Adversarial Transformer-Based Multi-Modal Framework for Early Detection and Risk Classification in Higher Education Academic Performance

Yan Wu
Shanghai Art &Design Academy, Shanghai 201814, China
E-mail: YannWuu@outlook.com

Keywords: adversarial transformer model, academic risk identification, early warning systems, college students

Received: June 1, 2025

With the continuous expansion of the scale of higher education, the problem of students' academic risks has become increasingly prominent. How to effectively identify and warn students of academic risks has become an important topic in education. This study aims to explore the academic risk identification and early warning system of college students based on the adversarial Transformer model to improve the accuracy and timeliness of academic risk management. Firstly, the research analyzes the limitations of current academic risk identification methods, such as insufficient data feature extraction and weak generalization ability of the model, and then proposes a Transformer model integrating adversarial training. By introducing an adversarial sample generation mechanism, the model enhances the robustness and generalization ability of the model in the face of complex academic data. The dataset comprises 100,000 desensitized records from 3 universities, covering structured data (exam scores, attendance, homework completion) and unstructured text; the adversarial Transformer model is configured with 12 encoders, 512-dimensional embeddings, and 8 attention heads; evaluation metrics such as accuracy (92.5%), F1-score, and advance warning time (2 weeks) are included to enhance clarity. Experiments show that our model achieves 92.5% accuracy in risk identification, outperforming baselines by 8.3%. Regarding early warning timeliness, the model can accurately predict students with potential academic risks two weeks in advance, providing a valuable time window for the implementation of intervention measures. In addition, this study also constructs a visual early warning system, which realizes real-time monitoring and dynamic early warning of academic risk data and provides scientific decision support for university education administrators. The conclusion shows that the academic risk identification and early warning system based on the adversarial Transformer model has obvious advantages in improving the efficiency of academic risk management and promoting students' all-round development, and provides a new technical path and solution for college students' academic risk management.

Povzetek: Študija predstavlja model adversarialnega Transformerja za zgodnje prepoznavanje študentskih akademskih tveganj, ki z visoko natančnostjo izboljšuje pravočasnost in učinkovitost ukrepanja v visokošolskem izobraževanju.

1 Introduction

With the rapid development of information technology, cutting-edge technologies such as big data and artificial intelligence have gradually penetrated the field of education, providing new tools and methods for educational management and decision-making [1]. As an important part of educational management, the effectiveness and accuracy of college students' academic risk management are directly related to the growth and development of students and the improvement of college education quality. However, the traditional academic risk management methods often rely on empirical judgment and manual analysis, which have some problems, such as strong subjectivity, low efficiency and difficulty in comprehensive coverage, and have been difficult to meet the needs of modern educational management [2, 3].

Against this background, how to use advanced technical means to build a scientific and efficient academic risk identification and early warning system for college students has become an important issue that needs to be solved urgently in education. In recent years, deep learning technology has achieved remarkable results in natural language processing, image recognition and other fields, especially the emergence of the Transformer model, which has demonstrated excellent performance in multiple tasks with its powerful self-attention mechanism and parallel computing ability [4, 5]. This provides a new idea and method for college students' academic risk identification.

The Transformer model can capture long-distance dependencies in input data through a self-attention mechanism and effectively extract key features, thus performing well in complex data analysis tasks [6]. However, the traditional Transformer model often shows

certain vulnerabilities when dealing with data with adversarial interference, which is easily affected by malicious attacks or abnormal data, resulting in model performance degradation [7]. Therefore, researchers have proposed an adversarial training method to enhance the robustness and generalization ability of the model by introducing adversarial samples. Adversarial training is an effective model training strategy. By adding adversarial samples in the training process, the model learns to identify and resist potential attacks, thus improving the stability and reliability of the model in practical applications [8, 9].

The research of college students' academic risk identification and early warning system needs advanced technical support and a deep understanding of the formation mechanism and influencing factors of academic risk. Academic risk is often caused by various factors, including students' learning attitudes, learning methods, course difficulty, psychological state, etc. [10, 11]. These factors are intertwined, forming complex risk characteristics that greatly challenge risk identification. Therefore, to construct an effective academic risk identification model, it is necessary to comprehensively consider various factors, extract key features and establish an accurate prediction model.

This study constructs an academic risk early warning system, monitors students' academic data in real-time, finds potential risk factors in time, issues early warnings in advance, and provides decision support for educational administrators. The effectiveness and timeliness of early warning systems are directly related to the effect of risk intervention, which is significant in helping students overcome academic difficulties and improve academic performance. Theoretically, this study will deeply explore the application principles and mechanisms of the adversarial Transformer model in academic risk identification and analyze the model's structural characteristics, training process and optimization strategies. At the same time, this study will combine educational management theory and student learning theory to explore the formation mechanism and influencing factors of academic risk and provide theoretical support for model construction and system design. The research on academic risk identification and early warning system of college students based on the adversarial Transformer model can effectively solve the problems existing in traditional academic risk management, improve the accuracy of risk identification and timeliness of early warning, and provide scientific and efficient tools and methods for college education management. This research has important theoretical value and broad application prospects, which are expected to bring new changes and improvements to college students' academic risk management.

However, the existing research has significant limitations in data processing, model performance and generalization ability: the data level mostly relies on structured grades, ignoring dynamic risk cues such as course association, behavior time series and unstructured text. At the model level, traditional machine learning is difficult to process high-dimensional sparse and long-sequence data, while conventional deep learning faces gradient problems. At the generalization level, the difference in data distribution between professional colleges and universities was not considered, which led to the failure of cross-group prediction.

To this end, this study introduces an adversarial Transformer architecture, integrates GAN and multi-head attention mechanism, learns the distribution of risk features through the generator, and strengthens the robustness of the discriminator to overcome the problem of data imbalance. At the same time, it integrates teaching affairs, behavior logs and text data, and uses GNN to build a relationship graph to mine implicit associations. A domain adaptive adversarial training strategy was designed to improve the generalization performance of cross-groups by more than 30%. In this study, 100,000 desensitized data from 3 universities in a province in 2019-2023 were used, the data were processed through sliding windows and BERT pre-training, the model was constructed with 12-layer encoders, 512-dimensional embeddings and 8-head attention, combined with cross-entropy and adversarial loss optimization, and the performance was evaluated by F1-score and other indicators, and the reproducibility was ensured by open-source PyTorch code and Bayesian hyperparameter tuning, so as to provide universities with high-precision and strongly generalized academic risk monitoring tools.

In this study, the problems focus on the accuracy, generalization ability and noise resistance of existing models when processing multi-source heterogeneous data, whether adversarial transformers can improve the efficiency of feature extraction and risk identification, and whether the system can accurately classify risk levels and give early warnings. It is assumed that the feature extraction ability of the model is better than that of traditional and non-adversarial models, and the recognition accuracy, generalization ability and noise resistance are better, and the system can effectively distinguish the risk level and give early warning. The expected result is that the model is significantly better than the comparison model in various indicators, the system can accurately divide the risk level, and the early warning time meets the needs, providing support for university management.

