



Article Week-Wise Student Performance Early Prediction in Virtual Learning Environment Using a Deep Explainable Artificial Intelligence

Hsing-Chung Chen ^{1,2,*}, Eko Prasetyo ^{1,3}, Shian-Shyong Tseng ^{4,*}, Karisma Trinanda Putra ^{1,5}, Prayitno ^{1,6}, Sri Suning Kusumawardani ⁷ and Chien-Erh Weng ^{8,*}

- ¹ Department of Computer Science and Information Engineering, Asia University, Taichung City 413, Taiwan; eko.prasetyo@umy.ac.id (E.P.); karisma@ft.umy.ac.id (K.T.P.); prayitno@polines.ac.id (P.)
- ² Department of Medical Research, China Medical University Hospital, China Medical University, Taichung City 404, Taiwan
- ³ Department of Information Technology, Universitas Muhammadiyah Yogyakarta, Bantul 55183, Indonesia
- ⁴ Department of M-Commerce and Multimedia Applications, Asia University, Taichung City 413, Taiwan
- ⁵ Department of Electrical Engineering, Universitas Muhammadiyah Yogyakarta, Bantul 55183, Indonesia
- Department of Electrical Engineering, Politeknik Negeri Semarang, Semarang 50275, Indonesia
 Department of Electrical and Information Engineering, Universitas Gadjah Mada, Yogyakarta 55281,
- Indonesia; suning@ugm.ac.id ⁸ Department of Telecommunication Engineering, National Kaohsiung University of Science and Technology,
- Kaohsiung 811, Taiwan
 Correspondence: cdma2000@asia.edu.tw (H.-C.C.); sstseng@asia.edu.tw (S.-S.T.); ceweng@nkust.edu.tw (C.-E.W.)

Abstract: Early prediction of students' learning performance and analysis of student behavior in a virtual learning environment (VLE) are crucial to minimize the high failure rate in online courses during the COVID-19 pandemic. Nevertheless, traditional machine learning models fail to predict student performance in the early weeks due to the lack of students' activities' data in a week-wise timely manner (i.e., spatiotemporal feature issues). Furthermore, the imbalanced data distribution in the VLE impacts the prediction model performance. Thus, there are severe challenges in handling spatiotemporal features, imbalanced data sets, and a lack of explainability for enhancing the confidence of the prediction system. Therefore, an intelligent framework for explainable student performance prediction (ESPP) is proposed in this study in order to provide the interpretability of the prediction results. First, this framework utilized a time-series weekly student activity data set and dealt with the VLE imbalanced data distribution using a hybrid data sampling method. Then, a combination of convolutional neural network (CNN) and long short-term memory (LSTM) was employed to extract the spatiotemporal features and develop the early prediction deep learning (DL) model. Finally, the DL model was explained by visualizing and analyzing typical predictions, students' activities' maps, and feature importance. The numerical results of cross-validation showed that the proposed new DL model (i.e., the combined CNN-LSTM and ConvLSTM), in the early prediction cases, performed better than the baseline models of LSTM, support vector machine (SVM), and logistic regression (LR) models.

Keywords: imbalanced data distribution; explainable student performance prediction; feature importance

1. Introduction

The World Health Organization (WHO) has advised all nations to restrict face-to-face contact to prevent the transmission of COVID-19. This recommendation applies to the formerly conducted classroom education, which has transformed into online teaching and learning through a virtual learning environment (VLE) such as Moodle [1,2]. The abrupt



Citation: Chen, H.-C.; Prasetyo, E.; Tseng, S.-S.; Putra, K.T.; Prayitno; Kusumawardani, S.S.; Weng, C.-E. Week-Wise Student Performance Early Prediction in Virtual Learning Environment Using a Deep Explainable Artificial Intelligence. *Appl. Sci.* 2022, *12*, 1885. https:// doi.org/10.3390/app12041885

Academic Editor: Vincent A. Cicirello

Received: 31 December 2021 Accepted: 8 February 2022 Published: 11 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/). transition from classroom to online teaching and learning created a sharp rise in the student failure rate caused by the behavior problems and their interactions in the VLE, for instance, some students lack of their ability in accessing learning material, assignment submission, and interaction during a discussion in the VLE. Assisting students in engaging with the VLE system is one of the critical factors for graduating from online courses [3]. The lack of interaction with classmates and teachers is one of the obstacles in this learning transition [4]. Diminished engagement due to a quick transition to online learning during a pandemic was another challenge [5]. Moreover, the investigation of student access to course materials during digital education was reported in [6]. Then, the effectiveness of a distance learning system was assessed using a comprehensive methodology in [7]. Therefore, early prediction of student learning performance and analysis of initial student behavior in the VLE are crucial to reduce the high failure rate in online courses during the pandemic.

The advancement of machine learning technology has received widespread attention for forecasting student performance in earlier weeks of the teaching and learning process in the VLE. The VLE records student interaction data such as course view, assignment submission, resource view, quiz, and discussion. This VLE student data interaction was employed to develop predictive models for various reasons, including predicting student performance. For instance, a multi-layer perceptron (MLP) architecture was proposed to predict student learning performance using the VLE database [8]. This early prediction scheme enables teachers to intervene with an at-risk student with a high failure rate in the online course. Furthermore, the machine learning analytics model can provide teachers with statistical insight to aid teaching and learning.

Nevertheless, existing research on student learning performance prediction only focuses on improving model prediction performance. This research goal will suffer from the weakness of temporal feature challenges and imbalanced data set distribution, resulting in poor performance and model prediction explainability in the prediction scenario of early student learning performance. Firstly, the temporal feature challenges refer to a week-byweek student learning performance prediction. For instance, the model could predict the student learning performance as early as the sixth week within 16 weeks of the teaching and learning process. However, the research on student performance prediction is ineffective without the student's activity data set captured in a whole week [8,9]. Secondly, the VLE data set has a highly imbalanced class data distribution between the positive and negative classes. Imbalanced class data distribution represents a data set in which some classes have a significantly larger sample size than others. For instance, the number of samples for failed students from the course was less than 3% compared to those who passed the course. However, a model developed with an imbalanced label data distribution may significantly degrade the prediction performance [9,10]. Finally, as for model prediction explainability, the output of model prediction only showed the probability score of students passing or failing. This probability score is complex for the teacher to explain the results to his or her students. The model needs to explain its prediction results based on the importance of student engagements in the VLE. Thus, allowing the teacher to intervene based on the model explainability results is essential. Nevertheless, the research focused on predicting student learning performance in superiority on one of these issues rather than considering them in their entirety. For instance, previous studies proposed a hybrid sampling strategy to improve the model performance with an imbalanced VLE data set [9,11,12], while the others focused on machine learning performance by investigating various model architectures such as the decision tree, naïve Bayes, and MLP on student learning performance [8,13,14].

