of student u on time t is the same as vector (5). To ignore the influence of regular behavior, we can also only keep the first item of vector (5).

The target output of the model is the actual online time of student u on time t: $\widehat{y}_u(t)$.

The structure of the network is shown as Figure 1.

Using the symbol we mentioned, the output of the public hidden layer is

$$b_c = f_c\left(V_c x_u(t) - \lambda_c\right). \tag{8}$$

The output of the private hidden layer is

$$b_u = f_u\left(V_u x_2 - \lambda_u\right), \tag{9}$$

and the output of the network is

$$y_u(t) = f_o\left(W\left(\text{concat}\left(b_c, b_u\right)\right) - \theta\right), \tag{10}$$

where f_c, f_u, and f_o are the activation functions of the public hidden layer, private hidden layer, and the output layer and θ is the threshold of the output layer. The network will update for every input, and the loss function we use is the mean square error:

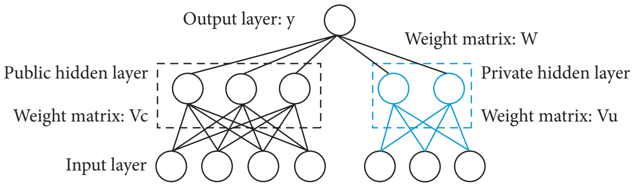


Figure 1: Neural network internet addiction model.

$$E = \frac{1}{2}\left(\widehat{y}_u(t) - y_u(t)\right)^2, \tag{11}$$

where $\widehat{y}_u(t)$ represents the actual online time of some student u on time t and $y_u(t)$ is the output of the whole model.

4.2.2. Internet Addiction Calculation. After the neural network training is completed, the sum of the contribution that internet addiction gives to the private hidden units is the value of students' internet addiction levels. We will calculate the internet addiction value as below:

$$a_u = \sum_{i=1}^{q_u} V_{uij}, \tag{12}$$

where q_u stands for the number of private hidden layer units. j is the corresponding index of internet addiction in the private part input vector, and here the index is one. V_u is the matrix, which connects the input layer and hidden layer of the private part. V_{uij} represents the i-th row and the j-th column value of the matrix V_u.

4.3. Clustering-Based Internet Addiction (CIA) Model. In this section, we develop a clustering-based method to calculate students' internet addiction value, which takes the similarity among students' behavior into consideration.

4.3.1. Internet Addiction Calculation. As the smartphone becomes an indispensable part of students' daily life, even the one not addicted to the internet will spend some time online, maybe for fun or just for killing time. However, those who are addicted to the internet heavily will spend much more time online than those who are not addicted to the internet. So, we believe that there is a normal online time corresponding to students' behavior, and those who spend more time online tend to be internet addicts. And the more time online compared with normal online time, the heavier the internet addiction level is. Therefore, here comes our online time prediction formula:

$$y_u(t) = n_u(t) + a_u, \tag{13}$$

where $y_u(t)$ represents the duration student u spends online at time t. $n_u(t)$ refers to the normal online time for student u at time t. a_u here is the internet addiction level of student u. Our task is to find the value of a_u that will minimize the loss function, that is,

$$\arg\min_{a_u} \sum_{u\in U}\sum_{t\in T}\left(y_u(t) - n_u(t) - a_u\right)^2 + \mu\sum_{u\in U} a_u^2. \quad (14)$$

The item $\mu\sum_{u\in U} a_u^2$ is used to adjust the weight between the normal online time and internet addiction.

4.3.2. Normal Online Time. Due to this, the students have different activities every day, and normal online time differs from behavior to behavior. To find the normal online time $n_u(t)$ of student u at time t, we first need to find those who behave similarly with student u at time t. The average online time of those who behave similarly with student u at time t is approximately equal to the normal online time. That is,

$$n_u(t) = \frac{\sum_{(u',t')\in S} y_{u'}(t')\mathrm{sim}\left(x_u(t), x_{u'}(t')\right)}{\sum_{(u',t')\in S}\mathrm{sim}\left(x_u(t), x_{u'}(t')\right)}, \quad (15)$$

where $x_u(t)$ represents the behavior vector of student u at time t. S stands for the similar behavior set, and $x_{u'}(t')$ is one similar behavior vector of $x_u(t)$, which means the behavior vector of student u' at time t' is similar with that of student u at time t. Students from different departments may behave differently because of the discipline characteristic, which will lead to a slight difference in normal online time. For example, students from the software engineering department may tend to spend more time online than the other students. So, we also take the profile information into consideration, the symbol $x_u(t)$ here is equal to the vector $x_u(t)$ in Section 4.1. And the formula sim (a, b) is the similarity value of vector a and vector b.

Considering the calculation amount, we do not compare every behavior of all students at all the time. Instead, we first aggregate students' behavior into k categories. When we need to find the similar behavior set S of the behavior vector $x_u(t)$, we first find the category the behavior vector $x_u(t)$ belongs to; let us assume the category is c, and then we start to calculate the similarity between $x_u(t)$ and all the other behavior vectors $x_{u'}(t')$ in the category c. Finally, we keep those behavior vectors that have similarity greater than a threshold in the set S; based on which, we will get the normal online time $n_u(t)$.

4.4. Model Comparison. The idea of LIA and NIA is direct, and the target of these two models is to find how students' behavior and internet addiction level influence their daily online time. The LIA model is easier to train because it has fewer parameters than the NIA model. Though the NIA model is much more powerful, it is hard to train the network as there are so many parameters.