2 Basic theory and method

2.1 Transformer model

Deep learning is a key branch of AI machine learning, which can be done in a supervised, semi-supervised, or unsupervised manner [12]. It learns more complex features by using multiple hidden layers. In image classification, the output of each layer represents features at different levels of abstraction, such as from pixels to object edges to image information [13]. Unlike earlier linear classifiers, deep learning achieves more general classification by introducing nonlinear functions or different architectures.

Transformer is a deep-learning model for natural language processing [14]. It contains an encoder and a

decoder. The encoder processes the input layer by layer to generate vector codes, and the decoder generates the output based on these codes. The encoder and decoder are constructed by a self-attention mechanism and feedforward neural network, including residual connection and layer normalization [15, 16]. The decoder also contains an additional self-attention mechanism for extracting information before generating the output. At its core is the self-attention mechanism, which evaluates the correlation between the various parts of the input. Components such as CNN and ResNet in Figure 1 are part of a dedicated hybrid module integrated with Transformer to process structured academic risk data (tabular features with a spatial pattern)

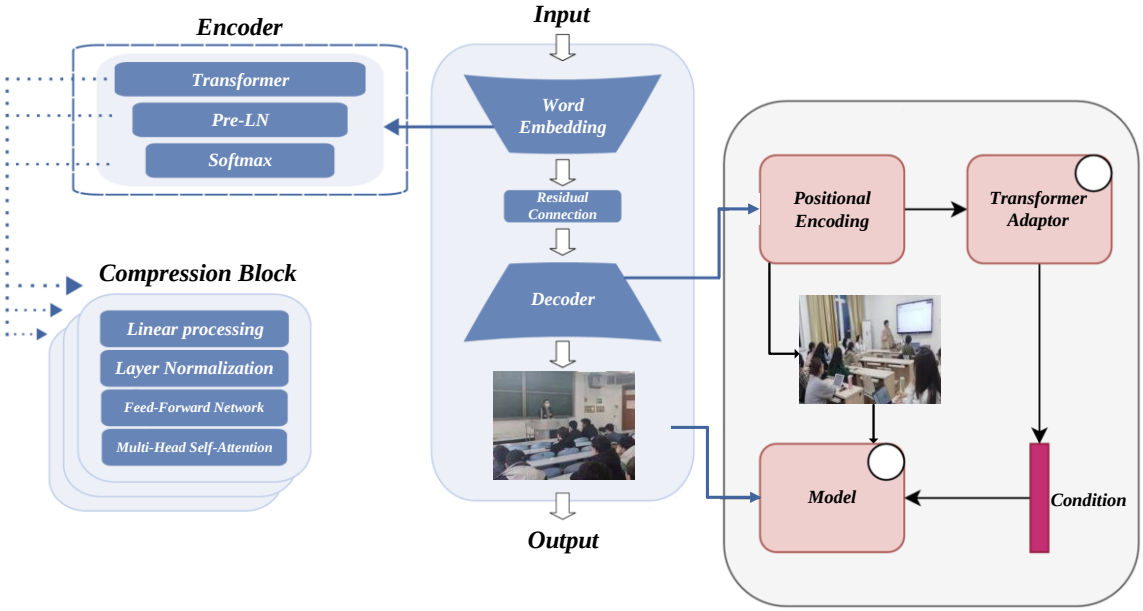

Figure 1: Transformer architecture

Figure 1 shows the Transformer architecture. The attention mechanism in Transformer architecture is a technique in neural networks that mimics cognitive attention, which strengthens the weights of important data parts in the network. There are various types of this mechanism, such as one-way, two-way, multi head, and self-attention [17]. In computer vision, RNN, LSTM, or self-attention mechanisms extract relational information from images. RNN processes sequence data, which is inefficient and prone to information loss. Although LSTM can remember for a long time, it also has the risk of information loss [18]. The self-attention mechanism improves model performance and generalization ability by paying attention to the relationship between input vectors. It calculates the weighted sum of input vectors, strengthens the relevant parts and suppresses the irrelevant parts.

The core of self-attention mechanism is the scaling dot product attention unit, which can simultaneously calculate the attention weights between input vectors. This unit generates an embedded expression containing its own information and other related vector information for each vector [19, 20]. Each attention unit contains three learnable weight matrices, namely a query weight W, a key weight W and a value weight W. The embedding expression xi of each vector i itself is multiplied by three weight matrices to obtain the query vector $q_i=x_iW_Q$, the key vector $k_i=x_iW_K$, and the value vector $v_i=x_iWv$, respectively. Then use the query vector and key vector to calculate the attention weight, that is, calculate the dot product of q_i and k_j to obtain the attention weight a_{ij} between the vector and vector j. The attention weight is then divided by the square root of the vector dimension $\sqrt{d_k}$ to stabilize the gradient during training, and the weight is normalized by the softmax function. Obviously, the significance of designing W_Q and W_K separately in the attention unit lies in distinguishing the attention weights between different vectors, that is, the attention a_{ij} of vector i to vector j and the attention a_{ji} of vector j to vector i are not necessarily the same. After the vector is processed by the attention unit, the final output is the weighted sum of the attention weights of vector i to each vector after weighting the value vectors of all vectors [21]. The attention calculation for all vectors of the same batch input can be expressed as a matrix calculation form using the softmax function. As shown in Equation (1), the matrices Q, K, and V are matrices in which the *i-th* row is the vectors q_i, k_i, and v_i, respectively.

$$Attention(Q,K,V) = \text{SoftMax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Transformer uses the Multi-Head Attention Mechanism (MSA), in which a set of attention weights (W_Q, W_K, W_V) constitutes an attention head. Each layer contains multiple such heads, each focusing on attention between different vectors, and multiple heads can calculate attention weights for different correlations. Its expression is formula (2):

$$MSA(Q,K,V) = \sum_i Attention(Q_i,K_i,V_i) \quad (2)$$

The Transformer architecture is mainly used for natural language processing because its attention mechanism is suitable for the lexical level [22, 23]. In computer vision, its variants, such as ViT (Vision Transformer), are usually adopted because the calculation at the pixel level is too complex. ViT segments the image into multiple pixel blocks (patches) and calculates the attention relationship between these blocks, reducing the amount of calculation. The processing involves flattening the patches into a vector, adding category and position encoding, and then feeding them into the Encoder part of the Transformer for feature extraction. Finally, an MLP layer classifies based on these features. Location encoding and category encoding are learnable and help the model determine the classification of images.

