applied sciences

MDPI

Article

FocalGAN-SEE: A Generative Adversarial Framework with SE Enhanced Encoder for Hard Sample Identification in Online Learning Early Warning

Jia Liu [1] and Mingyan Zhang [2,3,*]

[1] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China; skyy8484@mails.ccnu.edu.cn
[2] College of Education, Zhejiang Normal University, Jinhua 321004, China
[3] Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua 321004, China
* Correspondence: mingyanzhang@zjnu.edu.cn

Abstract

Online learning has become a mainstream instructional modality worldwide. However, early warning models suffer from three critical deficiencies in hard sample identification: temporal feature dependency, class imbalance, and weak feature extraction capability for boundary-ambiguous samples. This paper proposes an early warning framework, FocalGAN-SEE, which comprises a Focal Generative Adversarial Network (FocalGAN) that generates hard boundary samples to alleviate class imbalance, and an SE attention-enhanced encoder (SE Encoder) that adaptively amplifies discriminative features of hard at-risk samples through temporal channel recalibration. Unlike standard generative models, the FocalGAN incorporates a focal hard-sample loss that drives the generator to produce boundary-ambiguous samples, thereby enhancing the classifier's robustness to hard samples. We compared the proposed architecture with attention-enhanced autoencoders, deep generative models, and oversampling-based models currently used in online learning early warning systems. The results show that the proposed FocalGAN-SEE achieves superior performance across accuracy, precision, recall, F1-score, and AUC, and can identify more hard at-risk samples in earlier course modules. Data visualization shows that the focal loss successfully guides the model to focus on the hard region with the highest overlap in the feature space, and FocalGAN-SEE can detect more hard samples at ambiguous boundaries. Finally, we adopt SHAP to interpret the prediction logic of the model, so as to assist teachers in implementing personalized teaching interventions. This paper provides a principled solution to hard at-risk sample recognition in imbalanced sequential data and offers practical value for online course management and personalized interventions.

Keywords: online learning early warning; at-risk student; hard sample in education; deep generative model; SE attention-enhanced encoder

Academic Editor: Nicola Magnavita

Received: 5 August 2026
Revised: 18 October 2026
Accepted: 2 November 2026
Published: 4 November 2026

Copyright: © 2026 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.

1. Introduction

Online learning enables students to engage in educational activities anytime and anywhere, and has become a mainstream learning modality in the world [1]. Early warning systems can accurately predict student academic performance with online learning data, thereby providing instructors with critical information for timely instructional in-

terventions and ultimately enhancing students' online learning performance. Consequently, early warning has become a prominent topic within the field of learning analytics [2].

Online learning early warning systems increasingly employ machine learning algorithms to identify as many at-risk students as possible at the early stage of the course [3]. Such students become the focus of teachers' subsequent attention, and personalized intervention measures are adopted to help them avoid academic failure by the end of the teaching period [4].

For the identification of at-risk students, one key challenge arises from hard-sample cases. Hard at-risk samples refer to at-risk students with ambiguous, heavily overlapping behavioral features that hinder reliable model discrimination. The hard-sample problem frequently arises from data noise, limited dataset size, and model-capacity constraints [5]. Normal at-risk samples refer to at-risk students with distinct, well-separated behavioral features that can be readily captured by predictive models; however, hard at-risk samples exhibit obscure features and account for a relatively small proportion of the dataset, making them difficult for models to identify correctly. Consequently, these students cannot be precisely targeted by instructors.

A distinctive feature of education is that no student can be overlooked; hence, identifying at-risk students as comprehensively as possible is the primary requirement of instructors. Nevertheless, most existing early warning models pursue only the overall recognition rate of at-risk students, failing to distinguish between simple at-risk samples and hard at-risk samples, with the vast majority of model evaluation metrics being contributed by simple samples. Therefore, hard sample recognition has become a critical breakthrough for improving the performance of early warning models.

Online learning early warning systems face three major challenges in identifying hard-sample at-risk students: temporal feature dependency, class imbalance, and weak feature extraction capability for boundary-ambiguous samples. Hard samples are commonly characterized by ambiguous features, overlapping class boundaries, and strong noise interference, which prevent models from effectively learning hard sample features and result in poor generalization performance [6,7]. Hard at-risk students constitute a minority class in the classroom, and the lack of sufficient training samples further exacerbates the difficulty of their identification. In the field of machine learning, the core principle of hard sample mining is to focus on samples with ambiguous features and indistinguishability during model training, employing strategies such as differential weight assignment, adaptive sampling, and dynamic loss adjustment to optimize the model's feature representation capability, mitigate the model's bias toward simple samples, and enhance generalization performance [6]. Methods for hard sample recognition have received research attention across multiple industries, including computer vision [8], graph learning [9], bio-signal processing [10], and autonomous driving [11]. However, in educational contexts—particularly in student risk early warning tasks for online learning—research on the identification of hard-sample at-risk students remains notably scarce, and related technical approaches have yet to be fully explored.

Architectures such as generative models in deep learning and autoencoders incorporating attention mechanisms provide feasible approaches for addressing hard sample recognition: deep generative models learn the feature distribution of minority-class samples and generate synthetic samples to balance the class proportions in the training set, enabling classifiers to learn richer minority-class features and improve feature learning effectiveness [3]. Although some researchers have employed GANs [7] or VAEs [12] to generate minority-class samples to address the class imbalance problem in early warning systems, the feature learning process still lacks more precise guidance strategies, the

models are merely optimized to match the global data distribution and fool the discriminator, without any guiding mechanism oriented toward classification decision boundaries. The early warning system's evaluation metrics are also determined by the classifier's ability to learn samples from different categories. Autoencoders integrated with attention mechanisms are capable of screening key features and assigning corresponding weights; they realize key feature enhancement and noise reduction via feature reconstruction, thereby boosting overall classification performance. Existing research has adopted various attention mechanisms together with autoencoders in early warning tasks [4,13], offering valuable references for further advancing the identification performance of hard samples in early warning systems.

To address the three core challenges in identifying hard at-risk students in online learning—temporal feature dependency, class imbalance, and weak feature extraction capability for boundary-ambiguous samples—this paper proposes an early warning framework named FocalGAN-SEE (Focal Generative Adversarial Network with Squeeze-and-Excitation Enhanced Encoder), which integrates a Focal Generative Adversarial Network with an SE attention-enhanced encoder module. The framework aims to improve early warning metrics, particularly the identification of hard-sample at-risk students. The main contributions of this paper are as follows:

(1) We propose the FocalGAN-SEE architecture, which comprises three core modules—the Focal Generative Adversarial Network (FocalGAN) module, the SE attention-enhanced encoder module, and the pre-trained hardness evaluator module. The FocalGAN module learns the latent feature distribution of minority-class students and generates synthetic samples to address the class imbalance problem. Unlike standard GANs that treat all generated samples equally, our FocalGAN incorporates a focal hard-sample loss that explicitly drives the generator to produce confusion-inducing samples whose classifier predictions lie near the decision boundary, thereby enabling the classifier to obtain more challenging boundary-adjacent samples. The SE Encoder module introduces a Squeeze-and-Excitation attention mechanism after the LSTM encoder layer, enabling temporal dependency feature modeling and adaptive channel-wise feature recalibration, while simultaneously suppressing irrelevant features and strengthening the discriminability of hard samples. The pre-trained hardness evaluator—which is the SE-enhanced encoder-classifier pre-trained on the original dataset—provides stable gradient feedback to guide the generator toward producing boundary-ambiguous samples, with its parameters frozen during GAN training to ensure consistent evaluation criteria. The entire system employs gradient tape-based end-to-end backpropagation, ensuring complete gradient flow from the hardness evaluator back to the generator.

(2) We evaluated the classification ability of the core modules—SE attention-enhanced encoder (SE Encoder) on two student datasets derived from real-world online learning scenarios, demonstrating the superiority of SE Encoder on the classification task. The multi-head attention, self-attention, Bahdanau attention, weighted sum attention, local attention, cross-attention, and channel attention were used as baselines. The results indicate that the SE Encoder achieves better prediction performance than all baseline models.

(3) We compared the predictive metrics of FocalGAN-SEE against deep generative models and sampling models commonly used in the early warning domain. Deep generative models (Variational Autoencoder, GANs and their variants) and oversampling models (SMOTE, BSMOTE and ADASYN) commonly used in early warning are set as baselines. Experimental results demonstrate that the proposed architecture achieves higher accuracy, precision, recall, F1-score, and AUC at earlier

stages of learning than all baseline models, and is capable of identifying a greater number of hard at-risk samples.

(4) Through PCA dimensionality reduction and data visualization, we analyzed the feature differences in identified at-risk students between FocalGAN-SEE and commonly used baseline models in the early warning domain. Additionally, we employed the SHAP method to interpret the key variables affecting academic performance. The results indicate that the proposed architecture can identify at-risk students with a wider variety of feature patterns and exhibits superior recognition performance for hard samples located at ambiguous decision boundaries. The key variables contributing to students' academic risks were identified, thereby assisting instructors in conducting pedagogical interventions.

2. Related Work

2.1. Early Warning for Online Learning

Early warning systems monitor various student indicators during the learning process to identify potential academic risks and enable timely interventions [14]. These systems mainly rely on machine learning models to identify at-risk students, enabling instructors to provide personalized support [15].

Researchers have summarized that the machine learning models for academic performance prediction have progressed through several stages: the traditional machine learning model stage, the basic deep learning model stage, the sequential recurrent neural network model stage, and the attention-mechanism-based time-series prediction stage [16].

In the traditional machine learning model stage, prediction models primarily relied on algorithms such as decision trees (DT), support vector machines (SVM), and random forests (RF). Al Hashmi et al. evaluated the application of established machine learning (ML) classifiers as a regional case study, and compared XGBoost with random forest and Naïve Bayes models, achieving relatively good predictive metrics [17].

In the basic deep learning model stage, deep learning models such as DNNs achieved higher predictive performance and attracted considerable research attention. Li et al. [18] used a deep neural network as an early warning model, extracting informative data as features with corresponding weights to help teachers carry out instructional interventions. The traditional static machine learning models excel in small-sample adaptability and training speed, but they overlook the temporal continuity between learning behaviors.

In the sequential recurrent neural network model stage, recurrent neural networks (RNNs) and LSTMs were employed to process learners' sequential data. Katib et al. [19] proposed a variable-reduction-based optimal deep recurrent neural network (VR-ODRNN) approach for student performance and course outcome analysis. Vives et al. [20] explored the efficiency of Long Short-Term Memory (LSTM) networks in the field of Educational Data Mining (EDM) to predict the academic performance of students during the seventh, eighth, twelfth, and sixteenth weeks of the academic semester, which allowed teachers to identify students at risk of failing the course. These models are capable of analyzing the temporal continuity of learning behaviors and capturing temporal dependencies. Nevertheless, they treat behaviors at all time steps with equal weight and also suffer from poor interpretability.

Regardless of the method used, the key to improving early warning is to identify as many at-risk students as possible while ensuring accuracy. Models exhibit significant differences in their ability to identify at-risk students with different feature distributions. At-risk students with highly discriminative features are usually easy for most models to

identify correctly. Therefore, the key to further improving early warning performance lies in identifying more hard samples, namely at-risk students whose behavioral features are ambiguous, heavily overlap with those of successful students, and are difficult for models to discriminate reliably.

Identifying hard-sample at-risk students faces three major challenges: temporal feature dependency, class imbalance, and weak feature extraction capability for boundary-ambiguous samples. From the literature of the past three years (2023 to present), a notable trend can be observed: an increasing number of studies have introduced attention mechanisms and deep generative models into early warning tasks. On the one hand, attention mechanisms can assign learnable weights to different time steps or features, thereby highlighting information that is more important for prediction and modeling long-range dependencies between arbitrary time steps. This facilitates educators and researchers in understanding the model's decision-making logic for filtering key learning behaviors and is widely used to uncover the intrinsic relationships among features for performance prediction. For instance, Fazil et al. [21] proposed a model named MultIFAR that infused multi-dimensional information representing different aspects of student behavior with an attention-driven deep learning model integrating bidirectional long short-term memory and convolutional networks to learn student representation efficiently. Qiu et al. [5] employed a sparse attention mechanism to automatically identify the most discriminative features of students, and then processed these extracted learning features by a deep neural network, which incorporates a Cost-Sensitive Combined Loss function. Singh et al. [22] proposed a method employing the Gated Attention-Based Diffused Graph Convolutional Network (GADGCN), which integrates graph convolution and gating operations to capture complex spatiotemporal dependencies, and dynamically assigns attention weights to node hidden representations. When combined with an autoencoder, it can also learn more discriminative latent representations, so as to enhance feature representation capability, filter noisy behavioral signals, and focus on critical learning patterns [13,23].

On the other hand, deep generative models (such as Generative Adversarial Networks and Variational Autoencoders) are capable of learning the latent data distribution of minority-class samples and generating synthetic samples to mitigate class imbalance at the data level. Specifically, GANs approximate the real-world data distribution via adversarial training between generators and discriminators, while VAE learns latent-variable representations through variational inference. Such methods not only help balance class proportions but also expand the decision regions of minority classes to a certain extent and improve classification boundaries. For instance, Du and Yang [12] proposed the LVAEPre framework, which leverages variational autoencoders to model online learning trajectories and enhance the recognition of at-risk students. Classic oversampling methods such as SMOTE and ADASYN are commonly adopted as baseline approaches. To tackle the trade-off issue where conventional GANs yield high precision yet unsatisfactory recall in early warning tasks, Zhang et al. integrated GANs with oversampling algorithms through a blending strategy, achieving simultaneous improvements in both precision and recall [3]. Duan et al. put forward TVAE-GAN, a generative model designed for early warning of high-risk students in basic education, which achieves better predictive performance for at-risk students [7]. However, prior work only conducts feature learning and sample generation targeting the overall group of at-risk students, without specially addressing the recognition of hard samples. This limitation results in a large number of borderline and extremely hard samples remaining undetected. For student early warning tasks, one core objective is to boost overall prediction performance, and data imbalance mitigation strategies together with the classifier serve as two pivotal

components that determine final prediction outcomes.

Most traditional deep learning-based early warning models suffer from inscrutable internal decision-making logic. Instructors and administrators are unable to identify the root causes of students' academic risk from the prediction results, making it difficult to formulate precise intervention plans. In this context, the integration of machine learning interpretability methods into early warning systems has become increasingly important [24]. Explainable tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been extensively applied in early warning models [25,26]. In current research on early warning systems, it has become a common practice to adopt interpretability techniques such as SHAP to analyze the reasoning behind model predictions and generate actionable insights for instructional interventions. Huang et al. [27] adopted machine learning and the SHAP method to analyze 34,968 samples and 151 features from six East Asian education systems within the PISA 2022 dataset, aiming to pinpoint the crucial factors that affect middle-school students' mathematical literacy. Qiang et al. [28] identified key features influencing students' mathematics performance by SHAP and revealed that the original correlations learned by the ANN were inconsistent with established educational domain knowledge. Kala et al. [29] applied both SHAP and LIME methods to determine the contribution of different features to the DNN model's predictions.

2.2. Hard Sample Mining

Hard Sample Mining (HSM), as a core technique for enhancing the discriminative capability of machine learning, contrastive learning, and clustering models, has now established a comprehensive research framework across multiple domains, including computer vision [8], graph learning [9], bio-signal processing [10], and autonomous driving [11]. Its core principle is to focus on samples that are difficult to distinguish and have ambiguous features during model training, optimizing the model's feature representation capability through strategies such as differential weighting, adaptive sampling, and dynamic loss adjustment, thereby mitigating the model's bias toward easy samples and improving its generalization ability [6].

Tao et al. [30] proposed a foreground association-driven Siamese network with hard sample optimization (HASNet) to deal with the imbalance problem in remote sensing image change detection. Wang et al. [31] proposed a low-frequency attenuation supervised contrastive learning method (SupCon-LFA) to construct hard positive/negative samples for mammogram classification. Sun et al. [32] proposed a self-paced learning strategy with easy sample prior based on confidence (SPL-ESP-BC) to mitigate the negative influence of hard samples for flying bird object detection.

In contrastive learning and clustering tasks, hard samples are primarily divided into two categories: first, hard negative samples [33], which are samples belonging to different classes/clusters but exhibiting extremely high feature similarity; and second, hard positive samples [8], which are samples belonging to the same class/cluster but with significant feature differences. Both types of samples reside at the boundary regions of classes or clusters and are critical factors leading to degraded classification and clustering accuracy. Early machine learning models generally adopted equal-weight training for all samples, with a large number of easy samples dominating the loss computation, preventing the model from fully learning the fine-grained features of boundary samples. Hard sample mining techniques were proposed precisely to address this challenge [6].