The idea of the CIA conforms to our intuitive thinking that those who spend more time online than the normal time are more likely to be addicted to the internet. However, it is hard to find the normal online time. In this paper, we calculate the normal online time of a student u on a specific day t by averaging the online time of those students who behave similarly with u on t. The correctness of internet addiction calculation may be influenced by the precision of the clustering results.

5. Experiments

5.1. Data Description. Our data come from a Chinese college, including students' consuming records in the school restaurant and internet access records. Besides, they also include the personal attributes information of students such as department, gender, and age.

The consumption records consist of students' profiles, time, place, and amount of one consumption. Students have various consumption behaviors like normal dining, snack, shower, and deposit. Here we consider deposit is a special behavior, which is saving money to the school card. The behavior category can be identified through the place where consumption behavior takes place. For example, consumption in the school restaurant must be normal dining behavior, and consumption in the bathhouse must be shower behavior. Therefore, we first divide the places into different categories and then extract the consuming amount on dining, snack, shower, deposit, and total consuming amount per hour from the consuming records. We also count students' daily consumption frequency.

Besides, students can access the internet using campus Wi-Fi only when they get authenticated. Based on the authentication record, we extract the time student accesses the campus Wi-Fi per hour. And such time is approximate to the time they spend on the campus. Similarly, at each time when a student visits a website, a connection record is generated. When the visit is completed, there will be a disconnected record. Based on these records, we can extract the student's actual online time and the average gap between two internet access per day. After feature extraction, combining the daily consuming behavior and online behavior (actual online time is excluded), the behavior of a student in a day can be represented as a vector. We also represent every student with the one-hot method using their profile information.

Due to some reasons, we do not have students' internet access records in the dormitory and library. It is considered that students' activities are mainly centralized around classrooms and canteens as well as some college student activity centers. In class, students need to listen to the teachers most of the time, and at the restaurant, they always play with a phone to kill time. Therefore, the actual online time we extract is mainly about the entertainment. Intuitively, the entertainment time is suitable to be used to calculate the internet addiction level.

We choose the records of undergraduate students enrolled in 2016 and 2015 from September 1, 2018 to November 11, 2018. After dropping students with record number less than 35 days, there are 3767 students. The first 50 records are used for training, and the left records are used for testing. Students' profile representation and daily behavior vector are shown in Table 1.

Table 1: Features used in experiments.

Type	Feature	Dimension	Representation
Profile	Gender	2	One hot
	Department	61	
	Age	8	
Consumption	Dining amount	24	Statistical value
	Snack amount	24	
	Shower amount	24	
	Deposit amount	24	
	Total amount	24	
	Frequency	1	
Internet	Wi-Fi access time	24	Statistical value
	Internet access gap	1	

5.2. Internet Addiction Calculation. LIA, NIA, and CIA models can be used to study the internet addiction level by predicting students' online time every day. To show the correctness of our models, we conduct several experiments.

For LIA and NIA models, we conduct three experiments. The first experiment removes the internet addiction and regular behavior part of LIA and NIA models and predicts students' daily online time using students' behavior data and profile information, which is considered as a baseline. The second experiment only takes internet addiction into consideration. For LIA, it means using the naive LIA model, and as for NIA, it means there is only one item of the input of the private part. The last experiment takes internet addiction and regular behavior into consideration. For LIA, it means using LIA with regular behavior model, and for NIA, it means there is 8 items of the input of the private part. For the CIA model, we conduct two experiments: the first experiment uses the average online time in the similar behavior set S as the prediction, which is considered as a baseline. And the other experiment first calculates the internet addiction value of each student using equation (14) and then predicts students' online time using their neighbors' actual online time and the internet addiction value a_u by equation (13).

For the linear model, the value of λ is set to 0.6, and μ is set to 0.4. For the neural network model, the activation function of the hidden layer is $f(x) = x$, and the activation function of the output layer is $f(x) = \tanh(x)$. In addition, the number of public hidden layer units is 10, and the number of private hidden layer units is 2. The learning rate is set to 0.01, and the number of the epoch is 40. Note that for the third experiment of the NIA model, we set the learning rate to 0.05 which will get the best prediction accuracy. For the clustering-based model, the threshold is set to 0.7 and the cluster number is set to 50. The MSE performance of each method is shown in Table 2.

Table 2: Regression results.

Model	Feature		
	ia−	ia	ia+
LIA	0.000056	**0.000048** (14.3%)	0.000050 (10.7%)
NIA	0.000092	**0.000083** (9.8%)	0.000086 (6.5%)
CIA	0.000138	**0.000127** (8.0%)	No such condition

"ia−" refers to the baseline experiment; "ia" represents the second experiment; "ia+" stands for the third experiment.

From the results in Table 2, we know that no matter which model, the prediction accuracy increases with our internet addiction assumption. Such results guarantee the correctness of our internet addiction assumption. However, for the LIA and NIA models, adding the assumption of regular behavior, the accuracy does not improve compared to the results without such an assumption. One possible reason is that there is some volatility in students' behavior; however, LIA and NIA are not able to model it. Generally, the results of the neural network model and clustering-based model are worse than that of the linear model. Maybe it is because the linear model is strong enough to represent the relationship between students' behavior, internet addiction, and online time. And there are too many parameters in the neural network model, which is not easy to train. Though clustering students into several categories before calculating the similarity will reduce the computing complexity, the prediction results depend on the clustering results, and that may cause some error. The bias of the clustering results may be a reason leading the worst prediction accuracy of the CIA model.