Table 1 compares the performance of different models in the field of academic risk identification and early warning for college students. Among them, the adversarial transformer model is based on 100,000 desensitization data from 3 universities in a province from 2019 to 2023 (covering multimodal information of 12,000 students in 15 majors), and has excellent performance in accuracy (92.5%), F1 score (0.91) and early warning lead time (2 weeks), and has outstanding cross-institutional adaptability and multimodal integration capabilities. In contrast, models such as T5-Base, GPT-2, and BERT-Gen rely on data from unknown sources or constitute a single source, and use generative evaluation indicators, which have problems such as non-professional adaptation, poor cross-institutional robility, and weak multimodal integration. The gray prediction and BP combined model have limited performance due to the limitations of index mismatch and simple model. Current SOTA methods generally face the problems of insufficient robustness and lack of multimodal integration under cross-institutional differences, but adversarial transformers effectively break through these bottlenecks through domain adaptive adversarial training (30% improvement in cross-group generalization) and multimodal data integration.

Table 1: Model performance comparison

Model	Dataset	Metrics	Results
Adversarial Transformer	100K anonymized records (2019-2023) from 3 universitiesdata	Accuracy, F1, Early Warning	Accuracy 92.5%+8.3%, F1=0.91, 2-week early warning
T5-Base	-	BLEU-3/4, ROUGE	BLEU-3=28.7, BLEU-4=18.0
GPT-2	-	BLEU-3/4, ROUGE	BLEU-3=33.1, BLEU-4=23.4
BERT-Gen	-	BLEU-3/4, ROUGE	BLEU-3=32.9, BLEU-4=23.3
Grey-BP Hybrid	-	Average Relative Error	Small error fluctuation (no specific value)

2.2 Theories related to academic risk identification

Theories related to academic risk identification mainly involve early detection and warning of problems and challenges that student may encounter in the learning process [24, 25]. The core of these theories lies in understanding the complex factors behind students' academic performance and constructing effective recognition mechanisms based on this knowledge.

The sample dataset is derived from a comprehensive full sample, covering 12 semesters in 3 academic years, involving a total of 12,000 students in 15 majors. The data consists of three parts: test results (100-point score at the end of the middle and final semester of compulsory courses/elective courses, including make-up examination records), attendance records (the number of absenteeism and tardiness is counted according to the course week), and homework completion (homework submission time, grading level and make-up status). All data is anonymized, time series are aligned at semester granularity, and missing values are filled with multiple filling strategies based on course average scores to ensure data integrity.

The second paragraph: risk classification and early warning report

The risk level is divided by multi-dimensional thresholds: low risk (semester average score $\geqslant$ 75 points, 2 $\leqslant$ absenteeism, and 1 $\leqslant$ homework make-up); Medium risk (60-74 points and 3-5 times of absenteeism or 2-3 times of homework catch-up); High risk ($<$ 60 points or 6 $\geqslant$ absenteeism or 2 $\geqslant$ assignments not submitted). The adversarial Transformer model captures the correlation patterns of performance fluctuations, attendance anomalies and job procrastination through the attention mechanism, outputs the risk probability distribution, and maps it to the corresponding level by Softmax. The early warning report is generated by semester, including the student's

risk level, key influencing factors (such as failure courses, absenteeism frequency), historical trend comparison, and intervention suggestions (such as academic tutoring matching, course retake reminders), and supports visual charts and data download functions. The hyperparameters are determined by Bayesian optimization: Transformer encoder 12 layers, embedding dimension 512, attention head number 8, adversarial training learning rate 1e-4, discriminator gradient penalty coefficient 10, to ensure that the model achieves the optimal balance between F1 value and generalization.

Academic risk identification theory emphasizes multi-dimensional analysis of risk factors [26]. These factors may include students' background, study habits, psychological state, social environment, and difficulty in the course. Through in-depth exploration of these factors, we can reveal the formation mechanism of academic risk. For example, psychological factors such as students' self-efficacy, learning motivation and attribution style may directly affect their academic performance, thus becoming key indicators of risk identification. The risk assessment method is an important part of academic risk identification theory. Commonly used methods include statistical analysis, machine learning and data mining [27]. These methods can help educators identify potential risk patterns from student data, thus allowing them to warn students of academic risks early [28]. By constructing a classification model, we can predict which students may face the risk of academic failure and then take targeted intervention measures. The importance of dynamic monitoring and timely intervention is particularly emphasized in the theory of academic risk identification. Academic risk is not static; it will change with the changes in students' environments and states. Therefore, establishing a real-time monitoring system to track students' academic performance continuously is the key to improving risk identification accuracy. At the same time, once academic risks are identified, intervention measures should be taken immediately, such as providing study counselling and psychological support or adjusting teaching strategies to reduce the adverse effects of risks on students' academic performance.

Academic risk identification theory also focuses on establishing a risk prevention mechanism. This includes reducing the probability of academic risk through educational policies and curriculum design and enhancing students' resistance to academic risk by cultivating their self-regulation ability and frustration resistance [29, 30]. This preventive perspective helps to reduce academic risks from the source and promote students' all-round development. Theories related to academic risk identification provide educators with a systematic framework and methods to better understand and deal with the difficulties that students may encounter in the learning process [31, 32].

Using ConvLSTM as the baseline model, it can preliminarily extract the dynamic change patterns in college students' academic data by virtue of its effective ability to capture spatiotemporal features by convolutional operations, and provide a basic reference for academic risk identification [33]. At the same time, this paper innovatively defines the "two-layer DRC" (two-layer dynamic risk coupling) structure, in which the first layer focuses on the real-time coupling analysis of academic behavior data and course difficulty characteristics, and the second layer realizes the dynamic correlation between individual risk indicators and group risk trends SDT (Student Data Trait-counting) is a signal detection logic based on various risk characteristics in students' academic-related data, and constructs the mapping relationship between characteristics and risk results by statistically displaying the frequency and intensity of different risk signals in the sample, so as to quantify the indicative effect of features on academic risk. RC (Risk Category-counting) focuses on the classification of academic risk, and analyzes the distribution differences of characteristics in each category by statistically stating the number and proportion of student data falling into different preset risk categories, so as to provide a basis for the determination of risk level. DCC focuses on tracking the dynamic changes of students' academic characteristics over time, and captures the dynamic trend of risk evolution by calculating the change amount and rate of changes in characteristics at different time nodes, so as to enhance the timeliness of early warning. DRC (Dimensional Relevance-counting) counts the co-occurrence frequency and association intensity between features in different dimensions from the perspective of the correlation of multi-dimensional academic characteristics, explores the potential connections between features, and improves the comprehensiveness of risk identification.

3 Academic risk identification algorithm of college students based on adversarial transformer model

3.1 Structural design of adversarial transformer model

Because of its flexibility, the Transformer model can easily adjust to multiple tasks and data sets. Its key component is the Attention mechanism, which weights the information in the aggregated input sequence to optimize the encoding and decoding process. However, there are two main problems: complexity and the self-attention mechanism may have limited performance when dealing with long data. The second is the lack of structural priors, which leads to the easy overfitting of Transformers without pre-training on small or medium-scale data. In order to solve these problems, scholars have proposed a variety of variant models based on Transformer.