Currently, hard sample mining has evolved into multiple mainstream technical approaches, including loss-function-based optimization methods, sampling-strategy-based mining methods, probability-model-based hard sample discrimination methods, and

prototype-library-assisted mining schemes. Among these, loss-function-based optimization methods are the most commonly used. This approach was first applied in the field of object detection, where functions such as focal loss [34] and generalized focal loss [35] are employed to reduce the loss weights of easy samples and increase the training proportion of hard samples.

Nevertheless, the concept of hard sample mining has barely been adopted in the existing research on early warning for online learning, leaving considerable room to further boost the prediction metrics and generalization capability of current early warning models. Student risk samples generated in online learning scenarios inherently carry typical attributes of hard samples, yet this issue has long been overlooked by relevant studies, resulting in an evident research gap.

First, online learning datasets inherently suffer from class imbalance and vague class boundaries, and numerous students with latent academic risks constitute representative hard samples in early warning tasks. On online learning platforms, the vast majority of successful students exhibit standardized learning behaviors and stable academic performance, while only a small fraction of at-risk students shows distinctive risk features. In contrast, borderline students with moderate risks account for a large proportion of the dataset; their behavioral characteristics highly resemble those of normal students, making them typical hard negative samples. Meanwhile, some students with identical risk levels demonstrate drastically different behavioral trajectories due to disparities in prior academic foundations and learning styles, which creates hard positive samples. Conventional early warning models adopt uniform training strategies and overfit the behavioral patterns of the majority of normal students, failing to capture subtle abnormal signals from borderline at-risk students. This ultimately leads to a phenomenon where models achieve accurate detection of obvious high-risk students yet misjudge a large number of latent-risk students, degrading the overall effectiveness of early warning systems.

Second, existing early warning algorithms for online learning lack dedicated optimization strategies tailored to hard samples. As reviewed above, all types of early warning models—including traditional machine learning algorithms, deep learning architectures, and multi-feature hybrid models—take overall classification accuracy as their sole optimization objective, without incorporating weighting, sampling, or loss adjustment schemes derived from hard sample mining. When processing borderline students with ambiguous feature boundaries, such models struggle to extract discriminative features for reliable classification.

Third, hard sample mining bears profound practical value for online learning early warning from an application perspective. The core goal of online education is to deliver proactive interventions and prevent academic failure at an early stage. Compared with severely at-risk students with obvious abnormal features, borderline hard samples represent the primary targets for instructional intervention. If such borderline students can be detected timely and provided with personalized tutoring, teachers are able to help them recover to satisfactory academic status. State-of-the-art mainstream early warning models can only precisely identify overt high-risk groups while lacking reliable predictive power for borderline students.

Hard sample mining has achieved notable success across multiple technical domains, providing a research foundation for mining hard-sample at-risk students in early warning problems within the educational context. Most existing studies are designed for specific data structures such as vision, graph, and time-series data [36,37], and thus require adaptation to address the temporal behavioral data characteristics of educational scenarios. In summary, hard sample mining has already matured as a technique for en-

hancing boundary sample identification, alleviating data imbalance, and strengthening model generalization across various machine learning tasks. However, this approach has not yet been applied to the field of online learning early warning. Conducting specialized mining and model optimization targeted at borderline-risk students—who represent a typical hard sample category in online learning scenarios—constitutes a critical research direction for overcoming the current bottlenecks in early warning technology and improving both the accuracy and practical value of early warning systems. This also serves as one of the core research entry points of this paper.

2.3. Attention Mechanisms and Autoencoder in Education Application

With the rapid growth in the scale and dimensionality of educational data, traditional modeling methods struggle to capture the high temporal complexity and heterogeneity of student behaviors. Attention mechanisms, due to their ability to focus on salient information and improve both modeling efficiency and generalization [38], are widely adopted in educational behavior modeling. Different types of attention mechanisms offer distinct advantages. Bahdanau attention models dependencies between sequences through explicit alignment [39] and has been applied to educational tasks such as image captioning [40]. Self-attention offers global modeling capacity and parallelization, making it suitable for multimodal educational content understanding [41]. Local Attention and Channel Attention were used for short-term behavior focus and feature selection [42,43], while Cross Attention performs well in aligning visual and textual information in interactive teaching contexts [44]. Autoencoders (AEs) are unsupervised models capable of learning compressed representations, widely used for latent variable modeling and anomaly detection in educational data. Li et al. [45] reviewed major AE variants and their performance in educational contexts. Wen et al. [46] introduced temporal encoding to model long-term regression risks. Autoencoders are also used in anomaly detection based on reconstruction error [47]. Berahmand et al. [48] systematically reviewed the evolution of autoencoders in feature extraction and generative learning for education.

Online learning data vary in granularity, semantic level, and temporal span, making student state modeling challenging [46]. Moreover, online learning students exhibit phased and abrupt behavioral changes. Rapid fluctuations and frequent anomalies increase the risk of misjudgment in early warning systems without strong feature selection capabilities. Effectively leveraging multi-source data is essential. The SE mechanism recalibrates feature channels through compression and excitation, enhancing discriminative power by emphasizing informative dimensions. Hu et al. [49] first proposed the SE (Squeeze-and-Excitation Attention) module for dynamic channel-wise weighting in CNNs. Wang et al. [50] developed a lightweight ECA-Net variant for efficient time-series processing. Zhou et al. [51] designed a three-channel SE network that showed superior channel awareness in high-level signal recognition. Huang et al. [52] incorporated SE into sleep stage modeling, improving accuracy and stability in cross-phase identification. The channel selection capabilities with multi-scale inputs of SE have also been verified in medical image analysis [53,54].

Researchers have begun integrating attention mechanisms with autoencoders to form more adaptive modeling frameworks. For knowledge tracing, Ai and Zhang [23] combined convolutional AEs with attention-based GRU to focus on key problem-solving behaviors. Naidu and Xiong [55] designed an ABCD model that uses attention to adjust AE encoding for improved robustness. Sick and Engel [56] developed an AGMA framework combining masking and attention for high reconstruction performance in educational image tasks. Abinaya et al. [40] used Bahdanau Attention to guide autoencoder in

key information selection for image captioning. Zheng et al. [57], proposed a graph attention autoencoder to model student social interactions in latent space. These methods demonstrate the complementary strengths of autoencoder and attention in feature extraction and relational modeling. The integration of attention and AEs has become a key approach in early warning systems. These hybrid models enhance temporal dependency modeling and anomaly detection in complex, high-dimensional behavioral data.

The success of SE attention and the autoencoder in other fields offers a valuable reference for modeling multi-channel behavioral sequences in education. This study investigates the integration of SE attention with AEs to analyze highly temporal and noisy educational log data, to provide a more effective solution for early warning in online learning.

Through a systematic review of academic performance prediction models employing attention mechanisms, the following attention mechanisms are frequently adopted as baseline models or key modules within prediction models: multi-head attention, self-attention [13], Bahdanau attention [39], weight-sum attention [58], local attention [59], cross attention, and channel attention [60]. In this paper, we adopt these attention mechanisms as baseline models for comparison with the SE attention mechanism, and combine the standard autoencoder with each of these attention mechanisms to validate the effectiveness of the proposed architecture in early warning tasks.

We have analyzed the development of online learning early warning models, from classical machine learning to attention-augmented deep networks and deep generative approaches. Existing studies have separately tackled temporal dependency modeling, class imbalance mitigation, and feature extraction tasks, while explainable AI techniques further support educators in understanding model predictions. Nevertheless, prior early warning research rarely introduces hard sample mining to explicitly optimize boundary at-risk students, despite its proven effectiveness in computer vision and other machine learning domains. Attention mechanisms and autoencoders, especially SE attention-enhanced autoencoder architectures, offer powerful tools to extract discriminative temporal features and suppress noise, and deep generative models are able to synthesize minority-class samples for rebalancing datasets. However, few existing works integrate these components together with hard-sample oriented optimization to jointly address the three core challenges for hard at-risk sample recognition. Inspired by hard sample mining and the complementary strengths of SE attention, autoencoders, and generative models, this paper proposes the FocalGAN-SEE framework, which embeds focal hard-sample loss into a generative network and adopts an SE attention-enhanced encoder to strengthen the representation of boundary hard samples for online learning early warning.

3. Methods

To address the three core challenges in identifying hard at-risk student samples for academic performance early warning tasks—namely, temporal feature dependency, class imbalance, and difficulty in feature learning, we propose a Focal Generative Adversarial Network with SE-Enhanced Encoder (FocalGAN-SEE). Figure 1 shows the proposed structure of FocalGAN-SEE framework, which consists of three core components: (i) a Generative Adversarial Network (GAN) [61] with focal hard-sample loss for boundary minority sample synthesis, (ii) a Squeeze-and-Excitation (SE) [49] enhanced temporal encoder (SE Encoder) to extract discriminative student behavior representations, and (iii) an adaptive hard-sample mining strategy guided by a frozen pre-trained classifier.

The overall pipeline proceeds as follows: Given sequential behavioral data $X \in \mathbb{R}^{T \times D}$ of length T with feature dimension D, the framework first pre-trains an SE-en-

hanced encoder-classifier $C(\cdot)$ on the original imbalanced dataset. This pre-trained classifier serves as a hardness evaluator during subsequent GAN training. Specifically, for each synthetic sample X_{fake} generated by the generator G, the classifier outputs a prediction probability $C(X_{\text{fake}}) \in [0, 1]$. The closer this probability is to 0.5, the closer the sample lies to the classifier's decision boundary, and thus it is identified as a hard sample. During adversarial training, the generator is explicitly optimized to minimize a focal hard-sample loss $\mathcal{L}_{\text{hard}}$, which targets a value of 0.5, driving the generator to produce confusion-inducing samples that satisfy $C(X_{\text{fake}}) \approx 0.5$. This enables the classifier to obtain more challenging boundary-adjacent samples. Meanwhile, the parameters of the pre-trained classifier C are frozen during GAN training to provide stable gradient feedback, ensuring consistent hard-sample evaluation criteria. Finally, the generated hard samples are combined with the original data to train the final classifier for at-risk prediction. The entire system employs gradient tape-based end-to-end backpropagation, ensuring complete gradient flow from the classifier back to the generator, enabling the generator to continuously refine its outputs based on classifier feedback.

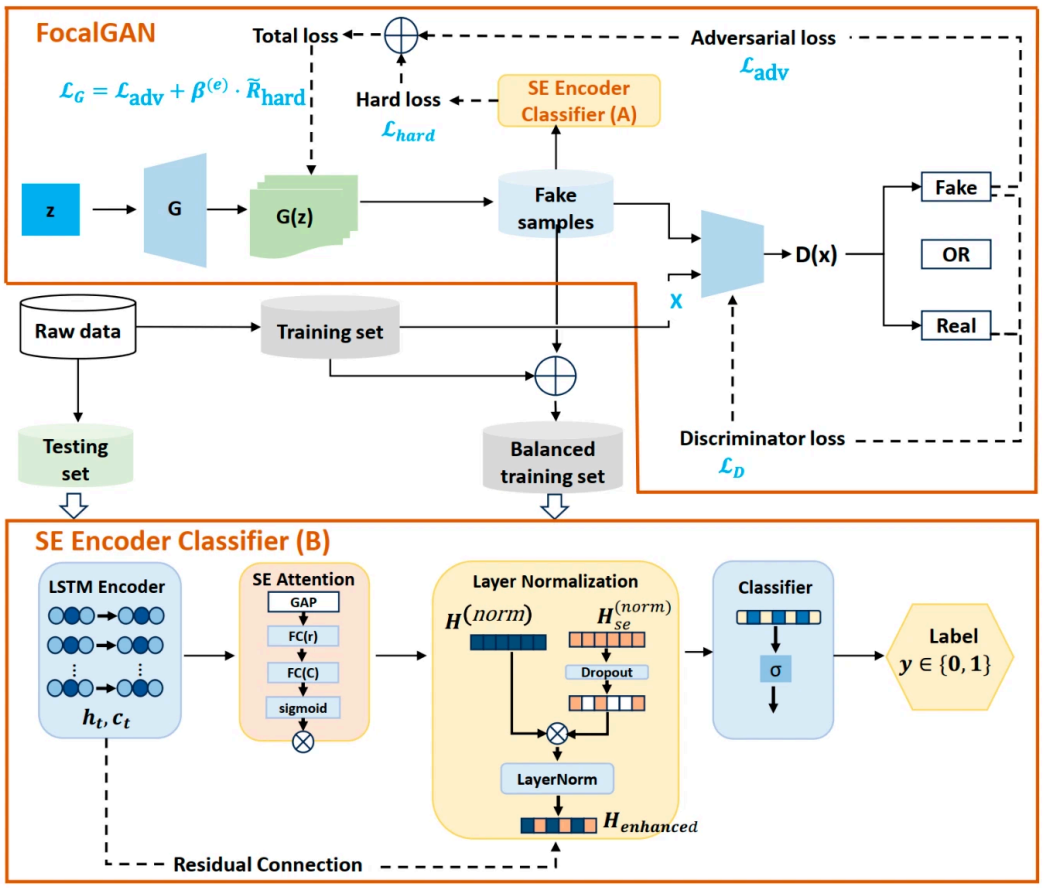


Figure 1. The proposed structure of FocalGAN-SEE framework.

3.1. Problem Definition

This study formulates online learning early warning as a binary temporal classification task based on students' learning behavioral data. Let the learning behavior time-series dataset be denoted as $\mathcal{D}=\left\{\left(X_i,y_i\right)\right\}_{i=1}^{N}$, where each input sample $X_i \in \mathbb{R}^{T\times D}$ represents the time-series feature matrix of a student in the course module of teaching period, where T denoting the time step length and D denoting the feature dimension at each time step. The binary label $y_i \in \{0, 1\}$: $y_i = 1$ represents academically at-risk student (i.e.,

the minority class requiring early warning), and $y_i = 0$ represents academically normal student (the majority class). The dataset inherently suffers from class imbalance—the number of at-risk samples is smaller than that of normal samples, posing a significant challenge to the model's ability to recognize the minority class. The objective of this study is to learn a mapping function $f: X \to \hat{y}$ that enables the model to accurately predict students' end of teaching period academic risk status based on their early-stage temporal behavioral features.

Our objective is twofold: (1) generate high-quality synthetic samples for the minority class to alleviate imbalance, and (2) enhance the classifier's robustness by exposing it to hard samples near the decision boundary.

For evaluation purposes on real-world test samples, we define hard at-risk samples via predicted probability output from the final retrained classifier (trained on the augmented training set, distinct from the frozen pre-trained hardness-evaluator classifier used inside FocalGAN). Given predicted probability $\hat{y} \in [0, 1]$ for each test sample:

Samples whose predicted probability falls within [0.35, 0.65] are treated as boundary-ambiguous hard samples, while others are easy samples.

Note that the frozen pre-trained classifier $C(\cdot)$ inside FocalGAN is only used to provide gradient signals for generator training and will never be applied to test-set inference or hard-sample labeling. Hard-sample categorization for evaluation relies exclusively on the final classifier retrained on the augmented dataset. This strict separation avoids circular reasoning between hard-sample definition and performance evaluation.

3.2. SE Encoder Classification Module

The SE-enhanced encoder-classifier serves two critical roles in our framework. First, it is pre-trained on the original imbalanced dataset to function as a hardness evaluator for the FocalGAN, providing stable gradient signals that guide the generator toward producing boundary-ambiguous samples. Second, after data augmentation via FocalGAN, it is retrained on the balanced dataset as the final classifier to predict at-risk probabilities for test samples.

To integrate temporal feature extraction with adaptive channel-wise recalibration, we construct the SE Encoder classification module by combining an LSTM encoder with a Squeeze-and-Excitation attention mechanism (SE attention). The core design rationale is threefold: (i) the LSTM encoder captures temporal dependencies in sequential learning behaviors through its gated recurrent structure; (ii) the SE attention mechanism adaptively recalibrates the importance of temporal feature channels, enhancing discriminative representations while suppressing irrelevant ones; and (iii) a classification head directly outputs academic risk probabilities via a sigmoid-activated dense layer. This design unifies feature extraction, temporal channel enhancement, and risk discrimination within a single end-to-end trainable framework.

3.2.1. LSTM-Encoder

The encoder maps the input sequence $\mathbb{R}^{T \times D}$ to a fixed-length representation vector $\mathbb{R}^{L}$, composed of three sequential transformations.