5.3. Internet Addiction Verification. In this section, we conduct some experiments to verify the correctness of the methods we propose. First, we show the consistency of the internet addiction value we calculated using the models we proposed and the value evaluated through the psychological scale. And then, we devise regression and classification tasks to verify the critical role the internet addiction value we calculated plays on daily online time prediction task.

5.3.1. Comparison with the Psychological Scale. In psychology, researchers usually use the internet addiction scale to measure if people are addicted to the internet. Therefore, we use a questionnaire to test if a student is an internet addict and compare the results calculated by the questionnaire with that by our method.

In consideration of the national condition of China, we choose the internet addiction scale devised by professor Fan [22], which is widely used in Chinese psychological researches. As the situation today is not exactly the same with that several years ago, we cut some questions on that scale and only keep five necessary questions. And we use 4 points Likert scale to measure the degree of each question. See Table S1 in the Supplementary Material section for the details of the scale we used.

After giving the questionnaire to students, we retrieve 128 questionnaires, which are enough to analyze students' internet addiction levels in the psychological field. The students who complete the questionnaire consist of 78 males and 50 females, and there are around 81 students in grade 3 and 47 students in grade 4, which shows the samples are evenly distributed.

To show the effectiveness of the new scale we use, we calculate the reliability and validity of our scale, which are two dimensions to test if a scale is credible to use in psychology. The reliability and validity of our scale are 0.789 and 0.731 separately. The higher the value of the reliability and validity is, the better the scale is, and 0.7 means that our scale is credible enough to test the internet addiction.

On the principle of voluntariness, we did not force students to write down their student id or name. Since there are only 39 students who volunteer to give us their student id, we mainly compare those students' results judged by the psychological scale and that by our methods. There are five questions on our scale. Because we use 4 points Likert scale to measure, the total grade is 20. The greater the grade a student gets, the more likely this student is addicted to the internet. We define those whose grade less than ten is not addicted to the internet, and the others are internet addicts. As for the results calculated by the LIA model, we consider those whose value greater or equal than 0.45 is addicted to the internet. The threshold of NIA and CIA models is set to 0.5 and 0.35 separately. 0.45, 0.5, and 0.35 are approximate to the average value of the corresponding method. We use F1 score to evaluate the consistency between the results of the LIA model, the NIA model, the CIA mode, and the psychological scale. The results are shown in Table 3.

From Table 3, we see that all the internet addiction values calculated through these three models are consistent with the results evaluated from the psychological scale. Particularly, though the CIA model performs poorly in the internet addiction calculation task, comparing with the NIA model, the internet addiction value of the CIA model is more consistent with the psychological scale results. Such results show the correctness of our methods and give us a clue that the relationship among behaviors is an important factor when calculating the internet addiction value.

5.3.2. *Online Time Prediction.*

5.3.2. Online Time Prediction. Based on our assumption, internet addiction is a hidden variable, which will influence students' daily time online. Therefore the learned internet addiction value should be a useful feature to predict students' online time. We devise two tasks to verify the correctness of our learned internet addiction value.

The aim of the regression task is to predict students' daily online time. The baseline experiment takes the daily behavior vector and the profile information as the input. The contrast experiment predicts the daily online time using students' internet addiction value, daily behavior vector, and profile information. For the classification task, it is similar to the regression task. First, the records are divided into two parts: one part with online time greater than or equal to the average online time, the other part with an online time less than the average online time. The aim of the classification task is to predict which part online time belongs to. The experiment settings are the same as the regression task. The methods used in the regression task and classification task consist of the decision tree (DT), support vector machine (SVM), k-nearest neighbors (KNN), random forest (RF), gradient boosting decision tree (GBDT), bagging and extremely randomized trees (ET).

MSE is used as the evaluation method for the regression task, and F1 score for the classification task. The results are shown in Tables 4 and 5.

Table 3: F1 score between the results of our methods and psychological scale.

Model	LIA	NIA	CIA
F1 score	0.71	0.63	0.71

From Table 4, we observe that, for the regression task, the SVM model gets a huge mean square error. One possible reason may be that it is not suitable for this task, so we will ignore the SVM results in the discussion below. After adding the internet addiction value calculated by LIA and CIA models, all the prediction accuracies lift. And after adding internet addiction value calculated by NIA, although the promotion of prediction accuracy is not as remarkable as that of adding the value calculated by LIA or CIA models, most of the methods still get some promotion.

For the classification task, no matter which internet addiction value is added to the behavior vector, except for the effect of the SVM method has not changed, the effect of all the other methods has evidently been improved.

Generally speaking, after adding the internet addiction value calculated by LIA, NIA, or CIA, both regression and classification tasks get a remarkable promotion, which shows the effectiveness of the internet addiction value learned by the models we propose.

5.4. Internet Addiction Analysis. To show the internet addiction situation in college, we analyze the distribution of internet addiction and the differences of the internet addiction level among different groups such as different gender and department. Because the naive LIA model has the best prediction accuracy when studying students' internet addiction value and the value learned through the naive LIA model is the most consistent with the psychological results, the following analysis is based on the value calculated by the naive LIA model.