To address the issues mentioned earlier, an intelligent predictive framework for explainable student performance prediction (ESPP) is proposed in this paper. It is a novel approach that addresses the above issues holistically. First, the proposed framework built a week-wise students' activity data set based on student interaction via clickstream activities in the VLE during the COVID-19 pandemic. After constructing the new week-wise students' activity data set, a hybrid sampling method with a synthetic minority over-sampling technique (SMOTE) and random under-sampling (RUS) was employed to balance the data set. Then, the prediction model was constructed by using a convolutional neural network (CNN) and long short-term memory (LSTM) layers for student performance prediction. By integrating the convolutional and LSTM mechanisms, the proposed deep learning (DL) model could accurately extract the spatiotemporal features of students' activities' data. Finally, the explainability of the best prediction model will be generated by visualizing the students' activity features maps and the feature importance contributing to the model. The importance features are visualized by using the local interpretable model-agnostic explanations (LIME).

In summary, the main contributions of our studies are four-fold. (1) Firstly, a new data set for student learning performance prediction was proposed. The data set was student interaction in the VLE collected during the COVID-19 pandemic arranged in a week-wise timely manner. (2) Secondly, the effectiveness of the proposed models was evaluated and compared with baseline models such as logistic regression (LR), support vector machine (SVM), and LSTM when dealing with spatiotemporal features in predicting student learning performance. (3) Thirdly, an insight into enhancing the school environment was introduced by supporting decision makers in implementing early intervention techniques to reduce the failure rate. For instance, the proposed method accurately forecasted student learning performance early in the sixth week with an accuracy of 0.91. This result could help the student improve his or her performance as early as possible. (4) Finally, an explainable DL model was provided by visualizing and analyzing individual predictions, the importance of features, and identifying typical predictions. This explainable instrument enables human and DL model interactions to intervene in the student failure rate. Furthermore, the teachers and students can focus on which features contribute to the poor performance. Thus, these assessments could be used to enhance the user experience in the VLE.

After a brief introduction, this paper is structured as follows. Section 2 summarizes a review of the present literature on student learning performance prediction and explainable machine learning. Then, the data set and proposed methodologies are explained in Section 3. Section 4 presents the findings. Finally, Sections 5 and 6 discuss the results and summarize the current work, respectively.

2. Related Works

Various studies have considered the issue of predicting students' academic achievement as a regression problem [15] or a classification problem [16]. In the first category, the learners' perspective results were forecasted, while in the second one, the learners' outcome was estimated as a pass, fail, or dropout. In addition to using various data analytic techniques, factors affecting student performance were also identified to predict student performance [17]. Moreover, the study areas of predicting student performance can be evaluated from several viewpoints, including prediction of early withdrawal from ongoing courses, analysis of inherent aspects influencing their achievement, and the application of statistical methods to evaluate student achievement. Early prediction is a relatively new concept in this area. It encompasses techniques for assessing students in real time with the intention of retaining them by providing appropriate procedures and interventions and then monitoring and minimizing the failure rates.

Deploying machine learning methods in analyzing student learning patterns and predicting at-risk students were conducted in some studies [16,18]. Research in ESPP took a subsequent method to transform the class period into a series of weekly structures and to measure the learning achievement based on students' interaction with the VLE. Machine-learning methods were used to predict students at risk based on attendance, quizzes, and assignments, with the addition of mid-term exams in the ninth week [19]. In another study, various data mining techniques, consisting of the decision tree, naïve Bayes classifier, k-nearest neighbor, SVM, and multi-layer perceptron (MLP), were deployed to predict at-risk students. Several studies employed LR as the baseline and identified the best prediction modeling compared among them [20,21]. According to the result presented in [22], KNN was the most accurate algorithm to classify successful or unsuccessful stu-

dents and determine the performance metrics. Moreover, student engagement patterns were effective in capturing students' behavior and persuading a better impact on their performance [23].

According to existing literature, deep learning on learning analytics to predict student learning success is still in its early stages. Deep learning is a computational method composed of several processing layers to study data representation using several levels of abstraction [24]. Student interaction with the VLE was captured to predict his or her learning performance deployed on a recurrent neural network (RNN) model. This DL approach outperformed the machine learning baseline methods [16,25]. Additionally, students' learning success was predicted by using their attendance and behavior over log data information [26]. By employing the RNN and LSTM models, this study could produce student engagement and interaction patterns in the VLE database. These DL methods were more effective in the early prediction of grades than conventional regression methods. However, the more advanced the prediction model, the lower the interpretability. This area of research has a small contribution in terms of the explainability of the results. Thus, this study provides the explainability approach on an ESPP framework using the LIME explainability model as proven by [27,28].

3. Methods

3.1. Implementing Teaching Design in the VLE System

The study case of this research was the Digital Transformation course at Gadjah Mada University in Indonesia. During the COVID-19 pandemic, this course has been held online using a VLE system. This research study was initiated by collecting teaching materials and questionnaires from student in the previous academic year. Then, the VLE system was designed and tested from September 2020 to January 2021. To achieve the aim of the study, an ESPP framework was proposed. The development of the ESPP framework was divided into three major phases, as illustrated in Figure 1.