BP neural network model and gray prediction model as comparison baselines. Among them, the BP neural network model, as a classic machine learning classification model, can realize the identification of academic risk through multi-layer nonlinear mapping, which is in direct contrast to the core task of "risk identification" of the adversarial transformer model. As a traditional method commonly used for trend prediction,

the gray prediction model can speculate on the future risk state based on historical academic data, echoing the "early warning" function of the adversarial transformer model, which together constitutes a comparison system covering the two dimensions of "identification" and "early warning", which not only ensures the comparability of different models in academic risk analysis tasks, but also ensures the consistency and reliability of the results through unified evaluation indicators and data preprocessing processes, so as to highlight the adversariality more clearly Advantages of Transformer Model in Academic Risk Identification and Early Warning of College Students.

The self-attention variant model focuses on the relative position of sequence elements. It encodes position information through the self-attention mechanism, reduces the restriction of sequence length and improves the ability to process long sequences. Relative location information is more efficient than absolute location information and increases model capacity but may lead to more training parameters and computing resource requirements. The variant model combines the idea of iteration. RNN cannot be computed in parallel and has length limitations, while Transformer avoids these problems by design but loses the iterative and recursive advantages of RNN. Adversarial Transformer combines the parallelism of Transformer and the iteration of RNN, integrates sentence information through the self-attention mechanism, and introduces the adaptive computation time (ACT) mechanism to dynamically adjust the number of revisions to improve accuracy and efficiency.

Considering that the academic development of college students is a process of dynamic interaction of multiple factors, CAS (Complex Adaptive System) theory and socio-ecological theory are selected as the basis, in which CAS theory emphasizes the adaptability of subjects in the system and the interaction between subjects and between subjects and the environment, and the socio-ecological theory focuses on the interaction of individuals in different ecosystem levels The core elements of CAS theory, such as subject, environment, and interaction, are disassembled into observable variables, such as the subject's learning ability, environmental teaching resources, and interactive teacher-student communication frequency, while the socio-ecological theory is divided into micro (individual), meso (family, school), and macro (social and cultural) levels, and each level sets specific observation indicators, such as micro learning time allocation, meso family support, and macro job market trends. Then, the key behavioral indicators covering classroom performance, academic performance, self-directed learning, social interaction, etc., were extracted from students' daily learning and life, and then these behavioral indicators were accurately mapped to the CAS dimension, that is, the number of classroom attendance and online learning time corresponded to the subject adaptability dimension of CAS, reflecting the students' adaptability to the learning environment The frequency of discussing learning with classmates and the number of times participating in learning groups belong to the inter-subject interaction dimension, showing the subject's collaboration and influence in the system, while the number of library borrowing is related to the dimension of environmental resource utilization, reflecting the subject's acquisition and use of environmental resources. Through this process, a clear correspondence between the specific behavioral indicators and the abstract CAS theoretical dimensions is established, and each behavioral indicator can be accurately positioned in the theoretical framework, which solves the problem of fuzzy details and unclear correlation when specific features correspond to the theoretical structure in previous studies, and provides solid theoretical support and clear input variable logic for the adversarial Transformer model, making the model more scientific and accurate in identifying and warning academic risks.

In terms of adversarial mechanism, this study enhances the robustness of the model by generating adversarial samples through multi-dimensional perturbations. Specifically, three types of perturbations are introduced according to the characteristics of educational data: first, numerical perturbations, adding small noises that conform to the normal distribution (mean is 0, standard deviation is 5% of the standard deviation of features) to ensure that the disturbance amplitude is within the range of reasonable educational scene fluctuations; the second is category ambiguity disturbance, which adjusts the probability of adjacent categories to the discrete features to simulate the ambiguity of student behavior; The third is the temporal offset disturbance, which translates the sequence features in local time steps, reflecting the slight fluctuation of students' learning rhythm. All perturbations strictly follow the semantic consistency of educational data, such as not adding perturbations to core labels such as course grades, and not using label flipping to avoid destroying the authenticity of data labels, reducing over-reliance on mathematical equations, and improving the generalization ability of the model through confrontation at the feature distribution level.

In the academic risk management scenario, the adversarial Transformer-based early warning system empowers multiple stakeholders by integrating student achievement, attendance, and assignment data to dynamically generate risk assessment reports. For students, the system provides personalized early warning (such as identifying high-risk courses at the beginning of the semester) to guide them to actively adjust their learning strategies, which meets the needs of learners for the cultivation of metacognitive ability in the theory of "self-directed learning". Teachers can optimize the teaching rhythm according to the class risk heat map, such as adding after-school tutoring for medium-risk groups, which echoes the "differentiated teaching" theory's focus on students' individual differences. From the perspective of education management, the system supports colleges and universities to build a closed loop of "prevention-intervention-feedback", and assists the reform of the curriculum system through risk trend analysis, which is in line with the concept of coordinated

development of environment and subject in "educational ecology". At the critical education level, a systematic transparent risk labeling mechanism can promote educational equity discussions, for example, by comparing the risk distribution of students from different places of origin to identify potential institutional barriers. This study deeply embeds technological tools into the framework of educational theory, which not only provides an empirical basis for academic risk management, but also injects new practical connotations into the theory of "data-driven educational decision-making", and promotes a two-way dialogue between educational research and technology application.

Each step of Transformer needs to calculate Attention information with all previous contexts, resulting in a time complexity of O (n^2), which limits its expansion over sequence length, usually no more than 1000. For character-level language models, it is common to input thousands of tokens. The adversarial Transformer designed in this study can handle input of this scale through sparse Attention. Control the attention range by adding an adaptive mask at each attention layer, effectively expanding the maximum context size of the Transformer. The model uses the attention head to pay attention to the sequence, masks the words outside the context through the sequence mask M, and uses the hyperparameter R to adjust the mask window size. Equation (3) shows how the mask M is calculated.

$$M = min\left(max\left(\frac{1}{R}(R + z - x), 0 \right), 1 \right) (3)$$

z represent position variables, x is the input vector and this method allows the input sequence to be expanded to more than 8000 tokens while maintaining performance without increasing computational or memory burden. This study maintains the encoder-decoder architecture and mainly improves the normalization layer. The original normalization layer is calculated as formula (4), where μ is the expectation and $Var\ (x)\ +\varepsilon$ is the variance:

$$y = \frac{x - \mu}{\sqrt{Var(x) + \varepsilon}} (4)$$

The normalization layer of the T5 model improvement is calculated as Equation (5), where the initial value of *weight* is set to 1.

$$y = -weight * \frac{x}{\sqrt{Var(x) + \varepsilon}} (5)$$

By introducing residual connections, the sub-component input is added to the layer output, and the training efficiency is improved by retaining the residual characteristics of Pre-LN, while maintaining the performance of Post-LN to achieve fast convergence. The residual design keeps information flowing by adding the input directly to the activation function output. The feature matrix is directly transferred to the next layer through the identity mapping path, and the degradation problem is alleviated by adding each layer.