5.4.1. Internet Addiction Distribution. Figure 2(a) illustrates the number of students with respect to the calculated internet addiction value. The greater the internet addiction value is, the more serious students' addiction towards the internet is. We observe that internet addiction distribution is similar to a normal distribution. To show the distribution of the internet addiction value clearly, we delete the value greater than 0.7 or less than 0.2, which is shown in Figure 2(b). If we define internet addiction less than 0.45 is normal, from Figure 2(b), we observe that most of the students are addicted to the internet with different levels.

5.4.2. Internet Addiction Differences among Groups. To reveal the differences in internet addiction between genders, we count the average internet addiction value of different genders. And we also count the average online time of different genders. Figure 3 shows that girls spend more time on the internet than boys. However, boys are more addicted to the internet than girls. Such a result is consistent with the finding in the psychological field. Wei et al. investigated the

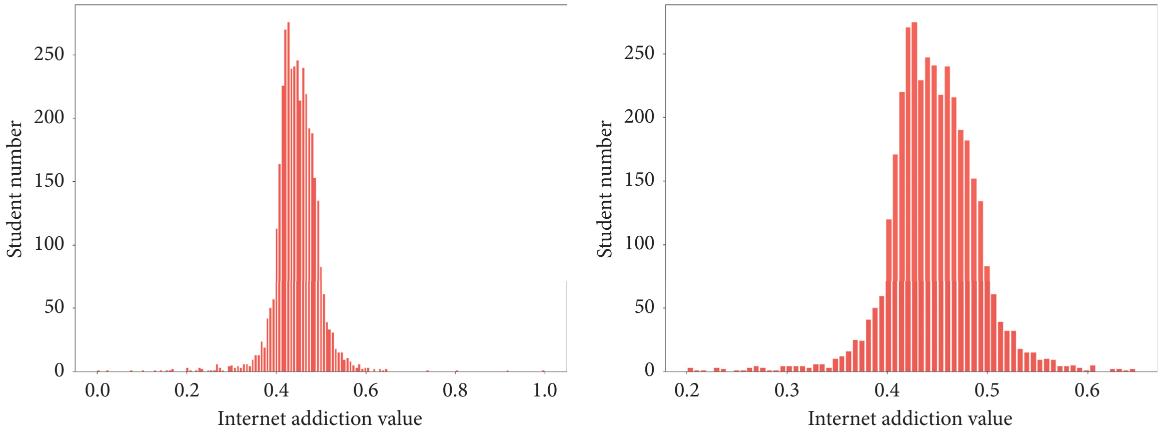
Table 4: Regression task.

Feature model	ia–	ia (LIA)	ia (NIA)	ia (CIA)
DT	0.000076	**0.000061** (19.7%)	0.000072 (5.3%)	0.000064 (15.8%)
SVM	0.004114	0.003636 (11.6%)	0.003677 (10.6%)	**0.003024** (26.5%)
KNN	0.000065	**0.000064** (1.5%)	0.000066 (−1.5%)	**0.000064** (1.5%)
RF	0.000040	0.000039 (2.5%)	0.000040 (0%)	**0.000038** (5%)
GBDT	0.000042	**0.000039** (7.1%)	0.000042 (0%)	0.000040 (4.8%)
Bagging	0.000041	**0.000039** (7.3%)	0.000041 (0%)	**0.000039** (7.3%)
ET	0.000068	**0.000057** (16.2%)	0.000065 (4.4%)	0.000065 (4.4%)

"ia–" refers to the baseline experiment and "ia (LIA)" stands for the experiment with the internet addiction value learned by the naive LIA model, which gets the best results in the internet addiction calculation task using the LIA model. "ia (NIA)" represents the experiment with the best internet addiction value learned by the NIA model without regular behavior consideration, which gets the best results in the internet calculation task using the NIA model. Similarly, "ia (CIA)" refers to the experiment with the internet addiction value learned from the clustering-based model.

Table 5: Classification task.

Feature model	ia–	ia (LIA)	ia (NIA)	ia (CIA)
DT	0.960643	0.997667 (3.9%)	0.997989 (3.9%)	**0.998899** (4.0%)
SVM	0.960773	0.960773 (0%)	0.960773 (0%)	0.960773 (0%)
KNN	0.959270	0.970958 (1.2%)	**0.981605** (2.3%)	0.973924 (1.5%)
RF	0.967783	0.978654 (1.1%)	0.979268 (1.2%)	**0.981382** (1.4%)
GBDT	0.959812	0.961584 (0.2%)	0.960936 (0.1%)	**0.962591** (0.3%)
Bagging	0.965652	0.998827 (3.4%)	0.998481 (3.4%)	**0.999378** (3.5%)
ET	0.958128	0.966017 (0.8%)	**0.970712** (1.3%)	0.969175 (1.2%)



(a)

(b)

Figure 2: Internet addiction distribution. (a) Number of students with different levels of addiction. (b) Some students with the value greater than 0.7 or less than 0.2 are deleted.

internet addiction situation of the college student in Hubei Polytechnic University using questionnaires.

They point out that boys are usually not good at communication, and therefore, the communication in real life is not enough to meet their actual communication needs. The way of communication with the network as the medium is easier to control; that is, they can improve the quality and quantity of communication in this way, which meets their needs of communication. Besides, Girls are better than boys in time management ability and deal with network use time more reasonably. So boys are more addicted to the internet than girls [23]. The consistency with the findings of psychology further proves the correctness of the internet addiction value we learned.