First, phase 1 aimed to determine the main elements in the VLE system focusing on four core areas below. The first core area is key subjects. The second core area is life and career skills. The third core area is learning and innovation skills. Finally, the fourth core area is information, media, and technology skills. The previous study suggested three main elements: learning activities, learning resources, and learning support [29]. Each element was carefully explained and managed using several learning objects, as displayed in Figure 2. These learning objects are valuable features that can be used as input of the prediction model. Then, the VLE system was developed and tested. After that, this VLE system could be repurposed by adding and revising the content to improve the student experience during online learning for the next academic years. Meanwhile, phase 2 aimed to extract the data from the VLE database to create meaningful insights regarding student performance. This phase involved data extraction, cleaning, anonymization, and labeling.

Finally, phase 3 showed the performance evaluation results of the baseline models of machine learning (ML) and DL in classification tasks, especially in student performance early prediction. In addition, this work provides an interpretability analysis by using the LIME method in this paper. The output of the prediction model is the students' final course grade category, i.e., pass or fail. The framework will extract the trained data set from the VLE repository. A hybrid-sampling method was implemented to the data set in order to reduce the overfitting and misclassification of imbalanced data distribution in multi-class prediction tasks. Then, the proposed model was designed by using several traditional modern machine learning classifiers as the baseline model for evaluating the performance by using their performance metrics. Finally, the importance features of the results were visualized by using the XAI method via LIME. In the following sections, a brief description of the results is presented. The works mentioned above were developed using Python programming, Tensor flow, and LIME libraries.



Figure 1. The proposed framework of ESPP. An Intelligent Framework for Explainable Student Performance Prediction consisted of three phases: (1) Implementing Teaching Design in VLE System; (2) Dataset Acquisition, Cleaning, Anonymization, and Labeling; (3) Data Pre-processing, AI Modeling, Performance Analysis, and Explainable AI Component.



Figure 2. Teaching design in a virtual learning environment (VLE) system involves relationships among the major elements including learning activities, learning resources, and learning support. Learning features are denoted as F1, F2, F3, F4, F5, F6, F7, and F8, corresponding to the assignments, files, forums, homepages, labels, pages, quizzes, and URLs, respectively.

This proposed framework simultaneously predicts student performance and generates the interpretability from the student side and the VLE point of view. The interpretability technique is useful for explaining what the results had been shown. In addition, it could explain why the results will follow the predicted results at the end of the semester. On the other hand, from the point of view of the VLE system, the interpretability technique becomes a general evaluation material to improve the features that do not play a role in balancing the learning process so that the VLE system can be revised and improved for learning in the next academic year.

3.2. Designing the ESPP Data Set

The data set was recorded from a 16-week, fully online "Digital Transformation" course at Gadjah Mada University in Indonesia from September 2020 to January 2021. The course was conducted through Moodle, a learning management system containing eight learning objects. The first materials were released at the beginning, updated every week, and available to students until the end of the course. The data set consisted of two different types: data on student achievements during the course and interaction data on students accessing materials and doing activities in the VLE. The learning process was designed using three major learning elements and was implemented using eight learning features as described in Figure 2 including assignment (F1), file (F2), forum (F3), homepage (F4), label (F5), page (F6), quiz (F7), and URL (F8). During the learning process, students navigated through the website to read the module, discuss in a forum, send files for assignments, or complete quizzes. The clickstream data (i.e., log of users when navigating in VLE webpages) were recorded. These were valuable data that could be used to generate insights regarding student performance in the final week of the course. More than 202,000 logs of 977 students for this data set were obtained and used as inputs for the prediction model. Meanwhile, final course scores were formerly maintained alphabetically. However, this prediction study converted the final scores into a categorical binary representation. The classification was conducted by differentiating students with the "passed" (1) if the final score \geq 50. In contrast, the "failed" and "withdrawn" (final score < 50) were combined in the class "0," indicating at-risk students.

3.3. Prediction Model Based on the LSTM Network

An LSTM-based model and its derivative models (i.e., CNN-LSTM and Conv-LSTM) were applied in this study. Specifically, the LSTM can be employed to extract temporal patterns in nonlinear time-series data [30]. The LSTM utilized two gates (i.e., forget gate and input gate) to control the value of the cell state Y. The forget gate f determined the value of the current cell state by retaining the previous cell state, while the input gate *i* determined the value of the current cell state by maintaining the network input. First, the forget and the input gates calculated the current input x_t , previously hidden state h_{t-1} , weight w, and bias b using a 'sigmoid' activation function γ .

$$f_t = \gamma \Big(w_f \cdot (h_{t-1}, x_t) + b_f \Big) \tag{1}$$

$$i_t = \gamma(w_i \cdot (h_{t-1}, x_t) + b_i) \tag{2}$$

Next, Equation (3) formulated a new candidate value of cell state Y_t , i.e., calculated using a 'tanh' activation function θ . Finally, the old cell state Y_{t-1} was updated to the current cell state Y_t using Equation (4).

$$\widetilde{Y}_t = \theta(w_c \cdot (h_{t-1}, x_t) + b_c)$$
(3)

$$Y_t = f_t \cdot Y_{t-1} + i_t \cdot \widetilde{Y}_t \tag{4}$$

Three models, including LR, SVM, and LSTM, were selected as the baseline models for performance comparison because they performed well in time-series data prediction [31,32]. The goal was to evaluate which model could generate the highest accuracy in classifying

at-risk students in the early week of the semester. The architectures of the LSTM, CNN-LSTM, and Conv-LSTM are depicted in Table 1. The models were constructed using several deep hidden layers (HL) with a number of trainable parameters integrating two main components: feature extraction and predictor. First, the baseline LSTM contained three layers, i.e., an LSTM layer, a dropout layer, and a fully connected network (FCN) or dense layer that propagated the inputs to the outputs. The number of neurons in each layer was equal to the input sample, and the number of outputs was two, representing the categorical label (i.e., "Fail" and "Pass"). Second, CNN-LSTM consisted of the CNN layer combined with LSTM as feature extraction and the FCN layer as the classifier. The feature extraction module was constructed using two layers of CNN1D, a pooling layer and a flatten layer. The output of the feature extraction module was in one dimension. Then, an LSTM layer (with 16 LSTM cells) was utilized. On the classifier, an FCN layer was employed. Because of the relatively small sample size of educational data (i.e., covering 16 weeks), the more sophisticated deep CNN architecture could not further improve the prediction performance. Thus, this model was proposed with a simplified concept. Third, the design of the Conv-LSTM model was detailed as follows. A feature extraction module was employed using a ConvLSTM2D layer. Subsequently, a flatten layer was applied. Finally, the FCN layer was implemented. This design provided a simple layer construction with a moderate number of parameters while maintaining good accuracy.