The encoder maps the raw input sequence $\boldsymbol{X} \in \mathbb{R}^{T \times D}$ to a sequential hidden state $\boldsymbol{H} \in \mathbb{R}^{T \times h}$, where h is the fixed LSTM hidden dimension. We denoted the sequential transformations as:

$$\boldsymbol{H} = \Phi(\mathbf{X}) \tag{1}$$

where Φ denotes a sequential transformation parameterized by learnable weight matrices. The raw sequential hidden states are normalized via layer normalization Γ(·) with ϵ for numerical stability:

$$\boldsymbol{H}^{(\text{norm})} = \Gamma(\boldsymbol{H}) = \gamma \odot \frac{\boldsymbol{H} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{2}$$

where $\boldsymbol{\mu} \in \mathbb{R}^h$ and $\boldsymbol{\sigma}^2 \in \mathbb{R}^h$ are the mean and variance computed across the temporal dimension for each feature channel; $\boldsymbol{\gamma} \in \mathbb{R}^h$ and $\boldsymbol{\beta} \in \mathbb{R}^h$ are learnable scaling and shifting vectors.

3.2.2. Time Series SE Attention Layer

The SE attention layer consists of three core operations performed sequentially: squeeze, excitation, and feature reweighting and residual fusion operations, with fixed reduction ratio r = 4 in implementation.

1. Squeeze operation: it aggregates temporal information via global average pooling to produce a channel descriptor:

$$\boldsymbol{s} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{h}_t^{(norm)} \tag{3}$$

where $\boldsymbol{h}_t^{(norm)} \in \mathbb{R}^h$ denotes the normalized hidden state at time step t.

2. Excitation operation: it captures channel-wise dependencies through a gating mechanism to generate per-channel modulation weights:

$$\boldsymbol{\mu} = \delta(\boldsymbol{W}_1\mathbf{s} + \boldsymbol{b}_1) \tag{4}$$

$$\boldsymbol{\alpha} = \sigma(\boldsymbol{W}_2\mathbf{s} + \boldsymbol{b}_2) \tag{5}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{(h/r)\times h}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{(h/r)\times h}$ are projection matrices; r is the reduction ratio controlling the bottleneck dimension; $\delta(\cdot)$ denotes the ReLU activation function; and $\sigma(\cdot)$ is the sigmoid function.

3. Feature Reweighting and Residual Fusion: attention weights $\boldsymbol{\alpha}$ are broadcast along the temporal dimension to scale normalized hidden states element-wise:

$$\boldsymbol{H}_{se}^{(norm)} = \boldsymbol{H}^{(\text{norm})} \odot sexpand_dims(\boldsymbol{\alpha}, axis = 1) \tag{6}$$

The final enhanced representation $\boldsymbol{H}_{enhanced}$ is obtained via a residual connection that combines the normalized hidden states with the channel-wise scaled features, followed by layer normalization:

$$\boldsymbol{H}_{enhanced} = \Gamma\left(\boldsymbol{H}^{(\text{norm})} + \boldsymbol{H}_{se}^{(norm)}\right) \tag{7}$$

where $\odot$ denotes element-wise multiplication; and Γ(·) is the layer normalization function. The residual connection preserves the original temporal information while allowing the SE module to adaptively amplify or suppress specific feature channels.

4. Sequence Representation Aggregation: the final encoding $\boldsymbol{e} \in \mathbb{R}^h$ for classification is extracted from the last time step of the enhanced sequence:

$$\boldsymbol{e} = \boldsymbol{H}_{enhanced}[T,:] \tag{8}$$

3.2.3. Classification Head

The binary prediction layer maps embedding $\boldsymbol{e}$ to student at-risk probability:

$$\hat{y} = f(\boldsymbol{e}) = \sigma\left(\boldsymbol{w}_c^{\top}\boldsymbol{e} + b_c\right) \tag{9}$$

where $\boldsymbol{w}_c \in \mathbb{R}^h$ is the learnable weight vector of the classification layer, $b_c \in \mathbb{R}$ is the learnable bias term, and $\sigma(\cdot)$ is the sigmoid function. The output $\hat{y} \in [0,1]$ represents the predicted probability that the student is at-risk.

3.2.4. Pre-Training Objective

The encoder-classifier is pre-trained on the original training dataset by minimizing the binary cross-entropy loss:

$$\mathcal{L}_{\text{pre}}(\Theta_C) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log\left(1-\hat{y}_i\right)\right] \tag{10}$$

where N is the total number of training samples; $y_i \in \{0, 1\}$ is the true label for the i -th sample; $\hat{y}_i \in [0, 1]$ is the corresponding predicted probability; and Θ_C denotes all trainable parameters of the encoder and classification head. After pre-training, Θ_C is frozen to provide stable hardness signals during GAN training.

3.3. The Architecture of FocalGAN

FocalGAN is constructed to synthesize hard boundary minority student behavior sequences, composed of a generator with Temporal Attention Block (generator G), BiLSTM discriminator D, and frozen pre-trained classifier C. The latent noise dimension is set to z_{dim}.

3.3.1. Generator with Temporal Attention Block

The generator $G: \mathbb{R}^{z_{dim}} \to \mathbb{R}^{T \times D}$ maps latent noise $\boldsymbol{z} \sim \mathcal{N}\left(0, \boldsymbol{I}_{z_{dim}}\right)$ to a synthetic time series. Its transformation can be expressed as:

$$G(z) = \mathcal{F}_{\text{BiLSTM}}(\mathcal{A}(\mathcal{F}_{\text{conv}}(\mathcal{F}_{\text{dense}}(z)))) \tag{11}$$

where $\mathcal{F}_{\text{dense}}$ denotes two fully connected layers with ReLU activation; $\mathcal{F}_{\text{conv}}$ denotes a 1D convolutional transformation with ReLU activation, kernel size = 2, same padding; and $\mathcal{F}_{\text{BiLSTM}}$ denotes a bidirectional LSTM transformation with 64 hidden units returning full sequential outputs; $\mathcal{A}(\cdot)$ denotes temporal attention block mapping function, which takes convolution feature sequence as input, calculates adaptive time-step importance weights via self-attention and residual connection, and outputs attention-enhanced temporal sequence;

After extracting local temporal patterns via 1D convolution, we adopt a lightweight temporal attention block to identify informative timestamps from student behavioral sequences. Given input sequence $\boldsymbol{X} \in \mathbb{R}^{T \times D}$, we first calculate unnormalized importance scores for each time step:

$$\boldsymbol{score} = \tanh(\boldsymbol{X}\boldsymbol{W}_a + \boldsymbol{b}_a) \tag{12}$$

where $\boldsymbol{W}_a \in \mathbb{R}^{D\times 1}$ is a learnable projection matrix that compresses per-timestep D-dimensional features into a single scalar importance value, and $\boldsymbol{b}_a \in \mathbb{R}$ is the corresponding bias term. The tanh activation restricts scores to the range [−1, 1] to stabilize attention distribution. Then softmax activation normalizes the scores across all time steps to obtain valid attention weights summing to one:

$$\boldsymbol{A} = \text{softmax}(score, axis = 1) \tag{13}$$

where $\boldsymbol{A} \in \mathbb{R}^{T\times 1}$. The original sequence is reweighted element-wise by attention weights to amplify informative behavioral timestamps:

$$\widetilde{\boldsymbol{X}} = \boldsymbol{X} \odot \boldsymbol{A} \tag{14}$$

$\widetilde{\boldsymbol{X}}$ denotes attention-weighted sequence, fused with raw input via residual connection.

Finally, a residual connection is adopted to avoid information loss caused by pure attention weighting:

$$\mathcal{T}(\boldsymbol{X}) = \widetilde{\boldsymbol{X}} + \boldsymbol{X} \tag{15}$$

where $\mathcal{T}(\boldsymbol{X})$ denotes the output of the temporal attention block, which is then fed into the subsequent BiLSTM layer.

3.3.2. BiLSTM Discriminator

The discriminator D: $\mathbb{R}^{T\times D} \rightarrow [0, 1]$ distinguishes real student behavior sequences from generator-synthesized samples, it is defined as:

$$D(\boldsymbol{X}) = \sigma\left(\boldsymbol{w}_d^{\top} \cdot \text{Pool}\left(\boldsymbol{H}_d\right) + b_d\right) \tag{16}$$

where $\boldsymbol{H}_d = \Phi_d(\boldsymbol{X}) \in \mathbb{R}^{T\times h_d}$ denotes the hidden state sequence produced by a bidirectional LSTM with h_d hidden units; a dropout layer with rate 0.3 is inserted after BiLSTM for regularization. $\text{Pool}\left(\boldsymbol{H}_d\right) = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{h}_{d,t}$ denotes global average pooling; $\boldsymbol{w}_d \in \mathbb{R}^{h_d}$ and $b_d \in \mathbb{R}$ are output layer weights.

3.3.3. Adversarial Loss

The discriminator loss optimizes binary classification between real minority samples and synthetic sequences; the adversarial loss of the discriminator is:

$$\mathcal{L}_D = \mathbb{E}_{\boldsymbol{X}\sim p_{\text{real}}}\left[-\log D(\boldsymbol{X})\right] + \mathbb{E}_{z\sim p_z}\left[-\log\left(1 - D\left(G(\boldsymbol{z})\right)\right)\right] \tag{17}$$

The generator adversarial loss aims to fool discriminator by generating realistic sequences:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\boldsymbol{z}\sim p_z}\left[-\log D\left(G(\boldsymbol{z})\right)\right] \tag{18}$$

where p_{real} denotes the data distribution of real samples; $p_z = \mathcal{N}(0, \boldsymbol{I}_{z_{dim}})$ denotes the prior distribution of the latent noise, and $\mathbb{E}[\cdot]$ denotes the expectation operator.

3.3.4. Focal Hard-Sample Loss with EMA Normalization

We define the hardness measure for a generated sample as its proximity to the classifier's decision boundary. Let $C(\cdot) = f(\mathcal{E}(\cdot))$ denote the pre-trained classifier with frozen parameters, where $\mathcal{E}$ is the encoder defined in Section 3.3.1 and f is the classification head defined in Section 3.3.2. The raw data hard-sample loss is defined as:

$$\mathcal{L}_{hard}^{(raw)} = \mathbb{E}_{z \sim p_z}\left[\ell_{\text{CE}}\left(0.5, C\left(G(z)\right)\right)\right] \tag{19}$$

where $\ell_{\text{CE}}(a,b) = -[a \log b + (1-a)\log(1-b)]$ is the binary cross-entropy function; 0.5 is the target value representing the decision boundary; and $C(G(z)) \in [0, 1]$ is the classifier's prediction for the generated sample. Minimizing $\mathcal{L}_{hard}^{(raw)}$ pushes generator outputs toward the ambiguous decision boundary to produce hard samples. To resolve magnitude mismatch between $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{hard}^{(raw)}$, we normalize raw hard loss via exponential moving average (EMA):

$$\widetilde{\mathcal{L}}_{hard} = \frac{\mathcal{L}_{hard}^{(raw)}}{\mu_{hard} + \epsilon} \tag{20}$$

ϵ is a constant value. EMA buffers for two loss terms are updated each training epoch with momentum ρ.

$$\mu_{\text{adv}}^{(e)} = \rho \cdot \mu_{\text{hard}}^{(e-1)} + (1-\rho) \cdot \mathcal{L}_{hard}^{(raw,e)} \tag{21}$$

where $\rho \in [0, 1]$ is the momentum coefficient; e denotes the current training epoch; $\mu_{\text{hard}}^{(e)}$ is the EMA value at epoch e.

Initial EMA values: $\mu_{\text{adv}}^{(0)} = 0$, $\mu_{\text{hard}}^{(0)} = 1$. The total generator loss combines adversarial loss and weighted normalized hard loss:

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \beta^{(e)} \cdot \widetilde{R}_{\text{hard}} \tag{22}$$

The composite generator loss $\mathcal{L}_G$ consists of two additive terms: the adversarial loss $\mathcal{L}_{\text{adv}}$ and the weighted hard-sample reward term $\beta^{(e)} \cdot \widetilde{R}_{\text{hard}}$. Here, $\widetilde{R}_{\text{hard}} = -\widetilde{\mathcal{L}}_{\text{hard}}$, where $\widetilde{\mathcal{L}}_{\text{hard}}$ represents the EMA-normalized hard loss. The optimization objective is to minimize $\mathcal{L}_G$. Minimizing $\mathcal{L}_G$ requires minimizing $\mathcal{L}_{\text{adv}}$ for realistic sample synthesis while maximizing the reward term $\widetilde{R}_{\text{hard}}$. Since $\widetilde{R}_{\text{hard}}$ increases when generated samples are close to the classifier decision boundary $C(G(z)) \approx 0.5$, this objective guides the

generator to produce boundary-ambiguous hard at-risk student samples. $\beta^{(e)}$ is a positive adaptive weighting coefficient that balances the magnitude of the adversarial loss and the hard-sample reward.

3.3.5. Adaptive Balancing Coefficient Scheduling

The time-variant weight $\beta^{(e)}$ consists of cosine annealing base term and magnitude scaling factor to dynamically balance two loss components:

$$\beta^{(e)} = \beta_{\text{base}}^{(e)} \cdot clip\left(\frac{\mu_{\text{adv}}^{(e-1)}}{\mu_{\text{hard}}^{(e-1)} + \epsilon}, \beta_{\min}, \beta_{\max}\right) \tag{23}$$

where ϵ is a small constant preventing division by zero; $\beta_{\min} = 0.1$ and $\beta_{\max} = 10.0$ are clipping thresholds; μ_{adv} is the EMA of $\mathcal{L}_{\text{adv}}$; and μ_{hard} is the EMA of $\mathcal{L}_{\text{hard}}$. The base coefficient follows cosine annealing over total GAN training epochs $E_{\text{gan}} = 300$.

$$\beta_{\text{base}}^{(e)} = \beta_{\text{init}} + (\beta_{\text{final}} - \beta_{\text{init}}) \cdot \frac{1 + \cos(\pi \cdot e / E_{\text{gan}})}{2} \tag{24}$$

with $\beta_{\text{init}} = 0.05$ and $\beta_{\text{final}} = 0.5$ in our implementation; e is the current epoch; and E_{gan} is the total number of GAN training epochs.

3.4. End-to-End Adversarial Training Procedure

The entire training process follows Algorithm 1, with gradient computation performed using automatic differentiation. The generator parameters Θ_G and discriminator parameters Θ_D are updated via the Adam optimizer with shared hyperparameters β_1, learning rate $\eta_G = \eta_D = 0.0002$. Standardized parameter update formulation:

$$\Theta_G \leftarrow \Theta_G - \eta_G \cdot \text{Adam}\left(\nabla_{\Theta_G}, \mathcal{L}_G \Theta_G\right) \tag{25}$$

$$\Theta_D \leftarrow \Theta_D - \eta_D \cdot \text{Adam}\left(\nabla_{\Theta_D}, \mathcal{L}_D \Theta_D\right) \tag{26}$$

where η_G and η_D are the learning rates for the generator and discriminator, respectively; and $\nabla_\Theta \mathcal{L}$ denotes the gradient of the loss with respect to the parameters. To prevent gradient explosion, we apply gradient clipping:

$$\tilde{g} = clip(g, -c, c) \tag{27}$$

where g is the raw gradient; $\tilde{g}$ is the clipped gradient; and c is the clipping threshold.

Algorithm 1 FocalGAN-SEE Training

Input: Training set $\mathcal{D}_{\text{train}}$, pre-training epochs E_{pre}, GAN epochs E_{gan}, warm-up epochs E_{warm}

Step 1. Pre-train encoder classifier C by minimizing $\mathcal{L}_{\text{pre}}$ on $\mathcal{D}_{\text{train}}$ for E_{pre} epochs; freeze all weights Θ_C

Step 2. Initialize generator G, discriminator D, EMA buffers $\mu_{\text{adv}} \leftarrow 0$, $\mu_{\text{hard}} \leftarrow 1$

Step 3. For each training epoch e from 1 to E_{gan}:

a. Randomly sample a mini-batch of real minority-class sequences $\{X_{\text{real}}\}$

b. Sample latent Gaussian noise $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{z_{dim}})$

c. Generate synthetic student sequences $\boldsymbol{X}_{\text{fake}} = G(z)$

d. Compute discriminator loss $\mathcal{L}_D$ and generator adversarial loss $\mathcal{L}_{\text{adv}}$

e. If $e > E_{\text{warm}}$:

- Calculate raw hard-sample loss $\mathcal{L}_{hard}^{(raw)} = \mathbb{E}\left[\ell_{\text{CE}}\left(0.5, C\left(\boldsymbol{X}_{\text{fake}}\right)\right)\right]$
- Update EMA moving averages for $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{hard}^{(raw)}$
- Compute adaptive balancing weight $\beta^{(e)}$ via cosine annealing and magnitude scaling
- Calculate normalized hard loss $\widetilde{\mathcal{L}}_{hard}$
- Total generator loss: $\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \beta^{(e)} \cdot \widetilde{R}_{\text{hard}}$

f. Else: Set generator loss as pure adversarial loss $\mathcal{L}_G = \mathcal{L}_{\text{adv}}$

g. Compute gradients $\nabla_{\Theta_G} \mathcal{L}_G$ and $\nabla_{\Theta_D} \mathcal{L}_D$ via automatic differentiation

h. Apply gradient clipping on generator gradients and update Θ_G, Θ_D using Adam optimizer

Output: trained generator G for minority hard sample augmentation; frozen classifier C for hardness evaluation

3.5. Baseline and Evaluation Metrics

Two categories of baseline models are adopted in this paper to validate the effectiveness of the SE Encoder classification module and the overall FocalGAN-SEE architecture, respectively.