Figure 4(a) illustrates the average internet addiction level of different departments. In general, except the internet addiction level of a few departments is extremely high, it fluctuates around 0.43. Furthermore, we statistically analyze the differences in internet addiction levels among students in different disciplines. In Figure 4(b), we can observe that there is no significant difference in internet addiction levels among students in different disciplines. The result is also consistent with the psychological finding in [23]. Experiments conducted by Wei et al. that demonstrate though there is some difference in the interpersonal health and time management ability among students in different disciplines, the difference is not significant. And the difference in internet addiction is not significant. The consistent result with

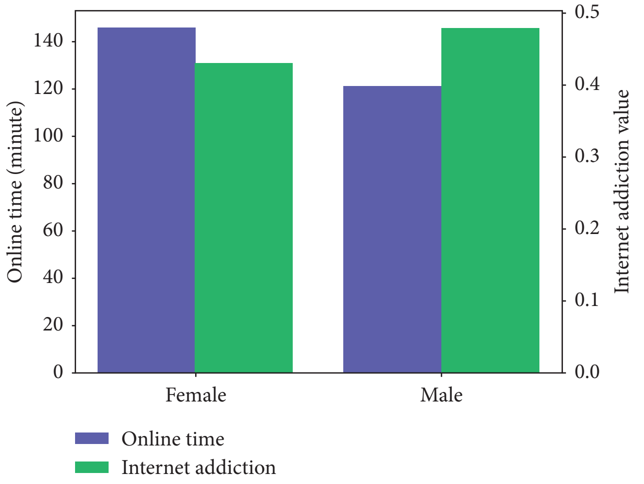


Figure 3: Differences of online time and internet addiction between different genders.

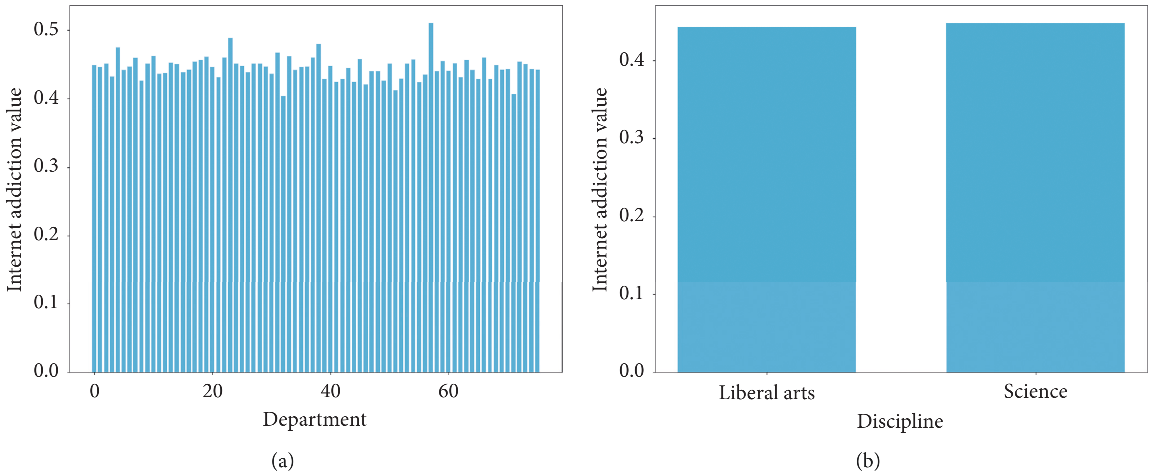


Figure 4: Differences of internet addiction among different departments and disciplines. (a) Department. (b) Discipline.

psychological findings is also evidence of the effectiveness of the internet addiction value we learned.

5.4.3. Effect of Internet Addiction on Online Time. The decision tree is a classical machine learning model. It is good at classification and regression tasks, and it is interpretable. Therefore, the decision tree model has plenty of applications in various fields [24–26]. To show the role internet addiction plays when predicting students' online time, we extract students' daily Wi-Fi access time, consuming amount, consuming frequency, average internet access gap, and actual online time. Then we conduct two binary classification experiments using classification and regression decision tree method: one predicts online time interval with daily Wi-Fi access time, consuming amount, consuming frequency, and average internet access gap, and the other predicts online time interval with daily Wi-Fi access time, consuming amount, consuming frequency, average internet access gap, and internet addiction value. Because the whole tree is too big to be put here, we select two representative branches. Note that all the values are normalized. The average value of the internet addiction value, consuming amount, consuming frequency, Wi-Fi access time, internet access gap, and online time is 0.45, 0.009, 0.044, 0.062, 0.004, and 0.015 separately.

From Figure 5(a), we know that Wi-Fi access time and average internet access gap are important features when predicting the online time. It is consistent with our intuitive thinking that less Wi-Fi access time and a long internet access gap will cause less online time. Figure 5(b) illustrates that after adding the internet addiction value, the value is critical for predicting daily online time. Particularly, in this branch, the relatively high internet addiction value is a reason leading to long online time.

5.4.4. Effect of Internet Addiction on Grade. The psychological research shows that internet addiction will damage students' study [1]. To show the bad influence of internet addiction and to verify the correctness of the internet addiction value we calculated, we do some statistics about the grades of those who are addicted to the internet and those who are not.