Table 1. An overview of the proposed LSTM, CNN-LSTM, and Conv-LSTM architectures provide an early prediction system for at-risk students.

Layer		LSTM	CNN-LSTM	Conv-LSTM
Parameters		d = 0.2; Y = 2; c = 8; $\alpha = 0.001; e = 200; b = 20$	d = 0.2; Y = 2; c = 8; r = 1; s = 2; o = 8; k = 3; p = 'same'; w = 2; $\alpha = 0.001; \varepsilon = 200; b = 20$	
Input		x [N, T, c]	x [N, s, t, c]	x [N, s, r, t, c]
Feature Extraction	HL_1	LSTM(N)	TimeDistributed(CNN1D(<i>o</i> , <i>k</i> , <i>p</i>))	ConvLSTM2D(o, k, p)
	HL_2	Dropout(<i>d</i>)	TimeDistributed(CNN1D(<i>o</i> , <i>k</i> , <i>p</i>))	Dropout(<i>d</i>)
	HL ₃	-	TimeDistributed(MaxPooling(w))	Flatten
	HL ₄	-	Flatten	-
	HL ₅	-	LSTM(N)	-
	HL ₆	-	Dropout(d)	-
Classifier		Dense(N)	Dense(N)	Dense(N)
Output		Dense(Y)	Dense(Y)	Dense(Y)
Number of trainable parameters		1906	3714	1978

The inputs of LSTM, CNN-LSTM, and Conv-LSTM models were denoted in the following order x [N, T, c], x [N, s, t, c], and x [N, s, r, t, c], respectively, where c, N, T, t, r, and s represent the number of features (channels), the number of samples taken by the prediction model to capture the temporal patterns, time-step, subsequence (columns), rows, and the number of samples, respectively. In both CNN-LSTM and Conv-LSTM architectures, the number of sequence data (i.e., time-step T) was split into subsequence t using a divider s; hence, $T = s \times t$. The number of samples N were 6, 8, 10, 12, and 16, indicating the week-wise data sequence. The detailed setting can be seen in the layer setting in Table 1. A constant number of features, dropout rate d, filter o, kernel size k, pooling window w, learning rate α , epoch ε , and batch size b were used. Padding parameter p was set to 'same'. These three designs were compared in terms of performance matrices using early sets and a full set of data, the efficient number of parameters, and the model interpretability.

3.4. Explainable AI Model

Explainable AI (XAI) was described specifically to exploit the local explainability of time series data [33]. Explainable AI collectively refers to methods that can exploit the interpretability of a given decision-making process, such as traditional and modern ML models. An enormous potential is shown from this branch of AI study that can unbox the modern ML 'black-box' model [34]. LIME is powerful because it provides accessibility and simplicity [35]. LIME inherits the basic idea of model agnosticism to explain any given supervised learning model by treating it as a 'black-box' separately. LIME provides a local explanation by weighting adjacent observations. Local explanations mean that LIME gives locally faithful explanations within the surrounding observations of the sample being explained. A local linear model based on the adjacent weighted observations was trained to achieve these explanations. LIME minimizes the objective function ξ using Equation (5), where *f* is the prediction model, *x* is the specific observations around *x*, and $\Omega(g)$ is the complexity of *g*, which is kept low.

$$\xi = \operatorname{argarg} L(f, g, \pi_x) + \Omega(g) \operatorname{argarg} L$$
(5)

The LIME explainer was trained using the LIME tabular explainer function. For each observation, LIME outputted importance values from data feature x [N, T, c] for each week-wise data. As week-wise individual observations were ineffectual, this study focused on the global observations of feature c with LIME. Nevertheless, the study also presents the local explanations and represents the inputs' data set, thus making the comparison more sensible.

4. Experiments

This study evaluated the performance of baseline ML models and LSTM-based models in classification tasks, especially in student performance early prediction. In addition, this work performed an interpretability analysis by using the LIME method. In the following sections, a brief description of the results is presented.

4.1. Performance Evaluation

For the first experiment, a real-world clickstream data set was collected containing records of 977 students who took an online course. There were eight features for each student in the data set and 16 monitoring weeks. The following eight features were used as data input for this experiment: assignment, file, forum, homepage, label, page, quiz, and URL. Meanwhile, the final student grade category (i.e., pass and fail) was employed as the classification label. Table 2 presents the statistical analysis of the realworld data set in terms of the normalized standard deviation (NSD), normalized mean (NM), skewness, and kurtosis. The NSD measured how close the data distribution was to the mean value. Skewness measured data asymmetry around the mean value. Normal distribution, ideally symmetric, had a zero skewness. Meanwhile, kurtosis indicated the distribution susceptibility to outliers. The ideal kurtosis value of the normal distribution was close to 3. Based on statistical analysis, the NSD and NM of the class label were the highest, with a value of 0.332847 and 0.873503, respectively. From here, the data set was imbalanced with most decisions going to the majority class (i.e., grade: pass). On the other hand, similar data characteristics were observed for the eight features. The features had a positive value of skewness. The kurtosis of the features also deviated from 3 when compared with the class label. Because the kurtosis was higher than 3, the distribution of the eight features had a heavier tail and a sharper peak than that of the normal distribution.

Instance	NSD	NM	Skewness	Kurtosis
Assignment	0.071788	0.016286	3.655578	16.921989
File	0.049630	0.006317	7.398770	70.913583
Forum	0.056196	0.016364	14.388217	247.806282
Homepage	0.114323	0.042051	9.775106	136.711158
Label	0.068407	0.011654	7.995445	78.579698
Page	0.080140	0.011738	8.452196	79.764197
Quiz	0.055256	0.006360	11.052219	143.488354
URL	0.054987	0.006173	12.353567	179.174638
Class label (target)	0.332847	0.873503	-2.246093	3.046236

Table 2. The evaluated data sets and their characteristics.