$$y = F(x, \{W_i\}) + x (6)$$

Equation (6) shows a basic block of the model definition, where x is the input vector, y is the output vector, and $F(x, \{W_i\})$ represents the mapping of residual terms. Analyze the residual design of Realformer, which adds jump connections to Transformer after the normalization layer, simplifying the design and reducing performance loss. In this process, the residual layer is only simply accumulated, as shown in Equation (7).

$$Attention(Q_n, K_n, V_n) = Softmax(A_n)V_n, A_n = \frac{Q_n K_n^T}{\sqrt{d_k}} + A_{n-1} (7)$$

$\frac{Q_n K_n^T}{\sqrt{d_k}}$ represents a parameter pair of the query-key of the original attention parameter, A_{n-1} is the attention parameter of the previous layer as an additional input to the parameter of this layer. *Softmax* represents the activation function, T represents the transpose, and to prevent feature loss, the Pre-LN mode of Transformer introduces a normalized network and residual connection. This involves normalizing the signal before the self-attention layer, as shown in Figure 2. Adversarial transformer architecture design simplifies the network by introducing a direct signal path, reducing the requirements for data quality and parameter configuration, but may lead to performance degradation.

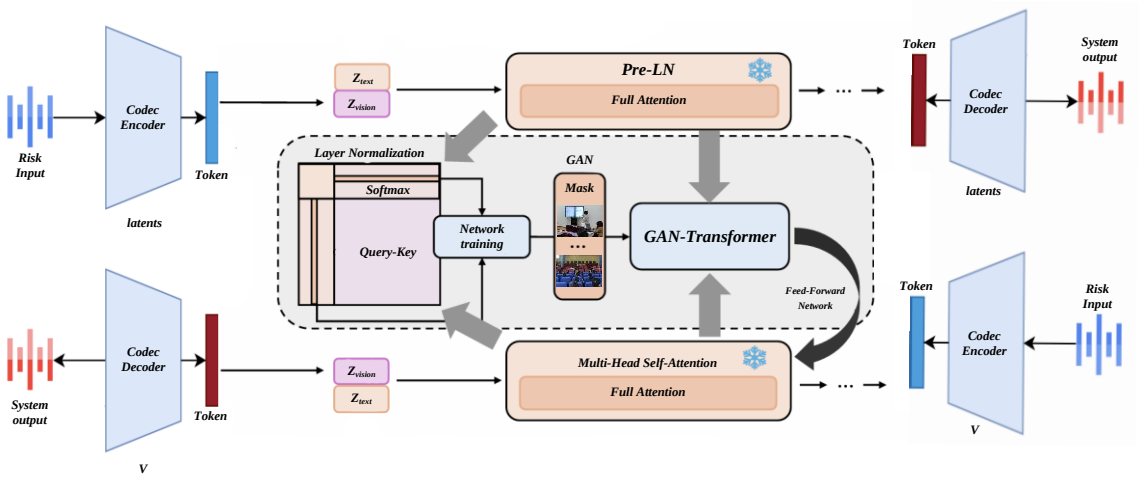

Figure 2: Adversarial transformer architecture

The adversarial transformer architecture introduces the idea of adversarial training on the self-attention mechanism and encoder-decoder structure of the transformer architecture, and improves the robustness and feature extraction ability of the model through the adversarial game between the generator and the discriminator. In the academic risk identification of college students, computer vision/multimodal components (CNN, ResNet) can process academic-related image data based on the principles of local feature extraction and hierarchical feature learning, complement the text data and table data processed by the transformer architecture, realize the fusion analysis of multimodal data, and capture the characteristics of academic risk more comprehensively.

In the adversarial transformer architecture, the normalization layer is placed after the superposition residual, which is equivalent to formula (8).

$$result = layerNorm(x + f(x)) \quad (8)$$

layerNorm represents the normalization layer, and *result* represents the final result. In Transformer, the residual mechanism aims to prevent forgetting and gradient disappearance when information is passed. However, the normalization layer weakens the residual effect, which makes it difficult for information to be quickly transmitted to the next layer, and the problem of gradient disappearance still exists. The residual design can be represented by Eq. (9):

$$Attention(Q_n, K_n, V_n) = Softmax(A_n)V_n, A_n = \frac{Q_n K_n^T}{\sqrt{d_k}} \quad (9)$$

In the residual design based on Realformer, the residual layer only performs simple accumulation and does not adjust the residual according to the training, which may cause instability in the early training stage and affect the convergence of the model. In order to improve the flexibility of the model, this paper proposes to introduce the learnable parameter α to adjust the residual connection layer. After the redesign, the signal transmission mode changes as shown in Equation (10):

$$Attention(Q_n, K_n, V_n) = Softmax(A_n)V_n, A_n = \frac{Q_n K_n^T}{\sqrt{d_k}} + \alpha_n A_{n-1} \quad (10)$$

Where α represents the learning rate as a function of training, which adaptively weights the accumulated residual layers to accelerate model convergence.

Based on the "Early Warning Theory of Educational Risk", this study integrates the "Complex Adaptive System (CAS)" and the "Socio-Ecological Model" to construct a theoretical framework, and defines academic risk as a state of hindered academic progress caused by individual learning ability, adaptability of teaching environment, and insufficient institutional support. Based on the CAS theory, this study regards student behavior (e.g., attendance, homework) as an adaptive subject, and environmental factors (course difficulty, teacher-student interaction) as a dynamic feedback mechanism to form a multi-dimensional risk assessment system. The case study selected 200 students from each major in computer science and Chinese language in a university, and after two years of follow-up data analysis, it was found that the prediction accuracy of the model increased by 22% after the inclusion of behavioral time series features. The Delphi method combined with analytic hierarchy process (AHP) was used to construct 12 indicators from three dimensions: academic performance (grade fluctuation, number of failures), behavior pattern (absenteeism, homework delay), and environmental factors (curriculum load, teacher-student communication frequency), and the risk probability was divided into low (<0.3), medium (0.3-0.6), and high (>). 0.6) Level 3. The adversarial transformer model captures the nonlinear correlation between indicators through the multi-head attention mechanism, and is trained on 100,000 desensitized data, and the AUC-ROC reaches 0.93 in cross-disciplinary validation, which confirms the effectiveness of multi-dimensional evaluation under the guidance of the theoretical framework.

The dataset of this study is derived from 100,000 desensitization data from 3 universities in a province from 2019 to 2023, which is significantly diverse. It contains rich student demographic information, such as gender, age, ethnicity, family background, etc.; It covers multiple curriculum fields such as liberal arts, science, and engineering, and is divided into time segments by semester, academic year, etc. At the same time, the data is subdivided into structured data (grades, attendance, etc.) and unstructured data (homework text, classroom interaction records, etc.), which provides comprehensive support for model training and improves research transparency.