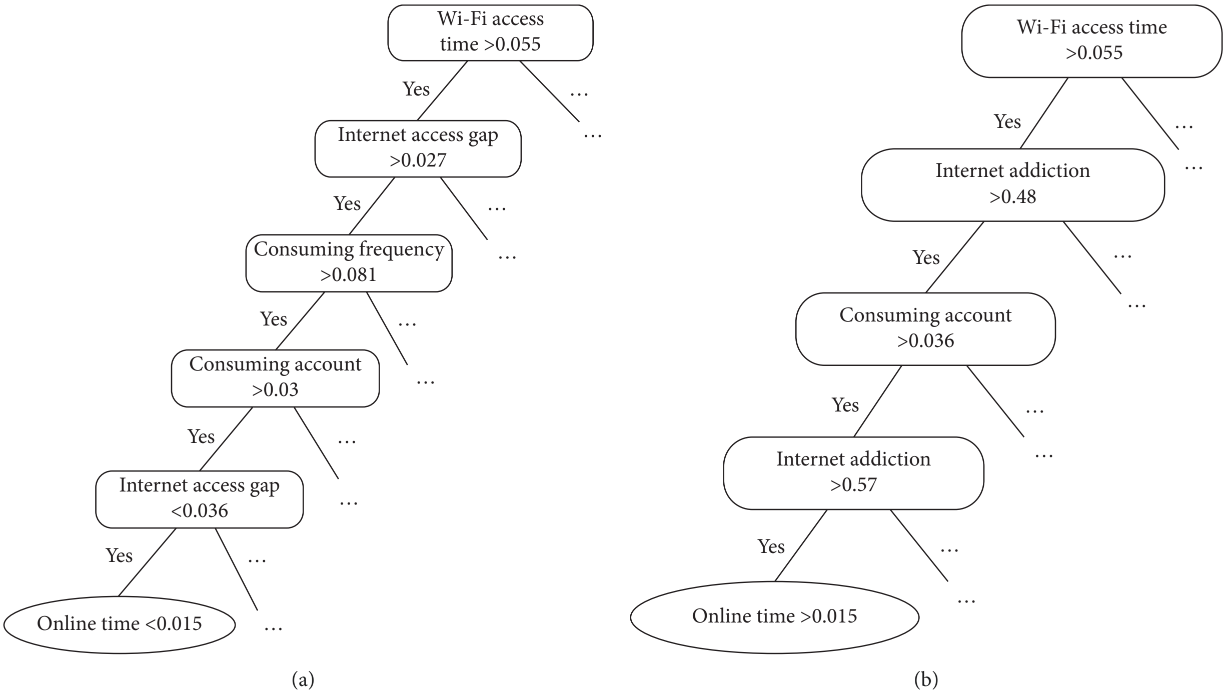


Figure 5: Decision tree with behavior and internet addiction value. (a) Prediction with behavior data. (b) Prediction with behavior data and internet addiction value learned through naive LIA.

As we mentioned before, there are only 39 students who volunteer to give us their student id, and one of them does not have any grade records, so the analysis of this part is mainly based on the grades of the remaining 38 students.

First, we define that students whose internet addiction values equal to or more than 0.45 are internet addicts, and the others are not. We divide students into two groups based on their internet addiction values. Then we calculate their average grade point of the second semester in 2018. At last, we count the average grade point and student number who failed at least one course of each group. The average grade of each student is calculated with the formula below:

$$G(u) = \sum_{c \in C(u)} \operatorname{cred}(c) * \operatorname{gp}(c), \tag{16}$$

where $G(u)$ refers to the average grade point of student u of the second semester in 2018, $C(u)$ stands for all the courses student u takes in this semester, $\operatorname{cred}(c)$ is the credit of course c, and $\operatorname{gp}(c)$ is the grade point of course c student u gets.

The analysis results are shown in Table 6.

Table 6: Classification task.

	Stu number	Average G	Failed stu number
Stu with ia	18	2.75	3
Stu without ia	20	3.21	1

"stu with ia" refers to the students who are addicted to the internet, and "stu without ia" refers to those who are not addicted to the internet. "Average G" stands for the average grade point of all the students in each group. "Failed stu number" is the number of students who fail at least one course in each group.

From this table, we see that almost half of the students are addicted to the internet. And the average grade point of students who are addicted to the internet is significantly lower than the normal students. There are more students who failed the exam in the internet addicts group than that in the other group. The statistics conform to the psychological findings that internet addiction has a bad influence on students' study. Such results further verify the correctness of the internet addiction value we calculated.

6. Conclusions

In this paper, we estimate college students' internet addiction levels quantitatively using their behavior data on the campus. Specifically, we define the internet addiction value as a hidden variable which will affect students' online time and formulate the problem as a regression problem.

Along this line, we first propose a linear internet addiction (LIA) model, which depicts the linear relationship among students' internet addiction level, behavior data, and time they spent online. To model the nonlinear relationship, we also provide a neural network internet addiction (NIA) model. Besides, we also develop a clustering-based internet addiction (CIA) model, which calculates the internet addiction based on the differences between students' actual online time and normal online time. These three models also take students' regular behavior and the similarity among students' behavior into consideration.

Finally, we conduct excessive experiments on a real-world dataset from a Chinese college, and the experimental results demonstrate the effectiveness of our model. The analysis results are consistent with some psychological findings, which also verify the correctness of the models we propose.