The first category consists of commonly used attention-augmented autoencoder models. The attention mechanisms in the early warning domain were used, including multi-head attention, self-attention [13], Bahdanau attention [39], weighted sum attention [58], local attention [59], cross-attention [62], and channel attention [60], which are combined with standard LSTM autoencoders and binary output to benchmark the classification performance of the SE Encoder.

The second category comprises commonly used imbalanced-data processing models in the early warning domain, including deep generative models—Generative Adversarial Networks (GANs) [61,63], Conditional GAN (CGAN) [64], Deep Convolutional GAN (DCGAN) [65], Least Squares GAN (LSGAN) [66] and WGAN [67]—as well as variational autoencoders (VAEs) [68], and oversampling-based models such as SMOTE [69–71]—to evaluate FocalGAN-SEE across multiple early warning metrics. All deep generative and oversampling models are connected to the same LSTM classifier

after learning at-risk student features and generating balanced training sets.

We employ stratified 5-fold cross-validation at the student level, preserving the original class distribution in each fold to ensure statistical robustness. The procedure for each fold is as follows:

(1) The minority class samples from the training fold are used to train FocalGAN, generating synthetic samples to balance the training set;

(2) A new SE Encoder-classifier C is trained from scratch on the augmented training set;

(3) The trained classifier is evaluated on the held-out test fold.

All performance metrics—including accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC)—are reported as the mean and standard deviation across the five folds:

Accuracy measures the model's overall ability to identify correctly both classes. Precision measures the proportion of samples predicted as at-risk students that are actually at-risk students. Recall measures the proportion of actual at-risk samples that are correctly identified by the model. F1-Score is the harmonic mean of precision and recall, providing a comprehensive measure of the model's effectiveness in identifying the minority class (at-risk students). AUC (Area Under the ROC Curve) measures the model's overall ranking ability across different thresholds and is relatively robust to class imbalance.

4. Results

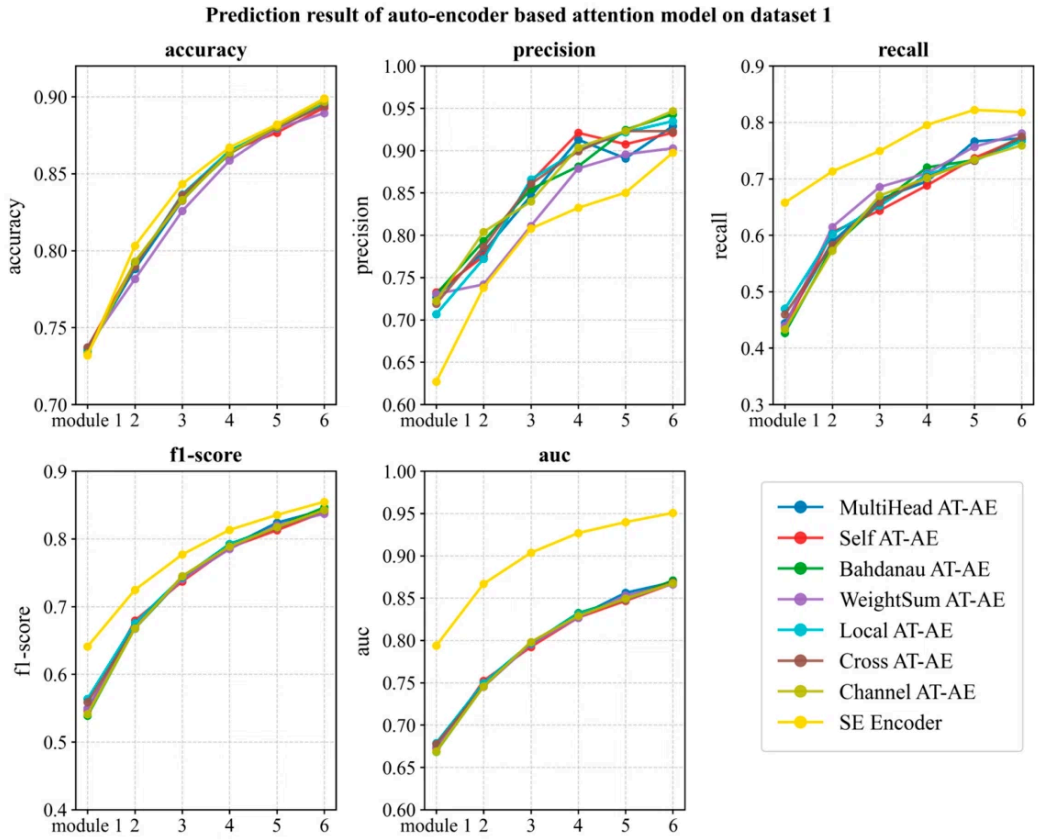
This paper first compares the early warning performance of the SE Encoder classifier with that of attention-augmented autoencoders on two real-world student datasets, to validate its superiority on the classification task. Subsequently, the early warning metrics of the proposed FocalGAN-SEE are evaluated on the two datasets. Finally, data visualization and the SHAP method are employed to analyze the characteristics of hard samples identified by the models, and interpret decision-making rationale in hard-sample prediction.

4.1. Datasets

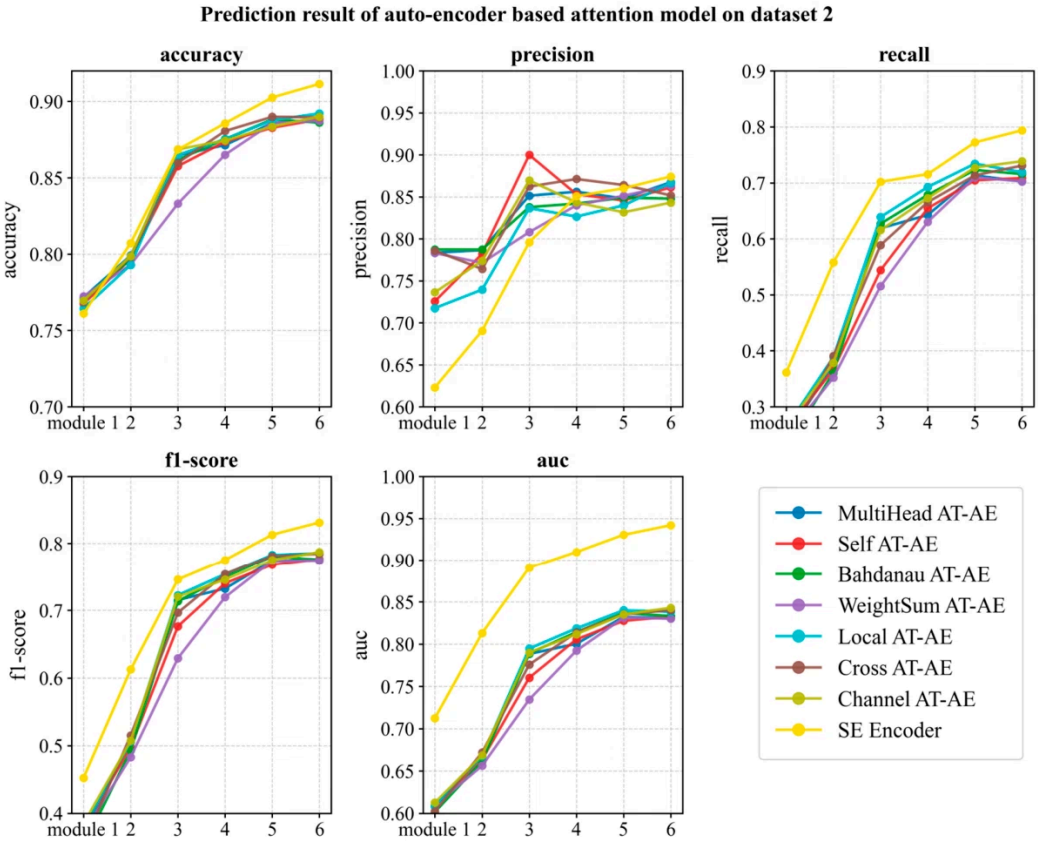
The students' learning behavior logs and academic scores used in this study were collected from a fully online course. After data preprocessing, we got two Datasets: Dataset 1 and Dataset 2, corresponding to the Fall and Spring teaching periods respectively. Each dataset includes data from the same 46 courses across course modules 1–11. Students' logs in each course module were coded into eight variables related to homework submission, testing, resource viewing, forum discussions, and time intervals of login and learning. Students' scores were dichotomized based on a threshold of 60 points: students' scores ≥ 60 were labeled as "normal" (label 0), and scores < 60 were labeled as "at-risk" (label 1). Finally, Dataset 1 contains 1,120,189 log entries from 13,526 students, with 8606 students (63.63%) classified as the success category, and 4920 (36.37%) as the at-risk category. Dataset 2 includes 810,031 logs from 9800 students, with 7103 (72.48%) in the success category and 2697 (27.52%) in the at-risk category.

4.2. Performance Evaluation of the SE Encoder Module

Figure 2 shows the prediction results of the proposed SE Encoder module and common attention-augmented autoencoder (AE-attention) models on Dataset 1 and Dataset 2. In order to leave sufficient time for teachers to conduct instructional interventions, early warning is typically carried out at the midpoint (1/2) of the teaching period [3]. This study only presents the prediction results of Course Module 6 and earlier modules.



(**a**) Results on Dataset 1



(**b**) Results on Dataset 2

Figure 2. The results of the SE Encoder and common attention-AE models on Dataset 1 and 2.

The results show that, as the course modules progress, the prediction metrics of all models on the two Datasets exhibit an increasing trend, indicating that a greater volume of log data brings higher predictive performance. The proposed SE Encoder module outperforms other attention-augmented autoencoder models in accuracy, recall, F1-score, and AUC, while achieving comparable precision. As the course progresses, the performance of SE Encoder improves consistently. On Dataset 1, SE Encoder achieves a recall of 76.75% by Course Module 3, meaning it can identify more than 3/4 of at-risk students, while other models achieve comparable levels of identification performance only by Course Module 5 or 6. On dataset 2, SE Encoder achieves a recall of 75.49% by Course Module 4 (3/4 of at-risk students), while other models achieve comparable levels of identification performance only after Course Module 6.

Therefore, SE Encoder achieves the highest metrics and provides instructors with an earlier intervention window, facilitating more timely efforts to transform at-risk students into successful ones.

4.3. Comparison of FocalGAN-SEE and Baseline for At-Risk Student Identification

We compare the performance of the proposed FocalGAN-SEE with commonly used deep generative models and oversampling models in current early warning research.

4.3.1. Early Warning Results on Dataset 1 and Dataset 2

The early warning results of the models on Dataset 1 and Dataset 2 are presented in Tables 1 and 2 respectively. All the data balancing modules (deep generative models and oversampling models) in the table are connected to the same LSTM predictor to perform the prediction task. For ease of presentation, we only list the names of the data balancing modules in the table.

Table 1. The early warning results of models on Dataset 1.

	Course Module 1					Course Module 2				
Method	**acc**	**pre**	**rec**	**f1**	**auc**	**acc**	**pre**	**rec**	**f1**	**auc**
GAN	0.6807	0.5505	0.6886	0.6109	0.7585	0.7615	0.6655	0.6976	0.6802	0.8299
VAE	0.7103	0.5975	0.6258	0.6111	0.7629	0.7491	0.6462	0.7030	0.6709	0.8207
BGAN	0.6438	0.5214	0.7620	0.6117	0.7597	0.7747	0.6743	0.7404	0.7051	0.8496
CGAN	0.6706	0.5385	0.7205	0.6144	0.7545	0.7433	0.6277	0.7274	0.6730	0.8219
DCGAN	0.6859	0.5594	0.6728	0.6085	0.7482	0.7667	0.6602	0.7449	0.6989	0.8490
LSGAN	0.6353	0.5012	**0.7659**	0.6049	0.7543	0.7517	0.6415	0.7238	0.6795	0.8277
WGAN	0.6924	0.5676	0.6669	0.6120	0.7580	0.7396	0.6222	0.7309	0.6715	0.8216
ADASYN	0.7375	0.6926	0.5026	0.5813	0.7723	0.7807	0.7115	0.6730	0.6909	0.8407
BSMOTE	0.7340	**0.7058**	0.4638	0.5587	0.7696	0.7790	0.7052	0.6762	0.6897	0.8414
SMOTE	0.7335	0.6600	0.5575	0.6035	0.7738	0.7791	0.6986	0.6917	0.6949	0.8414
Ours	**0.7484**	0.6523	0.6667	**0.6586**	**0.807** 0	**0.8005**	**0.7164**	**0.7492**	**0.7320**	**0.8724**
	Course module 3					**Course module 4**				
Method	**acc**	**pre**	**rec**	**f1**	**auc**	**acc**	**pre**	**rec**	**f1**	**auc**
GAN	0.7928	0.7066	0.7380	0.7215	0.8642	0.8321	0.7804	0.7498	0.7646	0.8964
VAE	0.7822	0.6823	0.7518	0.7152	0.8575	0.8234	0.7689	0.7368	0.7522	0.8868
BGAN	0.7667	0.6554	0.7734	0.7076	0.8579	0.8219	0.7457	0.7793	0.7613	0.8963
CGAN	0.8002	0.7217	0.7343	0.7278	0.8702	0.8330	0.7749	0.7657	0.7695	0.8996
LSGAN	0.8039	0.7265	0.7400	0.7331	0.8707	0.8358	0.7832	0.7618	0.7717	0.9012
WGAN	0.7865	0.6913	0.7492	0.7184	0.8596	0.8241	0.7580	0.7596	0.7587	0.8914

ADASYN	0.8182	0.7668	0.7220	0.7429	0.8794	0.8421	0.7953	0.7642	0.7790	0.9059
BSMOTE	0.8093	0.7358	0.7443	0.7395	0.8789	0.8374	0.7736	0.7829	0.7780	0.9082
SMOTE	0.8207	0.7675	0.7278	0.7471	0.8817	0.8444	0.8063	0.7549	0.7793	0.9068
Ours	**0.8437**	**0.7802**	**0.7951**	**0.7873**	**0.9096**	**0.8722**	**0.8281**	**0.8195**	**0.8235**	**0.9329**
	Course module 5					**Course module 6**				
Method	acc	pre	rec	f1	auc	acc	pre	rec	f1	auc
GAN	0.8498	0.8007	0.7844	0.720	0.9122	0.8640	0.8291	0.7902	0.8088	0.9255
VAE	0.8496	0.8194	0.7533	0.7845	0.9076	0.8603	0.8285	0.7789	0.8022	0.9211
BGAN	0.8497	0.7975	0.7894	0.7926	0.9157	0.8479	0.7735	0.8238	0.7976	0.9245
CGAN	0.8561	0.8121	0.7866	0.7990	0.9170	0.8666	0.8326	0.7945	0.8123	0.9285
DCGAN	0.8225	0.7337	0.8049	0.7674	0.9001	0.8430	0.7738	0.8128	0.7907	0.9166
LSGAN	0.8559	0.8220	0.7722	0.7960	0.9165	0.8677	0.8343	0.7980	0.8145	0.9293
WGAN	0.8505	0.8072	0.7740	0.7902	0.9106	0.8626	0.8297	0.7829	0.8055	0.9244
ADASYN	0.8625	0.8140	0.8079	0.8106	0.9243	0.8790	0.8522	0.8075	0.8291	0.9376
BSMOTE	0.8608	0.8128	0.8033	0.8076	0.9237	0.8714	0.8242	0.8232	0.8234	0.9342
SMOTE	0.8649	0.8275	0.7947	0.8106	0.9238	0.8800	0.8605	0.8010	0.8293	0.9370
Ours	**0.8887**	**0.8482**	**0.8470**	**0.8471**	**0.9464**	**0.9004**	**0.8684**	**0.8575**	**0.8624**	**0.9555**

Bold and underlined values denote the best and second-best results, respectively.

Table 2. The early warning results of models on Dataset 2.

