information

MDPI

Article

Random Forest with Sampling Techniques for Handling Imbalanced Prediction of University Student Depression

Siriporn Sawangarreerak [1] and Putthiporn Thanathamathee [2,*]

[1] School of Management, Walailak University, Nakhon Si Thammarat 80160, Thailand; siriporn.sa@wu.ac.th
[2] School of Engineering and Technology, Walailak University, Nakhon Si Thammarat 80160, Thailand
* Correspondence: putthiporn.th@wu.ac.th

Received: 14 September 2020; Accepted: 4 November 2020; Published: 5 November 2020

Abstract: In this work, we propose a combined sampling technique to improve the performance of imbalanced classification of university student depression data. In experimental results, we found that combined random oversampling with the Tomek links under sampling methods allowed generating a relatively balanced depression dataset without losing significant information. In this case, the random oversampling technique was used for sampling the minority class to balance the number of samples between the datasets. Then, the Tomek links technique was used for undersampling the samples by removing the depression data considered less relevant and noisy. The relatively balanced dataset was classified by random forest. The results show that the overall accuracy in the prediction of adolescent depression data was 94.17%, outperforming the individual sampling technique. Moreover, our proposed method was tested with another dataset for its external validity. This dataset's predictive accuracy was found to be 93.33%.

Keywords: depression prediction; imbalanced data; sampling techniques; feature selection; Patient Health Questionnaire-9 (PHQ-9)

1. Introduction

Depression is an important public health problem that occurs today and will do so in the future. It is one of the leading causes of "disability" worldwide and the most common mental disorder among college students [1]. A study in Thailand discovered that the prevalence of depression in adolescents was 14.9% [2]. Depression screening is the most important first step in assessing the symptoms and the severity of depression in people.

The observed depression is difficult to measure due to a condition within individuals who are sensitive and unaware of themselves being depressed. The diagnosis of depression in adolescents is necessary. The study found that adolescent depression is under-recognized by general doctors. Therefore, short and concise depressive screening tools are important to help doctors detect more students with depression at university.

The Patient Health Questionnaire-9 (PHQ-9) is a self-administered assessment. It is easy to understand for respondents. This tool is sensitive to the analysis and interpretation of results. It is also a tool that has few questions. Moreover, the questionnaire was assessed for both mental and physical symptoms [3]. Al-Busaidi et al. used PHQ-9 to screen for the prevalence of depressive symptoms among university students in Oman [4]. Levis et al. found that PHQ-9 was sensitive and accurate for screening to detect major depression [5].

Research related to prediction and analysis of the relevant factors of depression using data mining techniques has been carried out. For example, Chang et al. used ontologies and Bayesian networks

to build the terminology of depression and applied Bayesian networks to infer the probability of depression [6]. Thanathamathee applied the Adaboost decision tree and feature selection technique to predict adolescent depression. The results were useful for screening depression [7]. Ghafoor et al. used the association rule and a frequent pattern tree to extract the results of depression [8]. Hou et al. proposed many data mining techniques to predict depression in university students on the basis of their reading habits. The results found that logistic regression achieved the highest prediction accuracy at a lower relative error [9]. Li et al. employed a feature selection method called GreedyStepwise (GSW) on the basis of correlation feature selection (CFS) and K-nearest neighbor (KNN) for mild depressive detection [10]. Kim et al. proposed a simple unobtrusive sensing system to monitor the activities of elderly people living alone. The results showed that the neural network effectively detected normal and mild depression [11]. Gao et al. focused on machine learning to predict major depression disorder utilizing features derived from magnetic resonance imaging (MRI) data [12]. Alzyoud et al. developed a hybrid machine learning model to predict the level of depression for youth smokers [13]. They found that the age factor is an important attribute for predicting depression. Blessie and George employed naïve Bayes and KNN to extract data patterns to prevent mental illness and assist in mental health services [14]. Priya et al. applied different data mining techniques to predict anxiety, depression, and stress [15]. Random forest was the highest identifier for stress, and naïve Bayes was the highest identifier for depression.