Correspondingly, this imbalanced data set was investigated in terms of performance comparison, as depicted in Table 3. The imbalanced data set will become a major problem, for example, for an imbalanced data set with a 10% at-risk rate. Hence, a classification model may successfully make a prediction with a 0.9 accuracy for all students. However, in this case, the classification model failed to predict any at-risk students. To make detailed investigations, the baseline approaches of LR, SVM, and LSTM were evaluated. Then, a hybrid sampling technique was performed on the data set. This technique synthesized new examples from minority classes using the SMOTE algorithm. SMOTE worked by selecting minority examples being close in the feature space, drawing a line between the examples, and then determining a new sample at a point along that line. After that, a random under-sampling was performed on the majority class. This hybrid (i.e., SMOTE-RUS) sampling technique achieved better results, as stated in [36]. Table 3 signifies that the hybrid sampling method improved the performance of the prediction model, especially on precision, recall, and F1-score matrices, increasing to more than 90% than those without any sampling strategy. Meanwhile, accuracy was comparable for both with and without hybrid sampling.

Feature Extraction	Classification Model	Oversampling	Accuracy	Precision	Recall	F1-Score
None	LR	None hybrid SMOTE	0.920 0.910	0.775 0.900	0.955 0.915	0.830 0.910
None	SVM	None hybrid SMOTE	0.930 0.930	0.840 0.925	0.840 0.935	0.840 0.930
LSTM	FCN	None hybrid SMOTE	0.938 0.934	0.855 0.948	0.845 0.930	0.845 0.930

Table 3. Summary of the baseline model's performance using a 16-week-wise data set.

4.2. Characteristic of the Convolutional Neural Network Model

The second experiment explored whether the proposed convolutional deep neural network models (i.e., CNN-LSTM and ConvLSTM) could achieve better than baseline NN models (i.e., LSTM) for weight convergence in feedforward neural networks. In the LSTM approach, the 128 vector lengths with absolute values were constructed into onedimensional data for data recognition to identify at-risk students. On the other hand, in both the CNN-LSTM and the ConvLSTM approaches, the 128 vector lengths were composed into two-dimensional 16 \times 8 data to improve spatial feature recognition between adjacent data values.

An LSTM model architecture with an LSTM layer as feature extraction and a fully connected network (FCN) layer as a prediction model enabled the architecture to learn more complex long short-term representations for classification tasks. Each layer propagated the output of the previous layer as its input, providing comprehensive learning by training the weights to reach a convergence state. A dropout layer was used between HLs to avoid interdependency between the LSTM and Dense layers, reducing overfitting and underfitting during the training process. In the training phase, the data were computed at each epoch using several scenarios, i.e., from the first week to *i*-th week, where *i* = 6, 8, 10, 12, and 16. The week-wise sequences were converted to equal-length normalized input vectors. Then the vectors were padded into the input layer. A fixed Adam optimizer was applied with a dropout rate of 0.2 and categorical cross-entropy as the loss function. Multiple evaluations were performed to discover the optimal epoch, batch size, and learning rate, ranging from 10 to 300, 10 to 100, and 0.01 to 0.0001, respectively. The hyper parameter setting was decided after several experiments with the final epoch *e* = 200, the batch size *b* = 20, and the learning rate α = 0.001.

The results in Figure 3 illustrate the efficacy of the proposed (a) LSTM, (b) CNN-LSTM, and (c) Conv-LSTM models by representing the training improvement after 200 epochs. The figure presents three kinds of neural network models, i.e., LSTM, CNN-LSTM, and Conv-LSTM, to predict students' performance. Generally, on the left side of Figure 3, the loss accuracy progressively decreases from the early weeks to the late weeks. In contrast, on the right side of Figure 3, the accuracy progressively increases from the initial weeks to the final weeks. As a sufficient and conclusive data set could not be obtained in the initial weeks, especially in the first week, where the previous instance was unavailable, the experiment provided the loss and accuracy correction from week 6. Figure 3 exhibits the progress for accurate earlier prediction using the testing subset. On average, for three LSTM-based models, the accuracy scale ranged from 0.865 in the sixth week to 0.879 achieved in the eighth week up to 0.942 obtained in the last, 16th, week. A gradual improvement in the accuracy was observed since the 10th week, indicating the efficiency of the deployed model in learning the distinct multivariate patterns of students' performance with accuracy over 0.871, especially using both CNN-LSTM and Conv-LSTM. The figure depicts that the loss values continuously degrade, signifying a progressively smaller difference between the target class label and the predicted class label. In addition, overfitting became another challenge in the Al model. Therefore, how well the proposed models perform on a new data set cannot be known unless tested on actual data. To address this issue, a 10-fold cross-validation was performed on separate training and testing subsets, as illustrated in Figure 4. In general, the 10-fold cross-validation accuracy of CNN-LSTM and Conv-LSTM from week 6 to the final week was higher than those of the LSTM model. The longer the data set, the better the performance of the proposed models with an accuracy of more than 0.950 when using a 16-week-wise data set.

4.3. Early Prediction of At-Risk Students

The third experiment aimed to investigate the average performance of the proposed models compared to the baseline models when making an early six-week-wise prediction. Table 4 compares the performance of the models using the average values of the 10-fold cross-validation. The results revealed that the CNN-LSTM and the Conv-LSTM performed better than traditional machine learning models (i.e., LR and SVM) and regular LSTM models for all measurement matrices. On average, the LR, the SVM, the LSTM, the CNN-LSTM, and the Conv-LSTM had a recall rate of 0.64, 0.77, 0.84, 0.88, and 0.91, respectively. These results indicate that the LR and the SVM encountered difficulties in dealing with many input variables without vectorization. The classic machine learning methods had difficulties in extracting a more complex pattern. On the other hand, the LSTM-based model achieved better results with its massive learning capability. Specifically, the CNN-LSTM and Conv-LSTM models performed more superiorly than the LSTM model by 4.7% and 8.3% in the overall F1-score, respectively. Moreover, the Conv-LSTM model achieved the highest performance from all tested models with a moderate number of parameters.