	Course Module 1					**Course Module 2**				
Method	**acc**	**pre**	**rec**	**f1**	**auc**	**acc**	**pre**	**rec**	**f1**	**auc**
GAN	0.7303	0.5221	0.4635	0.4854	0.7032	0.7827	0.6464	0.4754	0.5449	0.7781
VAE	0.7319	0.5206	0.4594	0.4856	0.7030	0.7658	0.5879	0.4983	0.5389	0.7657
BGAN	0.6753	0.4374	**0.6203**	**0.5124**	0.7062	0.7388	0.5399	0.6277	0.5693	0.7872
CGAN	0.7082	0.4779	0.5257	0.4969	0.7056	0.7689	0.5994	0.5632	0.5712	0.7896
DCGAN	0.7286	0.5082	0.4891	0.4973	0.7129	0.7893	0.6455	0.5376	0.5824	0.7981
LSGAN	0.6771	0.4470	0.5591	0.4870	0.6897	0.7804	0.6125	0.5669	0.5860	0.7964
WGAN	0.7547	**0.5774**	0.4193	0.4834	0.7103	0.7895	0.6757	0.4757	0.5537	0.7968
ADASYN	0.7507	0.5752	0.4216	0.4809	0.7126	0.7715	0.5942	0.5476	0.5686	0.7869
BSMOTE	0.7380	0.5306	0.4616	0.4922	0.7132	0.7590	0.5619	0.5732	0.5667	0.7863
SMOTE	0.7459	0.5475	0.4442	0.4902	0.7125	0.7635	0.5688	0.5914	0.5789	0.7878
Ours	**0.7557**	0.5742	0.4479	0.5021	**0.7163**	**0.8234**	**0.6963**	**0.6370**	**0.6650**	**0.8417**
	Course module 3					**Course module 4**				
Method	**acc**	**pre**	**rec**	**f1**	**auc**	**acc**	**pre**	**rec**	**f1**	**auc**
GAN	0.7968	0.6583	0.5625	0.6053	0.8193	0.8442	0.7477	0.6552	0.6983	0.8674
VAE	0.7964	0.6385	0.6166	0.6243	0.8190	0.8155	0.6662	0.6655	0.6645	0.8501
BGAN	0.8383	0.7026	0.7212	0.7108	0.8746	0.8767	0.7988	0.7434	0.7685	0.9036
CGAN	0.8401	0.7348	0.6685	0.6984	0.8646	0.8683	0.8095	0.6956	0.7429	0.8942
DCGAN	0.8560	0.7778	0.6848	0.7242	0.8787	0.8669	0.7876	0.7245	0.7511	0.8960
LSGAN	0.8468	0.7398	0.6908	0.7134	0.8723	0.8800	0.8260	0.7219	0.7683	0.9030
WGAN	0.8473	0.7868	0.6118	0.6857	0.8673	0.8722	0.7887	0.7327	0.7594	0.9011
ADASYN	0.8036	0.6310	0.7011	0.6628	0.8494	0.8169	0.6442	0.7590	0.6955	0.8772
BSMOTE	0.8202	0.6735	0.6808	0.6754	0.8543	0.8333	0.6772	0.7560	0.7137	0.8832
SMOTE	0.8414	0.7330	0.6907	0.7066	0.8659	0.8600	0.7551	0.7378	0.7442	0.8905

Ours	**0.8793**	**0.8190**	**0.7227**	**0.7671**	**0.9034**	**0.9029**	**0.8609**	**0.7716**	**0.8137**	**0.9274**
	Course module 5					**Course module 6**				
Method	**acc**	**pre**	**rec**	**f1**	**auc**	**acc**	**pre**	**rec**	**f1**	**auc**
GAN	0.8721	0.8120	0.6974	0.7497	0.8963	0.8851	0.8533	0.7041	0.7710	0.9104
VAE	0.8481	0.7338	0.7100	0.7199	0.8793	0.8606	0.7434	0.7557	0.7491	0.8974
BGAN	0.8803	0.8016	0.7679	0.7801	0.9154	0.8783	0.7679	0.8009	0.7837	0.9223
CGAN	0.8906	0.8474	0.7445	0.7895	0.9180	0.8919	0.8361	0.7642	0.7961	0.9271
DCGAN	0.8858	0.8224	0.7623	0.7868	0.9172	0.8740	0.7713	0.7853	0.7749	0.9187
LSGAN	0.8828	0.8216	0.7449	0.7781	0.9129	0.8851	0.8208	0.7594	0.7851	0.9254
WGAN	0.8861	0.8302	0.7490	0.7845	0.9134	0.8909	0.8234	0.7731	0.7958	0.9263
ADASYN	0.8598	0.7253	0.7905	0.7563	0.9044	0.8605	0.719	0.8154	0.7631	0.9170
BSMOTE	0.8664	0.7404	0.7938	0.7659	0.9086	0.8720	0.7505	0.8083	0.7768	0.9188
SMOTE	0.8809	0.7929	0.7753	0.7820	0.9114	0.8806	0.7855	0.7838	0.7835	0.9161
Ours	**0.9170**	**0.8870**	**0.8016**	**0.8417**	**0.9423**	**0.9187**	**0.8621**	**0.8398**	**0.8505**	**0.9514**

Bold and underlined values denote the best and second-best results, respectively.

The results show that:

(1) FocalGAN-SEE (Ours in the tables) achieves superior predictive performance over all baseline models on both datasets. Specifically, at Course Module 1, FocalGAN-SEE already outperforms the baselines in terms of the comprehensive metrics accuracy and AUC; from Course Module 2 onward, FocalGAN-SEE achieves the highest accuracy, precision, recall, F1-score, and AUC among all models, demonstrating that the model identifies more at-risk students while maintaining robustness.

(2) On Dataset 1, FocalGAN-SEE is able to identify nearly 80% of at-risk students at Course Module 3, whereas the baseline models do not reach the 80% identification rate until Course Module 5 or even 6—meaning that FocalGAN-SEE achieves high-performance early warning predictions two modules earlier. On Dataset 2, FocalGAN-SEE identifies over 80% of at-risk students at Course Module 5, outperforming the baseline models by one module.

(3) Compared with SE Encoder alone, FocalGAN-SEE significantly improved early warning metrics, which indicates the contribution of the FocalGAN module.

(4) Given that the model achieves a recall of more than 0.8 on both datasets by Course Module 5 (indicating that 80% of at-risk students can be identified), Course Module 5 is selected as the early warning time point.

The computational costs of FocalGAN-SEE, GAN and its variants are compared for the early warning unit (Course Module 5) on Dataset 1. Table 3 lists the total training time and peak memory consumption of each model.

Table 3. Total training time and peak memory footprint of methods.

Method	**FocalGAN-SEE**	**GAN**	**BGAN**	**CGAN**	**DCGAN**	**LSGAN**	**WGAN**
Total training time (s)	157.6	60.2	113.67	41.0	44.1	41.7	167.3
Peak RSS(MB)	478.79	805.4	1323.97	813.80	1061.8	796.46	3076.18

The results show that our FocalGAN-SEE achieves the lowest peak memory usage among all compared GAN variants, at 478.79 MB. Its total training time is 157.62 s. Among all GAN variants compared baselines, CGAN and LSGAN have the shortest training time (~41 s), while WGAN requires the longest training time (167.30 s) and ex-

hibits extremely high memory consumption (3076.18 MB).

Although FocalGAN-SEE requires relatively longer total training time than most GAN variants, it achieves substantial memory reduction, which is valuable for deployment on resource-limited educational computing devices.

4.3.2. Cross-Teaching Period Generalization Performance Evaluation

The two datasets adopted in this study were collected from identical courses across two different teaching periods, with non-overlapping student populations between teaching periods. This characteristic enables a cross-teaching period generalization experiment: we train the model on the first-teaching period dataset (Dataset 1) and test its performance on data from the subsequent teaching period (Dataset 2), to evaluate the practical applicability of the model in real educational scenarios. Table 4 summarizes the metrics of baseline models and our proposed FocalGAN-SEE at Course Module 5.

Table 4. The results of cross-teaching period generalization experiment.

Method	acc	pre	rec	f1	auc
GAN	0.8613	0.8521	0.6003	0.7044	0.8867
VAE	0.8439	0.7751	0.6096	0.6824	0.8691
BGAN	0.8593	0.8554	0.5881	0.6970	0.8824
CGAN	0.8390	0.7575	0.6103	0.6760	0.8755
DCGAN	0.8436	**0.8870**	0.4946	0.6351	0.8726
LSGAN	0.8504	0.8742	0.5332	0.6624	0.8841
WGAN	0.8353	<u>0.8747</u>	0.4687	0.6103	0.8828
ADASYN	0.8416	0.7996	0.5666	0.6632	0.8811
BSMOTE	0.7871	0.7790	0.3163	0.4499	0.8772
SMOTE	0.7553	0.8637	0.1316	0.2284	0.8837
Ours	**0.8655**	0.7641	**0.7397**	**0.7517**	**0.8873**

Bold and underlined values denote the best and second-best results, respectively.

The results show that after replacing the test set with data from a new teaching period, the performance metrics of all models decline compared with those in Table 1 due to the increased divergence between student cohorts. Nevertheless, after being trained on data from the previous teaching period, our model can directly identify the largest number of at-risk students (73.97%) on the new-teaching period dataset, while achieving the highest accuracy and F1-score among all baseline models. Most GAN variants and oversampling methods suffer obvious performance degradation, especially in recall, because they cannot effectively learn discriminative features for borderline hard samples when facing cross-teaching period data distribution differences. Although some baseline models yield relatively high precision, their low recall means a large number of at-risk students are missed, which severely limits their practical value for early warning interventions.

4.3.3. Ablation Experiment

To evaluate the contribution of each core component in FocalGAN-SEE, we conducted ablation experiments by removing four key modules individually: (i) the hard-sample loss (w/o HardLoss), (ii) the adaptive beta scheduling (w/o AdaptiveBeta), (iii) the end-to-end gradient flow (w/o EndToEnd), and (iv) the SE Encoder (w/o SE Encoder). A baseline model without any of these four components (GANs with LSTM predictor in Tables 1 and 2) was also included for comparison.

As shown in Table 5, the full model equipped with all four modules achieves the best accuracy, precision, recall, F1-score, and AUC on both datasets. Compared with the baseline lacking all four components, the full model achieves F1 improvements of 17.65% and 12.27% on Dataset 1 and Dataset 2, respectively, which verifies that the combination of the proposed modules significantly boosts the identification performance of at-risk students under class imbalance.

Table 5. The result of ablation experiment.

Ablation Experiment on Data Set 1							
Order	**Component**	**Description**	**acc**	**pre**	**rec**	**F1**	**auc**
①	w/o HardLoss	Removing hard loss	0.8598	0.7997	0.8217	0.8101	0.9269
②	w/o AdaptiveBeta	Removing adaptive β	0.8560	0.785	0.8321	0.8078	0.9267
③	w/o EndToEnd	Removing end to end gradient propagation	0.8619	0.8077	0.8159	0.8113	0.9265
④	w/o SE Encoder	Removing SE Encoder	0.8601	0.7992	0.8234	0.8108	0.9281
⑤	Baseline	Removing ①②③④	0.8498	0.8007	0.7844	0.7200	0.9122
⑥	full	Removing none	**0.8887**	**0.8482**	**0.8470**	**0.8471**	**0.9464**
Ablation Experiment on Data set 2							
Order	**Module**	**Description**	**acc**	**pre**	**rec**	**F1**	**auc**
①	w/o HardLoss	Removing hard loss	0.8804	0.7886	0.7853	0.7842	0.9188
②	w/o AdaptiveBeta	Removing adaptive β	0.8831	0.8031	0.7764	0.7858	0.9193
③	w/o EndToEnd	Removing end to end gradient propagation	0.8786	0.7842	0.7894	0.7830	0.9162
④	w/o SE Encoder	Removing SE Encoder	0.8800	0.7881	0.7820	0.7824	0.9192
⑤	Baseline	Removing ①②③④	0.8721	0.8120	0.6974	0.7497	0.8963
⑥	full	Removing none	**0.9170**	**0.8870**	**0.8016**	**0.8417**	**0.9423**

Bold values denote the best results.

First, the hard-sample loss (HardLoss) serves as the most critical component. Removing HardLoss leads to a substantial drop in accuracy, precision, recall, F1-score, and AUC on both datasets. This result confirms our core motivation: the hard-sample loss guides the generator to learn ambiguous samples near the decision boundary and strengthens the model's ability to distinguish hard at-risk students.

Second, the adaptive β scheduler stabilizes multi-objective optimization. Without this adaptive balancing strategy, fixed weights fail to harmonize the magnitude of adversarial loss and hard-sample loss throughout training. The dynamic coefficient adjusted by EMA prevents one loss term from dominating the optimization process and retains the benefits of hard-sample regularization.

Third, end-to-end gradient propagation is an indispensable prerequisite for HardLoss. The variant w/o EndToEnd corresponds to an offline hard-sample evaluation scheme, where gradients cannot be back-propagated to the generator. The performance degradation demonstrates that offline hard sample screening is insufficient; embedding the SE Encoder into the computation graph and enabling end-to-end gradient flow is required to drive the generator to capture hard boundary patterns.

Fourth, the SE Encoder improves the discriminability of temporal representations. After removing the squeeze-and-excitation attention module, prediction performance decreases consistently. The SE mechanism adaptively recalibrates channel-wise temporal features and provides high-quality representations for risk classification and hard sample evaluation.

Moreover, simultaneous removal of all four modules yields the worst overall performance. The results reveal obvious complementarity among these modules: the SE Encoder extracts refined temporal features, end-to-end propagation builds the optimization pathway, adaptive β balances multiple loss terms, and HardLoss enhances the modeling of hard boundary samples. These components jointly alleviate the class-imbalance challenge in student early warning tasks.

It is worth noting that accuracy fluctuates mildly across all variants. This can be attributed to the imbalanced data distribution, where overall accuracy is dominated by majority-class ordinary students. Therefore, we prioritize F1-score and AUC for evaluation, as they provide more trustworthy measurements for minority at-risk student recognition.

To verify that the performance gains of FocalGAN-SEE do not stem merely from using the SE Encoder as the final classifier, we replaced the LSTM classifier with the SE Encoder for all baseline models. We observe that substituting the original LSTM classifier with the SE Encoder yields steady performance improvements across all baseline models. This confirms that the SE Encoder itself delivers more powerful feature representation and classification capabilities. Nevertheless, even adopting the same SE Encoder backbone, our FocalGAN-SEE still obtains better overall performance (F1-score and AUC) than all baseline models equipped with the SE Encoder. This finding demonstrates that the performance improvement of FocalGAN-SEE is not solely attributed to the SE Encoder. Instead, the performance benefits arise from the joint mechanism combining our specially designed focal hard-sample loss and the SE Encoder.

4.4. Comparison of At-Risk Student Characteristics Identified by Models

To further analyze the differences in the proposed architecture in identifying normal at-risk samples versus hard at-risk samples, this study analyzed the at-risk students identified by FocalGAN-SEE and baselines at Course Module 5 of Dataset 1.

Figure 3 shows that FocalGAN-SEE correctly identified the greatest number of hard samples (1968, 64.89% of hard at-risk samples) and normal samples (2890, 97.90% of normal at-risk samples), both exceeding those of the baseline models.

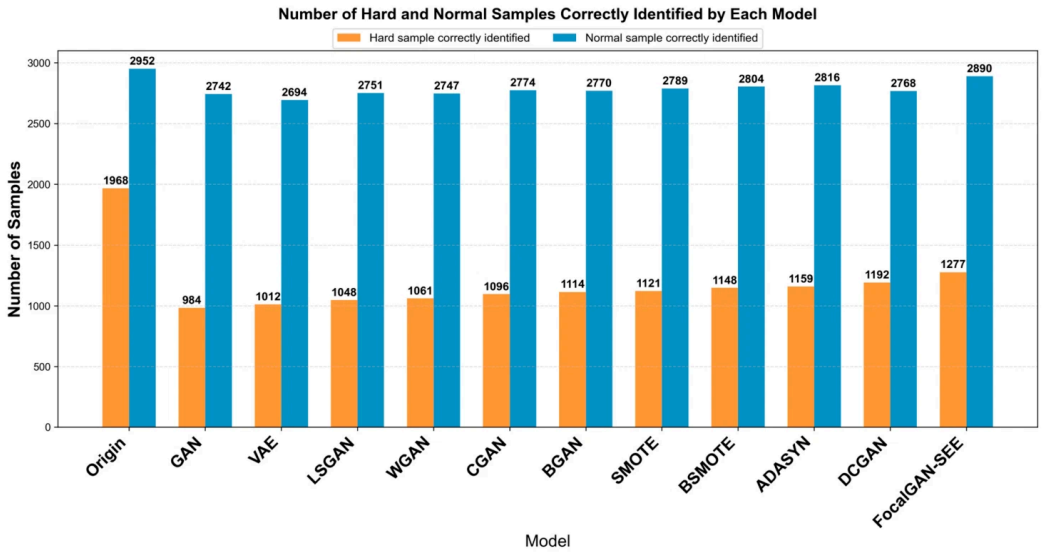


Figure 3. Number of hard and normal at-risk samples correctly identified by models.