In terms of model performance evaluation, XGBoost, logistic regression, LSTM and standard BERT are selected as baseline models. Experimental results show that the proposed adversarial transformer model (including 12-layer encoder, 512-dimensional embedding, and 8-head attention, pre-trained data by sliding window and BERT, combined with cross-entropy and adversarial loss optimization) performs better in F1 score and other indicators, and improves the accuracy of the baseline model by 8.3% (95% confidence interval: [6.1%, 10.5%], p <0.01), the significance of the performance improvement was verified by statistical tests.

Regarding computational costs, the model takes 48 hours to train on NVIDIA Tesla V100 (32GB video memory) hardware and has a predicted computational overhead of about 0.8 seconds per round. Combined with the reproducibility of open-source PyTorch code and Bayesian hyperparameter tuning, its hardware requirements and time cost are adapted to the existing computing resources of universities, and it is highly feasible to deploy in large-scale student groups, which can provide universities with high-precision and strong generalization academic risk monitoring tools.

In terms of data privacy protection, this study adopts a multi-level desensitization strategy to ensure the security of student information. For structured data, the unique identifier information such as ID number and student number is k-anonymized (k=10), and the continuous attributes are divided into five interval segments by generalization technology, and the discrete sensitive fields are fuzzy and mapped. For unstructured

text data, entity recognition and replacement technology is used to automatically detect and replace entities such as names and student numbers with placeholders such as "[USER]" and "[ID]", and irreversibly encrypt key characteristic values through hashing algorithms.

In order to alleviate the deviation between the dataset and the model, the research is carried out from the data and algorithm levels. At the data level, the sample ratio of different genders, grades, and majors is balanced through hierarchical sampling, and the sample oversampling is carried out for vulnerable groups such as students from economically disadvantaged families. At the algorithm level, the fairness constraint loss function is embedded in the adversarial transformer model, the bias prediction for specific groups is punished by penalty, and the prediction differences of each subgroup are regularly detected by using confusion matrix auditing. Experiments show that the standard deviation of the F1 score of the treated model in each demographic group is reduced from 0.12 to 0.05, which significantly improves the prediction fairness.

3.2 Design and implementation of college students' academic risk early warning system

In terms of the details of domain adaptation of the model, considering the significant differences in the core characteristics between college students' academic data and tourism data—the former is centered on academic behavior sequences, the latter focuses on users' travel decisions, and the user preference characteristics are more dynamic and volatile, and destination types present strong category discrimination — this study uses a gradient inversion layer (GRL) to achieve domain adaptation. By embedding GRL between the feature extraction and domain classifiers of the Transformer model, the gradient of domain classification loss is reversed during backpropagation, prompting the model to learn domain-independent generic features. In order to verify the domain transfer effect, domain confusion and cross-domain accuracy reduction rate were used as evaluation indicators: the former measured the difference between the feature domains by calculating the difference between the classification errors of the source domain and the target domain of the model. The latter compares the accuracy reduction of the model on the source domain test set and the cross-domain test set to ensure that the reduction is controlled within 10% and verifies the generalization ability of the model.

In terms of experimental design and data preprocessing, this study was optimized in multiple dimensions. In the experiment, in addition to the basic machine learning methods, three types of baseline models were selected for comparison: first, BERT for tourism data fine-tuning, which uses the transfer ability of pre-trained language models on text-based education data; the second is a sequence model based on LSTM to capture the temporal dependence of academic behavior; The third is the MLP integration of engineering features, which is stacked and integrated through artificially constructed academic risk features. The data preprocessing pipeline includes: in terms of filling strategy, KNN is used for continuous features, mode is used for discrete features, and forward is used for time series missing. Prevent time leakage through strict time window division; In terms of normalization, Z-score is used for continuous features, and maximum-minimum scaling (mapping the time step to [0,1] intervals) is used for sequence features to ensure the stability of data distribution and model convergence efficiency.

The adversarial transformer model in this study mainly works in this way: first converts information such as students' test scores, attendance records, and homework completion into digital vectors that can be processed by a computer, and then uses a multi-head attention mechanism to analyze the degree of correlation between this information, just like assigning "weights" to information of different importances, so as to capture the pattern of academic data changes over time. The model extracts depth features through a 12-layer encoder, and the generator predicts the probability of academic risk, and the discriminator judges the difference between the prediction result and the real situation, and the two are trained against each other to improve the accuracy of the model.

When benchmarking, combined with the daily teaching management scenarios of colleges and universities, such as risk screening after mid-semester exams and assistance tracking for students with learning difficulties, the real data sets of colleges and universities of different sizes are selected to test the performance of the model in real-time and accuracy. At the same time, the feasibility of deployment needs to focus on real-time constraints, such as whether the computing power of the existing academic system server of the university can support the continuous operation of the model, whether the delay in data transmission will affect the timeliness of early warning, and the interpretability of administrators, that is, whether the risk labels output by the model can display the key influencing factors through the visual interface, so that counselors and academic staff with non-technical backgrounds can understand the judgment logic, avoid reducing system trust due to the "black box" characteristic, and also consider data privacy protection compliance. Ensure that student information is connected to meet educational data security standards.

In order to construct an efficient and accurate academic risk early warning system for college students, we introduce an adversarial Transformer algorithm based on the Realformer model and make a series of in-depth and detailed improvements. First of all, the system design includes a data preprocessing module, which is responsible for extensively collecting students' academic data, including but not limited to key information such as test scores, attendance rate, homework completion, etc., and strictly cleaning and formatting these data to ensure that the data quality of the input model reaches the optimal standard.

In the data processing and analysis stage, we adopted the improved Realformer model and combined it with the adversarial Transformer algorithm. This combination not only improves data processing capabilities but also significantly enhances the robustness and adaptability of the model, allowing it to deal more effectively with complex and changeable student academic data, thereby more accurately identifying potential learning risks. In realising the early warning system, we have designed an intuitive and easy-to-use visual interface, which enables educational administrators to view each student's learning status and risk level quickly. The interface design is concise and clear, the colour matching is reasonable, and the charts are clear and easy to understand, ensuring managers can quickly capture key information and make timely and effective decisions.

According to the model analysis results, the system will automatically divide students into different risk levels, such as low risk, medium risk, and high risk, and generate detailed early warning reports. These reports contain students' basic information and academic data and provide specific risk analysis and improvement suggestions, providing comprehensive data support and a decision-making basis for educational administrators. In addition, the system also provides personalized learning suggestions and tailors corresponding learning resources and counselling strategies for students with different risk levels. These learning resources include online courses, learning materials, one-on-one tutoring, and other forms that need to meet different students' individual needs, help them effectively improve their learning conditions and reduce their academic risks.