Figure 3. Week-wise validation loss and accuracy of three neural networks-based approach: (**a**) LSTM, (**b**) CNN-LSTM, and (**c**) Conv-LSTM.



Figure 4. Boxplot of the 10-fold cross-validation of validation accuracy.

Table 4. Summary of the early prediction of model performance using a six-week-wise data set.

Feature Extraction	Classifier	Accuracy	Precision	Recall	F1-Score
Loss Function	LR	0.64	0.64	0.64	0.64
Loss Function	SVM	0.80	0.78	0.77	0.79
LSTM	FCN	0.85	0.84	0.84	0.84
CNN-LSTM	FCN	0.88	0.88	0.87	0.88
Conv-LSTM	FCN	0.91	0.91	0.91	0.91

4.4. Interpretability Analysis

The last experiment unveiled that both LSTM-based approaches could provide explainability of the results comparable to the traditional methods in the early warning prediction. However, due to the complexity of both models, it was hard to identify the important features of both models. However, several local features could be identified in the feature importance visualization in Figure 5. The visualized data set regions between the successful and at-risk students could be distinguished. Furthermore, the LIME method could describe that students should notice several areas because of the importance of these activities, which could strongly affect the final results. The students could use this explainability to improve and put their efforts into the mentioned activities; thus, they would be successful at the end of the course.



Figure 5. The visualized data set of two types of student performance with its interpretability: (**a**) 'Fail' and (**b**) 'Pass'.

5. Discussion

5.1. Findings from the ESPP Experiment

Our experiment revealed that the LTSM and its derivative model outperformed baseline models, including LR and SVM. The baseline ML models could not capture any at-risk students at the end of the sixth week in a 16-week course with F1-scores below 0.8. On the other hand, the LSTM-based models (i.e., LSTM, CNN-LSTM, and Conv-LSTM) could effectively extract the temporal patterns inside the VLE data set with F1-scores of 0.84, 0.88, and 0.91, respectively. The traditional ML experiment had poor validation performance due to serious overfitting during the training phase, as exhibited in Table 4, especially when the data set was highly imbalanced. From the aspect of interventions, the combination of convolutional and LSTM layers had greater potential in both research and implementation, especially in spatiotemporal data processing. Nowadays, more and more end devices such as IoT that connect to the network might be utilized to track students' learning in real time. This DL architecture can be used to generate intelligent predictive modeling that improves the VLE learning process in the future.

The major purpose of this study was to improve the performance in early prediction cases while maintaining the explainability of the prediction results. Two promising models were revealed from the experimental results, namely, CNN-LSTM and Conv-LSTM. CNN-LSTM has proven to produce higher accuracy but with a trade-off of an increasing number

of training parameters. On the other hand, the Conv-LSTM provided the highest performance but with a slightly increasing number of trained parameters. The total parameters used in the training phase for LSTM, CNN-LSTM, and Conv-LSTM were 1906, 3714, and 1978, respectively. The more parameters used in a predictive model, the longer the training process because all parameters were calculated in propagation learning. However, only an inference process was performed in the implementation phase. Thus, the difference in time execution could be minimized. If given the limited data set, traditional ML methods are insufficient to extract the temporal patterns inside the data set, resulting in overfitting learning. Therefore, this research supports the development of intelligent predictive VLE

with two proposed LSTM-based model architectures improving prediction performance.

5.2. Insights from the XAI Experiment

The current study used a well-characterized student performance data set to exploit modern ML methods to interpret the prediction results. Both types of modern ML methods (i.e., CNN-LSTM and Conv-LSTM) demonstrated the best and the second-best performances from the experiments. This work shows that modern DL techniques are not necessarily black boxes. Descriptive knowledge can be presented to explain the assessments of features regarding future results, especially in the context of the presented study, i.e., student performance early prediction. From the students' point of view, black-box understanding might be unacceptable because they need reasons to categorize them as at-risk students [37]. Therefore, an explainable prediction could improve student experience during learning in the VLE. Accuracy of the prediction model can be essential but explainability is more important [38]. Explainability is a must-have feature on AI-based prediction systems without limiting the accuracy. Nonetheless, explainability is a feature that supports the development of ESPP but does not replace the role and assessments made by teachers.

As depicted in Figure 5, the local explanation does not completely represent the global explanations. To simplify the local explanations, global feature importance needs to be calculated. The instructor can use the feature importance to improve the study experience in the VLE by assessing the ESPP globally. A basic idea to summarize local explanations provided by LIME is by calculating local predictions for many samples and assessing the mean value assigned to each feature importance. Figure 6 illustrates the feature importance rating derived from the explainability methods for the LSTM prediction model. The bar heights denote mean values and error bars indicate standard deviation over data samples. These values were normalized to the range 0 to 1. The model rated assignment consistently as the most important feature. The model pointed out forum, label, file, and homepages with a comparable importance score for important moderate features. For less important features, the results were insignificant for URL, page, and quiz.



Figure 6. The global feature importance from the data set involving recorded activities of all students.

Given the trade-off between performance and explainability, the general performance of modern ML techniques outperformed traditional ML techniques. The results revealed that the proposed CNN-LSTM and Conv-LSTM models provided comparable explainability while achieving higher accuracy than the baseline models. These promising predictive models have greatly accelerated the development of explainable VLE. However, as displayed in Figure 5, technical presentations of at-risk activities might be less sufficient. A translated at-risk activity in the easy-to-understand ranking (e.g., from "at-risk" to "fair" to "outstanding") might provide a better understanding. In addition, the security aspect, specifically in the federation of the VLE system, could be further analyzed as the federated learning approach could be used to securely train anonymous data sets [39]. This study also encourages future research to evaluate hybrid learning (i.e., online and offline learning) with detailed-class labeling, improving the user experience during the learning process.