Imbalanced data problems are often encountered in the real world, especially medical data, as there are many patients who are admitted to the hospital. However, the number of patients diagnosed with disease is small compared to the total number of patients. When these data are used to predict outcomes (by machine learning and data mining), the learning of the algorithm is affected. It is assumed that the data are drawn from the same distribution as the training data, presenting imbalanced data to the classifier and producing unacceptable results [16]. Sampling methods and synthetic data generation provide a balanced class distribution by adding or removing samples. Naïve random oversampling allows generating new samples by randomly sampling the current samples of the minority class with replacement. With regard to synthetic sampling, Chawla et al. proposed the synthetic minority oversampling technique (SMOTE) [17]. SMOTE generated synthetic data for the minority class by selecting some of the nearest minority neighbors using minority data and generated synthetic minority data along with the lines between the minority data and the nearest minority neighbors [18]. Borderline-SMOTE developed by Han et al. was used to generate synthetic data along the line between the borderline samples of the decision regions [19]. Random undersampling was used to undersample the data in the majority class by randomly picking samples with replacement. Tomek proposed the Tomek links method to find all examples in the majority class closest to the minority class. Then, these examples were removed from the majority class, thereby providing better borderline decisions for the classifier [20]. Wilson introduced the edited nearest-neighbors method, which applied a nearest-neighbor algorithm to delete the samples from the majority class that were near or around the borderline of different classes [21]. Moreover, Tomek proposed a new repeated edited nearest-neighbor method, extending the edited nearest-neighbor method by repeating it multiple times [22].

In predicting highly imbalanced disease data, Khalilia et al. presented a method to predict disease risks from highly imbalanced data, using random forest and combined repeated, random subsampling [16]. Bektas et al. applied feature selection and an oversampling method to predict imbalanced cardiovascular diseases [23]. Sharma and Verbeke handled imbalanced Dutch depression data by undersampling, oversampling, over–undersampling and random oversampling example (ROSE) techniques, and then applied extreme gradient boosting to classify mental illness [24].

According to the literature review, the following research questions are proposed:

RQ1. Which questions in PHQ-9 were involved in the depression prediction performance of university students who responded to this questionnaire? Is it necessary to use all 9 questions?

RQ2. Which oversampling or undersampling techniques should be used to achieve better depression prediction performance of university students who responded to this questionnaire?

RQ3. Is using only one dataset of depression enough to predict university student depression effectively?

This paper is organized as follows: Section 2 presents the proposed method; Section 3 summarizes the performance evaluation; the deployment and discussion are presented in Section 4; lastly, the conclusion is provided in Section 5.

2. Materials and Proposed Method

2.1. Feature Selection

Feature selection is a process of extracting the most relevant features from the dataset; it generates a subset of the original feature set to improve predictive performance with less processing time. Some of the original features that were not important are eliminated. In this paper, the filter approach method was applied. This is a way to calculate the weight or the relative value of each feature by selecting only the features that are important to keep.

2.1.1. Chi-Square

The chi-square $\left(x^2\right)$ test is used for categorical features in a dataset. To calculate the chi-square between each feature and the target in a descending order of weight values, the following equation can be used [25]:

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{1}$$

where O_i represents the observed value(s) and E_i represents the expected value(s).

2.1.2. Gini Index

The Gini index is a measure of the impurity of a dataset. A higher attribute weight denotes greater relevance. Suppose S is a dataset and the samples have k different classes $(C_i,\ i = 1,\ \ldots,\ k)$. S is divided into k subsets $(S_i,\ i = 1,\ \ldots,\ k)$, where S_i is a sample set that belongs to class C_i. s_i is the sample number of sets S_i. Then, the Gini index of set S is described as follows [26]:

$$Gini(S) = 1 - \sum_{i=1}^{k} P_i^2, \tag{2}$$

where P_i^2 is the probability, estimated using s_i/s. The minimum value of $Gini(S)$ is 0, indicating that the maximum useful information can be obtained.

2.1.3. Information Gain

Information gain is calculated using the entropy value, which is a measure of the difference or dispersion of the data. If the data are very different, the entropy value is high; on the other hand, if the data are very similar, the entropy value is low. Information gain is defined as follows [27]:

$$I = \sum_{j} p(j,x) \log \frac{p(j,x)}{p(j)p(x)}, \tag{3}$$

$$I = \sum_{j} p(j)p(x|j) \log \frac{p(x|j)}{p(x)},$$

where $p(j,x)$ is the joint distribution of class j and feature x.

2.2. Proposed Method

In this paper, we proposed seven major steps to handle the problem of imbalanced depression data, as outlined in Figure 1.