Data Availability

The behavior data used to support the findings of this study have not been made available due to privacy concerns.

Disclosure

It is an extension of the paper *Using Behavior Data to Predict the Internet Addiction of College Students* [27] which is published in the International Conference on Web Information Systems and Applications (WISA) in 2019.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Key R&D Program of China (no. 2017YFC0803700) and NSFC grants (nos. 61532021 and 61972155).

Supplementary Materials

Table S1. Internet usage survey of college student. (*Supplementary Materials*)

References

[1] Internet addiction disorder, 2019, https://en.wikipedia.org/wiki/Internet_addiction_disorder.

[2] N. Upadhayay and S. Guragain, "Internet use and its addiction level in medical students," *Advances in Medical Education and Practice*, vol. 8, pp. 641–647, 2017.

[3] Y. Xue, Y. Dong, M. Luo et al., "Investigating the impact of mobile SNS addiction on individual's self-rated health," *Internet Research*, vol. 28, no. 2, pp. 278–292, 2018.

[4] A. Fumero, R. J. Marrero, D. Voltes, and W. Peñate, "Personal and social factors involved in internet addiction among adolescents: a meta-analysis," *Computers in Human Behavior*, vol. 86, pp. 387–400, 2018.

[5] M. Z. Malak, A. H. Khalifeh, and A. H. Shuhaiber, "Prevalence of Internet Addiction and associated risk factors in Jordanian school students," *Computers in Human Behavior*, vol. 70, pp. 556–563, 2017.

[6] W. He, A. Qi, Q. Wang et al., "Abnormal reward and punishment sensitivity associated with Internet addicts," *Computers in Human Behavior*, vol. 75, pp. 678–683, 2017.

[7] Y. Zhang, X. Qin, and P. Ren, "Adolescents' academic engagement mediates the association between Internet addiction and academic achievement: the moderating effect of classroom achievement norm," *Computers in Human Behavior*, vol. 89, pp. 299–307, 2018.

[8] F. Zhao, Z.-H. Zhang, L. Bi et al., "The association between life events and internet addiction among Chinese vocational school students: the mediating role of depression," *Computers in Human Behavior*, vol. 70, pp. 30–38, 2017.

[9] K. Wang, Y. Jin, H. Wang, H. Peng, and X. Wang, "Personalized time-aware tag recommendation," in *Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 459–466, AAAI-18, New Orleans, LA, USA, February 2018.

[10] C. Guan, X. Lu, X. Li et al., "Discovery of college students in financial hardship," in *Proceedings of the 2015 IEEE International Conference on Data Mining*, pp. 141–150, IEEE, Atlantic City, NJ, USA, November 2015.

[11] H. J. Ye, D. C. Zhan, X. Li et al., "College student scholarships and subsidies granting: a multi-modal multi-label approach," in *Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 559–568, IEEE, Barcelona, Spain, December 2016.

[12] K. Wang, X. Zhao, H. Peng, and X. Wang, "Bayesian probabilistic multi-topic matrix factorization for rating prediction," in *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI*, pp. 3910–3916, New York, NY, USA, July 2016.

[13] Y. Zhu, H. Zhu, Q. Liu, E. Chen, H. Li, and H. Zhao, "Exploring the procrastination of college students: a data-driven behavioral perspective," in *Database Systems for Advanced Applications*, pp. 258–273, Springer, Cham, Switzerland, 2016.

[14] J. Peng, Y. Zhou, X. Sun et al., "Social media based topic modeling for smart campus: a deep topical correlation analysis method," *IEEE Access*, vol. 7, pp. 7555–7564, 2018.

[15] G. N. Burlak, J. Hernandez, A. Ochoa et al., "The use of data mining to determine cheating in online student assessment," in *Proceedings of the Electronics, Robotics and Automotive Mechanics Conference (CERMA'06)*, vol. 1, pp. 161–166, IEEE, Cuernavaca, Mexico, September 2006.

[16] S. Abdi, H. Khosravi, and S. Sadiq, "Predicting student performance: the case of combining knowledge tracing and collaborative filtering," in *Proceedings of the International Conference on Educational Data Mining*, Buffalo, NY, USA, July 2018.

[17] W. Liu, C. Y. Bao, and B. Wen, "Investigation on the Internet dependence of undergraduates and analysis of correlative causes," *Chinese General Practice*, vol. 13, no. 8, pp. 2485–2487, 2010.

[18] M. A. Albahar, "Skin lesion classification using convolutional neural network with novel regularizer," *IEEE Access*, vol. 7, pp. 38306–38313, 2019.

[19] A. Almuhareb, W. Alsanie, and A. Al-Thubaity, "Arabic word segmentation with long short-term memory neural networks and word embedding," *IEEE Access*, vol. 7, pp. 12879–12887, 2019.

[20] S. Al-Dahidi, O. Ayadi, M. Alrbai, and J. Adeeb, "Ensemble approach of optimized artificial neural networks for solar photovoltaic power prediction," *IEEE Access*, vol. 7, pp. 81741–81758, 2019.

[21] J. Yuan, Y. Jin, W. Liu, and X. Wang, "Attention-based neural tag recommendation," in *Database Systems for Advanced Applications (DASFAA)*, pp. 350–365, Springer, Cham, Switzerland, 2019.