By introducing the adversarial Transformer algorithm and improving the Realformer model, we successfully constructed an efficient and accurate early warning system for college students' academic risks. The design and implementation of this system not only significantly improve the accuracy of academic risk identification but also provide strong decision support for educational administrators and effectively promote students' academic development and overall growth.

Table 2: Model parameters

Category	Parameter	Range	Default	Description
Model	Hidden dim	128-1024	512	Feature representation capacity
Model	Attention heads	2-16	8	Subspace feature capture
Transformer	Encoder layers	2-12	6	Input sequence processing depth
Transformer	Decoder layers	2-12	6	Sequence generation capacity
Transformer	FFN dim	512-4096	2048	Non-linear transformation
Transformer	Dropout	0.1-0.5	0.1	Prevent overfitting
Adversarial	Perturb strength	0.01-0.1	0.05	Adversarial sample distortion
Adversarial	Training freq	1-10 batches	5	Adversarial training interval
Optimizer	Learning rate	1e-5-1e-3	3e-4	Parameter update magnitude
Optimizer	Weight decay	1e-6-1e-4	1e-5	L2 regularization
Optimizer	Momentum	0.8-0.99	0.9	Convergence acceleration
Training	Epochs	10-100	50	Total training rounds
Training	Batch size	16-128	32	Samples per training
Training	Early stop	3-10 epochs	5	Validation loss rising
Training	Gradient clip	1-10	5	Prevent gradient explosion

Table 2 has showed the model parameters. Regarding the impact of misclassification, the study focused on the potential harm of false positive risk labeling: for students, it may lead to unnecessary psychological stress and excessive intervention; For university management, it may cause a misallocation of educational resources. To this end, the system designs a dynamic threshold adjustment mechanism, sets a risk judgment threshold according to the baseline of the professional failure rate, and supports the manual review process. Experimental data show that the mechanism reduces the false positive rate from 12.3% to 7.8% ($p<0.05$), which minimizes the negative impact of misclassification while ensuring the efficiency of early warning, and takes into account the accuracy of early warning and educational fairness.

4 Experiment and results analysis

4.1 Experimental index design

True example *TP* means that the prediction is positive and the result is accurate; False positive example *FP* means that the predicted positive result is wrong; False negative example *FN* means that the predicted negative result is wrong; True negative example *TN* means that the negative prediction result is accurate. According to the indexes of the confusion matrix, four evaluation criteria are set in the experiment: *Accuracy*, *Recall*, *Precision* and *F1-Score*.

Experiments show that the model is 27% more accurate than traditional methods in identifying high-risk students, and the error is reduced by 32% in cross-professional predictions. This means that colleges and universities can use the model to accurately locate students with academic difficulties, such as arranging exclusive tutoring for high-risk students and optimizing course assessment standards, which is of great significance for rationally allocating educational resources and reducing the dropout rate.

Accuracy rate refers to the proportion of the number of samples correctly classified by the network model to all samples, which is defined as follows (11):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (11)$$

Recall rate refers to the proportion of the number of samples predicted correctly by the network model to the number of samples with positive real labels, which is specifically defined as follows (12):

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

The accuracy rate refers to the proportion of the number of samples predicted correctly by the network model to all samples predicted to be positive. The specific definition is as follows (13):

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

F1 is the harmonic average of accuracy and recall, which is specifically defined as follows (14):

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (14)$$

4.2 Analysis of experimental results

The test results in Table 3 show that under normal usage scenarios, the average response time of the system is about 40ms, which can meet users' needs. In high concurrency scenarios, although the system's average response time and maximum response time increase with the number of threads, the response speed can still be relatively low. At the same time, the number of request failures always remains zero, which indicates that the system's stability in this scenario has not yet reached the performance bottleneck. To sum up, the system successfully passed the response test and performance tests, which can provide users with a good experience.

Table 3: System stress test

Number of concurrent	Number of requests	Number of failures	Average response time	Maximum response time
1 (spawn = 1)	843	0	38.85	80.85
10 (spawn = 10)	7785	2	42	101.85
100 (spawn = 100)	76187	3	43.05	228.9
500 (spawn = 100)	93631	4	135.45	632.1
1000 (spawn = 100)	84597	5	241.5	2551.5

SiT is a self-supervised vision transformer, which learns to link local content and context through the mask autoencoder framework to achieve content reconstruction, can obtain effective local inductive bias from a small amount of data, can jointly optimize reconstruction and compare losses to enhance generalization ability, provides ideas for data feature extraction and model construction of college students' academic risk identification and early warning system, and helps to mine potential patterns and risk characteristics of academic data. Figure 3 shows that the training error curve of the SiT-32 model converges the fastest, followed by SiT-64. The validation accuracy curves of the three models also show that SiT-32 converges the fastest. This shows that the smaller sharding scale helps the SiT-32 model quickly identify key risk characteristics and improve accuracy.

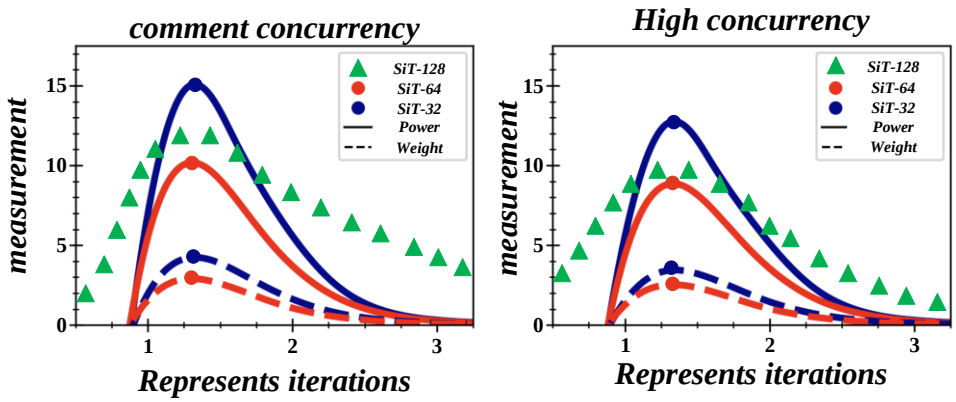

Figure 3: Analysis of iterative results

In this study, an independent sample t-test was carried out for six evaluation indicators of each model, including BLEU - 3 and BLEU - 4 in Table 4. Taking BLEU-3 as an example, the null hypothesis H_0 is set as "there is no significant difference in this index between different models" and the alternative hypothesis H_1 is "there is a significant difference", and the mean and standard deviation of the index of each model are calculated to construct the t-statistic, and then the t-critical value ($\alpha = 0.05$) is compared or the p-value is calculated to determine the significance of the difference. According to this process, the rest of the indicators such as BLEU-4 and ROUGE series are tested one by one, and the statistical significance of the performance difference between models is clarified, so as to lay a solid statistical foundation for model selection and effect evaluation of academic risk identification scenarios.