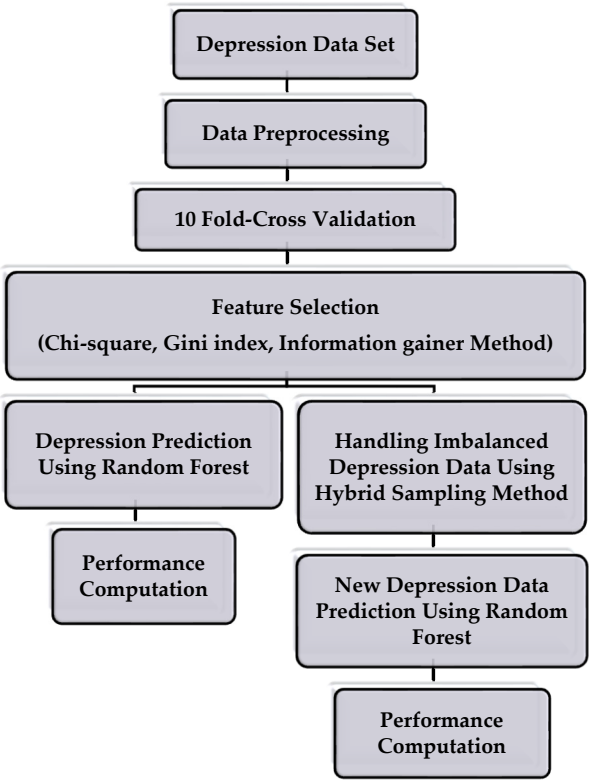

Figure 1. The proposed method to handle imbalanced depression data.

1. The depression data were divided into training data, for generating the prediction model for depression, and unseen testing data, for evaluating the performance of the model for prediction. For training data, 10-fold cross-validation was used to divide the data into 10 equally sized sets. Each set was in turn used as the test set, while the random forest classifier was trained on the other nine sets [28].
2. Chi-square, Gini index, and information gain were used to extract the principle feature items. These techniques were selected to automatically identify meaningful smaller subsets of feature items and still exhibit high performance in terms of the prediction of depression [29].
3. The reduced feature items of training data were obtained by oversampling the minority class using (1) random over-sampling, (2) SMOTE, and (3) Borderline-SMOTE. Then, the technique that balanced the data with the best performance in terms of prediction was selected.
4. Moreover, the reduced feature items of training data were also undersampled by (1) Tomek links, (2) edited nearest neighbors, and (3) repeated edited nearest neighbors. Then, the undersampling technique with the best performance in terms of prediction was selected.
5. The best-performing oversampling and undersampling techniques from steps 3 and 4 were used as a hybrid sampling approach. First, the random oversampling method was applied to generate new samples by randomly sampling the current training data with replacement. The minority class was oversampled, and a new balanced training dataset was created. Then, the Tomek links undersampling method was applied to reduce the number of each class by removing unwanted overlap between classes. This procedure was based on the assumption that random oversampling with replacement in the minority class was used for balancing data. Then, Tomek links were used

to remove the majority data closest to real or synthetic data in the minority class. Thus, the majority samples were deleted until a better boundary was provided for classifier decisions.

6. The new training data from step 5 were used to generate the depression prediction model with the random forest classifier.
7. Then, the unseen testing data were used to evaluate the performance of the predicted depression model.

3. Results and Performance Evaluation

3.1. Dataset Description

The data used in this paper were obtained from PHQ-9. It is a self-assessment questionnaire developed from the Diagnostic and Statistical Manual of Mental Disorders fourth edition (DSM-IV) [3]. It assesses depression by asking questions related to symptoms over the past 2 weeks, as shown in Table 1.

Table 1. Patient Health Questionnaire-9 depression assessment questions.

Question No.	During the Past 2 Weeks Including Today, How Often Have You Had These Symptoms?
Q1	Little interest or pleasure in doing things
Q2	Feeling down, depressed, or hopeless
Q3	Trouble falling asleep, staying asleep, or sleeping too much
Q4	Feeling tired or having little energy
Q5	Poor appetite or overeating
Q6	Feeling bad about yourself or that you are a failure or have let yourself or your family down
Q7	Trouble concentrating on things, such as reading the newspaper or watching television
Q8	Moving or speaking so slowly that other people could have noticed. Alternatively, being so fidgety or restless that you have been moving around a lot more than usual
Q9	Thoughts that you would be better off dead or of hurting yourself in some way

The scores for each question constitutes four levels: none (points = 0), some days or rarely (points = 1), quite often (points = 2), and almost every day (points = 3). The total score of all questions ranges from 0 to 27 points. There are 4 depression symptoms, and the detailed scores are described in Table 2.

Table 2. PHQ-9 scores.

Depression Symptoms	PHQ-9 Score	*N*	Proportion of Each Class (%)
No symptoms	0–6	821	53.04
Mild	7–12	567	36.63
Moderate	13–18	130	8.39
Severe	19–27	30	1.94

The data included 1549 examples collected from male and female undergraduate students of a university in the south of Thailand [29]. There were four classes of depression symptoms: class 0 (no symptoms), class 1 (mild symptoms), class 2 (moderate symptoms), and class 3 (severe symptoms).