6. Conclusions

The advancement of machine learning technologies has received great attention in the development of VLE for predicting student performance in earlier weeks of the teaching and learning process. In early prediction cases, traditional machine learning models fail to predict student performance due to insufficient data, imbalanced data sets, and a lack of understanding about how the models predict the results. This study provides an intelligent framework for explainable student performance prediction using two innovative prediction models called CNN-LSTM and Conv-LSTM to simultaneously improve the predictive models' performance and explainability. The prediction performance was evaluated by comparing the proposed CNN-LSTM and Conv-LSTM models with three baseline LSTM, SVM, and LR models, resulting in F1-scores of 0.91, 0.88, 0.84, 0.79, and 0.64, respectively. The proposed prediction models offer many improvements including a lower misclassification rate, a higher sensitivity rate, and explainability features for instructors to improve the VLE activities. There are two potential extensions of the research in the VLE. One possible extension is to conduct experiments on hybrid learning strategies with deep convolutional prediction models. The other extension is to inspect the student's performance if the instructor intervention is imposed during learning activities.

Author Contributions: Conceptualization, H.-C.C. and E.P.; methodology, E.P., K.T.P. and P.; software, K.T.P. and P.; validation, H.-C.C., S.-S.T. and S.S.K.; formal analysis, H.-C.C. and S.S.K.; investigation, E.P. and K.T.P.; resources, E.P. and S.S.K.; data curation, P. and S.S.K.; writing—original draft preparation, H.-C.C., S.-S.T. and E.P.; writing—review and editing, H.-C.C., E.P., S.-S.T. and C.-E.W.; visualization, K.T.P., P. and C.-E.W.; supervision, H.-C.C. and S.-S.T.; project administration, H.-C.C.; funding acquisition, H.-C.C. and C.-E.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology (MOST), Taiwan, under MOST Grant Nos. 110-2218-E-468-001-MBK, 110-2221-E-468-007, and 110-2218-E-002-044. This work was supported in part by the Ministry of Education under Grant No. I109MD040.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work was funded by the Ministry of Science and Technology (MOST), Taiwan, under MOST Grant Nos. 110-2218-E-468-001-MBK, 110-2221-E-468-007, and 110-2218-E-002-044. This work was supported in part by the Ministry of Education under Grant No. 1109MD040. This work was supported in part by Asia University, Taiwan, and China Medical University Hospital, China Medical University, Taiwan, under Grant Nos. ASIA-110-CMUH-22, ASIA-108-CMUH-05, ASIA-106-CMUH-04, and ASIA-105-CMUH-04. This work was also supported in part by Asia University, Taiwan, UMY, Indonesian, under Grant 107-ASIA-UMY-02. This study was also supported in part by Universitas Muhammadiyah Yogyakarta, Indonesia.

Conflicts of Interest: The authors declare no conflict of interest. Moreover, this article can be counted as the KPI record of Ph.D. degree for only the second author, E. Prasetyo.

References

- 1. Crawford, J.; Butler-Henderson, K.; Rudolph, J.; Malkawi, B.; Glowatz, M.; Burton, R.; Magni, P.A.; Lam, S. COVID-19: 20 countries' higher education intra-period digital pedagogy responses. *J. Appl. Learn. Teach.* **2020**, *3*, 1–20. [CrossRef]
- 2. Murphy, M.P.A. COVID-19 and emergency eLearning: Consequences of the securitization of higher education for post-pandemic pedagogy. *Contemp. Secur. Policy* 2020, *41*, 492–505. [CrossRef]
- Bin Mat, U.; Buniyamin, N.; Arsad, P.M.; Kassim, R. An Overview of Using Academic Analytics to Predict and Improve Students' Achievement: A Proposed Proactive Intelligent Intervention. In Proceedings of the 2013 IEEE 5th Conference on Engineering Education (ICEED), Kuala Lumpur, Malaysia, 4–5 December 2013; pp. 126–130.
- Zhou, J.; Zhang, Q. A Survey Study on U.S. College Students' Learning Experience in COVID-19. Educ. Sci. 2021, 11, 248. [CrossRef]
- 5. Walker, K.A.; Koralesky, K.E. Student and instructor perceptions of engagement after the rapid online transition of teaching due to COVID-19. *Nat. Sci. Educ.* **2021**, *50*, e20038. [CrossRef]
- 6. Guzsvinecz, T.; Szűcs, J. Using Analytics to Identify When Course Materials Are Accessed Relative to Online Exams during Digital Education. *Educ. Sci.* 2021, *11*, 576. [CrossRef]
- Barabash, O.; Musienko, A.; Hohoniants, S.; Laptiev, O.; Salash, O.; Rudenko, Y.; Klochko, A. Comprehensive Methods of Evaluation of Efficiency of Distance Learning System Functioning. Int. J. Intell. Syst. Appl. 2021, 13, 16–28. [CrossRef]
- Riestra-González, M.; del Puerto Paule-Ruíz, M.; Ortin, F. Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Comput. Educ.* 2021, 163, 104108. [CrossRef]
- Chen, H.-C.; Prasetyo, E.; Prayitno; Kusumawardani, S.S.; Tseng, S.-S.; Kung, T.-L.; Wang, K.-Y. Learning Performance Prediction with Imbalanced Virtual Learning Environment Students' Interactions Data. In *Innovative Mobile and Internet Services in Ubiquitous Computing*; Barolli, L., Yim, K., Chen, H.-C., Eds.; Lecture Notes in Networks and Systems; Springer International Publishing: Cham, Switzerland, 2022; Volume 279, pp. 330–340, ISBN 978-3-030-79727-0.
- 10. Wen, G.; Li, X.; Zhu, Y.; Chen, L.; Luo, Q.; Tan, M. One-step spectral rotation clustering for imbalanced high-dimensional data. *Inf. Process. Manag.* **2020**, *58*, 102388. [CrossRef]
- 11. Kraiem, M.S.; Sánchez-Hernández, F.; Moreno-García, M.N. Selecting the Suitable Resampling Strategy for Imbalanced Data Classification Regarding Dataset Properties. An Approach Based on Association Models. *Appl. Sci.* **2021**, *11*, 8546. [CrossRef]
- 12. Lin, I.; Loyola-González, O.; Monroy, R.; Medina-Pérez, M. A Review of Fuzzy and Pattern-Based Approaches for Class Imbalance Problems. *Appl. Sci.* **2021**, *11*, 6310. [CrossRef]
- 13. Sekeroglu, B.; Abiyev, R.; Ilhan, A.; Arslan, M.; Idoko, J.B. Systematic Literature Review on Machine Learning and Student Performance Prediction: Critical Gaps and Possible Remedies. *Appl. Sci.* **2021**, *11*, 10907. [CrossRef]
- 14. Zafari, M.; Sadeghi-Niaraki, A.; Choi, S.-M.; Esmaeily, A. A Practical Model for the Evaluation of High School Student Performance Based on Machine Learning. *Appl. Sci.* 2021, *11*, 11534. [CrossRef]
- 15. Alshanqiti, A.; Namoun, A. Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification. *IEEE Access* 2020, *8*, 203827–203844. [CrossRef]
- Aljohani, N.R.; Fayoumi, A.; Hassan, S.-U. Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment. Sustainability 2019, 11, 7238. [CrossRef]
- 17. Hone, K.S.; El Said, G.R. Exploring the factors affecting MOOC retention: A survey study. *Comput. Educ.* **2016**, *98*, 157–168. [CrossRef]
- 18. Wasif, M.; Waheed, H.; Aljohani, N.R.; Hassan, S.-U. Understanding Student Learning Behavior and Predicting Their Performance. In *Cognitive Computing in Technology-Enhanced Learning*; IGI Global: Hershey, PA, USA, 2019; p. 28. [CrossRef]
- Marbouti, F.; Diefes-Dux, H.; Strobel, J. Building Course-Specific Regression-Based Models to Identify At-Risk Students. In Proceedings of the 2015 ASEE Annual Conference and Exposition Proceedings, ASEE Conferences. Seattle, WA, USA, 14–17 June 2015; pp. 26.304.1–26.304.11.
- 20. Chui, K.T.; Fung, D.C.L.; Lytras, M.D.; Lam, T.M. Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Comput. Hum. Behav.* **2018**, 107, 105584. [CrossRef]
- 21. Marbouti, F.; Diefes-Dux, H.; Madhavan, K. Models for early prediction of at-risk students in a course using standards-based grading. *Comput. Educ.* 2016, 103, 1–15. [CrossRef]
- Fedushko, S.; Ustyianovych, T. Predicting Pupil's Successfulness Factors Using Machine Learning Algorithms and Mathematical Modelling Methods. In *Advances in Computer Science for Engineering and Education II*; Hu, Z., Petoukhov, S., Dychka, I., He, M., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2020; Volume 938, pp. 625–636, ISBN 978-3-030-16620-5.
- 23. Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.M.R. Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. *Comput. Intell. Neurosci.* 2018, 2018, 1–21. [CrossRef]
- 24. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521, 436–444. [CrossRef]
- 25. Hassan, S.-U.; Waheed, H.; Aljohani, N.R.; Ali, M.; Ventura, S.; Herrera, F. Virtual learning environment to predict withdrawal by leveraging deep learning. *Int. J. Intell. Syst.* 2019, *34*, 1935–1952. [CrossRef]
- Okubo, F.; Yamashita, T.; Shimada, A.; Ogata, H. A Neural Network Approach for Students' Performance Prediction. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, 13 March 2017; pp. 598–599.

- Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 2019, *58*, 82–115. [CrossRef]
- Kenny, E.M.; Keane, M.T. Explaining Deep Learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI. *Knowledge-Based Syst.* 2021, 233, 107530. [CrossRef]
- Khlaisang, J.; Songkram, N. Designing a Virtual Learning Environment System for Teaching Twenty-First Century Skills to Higher Education Students in ASEAN. *Technol. Knowl. Learn.* 2017, 24, 41–63. [CrossRef]
- Putra, K.; Chen, H.-C.; Prayitno; Ogiela, M.; Chou, C.-L.; Weng, C.-E.; Shae, Z.-Y. Federated Compressed Learning Edge Computing Framework with Ensuring Data Privacy for PM2.5 Prediction in Smart City Sensing Applications. *Sensors* 2021, 21, 4586. [CrossRef]
- Al-Hadeethi, H.; Abdulla, S.; Diykh, M.; Deo, R.C.; Green, J. Adaptive boost LS-SVM classification approach for time-series signal classification in epileptic seizure diagnosis applications. *Expert Syst. Appl.* 2020, 161, 113676. [CrossRef]
- Solgi, R.; Loáiciga, H.A.; Kram, M. Long short-term memory neural network (LSTM-NN) for aquifer level time series forecasting using in-situ piezometric observations. J. Hydrol. 2021, 601, 126800. [CrossRef]
- Yang, G.; Ye, Q.; Xia, J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf. Fusion* 2021, 77, 29–52. [CrossRef]
- 34. Chen, H.-C.; Damarjati, C.; Putra, K.T.; Chen, H.-M.; Hsieh, C.-L.; Lin, H.-J.; Wu, M.-Y.; Chen, C.-S. Pulse-line intersection method with unboxed artificial intelligence for hesitant pulse wave classification. *Inf. Process. Manag.* **2022**, *59*, 102855. [CrossRef]
- Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
- Mishra, N.K.; Singh, P.K. Feature construction and smote-based imbalance handling for multi-label learning. *Inf. Sci.* 2021, 563, 342–357. [CrossRef]
- 37. van der Waa, J.; Nieuwburg, E.; Cremers, A.; Neerincx, M. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artif. Intell.* **2020**, *291*, 103404. [CrossRef]
- Panigutti, C.; Perotti, A.; Panisson, A.; Bajardi, P.; Pedreschi, D. FairLens: Auditing black-box clinical decision support systems. *Inf. Process. Manag.* 2021, 58, 102657. [CrossRef]
- Prayitno; Shyu, C.-R.; Putra, K.T.; Chen, H.-C.; Tsai, Y.-Y.; Hossain, K.S.M.T.; Jiang, W.; Shae, Z.-Y. A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. *Appl. Sci.* 2021, 11, 11191. [CrossRef]