[22] F. Fan and Y. Bai, "A study on the internet dependence of college students: the revising and applying of a measurement," *Psychological Development and Education*, vol. 24, no. 2, pp. 187–203, 2005.

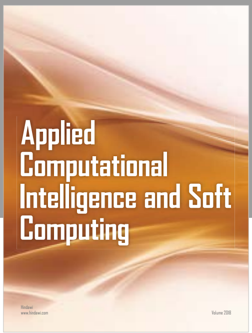
[23] Y. Y. Wei, G. S. Huang, Z. B. Xie et al., "Research on the relationship between students' internet dependence and loneliness by taking Hubei polytechnic university as an example," *Journal of Liuzhou Vocational & Technical College*, vol. 3, 2018.

[24] J. Cheng, G. Li, and X. Chen, “Research on travel time prediction model of freeway based on gradient boosting decision tree,” *IEEE Access*, vol. 7, pp. 7466–7480, 2018.

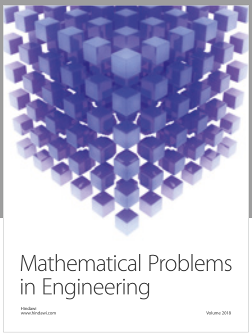
[25] M. B. B. Heyat, D. Lai, and F. I. K. Y. Zhang, “Sleep bruxism detection using decision tree method by the combination of C4-P4 and C4-A1 channels of scalp EEG,” *IEEE Access*, vol. 7, pp. 102542–102553, 2019.

[26] W. Kuang, Y.-L. Chan, S.-H. Tsang, and W.-C. Siu, “Fast HEVC to SCC transcoder by early CU partitioning termination and decision tree-based flexible mode decision for intra-frame coding,” *IEEE Access*, vol. 7, pp. 8773–8788, 2019.

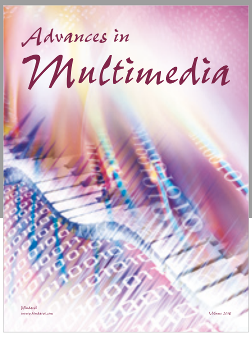
[27] W. Peng, X. Zhang, and X. Li, “Using behavior data to predict the internet addiction of college students,” in *Web Information Systems and Applications*, pp. 151–162, Springer, Cham, Switzerland, 2019.

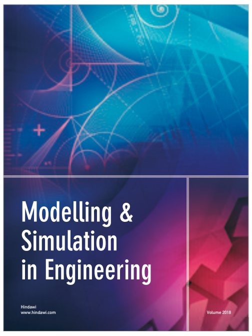

Applied Computational Intelligence and Soft Computing

The Scientific World Journal


Mathematical Problems in Engineering

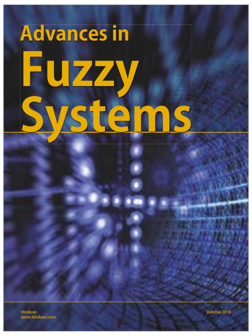
Journal of Engineering

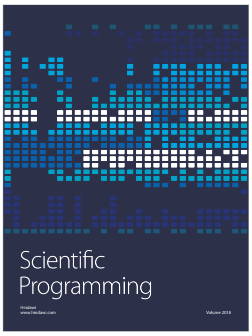

Advances in Multimedia

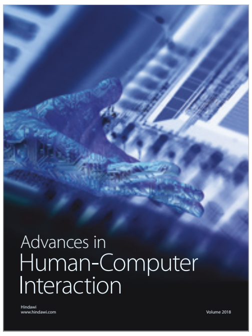

Modelling & Simulation in Engineering

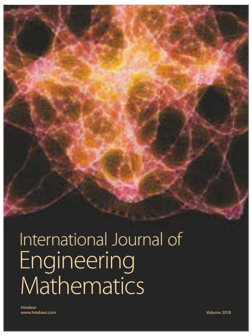
Advances in Artificial Intelligence

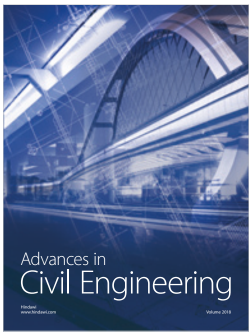
International Journal of Reconfigurable Computing

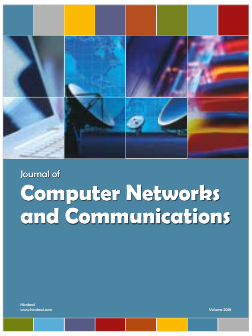

Advances in Fuzzy Systems


Scientific Programming


Advances in Human-Computer Interaction

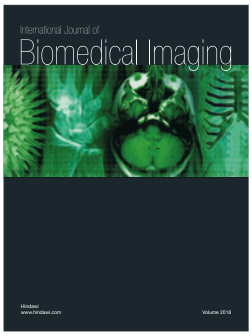

International Journal of Engineering Mathematics


Advances in Civil Engineering

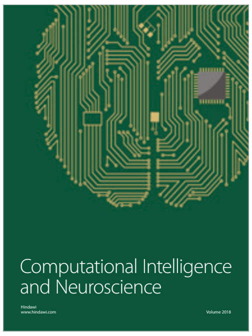

Journal of Computer Networks and Communications

Journal of Robotics

International Journal of Computer Games Technology


International Journal of Biomedical Imaging

Journal of Electrical and Computer Engineering


Computational Intelligence and Neuroscience