3.2. Ethical Consideration

Permission for the study was obtained from the Ethics Committee of Walailak University, Thailand (Protocol Number WUEC-18-060-01).

3.3. Assessment Matrices

In this paper, the performance of depression prediction was evaluated as a function of accuracy, precision, and recall. The confusion matrix in Table 3 represents a summary of the prediction results from the classification problem. The number of correct and incorrect predictions of each class are summarized with count values. {P, N} represent the positive and negative testing data, and {Y, N} represent the predicted results given by classifier for the positive and negative class [30].

Table 3. Confusion matrix.

	Actual Positive (P)	**Actual Negative (N)**
Predicted Positive (Y)	TP (true positives)	FP (false positives)
Predicted Negative (N)	FN (false negatives)	TN (true negatives)

According to Table 3, true positive (TP) is the number of correct predictions of a positive example, true negative (TN) is the number of correct predictions of a negative example, false positive (FP) is the number of incorrect predictions of a positive example, and false negative (FN) is the number of incorrect predictions of a negative example. The evaluation metrics used to assess the depression prediction from the datasets are defined below in Equations (4)–(7).

Accuracy:

$$\frac{TP + TN}{TP + FP + FN + TN'} \tag{4}$$

Precision:

$$\frac{TP}{TP + FP'} \tag{5}$$

Recall:

$$\frac{TP}{TP + FN'} \tag{6}$$

F-measure:

$$\frac{\left(1 + \beta^2\right) \times Recall \times Precision}{\beta^2 \times Recall + \ Precision}, \tag{7}$$

where β is a coefficient used to adjust the relative importance of precision versus the recall, which is usually set to 1. The F-measure is high when both recall and precision are high, indicating the goodness of a learning algorithm for the interest class [28,31].

3.4. Results

In this paper, all sampling techniques were included in the imbalance-learn python package [32]. To achieve the best performance in terms of prediction, the proposed method was applied several times to determine the best parameters for all sampling techniques with *k_neighbors* = 5. For the random forest classifier, we set the number of *random trees* to 100, and the *maximum depth* for all trees was built until it attained a value of 10.

The results achieved when using the random forest to classify the dataset with only seven feature items [29] are shown in Table 4. Questions no. 1 and 9 were less important according to all three statistical weight measures of the feature selection methods.

Table 4. Feature selection using chi-Square, Gini index, and information gain.

Chi-Square		Gini Index		Information Gain	
Feature Item	**Weight**	**Feature Item**	**Weight**	**Feature Item**	**Weight**
Q8	1	Q4	1	Q4	1
Q6	0.932	Q8	0.978	Q6	0.926
Q4	0.891	Q6	0.955	Q7	0.907
Q7	0.881	Q7	0.95	Q8	0.852
Q5	0.474	Q2	0.715	Q2	0.714
Q2	0.437	Q3	0.601	Q3	0.597
Q3	0.227	Q5	0.508	Q5	0.534
Q1	0.158	Q1	0.247	Q1	0.372
Q9	0	Q9	0	Q9	0

Table 5 shows the performance of depression prediction without handling the imbalanced techniques, where the accuracy achieved was 91.26%. It can be seen that the precision of the model in the prediction of class 3 (severe depression) was only 57.14% and the recall was only 66.67%. Furthermore, although the efficiency (precision) of this model could predict class 2 (moderate depression) with 84.62% accuracy, the prediction recall of this class was only 42.31%. This shows that this model could not effectively predict class 2 and class 3. Therefore, in this paper, oversampling and undersampling techniques were applied to handle the imbalanced data in order to improve the predictive performance of the minority class. Then, the best oversampling and undersampling techniques were applied together according to the research of Batista et al. [33], which indicated that, although oversampling minority class examples can balance class distributions, some skewed class distributions are not solved, which can lead to an overfitting classifier problem. Thus, the undersampling technique was applied as a data-cleaning method. Instead of removing only the majority class examples, examples from both minority and majority classes were also removed.

Table 5. Performance of prediction without using imbalanced techniques.

Technique	Class	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Random forest without sampling	0	91.26%	96.39%	97.56%	96.97%
	1		86.66%	94.69%	90.50%
	2		84.62%	42.31%	56.41%
	3		57.14%	66.67%	61.54%

The oversampling techniques used in this paper were random oversampling, SMOTE, and Borderline-SMOTE. For each technique, Table 6 presents the results in terms of accuracy, precision, and recall. The results indicate that random oversampling was the best oversampling technique for this data, with classes 2 and 3 predicted more efficiently in terms of precision and recall compared to other oversampling methods with an accuracy of 93.53%.