We utilize the PCA algorithm to reduce the dimensionality of all at-risk student features and visualize their distributions by a scatter plot.

Figure 4 shows the distribution of hard at-risk samples and normal at-risk samples (subplot 1), and the hard samples correctly identified or misclassified, normal samples correctly identified or misclassified by FocalGAN-SEE and the baselines are marked in subplots 2–12. According to the distribution of these four classes of at-risk samples, the figure can be divided into three regions:

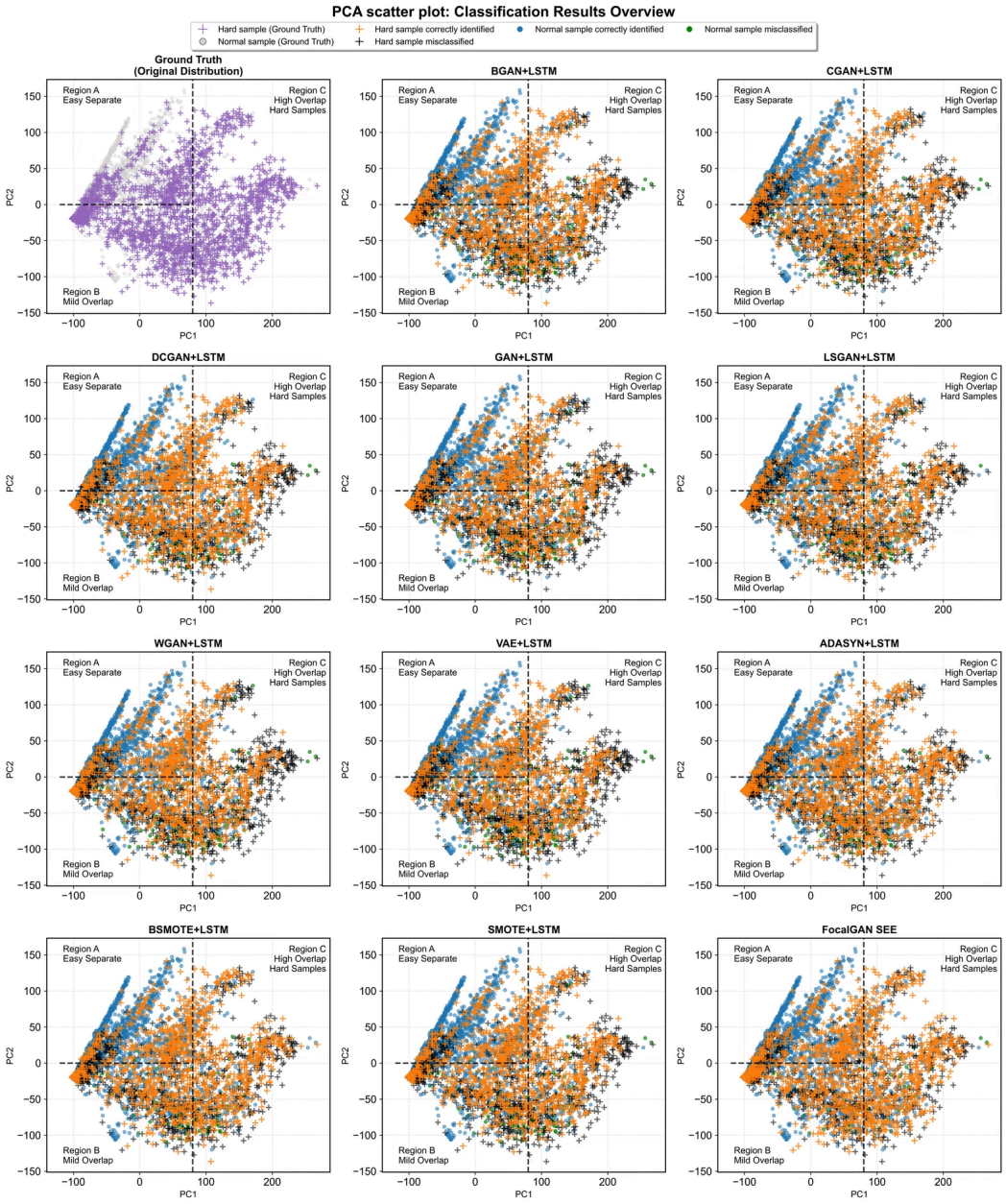


Figure 4. At-risk students identified by FocalGAN-SEE and baseline model (in course module 5).

Region A, which contains the largest number of normal at-risk samples and is relatively easy to distinguish;

Region B, where sample features exhibit moderate confusion, with a small degree of overlap between normal at-risk students and hard at-risk students, representing a moderate level of recognition difficulty;

Region C, a high-overlap area between hard samples and normal at-risk samples, where all hard samples are highly clustered and the four labeled classes are completely intermixed with blurred feature boundaries. This region serves as a critical area for differentiating model performance.

As shown in Figure 4, the performance of FocalGAN-SEE can be summarized as follows:

(1) It is capable of identifying the largest number of both hard at-risk samples and normal at-risk samples across a wide range of feature patterns. As observed in the figure, the correctly identified at-risk students (red crosses and blue dots) cover all three regions—A, B, and C—and are uniformly distributed throughout the entire feature space. This indicates that the focal loss successfully guides the model to focus on the hard region with the highest overlap in the feature space.

(2) The proposed model achieves the lowest total number of misclassified samples among all models, with the fewest misclassified hard at-risk samples and misclassified normal at-risk samples across all models.

(3) FocalGAN-SEE shows the best performance in distinguishing boundary samples: in Region C, where hard at-risk samples and normal at-risk students exhibit overlapping features and are most difficult to differentiate, the density of correctly identified hard at-risk samples produced by FocalGAN-SEE is substantially higher than that of other models, highlighting its outstanding discriminative ability for at-risk students near the confusion boundary.

(4) FocalGAN-SEE does not overly bias toward identifying normal at-risk samples; rather, it effectively balances the recognition of both hard at-risk samples and normal at-risk samples.

Although ADASYN + LSTM identifies a large number of at-risk students, the hard at-risk samples it identifies are mainly concentrated in Region A (easy to distinguish) and Region B (the transition zone), and the number of hard at-risk samples identified in Region C is far fewer than that of FocalGAN. This indicates that the samples generated by the oversampling strategy fail to cover the true hard distribution. While it maintains normal at-risk sample recognition, its improvement in hard at-risk sample recognition is limited. SMOTE + LSTM performs similarly to ADASYN, identifying far fewer hard at-risk samples in Region C than FocalGAN-SEE, suggesting that the new samples generated by linear interpolation cannot effectively cover the high-overlap region of hard at-risk samples. BSMOTE + LSTM performs the worst in hard at-risk sample recognition, with almost no hard at-risk samples identified in Region C. Its boundary sampling strategy appears to have a counterproductive effect—the newly generated samples instead bias the model toward predicting normal samples.

GAN + LSTM provides some benefit through generative augmentation, but excessive generation leads to blurred boundaries, resulting in a considerable number of misclassifications among normal at-risk samples. DCGAN + LSTM performs close to traditional resampling methods; it ranks second only to FocalGAN-SEE in the number of hard at-risk samples identified, but ranks only fifth in the number of normal at-risk samples identified. This suggests that DCGAN's generator architecture has some specificity for hard at-risk samples, but at the cost of sacrificing normal at-risk sample recognition capability.

LSGAN + LSTM exhibits performance highly similar to that of DCGAN + LSTM, also improving hard at-risk sample recognition at the expense of normal at-risk sample recognition, indicating that the least squares loss does not significantly improve in the hard at-risk sample recognition task. Although WGAN + LSTM's Wasserstein loss provides more stable training, it tends to identify samples in Region B (the transition zone) rather than the truly hard samples in Region C. BGAN is similar to WGAN, with the identified hard at-risk samples mainly concentrated in Region B rather than Region C. VAE's reconstruction loss prioritizes generation quality over classification performance,

resulting in a large number of hard at-risk samples in Region C being misclassified.

4.5. Interpreting Early Warning Results by SHAP

To interpret the decision-making logic of the proposed FocalGAN-SEE model for identifying hard at-risk samples, we adopt the SHAP framework to quantitatively decompose the feature contributions from three granularities: individual input variables, learning behaviors, and course modules. The results of FocalGAN-SEE in Course Module 5 are used for analysis.

The beeswarm SHAP plot (in Figure 5) reveals the marginal contribution distribution of each feature to the model's output. Discussion frequency in Course Modules 5,4 and resource viewing frequency in course module 4 show consistent positive SHAP values, indicating that higher interaction frequency in discussion forums and repeated access to course resources in module 5 and 4 significantly increase the likelihood that the model classifies students as at-risk. In contrast, time-interval features in Course Module 5 dominate the negative SHAP region: shorter time intervals between learning events correspond to lower risk probability, while long time gaps between enrollment, login, and assignments are typical risk signals captured by the FocalGAN-SEE model.

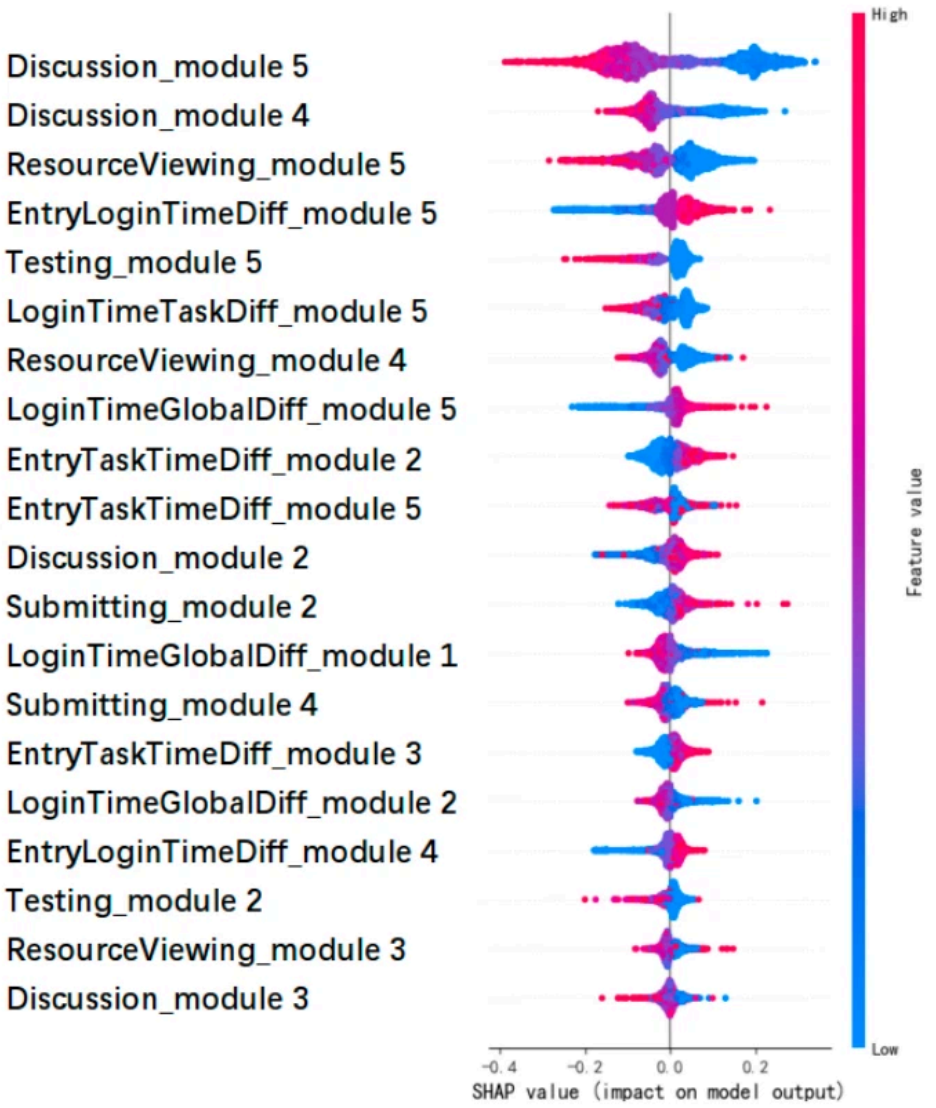


Figure 5. Variable importance via beeswarm SHAP plot.

The students' learning behaviors importance in Figure 6a shows that discussion frequency ranks first with a dominant contribution in prediction results, with resource viewing frequency forming the second core decision dimension, followed by testing frequency. The variable LoginTimeTaskDiff contributes least to the prediction results. This indicates that actively participating in discussion, viewing more course resources and taking course module tests in a timely manner can help improve students' learning performance.

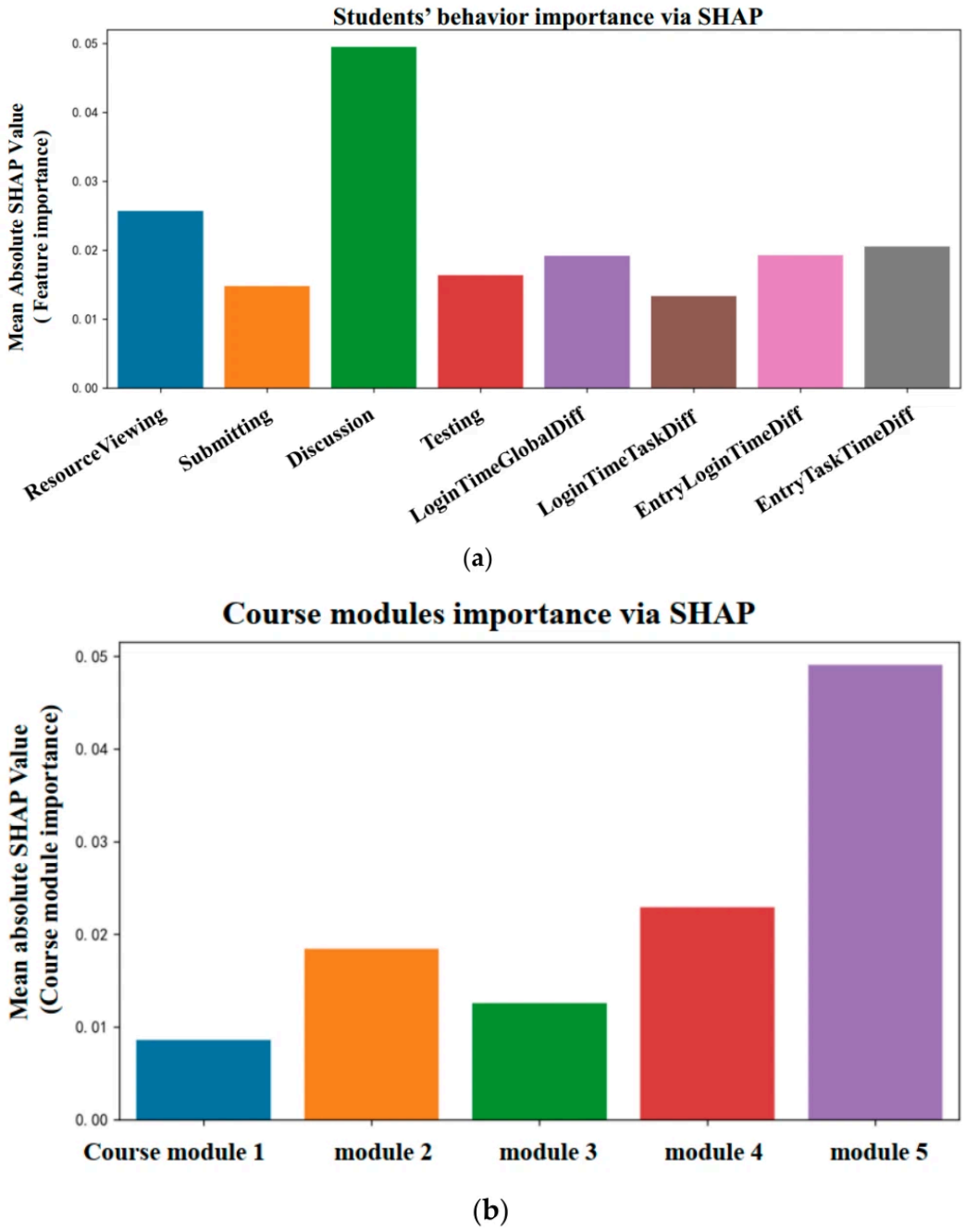


Figure 6. SHAP plot Students' learning behaviors importance (**a**) and course modules importance (**b**) via SHAP.