Moreover, we used undersampling techniques to balance the class distribution via elimination of the majority class examples. The undersampling techniques used were Tomek links, edited nearest neighbors, and repeated edited nearest neighbors. Table 7 shows the performance in terms of prediction using all three techniques. The results show that Tomek links produced predictions with the best accuracy of 91.59%. Therefore, we applied random oversampling together with the Tomek links technique to improve the efficiency of the depression prediction for both the majority class and the minority class. Table 8 shows the number of samples after using the combined sampling techniques, displaying almost the same or balanced fractions.

Table 6. Performance in terms of prediction with oversampling techniques used to handle imbalanced data. SMOTE, synthetic minority oversampling technique.

Technique	Class	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Random forest with random oversampling	**0**	**93.53%**	**97.53%**	**98.17%**	**97.85%**
	1		**88.00%**	**96.76%**	**92.17%**
	2		**86.08%**	**73.76%**	**79.45%**
	3		**83.33%**	**83.33%**	**83.33%**
Random Forest with SMOTE	0	92.56%	97.53%	96.34%	96.93%
	1		90.43%	92.04%	91.23%
	2		76.00%	73.08%	74.51%
	3		71.43%	83.33%	76.92%
Random Forest with Borderline-SMOTE	0	91.59%	97.50%	95.12%	96.30%
	1		87.60%	93.81%	90.60%
	2		77.27%	65.38%	70.83%
	3		66.67%	66.67%	66.67%

Table 7. Performance in terms of prediction with undersampling techniques used to handle imbalanced data.

Technique	Class	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Random forest with Tomek links	**0**	**91.59%**	**96.41%**	**98.17%**	**97.28%**
	1		**87.70%**	**94.69%**	**91.06%**
	2		**84.62%**	**42.31%**	**56.41%**
	3		**57.14%**	**66.67%**	**61.54%**
Random forest with edited nearest neighbors	0	88.35%	94.12%	97.56%	95.81%
	1		81.75%	91.15%	86.19%
	2		100.00%	19.23%	32.26%
	3		71.43%	83.33%	76.92%
Random forest with repeated edited nearest neighbors	0	87.70%	93.02%	97.56%	95.24%
	1		81.60%	90.27%	85.72%
	2		100.00%	15.38%	26.66%
	3		62.50%	83.33%	71.43%

Table 8. The number of samples after using hybrid sampling.

Depression Symptom	Number of Data	Fraction (%)
No symptoms	658	25.02%
Mild	658	25.02%
Moderate	658	25.02%
Severe	656	24.94%

To clarify the performance of the proposed prediction, a comparative study with random oversampling and the Tomek links technique is summarized in Table 9. The results show that our proposed method achieved the best performance in terms of depression prediction as a function of accuracy, precision, recall, and F-measure evaluation indicators.

Table 9. Performance of our hybrid sampling compared with random oversampling and Tomek links techniques.

Technique	Class	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Random forest with random oversampling	0	93.53%	97.53%	98.17%	97.85%
	1		88.00%	96.76%	92.17%
	2		86.08%	73.76%	79.45%
	3		83.33%	83.33%	83.33%
Random forest with Tomek links	0	91.59%	96.41%	98.17%	97.28%
	1		87.70%	94.69%	91.06%
	2		84.62%	42.31%	56.41%
	3		57.14%	66.67%	61.54%
Random forest combined with random oversampling and Tomek links	**0**	**94.17%**	**98.17%**	**98.17%**	**98.17%**
	1		**95.58%**	**97.35%**	**96.46%**
	2		**87.62%**	**89.47%**	**88.54%**
	3		**83.33%**	**83.33%**	**83.33%**

In addition, a significant independent *t*-test was applied to these measures, i.e., precision, recall, and F-measure, to evaluate and confirm the statistical significance of our performance results. The different performance results were assumed as statistically significant when $p < 0.05$. The evaluation indicates that our proposed method performed significantly better than random forest with random oversampling and random forest with Tomek links in terms of precision and F-measure, as shown in Table 10.

Table 10. Significance of precision, recall, and F-measure results using the proposed method and other sampling techniques.

Technique	*p*-Value (Proposed Method vs. Other Sampling Techniques		
	Precision	Recall	F-Measure
Random forest with random oversampling	**0.0029**	0.1117	**0.0090**
Random forest with Tomek links	**0.0045**	0.0737	**0.0310**

4. Deployment and Discussion

4.1. Research Questions and Proposed Solutions

4.1.1. RQ1. Which Questions in PHQ-9 Were Involved in the Depression Prediction Performance of University Students Who Responded to this Questionnaire? Is it Necessary to Use all 9 Questions?

We applied feature selection including chi-square, Gini index, and information gain [29] to identify significant features. As shown in Table 4, only seven important questions were found to influence the prediction of university student depression. Furthermore, the question with the least relevance to the prediction of depression was Q9 (weight of 0), consistent with Lotrakul et al.'s research showing that the least endorsed question item was thoughts of suicide (Q9) [34].

4.1.2. RQ2. Which Oversampling or Undersampling Techniques Should Be Used to Achieve Better Depression Prediction Performance of University Students Who Responded to This Questionnaire?

First, our research was conducted using oversampling. It was found that random oversampling was an effective technique for predicting the depression data, achieving an accuracy of 93.53%, as shown in Table 6. Next, we applied undersampling techniques to our dataset. The results in Table 7 indicate that the Tomek links technique produced the best prediction among other undersampling methods with an accuracy of 91.59%. However, the Tomek links technique was less effective in predicting class 2 and 3 depression, producing lower precision and recall values than the random

oversampling technique. Although the random oversampling technique showed good predictive performance, using only random oversampling for minority classes encountered problems of a skewed class distribution, leading to overfitting [35]. For this reason, our research applied a combination of random oversampling and Tomek links to handle our imbalanced depression data. Thus, Tomek links were also applied to remove the majority class examples. The proposed method showed an improved predictive performance for all classes, as shown in Tables 9 and 10. When considering the *p*-value, it was found that our model showed the best performance in terms of precision and f-measure among other sampling techniques.

4.1.3. RQ3. Is Using Only One Dataset of Depression Enough to Predict University Student Depression Effectively?

The dataset in this paper included 1549 examples to predict university student depression. Furthermore, 165 examples were collected from students in accounting courses over four years, unrelated to the original dataset, to test the performance of the model. The details of the examples are shown in Table 11.

Table 11. Details of new unseen data for external validity.

Depression Symptom	Number of Data	Fraction (%)
No symptoms	113	68.48%
Mild	42	25.45%
Moderate	9	5.45%
Severe	1	0.62%

The performance in terms of prediction is shown in Table 12. It can be seen that our proposed method was capable of predicting unseen depression data with an accuracy of up to 93.33%. The precision and recall remained satisfactory, and this model could also precisely predict severe depression. It can be concluded that our proposed method is effective in correctly predicting each class.

Table 12. Performance in terms of prediction of unseen new data for external validity.

Accuracy 93.33%					
	Actual 0	**Actual 1**	**Actual 2**	**Actual 3**	**Precision (%)**
Pred.0	112	7	0	0	94.12%
Pred.1	1	34	2	0	91.90%
Pred.2	0	1	7	0	87.50%
Pred.3	0	0	0	1	100%
Recall (%)	99.12%	80.95%	77.78%	100%	

Through the feature selection methods, we found that feature items Q8, Q4, Q6, and Q7 were the most important items in predicting depression. These features are related to cognition and physical change [33] when negative self-consciousness affects physical expression. It was found that physical symptom change always appears at the severe depression level [36]. When delving into items Q8, Q4, Q6, and Q7, it was evident that severe depression had a score level of 3 (almost every day) for these four items. It was seen that these four items must be considered or given special importance when screening for and predicting depression in university students. Moreover, the questions identified were consistent with the research of Lotrakul et al. [34], which showed that the most important item affecting depression was related to low energy, i.e., Q4 and Q8 in our paper. This is also consistent with the results from [25], whereby depressed patients mostly emphasized somatic symptoms and emotional symptoms, corresponding to Q6 and Q7 in our research.

Predicting university student depression is very important and must be the first step of primary care. Empirical research was conducted using data collected from the university in the south of

Thailand. The results show that our proposed method can be applied to correct identification of depression in a screening process. It can also reduce the time consumption and increase the accuracy of the assessment, while allowing practitioners to choose appropriate depression assessments that are sensitive and effective for screening of university students.

5. Conclusions

Sampling techniques are effective methods for resolving the class imbalance problem. The concept of the proposed method involved applying random oversampling to minority class examples and Tomek links to majority class examples, thereby removing unwanted overlap between classes. The proposed method generated a relatively balanced dataset, without losing significant information. This balanced dataset was classified using random forest. The predictive depression performance of the proposed method was evaluated in terms of accuracy, precision, recall, and F-measure. The experimental results show that our proposed method outperformed the individual sampling techniques. Moreover, our proposed method can be used for effectively predicting university student depression, which is the primary step in screening for depression symptoms.