The importance of course modules shown in Figure 6b indicates that course module 5 achieves far higher global SHAP importance than other modules, which contains high-weight risk-related behaviors including discussion frequency and testing behaviors, making it the core data source for the FocalGAN-SEE model to distinguish at-risk stu-

dents. Module 4 and Module 2 provide secondary auxiliary discrimination signals, while Module 3 contributes the least to model prediction due to weak behavioral differentiation between normal and at-risk students.

5. Discussion

The FocalGAN-SEE proposed in this paper achieves better predictive performance than baseline models on two real-world datasets, and enables earlier early warning. The main reasons are as follows: the architecture employs a HardLoss module to specifically augment the number of minority hard at-risk samples near the decision boundary; it uses end-to-end complete gradient backpropagation to establish an optimization pathway from classification supervision to the generator; it adopts adaptive β weight scheduling to balance the magnitude conflict between adversarial loss and hard loss, ensuring stable convergence over 300 epochs of GAN training; and finally, it leverages the SE Encoder to provide highly discriminative behavioral representations, assigning weights to each behavioral feature channel—suppressing ineffective noisy behaviors while amplifying key temporal patterns that distinguish at-risk students from normal students, with the residual structure preventing the attention mechanism from filtering out original behavioral information. The ablation experiment results demonstrate that each of the above modules plays an important role.

The oversampling-based methods generate new samples via linear interpolation among neighboring normal at-risk samples. They can only cover hard at-risk samples within moderately easy to distinguish regions (Region A) and overlapping regions (Region B), and fail to synthesize extremely hard at-risk samples located close to the classification boundary with heavily mixed features. SMOTE performs random interpolation and tends to generate intra-class samples. Similar to ADASYN, it struggles to identify hard samples residing in highly overlapping regions. Although BSMOTE focuses on boundary interpolation, its interpolation space is limited to the manifold neighborhood of original samples, making it difficult to construct novel temporal sequences of student learning behaviors. In contrast, FocalGAN-SEE leverages the hard confusion loss embedded with a classifier to directionally synthesize temporal behavior samples of students with ambiguous decision-boundary features, actively augmenting samples in high-overlap feature regions. The model correctly identifies 1277 hard samples, 118 more than ADASYN, which greatly remedies the deficiency of interpolation methods in recognizing extremely hard at-risk samples. Furthermore, samples generated by interpolation approaches lack sufficient feature diversity for normal at-risk students and tend to introduce noisy interpolated instances, which leads to frequent misclassification of ordinary students after model training. While strengthening the model's learning of hard minority samples, FocalGAGN-SEE balances the adversarial objective and hard-sample objective via adaptive loss weights. It does not sacrifice the discrimination ability of majority-class samples to boost minority-class recognition. The model achieves balanced recognition accuracy for both categories of samples.

General generative adversarial variants (GAN, BGAN, CGAN, DCGAN, LSGAN, WGAN) are merely optimized to match the global data distribution and fool the discriminator, without any guiding mechanism oriented toward classification decision boundaries. Their generated samples are concentrated at the center of minority-class distributions rather than ambiguous cross-class regions, which limits their ability to boost the recognition of hard samples. In contrast, the proposed FocalGAN integrates the pretrained SE Encoder into the gradient computation graph and introduces a confusion loss targeting a confidence threshold of 0.5 as the hard-sample constraint. This drives the generator to intentionally synthesize ambiguous temporal sequences near the decision

boundary, specially strengthening the model's discriminative power for hard at-risk students in high-overlap feature spaces. Compared with DCGAN, FocalGAN correctly identifies 85 additional hard samples, filling the gap in boundary sample recognition left by conventional GAN variants.

In terms of identifying normal at-risk samples, standard GANs only optimize for alignment between real and synthetic data distributions, which easily yields generated samples with distorted behavioral features and disturbs the model's learning of normal student temporal patterns. FocalGAN adopts adaptive EMA-normalized loss weights: it first fits the overall distribution of authentic student behaviors in the early warm-up stage, and then imposes hard-sample constraints in later training epochs. This guarantees the temporal authenticity of generated samples without impairing recognition performance for normal at-risk students. Consequently, FocalGAN detects far more normal at-risk samples than all compared GAN variants.

Combining comparative experiments across multiple course modules, hard-sample visualization, and SHAP interpretability analysis, the proposed FocalGAN-SEE temporal early warning model not only improves identification metrics for at-risk students at the algorithmic level but also delivers practical support for online course teaching management and personalized learning intervention. In traditional online early warning scenarios, numerous borderline students (hard samples) exhibit learning behaviors that overlap heavily with both normal and fully at-risk groups, making them easily overlooked by conventional warning systems. Such students have not yet displayed obvious signs of academic failure but have already developed latent learning risks, constituting the most critical group for targeted teaching intervention. Benefiting from its stronger capacity to capture hard samples, our model helps instructors and teaching administrators identify "students with hidden academic risks" at an earlier stage, shifting the intervention paradigm from passive remediation to refined proactive support and substantially improving the effectiveness of academic assistance.

We compared this study with the published literature within learning-analytics-driven academic risk prediction. Recent representative studies advance generative-based imbalance mitigation for educational prediction. Zhang et al. [3] constructed a G-B-Blending ensemble framework that combines generative models and oversampling models to leverage the respective advantages of high precision from GAN-based learners and high recall from B-SMOTE-based learners. Although their blending strategy improved overall F1-score and AUC and enabled earlier warning, the generative component only fitted the global distribution of all at-risk students. Boundary hard-sample at-risk students were captured passively through ensemble voting results, lacking explicit loss-level guidance for synthesizing ambiguous borderline instances. Tajdar et al. [72] utilized CTGAN to generate realistic minority-class samples for severely imbalanced educational datasets and adopted Integrated Gradients for model interpretation. Their approach obtained promising minority-class recall, yet CTGAN was trained to learn the overall statistical distribution of minority groups without special constraints to prioritize decision-boundary hard samples. Both aforementioned works treat all synthetic minority samples uniformly. Regarding class-imbalance mitigation for student early-warning tasks, prior studies have demonstrated that generative models and oversampling techniques can improve minority-class recall for at-risk student identification [3,7]. Most existing GAN- and VAE-based educational generative approaches focus on matching the global data distribution of at-risk samples, which enhances overall minority-class performance but offers limited targeted optimization for boundary-hard samples [7,12]. Consistent with those reports, our results also confirm that generative augmentation brings performance gains for imbalanced student datasets. Nevertheless, whereas previ-

ous works treat all synthetic minority samples equally, FocalGAN further introduces classification-boundary-aware hard-sample supervision to steer generation toward ambiguous borderline instances. This provides a complementary perspective to existing imbalance-mitigation research in education: instead of merely increasing the quantity of synthetic minority samples, our framework prioritizes the quality of boundary hard samples, which is particularly valuable for detecting latent borderline-at-risk students who are easily overlooked in practical teaching scenarios.

Existing educational prediction literature has widely adopted SHAP as a post hoc interpretability tool for unpacking black-box model outputs. Kesgin et al. [73] combined SMOTE resampling with SHAP and LIME to analyze student performance prediction on UCI datasets; their work mainly focused on global feature importance and model fairness evaluation for general student groups, without special interpretability analysis for borderline hard-sample students. Abukader et al. [74] adopted SHAP to interpret metaheuristic-optimized LightGBM, quantifying how attendance, study hours and parental involvement shape exam scores, yet their interpretability experiments were oriented toward overall regression prediction rather than risk-stratified hard-sample cases. For sequential learning data, Kalita et al. [75] built an LSTM-SHAP pipeline for disabled learners in virtual learning environments, leveraging SHAP to interpret temporal interaction features derived from LSTM. Nevertheless, their SHAP explanations targeted the whole test-set population and did not isolate feature patterns belonging to misclassified or boundary-ambiguous hard at-risk students. Most prior SHAP-based educational interpretability studies complete feature-level analysis on ordinary samples trained by vanilla or oversampling-augmented models. Different from the above works, our study applies SHAP interpretation to a generative-augmented sequential early-warning pipeline, and further focuses on revealing feature patterns of borderline hard-sample at-risk students. Even though synthetic generated samples are involved in model training, SHAP still yields reliable feature attribution results, which provides targeted evidence for teachers to understand latent-risk students who are easily overlooked by conventional models.

In this study, SHAP interpretability results quantify the correlation between learning behaviors and academic risks. Specifically, forum discussion frequency and course resource viewing frequency stand out as the most discriminative features for identifying at-risk students, while excessively long time intervals between login and assignment submission serve as typical risk indicators. Our SHAP-based observations align with previous research that highlights learning-activity metrics such as discussion participation and resource engagement as strong predictors of academic risk. At the same time, our analysis further reveals how these temporal behavioral features drive hard-sample identification within a GAN-augmented pipeline. Compared with existing literature, this work combines generative hard-sample augmentation and post hoc interpretability, showing how SHAP can help educators understand model outputs even when synthetic generated samples are included in model training. Based on these findings, instructors can provide clear, data-backed learning guidance for students: encouraging active participation in forum discussions instead of passive consumption of lecture videos; guiding continuous, consistent browsing of supplementary course resources; urging timely completion of modular quizzes; and advising students to avoid prolonged learning interruptions to maintain coherent study rhythms.

From a practical implementation perspective, deploying this framework for early-warning system in real-world educational institutions introduces several non-algorithmic challenges worthy of consideration. Although the dataset used in this research has been fully anonymized with all personally identifiable information removed, aggregated

student behavioral and learning-activity features still constitute sensitive educational data. Appropriate cybersecurity safeguards and strict access-permission controls are still required during data storage, transmission and model inference to prevent unauthorized access or misuse of student learning records. Second, as an online application supporting instructional decision-making, the early-warning platform may experience service downtime or outages caused by server malfunctions or network instability. Such interruptions can delay risk screening workflows and prevent educators from receiving timely early-warning feedback.

In terms of computational resources, our model requires a certain amount of hardware overhead during the training phase. Nevertheless, such requirements can generally be satisfied by the existing hardware resources available in most modern-day universities. After training is finished, the inference process for generating student risk warnings imposes only a lightweight computational burden.

Beyond technical infrastructure challenges, organizational and pedagogical obstacles also deserve attention. Practitioners need to build teachers' trust in model outputs and formulate reasonable usage specifications: model-generated risk predictions serve only as decision-support references. Pedagogical experts must retain human judgment, and students should not be simplistically categorized or labeled solely according to algorithmic warning scores. Successful practical deployment also requires alignment with existing teaching-management workflows within universities.

6. Conclusions

This study addresses the key challenges in identifying hard at-risk samples in online learning early warning systems—namely, temporal feature dependency, class imbalance, and weak feature extraction capability for boundary-ambiguous samples—and proposes an early warning framework FocalGAN-SEE (Focal Generative Adversarial Network with Squeeze-and-Excitation Enhanced Encoder), which integrates a Focal Generative Adversarial Network with an SE attention-enhanced encoder module.

The contributions of this work are fourfold. First, a focal hard-sample loss with a 0.5 confidence target drives the generator to produce samples that maximize classifier uncertainty, effectively expanding the minority distribution toward the decision boundary. Second, an adaptive balancing mechanism—combining cosine annealing with exponential moving average-based magnitude scaling—dynamically adjusts the weight between adversarial and hard-sample objectives, ensuring stable convergence throughout the training process while preventing either loss component from dominating the optimization. Third, the SE Encoder enhances temporal feature extraction through channel-wise recalibration with residual fusion, adaptively weighting behavioral features to amplify risk-discriminative patterns while preserving original sequential information. Fourth, end-to-end gradient propagation establishes a complete optimization pathway from classification supervision back to the generator, enabling direct and coherent optimization toward boundary-ambiguous sample generation.

Experimental evaluations conducted on two real-world online learning datasets spanning 46 courses across two teaching periods yield the following key findings:

(1) The SE Encoder module achieves significantly superior predictive performance compared to seven attention-mechanisms-augmented autoencoder models commonly used in early warning systems, demonstrating the effectiveness of the Squeeze-and-Excitation mechanism for educational time-series classification tasks.

(2) FocalGAN-SEE consistently achieves the highest accuracy, precision, recall, F1-score, and AUC across course modules throughout the teaching period, substantially outperforming interpolation-based methods (SMOTE, ADASYN, BSMOTE)

and generative models (GAN, DCGAN, LSGAN, WGAN, BGAN, VAE). The five-fold cross-validation results show that the full model achieves an F1-score of 0.8471 and an AUC of 0.9464 in the early warning course module, substantially improving over the baseline (F1: 0.7200; AUC: 0.9122).

(3) Ablation studies confirm each component's indispensable contribution: removing the hard loss, adaptive balancing mechanism, end-to-end propagation, or SE Encoder individually degrades F1 by 4.4%, 4.6%, 4.2%, and 4.3%, respectively. This indicates that all modules play an important role in the whole architecture.

(4) Visualization analysis reveals that FocalGAN-SEE uniquely excels in identifying hard at-risk samples in the high-overlap decision boundary region while maintaining strong normal sample recognition, effectively resolving the trade-off between minority-class sensitivity and majority-class specificity. SHAP interpretability further identifies forum discussion frequency and course resource browsing frequency as the most salient risk predictors, offering empirically grounded guidance for targeted pedagogical interventions.

In summary, FocalGAN-SEE advances the state-of-the-art in educational early warning systems by providing a principled solution to hard at-risk sample recognition in imbalanced sequential data. Beyond algorithmic improvements, the framework offers practical value for online course management and personalized intervention, enabling educators to transition from reactive remediation to proactive, precision-oriented academic support.

In the cross-teaching period generalization experiment, FocalGAN-SEE was trained on data from the first teaching period and achieved better predictive performance than baseline models when tested on the second teaching period. This result verifies the model's transferability and generalization performance across different teaching periods. The input of FocalGAN-SEE relies on learning behavioral logs, which are relatively easy to collect in both higher education and K-12 environments, and the model does not impose strict requirements on data formats. Benefiting from this feature, our method has the potential to be adapted to other educational contexts such as K-12 education and hybrid teaching. Nevertheless, we note that this potential remains to be further validated with dedicated datasets from K-12 and hybrid teaching scenarios, which we save as a limitation and direction for future work. Future work may also explore multimodal data fusion and adaptation to other sequential classification tasks with severe class imbalance, such as early learning performance warning based on multimodal data fusion of courses.

Author Contributions: J.L.: conceptualization, supervision, resources, writing—original draft, editing. M.Z.: methodology, formal analysis, editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Humanities and Social Science Fund of Ministry of Education of China, grant number 24YJC880170.

Institutional Review Board Statement: Ethical review and approval were waived for this study by the Human Research Ethics Committee of Zhejiang Normal University. The research involved the secondary analysis of pre-existing retrospective administrative data that had been fully anonymized and de-identified before access. Therefore, it does not meet the definition of human subject research and is exempt from further ethics committee review.I

Informed Consent Statement: Informed consent was waived due to the retrospective, fully anonymized nature of the dataset. No identifiable personal information was accessed, utilized, or reported during this study.

Data Availability Statement: The data obtained through the experiments are available upon request from corresponding author.