The limitation of this work is related to the depression prediction model using the PHQ-9 questionnaire. This research involved predicting depression among university students, discovering that only seven questions were important in depression prediction. Another dataset, e.g., involving patients of working age or the elderly, may use different prediction models for depression. Therefore, this model is only suitable for university students using PHQ-9.

Future studies will be carried out using other assessment methods; we also intend to use additional information provided by university students' social media posts to help improve the performance of depression prediction.

Author Contributions: Conceptualization, S.S. and P.T.; Methodology, S.S. and P.T.; Software, P.T.; Validation, S.S. and P.T.; Formal Analysis, P.T.; Investigation, S.S. and P.T.; Resources, P.T.; Data Curation, P.T.; Writing Original Draft Preparation, S.S. and P.T.; Writing Review & Editing, S.S. and P.T.; Visualization, S.S.; Supervision, P.T.; Project Administration, S.S.; Both authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This work was supported by the Division of Student Affairs, Walailak University. This study is part of the research project titled "Development of Care and Empowerment System for Students in Walailak University".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ebert, D.D.; Buntrock, C.; Mortier, P.; Auerbach, R.; Weisel, K.K.; Kessler, R.C.; Cuijpers, P.; Green, J.G.; Kiekens, G.; Nock, M.K.; et al. Prediction of major depressive disorder onset in college students. *Anxiety Depress. Assoc. Am.* **2019**, *36*, 294–304. [CrossRef] [PubMed]
2. Jatchavala, C.; Chan, S. Thai Adolescent Depression: Recurrence Prevention in Practice. *J. Health Sci. Med. Res.* **2018**, *36*, 147–155. [CrossRef]
3. Kroenke, K.; Spitzer, R.L.; Williams, J.B.W. The PHQ-9 Validity of a Brief Depression Severity Measure. *J. Gen Int. Med.* **2001**, *16*, 606–613. [CrossRef] [PubMed]
4. Al-Busaidi, Z.; Bhargava, K.; Al-Ismaily, A.; Al-Lawati, H.; Al-Kindi, R.; Al-Shafaee, M.; Al-Maniri, A. Prevalence of Depressive Symptoms among University Students in Oman. *Oman Med. J.* **2011**, *26*, 235–239. [CrossRef] [PubMed]
5. Levis, B.; Benedetti, A.; Thombs, B.D. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: Individual participant data meta-analysis. *BMJ Clin. Res.* **2019**, *365*, 1–11. [CrossRef] [PubMed]
6. Chang, Y.S.; Hung, W.C.; Juang, T.Y. Depression diagnosis based on ontologies and Bayesian networks. In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, 13–16 October 2013; pp. 3452–3457.