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

References

1. Miranda, E.; Aryuni, M.; Rahmawati, M.I.; Hiererra, S.E.; Sano, A.V.D. Machine learning's model-agnostic interpretability on the prediction of students' academic performance in video-conference-assisted online learning during the covid-19 pandemic. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100312. [Google Scholar] [CrossRef]
2. Du, X.; Yang, J.; Shelton, B.E.; Hung, J.-L.; Zhang, M. A systematic meta-review and analysis of learning analytics research. *Behav. Inf. Technol.* **2019**, *40*, 49–62. [Google Scholar] [CrossRef]
3. Zhang, M.; Wang, Y.; Hung, J.-L.; Wang, J.; Duan, C. An early warning method based on blending of deep generative model and oversampling model for online learning. *IEEE Access* **2025**, *13*, 72248–72268. [Google Scholar] [CrossRef]
4. Altalhi, A.H.; Ragab, M. Modelling of enhanced predictive analytics and decision-making approach using recurrent autoencoder with attribute subset reduction for students' academic performance. *Ain Shams Eng. J.* **2026**, *17*, 103999. [Google Scholar] [CrossRef]
5. Qiu, H.; Zhu, H.; Hu, X.; Wang, Y.; Xu, Y.; Li, W. Predicting Student Performance Using a Cost-Sensitive Deep Learning Network With Sparse Attention Mechanism. *IEEE Trans. Educ.* **2026**, *69*, 10–18. [Google Scholar] [CrossRef]
6. Liu, L.; Liang, Y.; Yan, X.; Huangfu, L.; Samtani, S.; Yu, Z.; Zhang, Y.; Zeng, D.D. Hard Sample Mining: A New Paradigm of Efficient and Robust Model Training. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *37*, 526–546. [Google Scholar] [CrossRef] [PubMed]
7. Duan, C.; Wang, Y.; Zhang, W.; Yu, Z.; Pei, Y.; Zhang, M.; Huang, Q. TVAE-GAN: A Generative Model for Providing Early Warnings to High-Risk Students in Basic Education and Its Explanation. *Information* **2026**, *17*, 356. [Google Scholar] [CrossRef]
8. Guan, R.; Tu, W.; Li, Z.; Yu, H.; Hu, D.; Chen, Y.; Tang, C.; Yuan, Q.; Liu, X. Spatial–spectral graph contrastive clustering with hard sample mining for hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [Google Scholar] [CrossRef]
9. Zhou, L.; Wu, M.; Xu, X.; Li, X.; Niu, C.; Wu, B. Multi-scale contrastive learning with cluster-aware hard sample mining for attributed graph clustering. *Appl. Soft Comput.* **2026**, *193*, 114758. [Google Scholar] [CrossRef]
10. Fan, H.; Gao, S.; Han, H.; Han, M.; Zhou, B.; Wang, Z. Pseudo Anomalies and Hard Sample Mining for Ventricular Arrhythmia Anomaly Detection. *IEEE J. Biomed. Health Inform.* **2026**, *30*, 8884–8896. [Google Scholar] [CrossRef] [PubMed]
11. Lin, C.; Shi, Y.; Wang, G.; Liu, S. Improving vision-language models through intra-modal contrastive learning-based hard sample mining. *Neurocomputing* **2025**, *652*, 131047. [Google Scholar] [CrossRef]
12. Du, X.; Yang, J.; Hung, J.-L. An integrated framework based on latent variational autoencoder for providing early warning of at-risk students. *IEEE Access* **2020**, *8*, 10110–10122. [Google Scholar] [CrossRef]
13. Chen, Y.; Wei, G.; Liu, J.; Chen, Y.; Zheng, Q.; Tian, F.; Zhu, H.; Wang, Q.; Wu, Y. A prediction model of student performance based on self-attention mechanism. *Knowl. Inf. Syst.* **2023**, *65*, 733–758. [Google Scholar] [CrossRef]
14. Jieyang, P.; Kimmig, A.; Dongkun, W.; Niu, Z.; Zhi, F.; Jiahai, W.; Liu, X.; Ovtcharova, J. A systematic review of data-driven approaches to fault diagnosis and early warning. *J. Intell. Manuf.* **2023**, *34*, 3277–3304. [Google Scholar] [CrossRef]
15. Lee, S.; Chung, J.Y. The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Appl. Sci.* **2019**, *9*, 3093. [Google Scholar] [CrossRef]
16. Leelaluk, S.; Tang, C.; Minematsu, T.; Taniguchi, Y.; Okubo, F.; Yamashita, T.; Shimada, A. Attention-based artificial neural network for student performance prediction based on learning activities. *IEEE Access* **2024**, *12*, 100659–100675. [Google Scholar] [CrossRef]
17. Al Hashmi, R.A.M.; Ozturk, I.; Elmehdi, H.M. Early detection of at-risk health sciences students: A machine learning-based predictive study using midterm grades. *BMC Med. Educ.* **2025**, *25*, 1651. [Google Scholar] [CrossRef] [PubMed]
18. Li, S.; Liu, T. Performance prediction for higher education students using deep learning. *Complexity* **2021**, *2021*, 9958203. [Google Scholar] [CrossRef]
19. Katib, I.; Ragab, M. Harnessing variable reduction approach with deep recurrent neural network for student's academic performance analysis. *Alex. Eng. J.* **2025**, *118*, 393–405. [Google Scholar] [CrossRef]
20. Vives, L.; Cabezas, I.; Vives, J.C.; Reyes, N.G.; Aquino, J.; Cóndor, J.B.; Altamirano, S.F.S. Prediction of Students' Academic Performance in the Programming Fundamentals Course Using Long Short-Term Memory Neural Networks. *IEEE Access* **2024**, *12*, 5882–5898. [Google Scholar] [CrossRef]

21. Fazil, M.; Albahlal, B.M. MultiFAR: Multidimensional information fusion with attention-driven representation learning for student performance prediction. *PLoS ONE* **2026**, *20*, e0333099. [Google Scholar] [CrossRef]
22. Singh, H.P.; Dubey, G.; Sheoran, K.; Dhand, G.; Singh, A. Socially enhanced graph network for early identification of at-risk students in virtual learning environments. *User Model. User-Adapt. Interact.* **2026**, *36*, 5. [Google Scholar] [CrossRef]
23. Ai, L.; Zhang, X. Attention model based on one-dimensional residual convolutional autoencoder and GRU for knowledge tracing. *Multimed. Tools Appl.* **2024**, *84*, 23725–23743. [Google Scholar] [CrossRef]
24. Hung, J.L.; Rice, K.; Zhang, M. Interpretable AI techniques unveil the factors and types of at-risk early warning: A case study in K-12 online learning. *Data Technol. Appl.* **2025**, *59*, 516–538. [Google Scholar] [CrossRef]
25. Kalita, E.; Alfarwan, A.M.; El Aouifi, H.; Kukkar, A.; Hussain, S.; Ali, T.; Gaftandzhieva, S. Predicting student academic performance using Bi-LSTM: A deep learning framework with SHAP-based interpretability and statistical validation. *Front. Educ.* **2025**, *10*, 1581247. [Google Scholar] [CrossRef]
26. 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. [Google Scholar] [CrossRef]
27. Huang, Y.; Zhou, Y.; Chen, J.; Wu, D. Applying Machine Learning and SHAP Method to Identify Key Influences on Middle-School Students' Mathematics Literacy Performance. *J. Intell.* **2024**, *12*, 93. [Google Scholar] [CrossRef] [PubMed]
28. Qiang, M.; Liu, Z.; Zhang, R. Explainable AI in education: Integrating educational domain knowledge into the deep learning model for improved student performance prediction. *Sci. Rep.* **2026**, *16*, 9515. [Google Scholar] [CrossRef] [PubMed]
29. Kala, A.; Torkul, O.; Yildiz, T.T.; Selvi, I.H. Early Prediction of Student Performance in Face-to-Face Education Environments: A Hybrid Deep Learning Approach With XAI Techniques. *IEEE Access* **2024**, *12*, 191635–191649. [Google Scholar] [CrossRef]
30. Tao, C.; Kuang, D.; Huang, Z.; Peng, C.; Li, H. HASNet: A Foreground Association-Driven Siamese Network With Hard Sample Optimization for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–16. [Google Scholar] [CrossRef]
31. Wang, C.; Jiang, W.; Li, C.; Yu, Z.; Liu, W.; Wang, X.; Liu, B. Diagnosis-driven hard sample generation: Low-frequency attenuation supervised contrastive learning for mammogram classification. *Multimed. Syst.* **2026**, *32*, 240. [Google Scholar] [CrossRef]
32. Sun, Z.-W.; Hua, Z.-X.; Li, H.-C.; Li, Y. Self-paced learning strategy with easy sample prior based on confidence for the flying bird object detection model training. *Signal Image Video Process.* **2025**, *19*, 95. [Google Scholar] [CrossRef]
33. Ren, T.; Zhang, H.; Wang, Y.; Ju, W.; Liu, C.; Meng, F.; Yi, S.; Luo, X. MHGC: Multi-scale hard sample mining for contrastive deep graph clustering. *Inf. Process. Manag.* **2025**, *62*, 104084. [Google Scholar] [CrossRef]
34. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017. [Google Scholar]
35. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012. [Google Scholar] [CrossRef]
36. Cordeiro, F.R.; Belagiannis, V.; Reid, I.; Carneiro, G. Propmix: Hard sample filtering and proportional mixup for learning with noisy labels. *arXiv* **2021**, arXiv:2110.11809. [Google Scholar]
37. Li, B.; Xie, S.; Xie, X.; Tang, H.; Scholz, M.; Zhang, Z.; Tian, Z. MFCR-SleepNet: A hard sample mining-based multi-scale feature contrastive representation learning model for sleep stage classification. *Neurocomputing* **2025**, *668*, 132399. [Google Scholar] [CrossRef]
38. Hao, Y.; Cao, H. A new attention mechanism to classify multivariate time series. In Proceedings of the 29th International Joint Conference on Artificial Intelligence, Yokohama, Japan, 11–17 July 2020. [Google Scholar]
39. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473. [Google Scholar]
40. Abinaya, S.; Deepak, M.; Alphonse, A.S. Enhanced Image Captioning Using Bahdanau Attention Mechanism and Heuristic Beam Search Algorithm. *IEEE Access* **2024**, *12*, 100991–101003. [Google Scholar] [CrossRef]
41. Huang, Z.; Liang, M.; Qin, J.; Zhong, S.; Lin, L. Understanding self-attention mechanism via dynamical system perspective. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023. [Google Scholar]
42. Qiu, J.; Liu, Y.; Chai, Y.; Si, Y.; Su, S.; Wang, L.; Wu, Y. Dependency-Based Local Attention Approach to Neural Machine Translation. *Comput. Mater. Contin.* **2019**, *59*, 547–562. [Google Scholar] [CrossRef]
43. Bastidas, A.A.; Tang, H. Channel attention networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019. [Google Scholar]

44. Lin, H.; Cheng, X.; Wu, X.; Shen, D. Cat: Cross attention in vision transformer. In *Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022*; IEEE: Piscataway, NJ, USA, 2022. [Google Scholar]
45. Li, P.; Pei, Y.; Li, J. A comprehensive survey on design and application of autoencoder in deep learning. *Appl. Soft Comput.* **2023**, *138*, 110176. [Google Scholar] [CrossRef]
46. Wen, X.; Juan, H. Early prediction of students' performance using a deep neural network based on online learning activity sequence. *Appl. Sci.* **2023**, *13*, 8933. [Google Scholar] [CrossRef]
47. Oluwasanmi, A.; Aftab, M.U.; Baagyere, E.; Qin, Z.; Ahmad, M.; Mazzara, M. Attention autoencoder for generative latent representational learning in anomaly detection. *Sensors* **2021**, *22*, 123. [Google Scholar] [CrossRef] [PubMed]
48. Berahmand, K.; Daneshfar, F.; Salehi, E.S.; Li, Y.; Xu, Y. Autoencoders and their applications in machine learning: A survey. *Artif. Intell. Rev.* **2024**, *57*, 28. [Google Scholar] [CrossRef]
49. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [Google Scholar]
50. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020. [Google Scholar]
51. Zhou, X.; Tu, G.; Zhu, X.; Zhao, D.; Zhang, L. A Three-Channel Improved SE Attention Mechanism Network Based on SVD for High-Order Signal Modulation Recognition. *Electronics* **2025**, *14*, 2233. [Google Scholar] [CrossRef]
52. Huang, Z.; Yang, Y.; Wang, Z.; Li, Y.; Chen, Z.; Ma, Y.; Zhang, S. MultiSEss: Automatic Sleep Staging Model Based on SE Attention Mechanism and State Space Model. *Biomimetics* **2025**, *10*, 288. [Google Scholar] [CrossRef] [PubMed]
53. Yu, X.; Zhu, D.; Guo, H.; Zhou, C.; Elhassan, M.A.M.; Wang, M. DASNet: A Convolutional Neural Network with SE Attention Mechanism for ccRCC Tumor Grading. *Interdiscip. Sci. Comput. Life Sci.* **2025**, *18*, 60–76. [Google Scholar] [CrossRef] [PubMed]
54. Liu, L.; Wu, K.; Wang, K.; Han, Z.; Qiu, J.; Zhan, Q.; Wu, T.; Xu, J.; Zeng, Z. SEU2-Net: Multi-scale U2-Net with SE attention mechanism for liver occupying lesion CT image segmentation. *PeerJ Comput. Sci.* **2024**, *10*, e1751. [Google Scholar] [CrossRef] [PubMed]
55. Naidu, S.M.; Xiong, N.; Ghomi, M.; Schuderer, J. Context-Driven Framework for Maintenance Decision Optimization in a Power Network. In *Proceedings of the 2024 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE), Dubrovnik, Croatia, 14–17 October 2024*; IEEE: Piscataway, NJ, USA, 2024. [Google Scholar]
56. Sick, L.; Engel, D.; Hermosilla, P.; Ropinski, T. Attention-guided masked autoencoders for learning image representations. In *Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 26 February–6 March 2025*; IEEE: Piscataway, NJ, USA, 2025. [Google Scholar]
57. Zheng, Z.; Chen, X.; Huang, M. Cluster-guided graph attention auto-encoder. *J. Supercomput.* **2025**, *81*, 470. [Google Scholar] [CrossRef]
58. Song, Y.; Yuan, L. Optimizing Academic Performance Prediction in Higher Education: A Big Data Analytics Framework Integrating CNN-LSTM and Attention Mechanisms. *J. Cases Inf. Technol.* **2026**, *28*, 1–20. [Google Scholar] [CrossRef]
59. Xie, Y. Student performance prediction via attention-based multi-layer long-short term memory. *J. Comput. Commun.* **2021**, *9*, 61–79. [Google Scholar] [CrossRef]
60. Zhang, M.; Zeng, Z.; Liu, Y.; Bai, J.; Hung, J.-L.; Duan, C. An Early Warning Method Combining Autoencoder and Local Attention Mechanism for K-12 Education. In *Proceedings of the 2025 5th International Conference on Educational Technology (ICET), Chongqing, China, 26–28 September 2025*; IEEE: Piscataway, NJ, USA, 2025. [Google Scholar]
61. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680. [Google Scholar] [CrossRef]
62. Hou, R.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Cross attention network for few-shot classification. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 4005–4016. [Google Scholar] [CrossRef]
63. Hjelm, R.D.; Jacob, A.P.; Che, T.; Trischler, A.; Cho, K.; Bengio, Y. Boundary-seeking generative adversarial networks. *arXiv* **2017**, arXiv:1702.08431. [Google Scholar]
64. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling tabular data using conditional GAN. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 7333–7343. [Google Scholar] [CrossRef]
65. Patil, A.; Venkatesh. DCGAN: Deep convolutional GAN with attention module for remote view classification. In *Proceedings of the 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), Bengaluru, India, 21–22 December 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 1–10. [Google Scholar]

66. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.K.; Wang, Z.; Smolley, S.P. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802. [Google Scholar]
67. Liu, Y.; Zhou, Y.; Liu, X.; Dong, F.; Wang, C.; Wang, Z. Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology. *Engineering* **2019**, *5*, 156–163. [Google Scholar] [CrossRef]
68. Kingma, D.P.; Welling, M. An introduction to variational autoencoders. *Found. Trends Mach. Learn.* **2019**, *12*, 307–392. [Google Scholar] [CrossRef]
69. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [Google Scholar] [CrossRef]
70. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887. [Google Scholar]
71. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008*; IEEE: Piscataway, NJ, USA, 2008; pp. 1322–1328. [Google Scholar]
72. Tajdar, S.; Dustmohammadloo, H.; Rakhmonov, R.; Mamadiyarov, Z.; Urazbaeva, D.; Tursunov, L.; Khamraev, A.; Safoyev, H.; Eldor, T. A Privacy-Preserving and Explainable Deep Learning Framework for Imbalanced Student Performance Prediction Using CTGAN-Based Data Augmentation. *Concurr. Comput. Pract. Exp.* **2026**, *38*, e70882. [Google Scholar] [CrossRef]
73. Kesgin, K.; Kiraz, S.; Kosunalp, S.; Stoycheva, B. Beyond Performance: Explaining and Ensuring Fairness in Student Academic Performance Prediction with Machine Learning. *Appl. Sci.* **2025**, *15*, 8409. [Google Scholar] [CrossRef]
74. Abukader, A.; Alzubi, A.; Adegboye, O.R. Intelligent System for Student Performance Prediction: An Educational Data Mining Approach Using Metaheuristic-Optimized LightGBM with SHAP-Based Learning Analytics. *Appl. Sci.* **2025**, *15*, 10875. [Google Scholar] [CrossRef]
75. Kalita, E.; El Aouifi, H.; Kukkar, A.; Hussain, S.; Ali, T.; Gaftandzhieva, S. LSTM-SHAP based academic performance prediction for disabled learners in virtual learning environments: A statistical analysis approach. *Soc. Netw. Anal. Min.* **2025**, *15*, 65. [Google Scholar] [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.