7. Thanathamathee, P. Boosting with feature selection technique for screening and predicting adolescents depression. In Proceeding of the Fourth International Conference on Digital Information and Communication Technology and its Applications (DICTAP), Bangkok, Thailand, 6–8 May 2014; pp. 23–27.
8. Ghafoor, Y.; Huang, Y.P.; Lui, S.I. An intelligent approach to discovering common symptoms among depressed patients. In *Soft Computing—A Fusion of Foundations, Methodologies and Applications*; Springer: Cham, Switzerland, 2015; Volume 19, pp. 819–824.
9. Hou, Y.; Xu, J.; Huang, Y.; Ma, X. A big data application to predict depression in the university based on the reading habits. In Proceeding of the 3rd International Conference on Systems and Informatics (ICSAI), Shanghai, China, 19–21 November 2016; pp. 1085–1089.
10. Li, X.; Hu, B.; Sun, S.; Cai, H. EEG-based mild depressive detection using feature selection methods and classifiers. In *Computer Methods and Programs in Biomedicine*; Elsevier: New York, NY, USA, 2016; Volume 136, pp. 151–161.
11. Kim, J.Y.; Liu, N.; Tan, X.T.; Chu, C.H. Unobtrusive Monitoring to Detect Depression for Elderly with Chronic Illnesses. *IEEE Sens. J.* **2017**, *17*, 5694–5704. [CrossRef]
12. Gao, S.; Calhoun, V.D.; Sui, J. Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neurosci. Ther.* **2018**, *24*, 1037–1052. [CrossRef]
13. Alzyoud, S.; Kharabsheh, M.; Mudallal, R. Predicting Depression Level of Youth Smokers Using Machine Learning. *Int. J. Sci. Technol. Res.* **2019**, *8*, 2245–2248.
14. Blessie, E.C.; George, B. A Novel approach for Psychiatric Patient Detection and Prediction using Data Mining Techniques. *Int. J. Eng. Res. Technol.* **2019**, *7*, 1–4.
15. Priya, A.; Garg, S.; Tigga, N.P. Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms. *Proced. Comput. Sci.* **2020**, *167*, 1258–1267. [CrossRef]
16. Khalilia, M.; Chakraborty, S.; Popescu, M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med. Inform. Decis. Mak.* **2011**, *11*, 1–13. [CrossRef]
17. Chawla, N.; Bowyer, K.; Hall, L.; Kegelmeyer, W. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
18. Thanathamathee, P.; Lursinsap, C. Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques. *Pattern Recognit. Lett.* **2013**, *34*, 1339–1347. [CrossRef]
19. Han, H.; Wang, W.; Mao, B. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In Proceedings of the 2005 international conference on Advances in Intelligent Computing, Hefei, China, 23–26 August 2005; pp. 878–887.
20. Tomek, I. Two modifications of CNN. In *IEEE Transactions on Systems, Man, and Cybernetics*; IEEE: New York, NY, USA, 1976; Volume 6, pp. 769–772.
21. Wilson, D.L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. In *IEEE Transactions on Systems, Man, and Cybernetics*; IEEE: New York, NY, USA, 1972; Volume 2, pp. 408–421.
22. Tomek, I. An Experiment with the Edited Nearest-Neighbor Rule. In *IEEE Transactions on Systems, Man, and Cybernetics*; IEEE: New York, NY, USA, 1976; Volume 6, pp. 448–452.
23. Bektas, J.; Ibrikci, T.; Ozcan, I.T. Classification of Real Imbalanced Cardiovascular Data Using Feature Selection and Sampling Methods: A Case Study with Neural Networks and Logistic Regression. *Int. J. Artif. Intell. Tools* **2017**, *26*, 1750019. [CrossRef]
24. Sharma, A.; Verbeke, W.J.M.I. Improving Diagnosis of Depression with XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset. *Front. Big Data* **2020**, *3*, 1–11. [CrossRef]
25. Zhai, Y.; Song, W.; Liu, X.; Lui, L.Z.; Zhao, X. A Chi-Square Statistics Based Feature Selection Method in Text Classification. In Proceedings of the IEEE 9th International Conference on Software Engineering and Service Science (ICSESS 2018), Beijing, China, 23–25 November 2018; pp. 160–163.
26. Park, H.; Kwon, H.C. Improve Gini-Index Algorithm to Correct Feature-Selection Bias in Text Classification. *IEICE Trans. Inf. Syst.* **2011**, *E94.D*, 855–865. [CrossRef]
27. Dhir, C.S.; Iqbal, N.; Lee, S.Y. Efficient feature selection based on information gain criterion for face recognition. In Proceedings of the International Conference on Information Acquisition, Jeju, Korea, 8–11 July 2007; pp. 523–527.
28. Guo, H.; Viktor, H. Learning from imbalanced data sets with boosting and data generation: The databoost-im approach. *SIGKDD Explor.* **2004**, *6*, 30–39. [CrossRef]

29. Sirisathitkul, Y.; Thanathamathee, P.; Aekwarangkoon, S. Predictive Apriori Algorithm in Youth Suicide Prevention by Screening Depressive Symptoms from Patient Health Questionnaire-9. *TEM J.* **2019**, *8*, 1449–1455.
30. Kaur, A.; Kaur, I. An empirical evaluation of classification algorithms for fault prediction in open source projects. *J. King Saud Univ. Comput. Inform. Sci.* **2018**, *30*, 2–17. [CrossRef]
31. Chen, S.; He, H.; Garcia, E. Ramoboost: Ranked minority oversampling in boosting. *IEEE Trans. Neural Netw.* **2010**, *21*, 1624–1642. [CrossRef]
32. Lemaitre, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 17.
33. Batista, G.; Prati, R.; Monard, M.C. A Study of the Behavior of Several Methods for Balancing machine Learning Training Data. *SIGKDD Explor.* **2004**, *6*, 20–29. [CrossRef]
34. Lotrakul, M.; Sumrithe, S.; Saipanish, R. Reliability and validity of the Thai version of the PHQ-9. *BMC Psychiatry* **2008**, *8*, 1–7. [CrossRef]
35. Prati, R.C.; Batista, G.E.A.P.A.; Monard, M.C. Data mining with imbalanced class distributions: Concepts and methods. In Proceedings of the 4th Indian International Conference on Artificial Intelligence (IICAI-09), Tumkur, India, 16–18 December 2009; pp. 359–376.
36. Rice, F.; Riglin, L.; Lomax, T.; Souter, E.; Potter, R.; Smith, D.J.; Thapar, A.K.; Thapar, A. Adolescent and adult differences in major depression symptom profiles. *J. Affect. Disord.* **2019**, *243*, 175–181. [CrossRef]