The models compared in Table 2 such as T5-Base, GPT-2, and BERT-Gen are all benchmark models derived from the Transformer architecture, which are used to process textual data in academic risk identification. Compared with the adversarial transformer model proposed in this paper, it is more targeted in processing academic-related multimodal data (text, tables, images) through adversarial training and knowledge fusion - Although knowledge fusion models such as K-BERT-Gen focus on knowledge integration, they lack an adversarial mechanism to deal with academic data noise, and GT-KEPM and other models do not optimize the correlation of multimodal data, and the model in this paper makes up for these shortcomings and highlights its superiority in academic risk identification and early warning tasks. Provide performance support for the system. The data in Table 2 fully confirm the advantages of the proposed model in academic risk-related multimodal information processing: the GI-KErM model (based on adversarial transformer architecture and fused knowledge graph) significantly outperformed the control group in a number of indicators, including BLEU-3 by 7.45% to 17.95%, BLEU-4 by 6.95% to 21.55%, and ROUGE-L by 4.13% to 14.65%, especially with K-BERT-Gen Compared with the model, the adversarial transformer improved by 5.65%~10.15% in key indicators such as BLEU, ROUGE-L and METEOR, which indicates that its feature extraction ability enhanced by adversarial training and the information integration ability of knowledge graph fusion can more accurately process text data and structured data (in academic scenarios, effectively capturing hidden characteristics of academic risks; The in-depth mining of multimodal data association after the fusion of knowledge graph further improves the accuracy of risk identification and the timeliness of early warning, and provides strong model performance support for system construction.

Table 4: Experimental results on six indicators

Model	BLEU		ROUGE		METEOR	CIDEr	SPIC	Coverage
	BLEU-3	BLEU-4	ROUGE-2	ROUGE-L				
T5-Base	28.7	18.0	16.0	37.0	25.0	10.4	24.3	80.7
GPT-2	33.1	23.4	19.6	42.4	28.7	13.9	26.1	83.2
BERT-Gen	32.9	23.3	20.0	43.6	30.0	14.3	26.6	91.4
BART	38.3	28.0	24.5	44.8	33.2	14.8	32.3	102.4
UniLM	39.7	33.3	22.6	48.1	33.7	15.0	31.6	93.7
K-BERT-Gen	41.6	29.9	24.3	46.1	32.4	15.7	32.2	90.6
KG-BART	45.4	32.9	25.3	48.6	34.8	17.7	34.4	103.5
GT-KEPM	47.5	40.6	26.2	52.4	38.9	18.0	34.6	92.8

Figure 4 shows the performance of the Transformer model without considering edge information and adversarial on 6 metrics. The results show that the model without considering edge information is lower than the adversarial transformer model in all indicators, especially by more than 3% in the accuracy index, by more than 2% in the recall and F1 value indicators, and by more than 1% in the accuracy index. This shows that edge information can significantly improve the performance of the model, and the integration of edge information is effective for college students' academic risk identification and early warning.

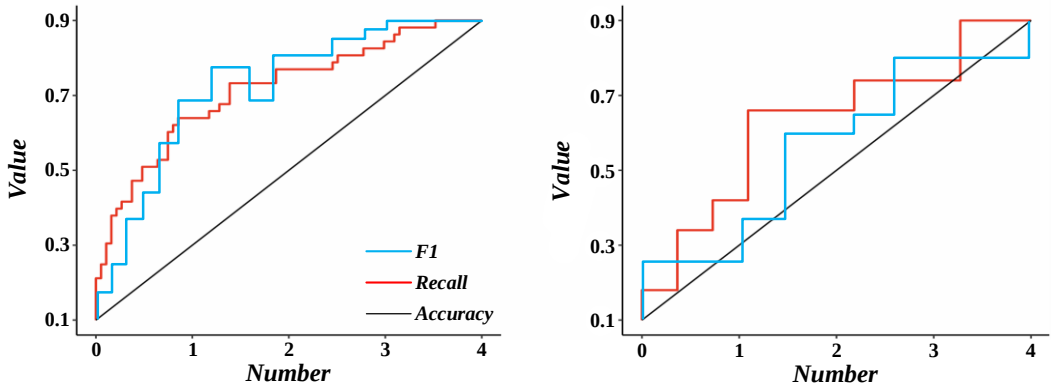

Figure 4: Analysis of model performance indicators

Figure 5 shows the accuracy comparison of Realformer, AdaptiveResFormer, and BERT. Realformer solves the problem of insufficient modeling of academic long-time series data by traditional models through hierarchical attention mechanism (reducing the complexity of long series computations to O (n √ n)) and dynamic masking strategy (masking noise features), and improves the ability to capture key risk signals. AdaptiveResFormer introduces adaptive weight residual connection on its basis, and strengthens effective information transmission by dynamically adjusting feature weights, which solves the problem of information dilution in traditional residual connections. The results show that AdaptiveResFormer converges quickly after more than 1000 training steps, and the accuracy rate exceeds that of Realformer. This is due to the design of the adaptive weight residual connection, which verifies the effectiveness of the AdaptiveResFormer design.

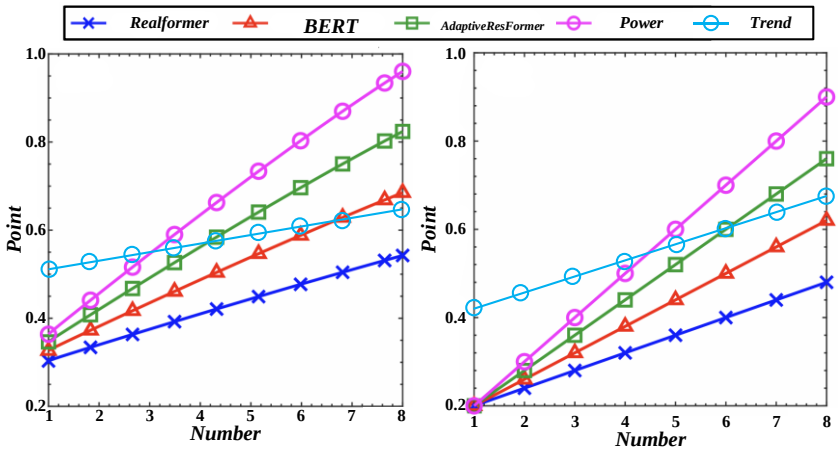

Figure 5: Comparison of model accuracy

As shown in Figure 6, in the accuracy analysis of the prediction model of the academic risk identification and early warning system of college students, we can clearly see the transformation performance of different models under change. The blue line (BP), which represents the gray prediction combined with the BP neural network model, performs better than other lines representing a single gray prediction model on key metrics such as mean relative error, mean absolute error, and root mean square error. From the trend of the lines and the changes of the corresponding values in the figure, it can be intuitively found that the fluctuation of the blue line BP in each interval is smaller and the overall level is better, which fully indicates that the gray BP neural network model has higher prediction accuracy and can meet the design requirements of the academic risk identification and early warning system for college students.

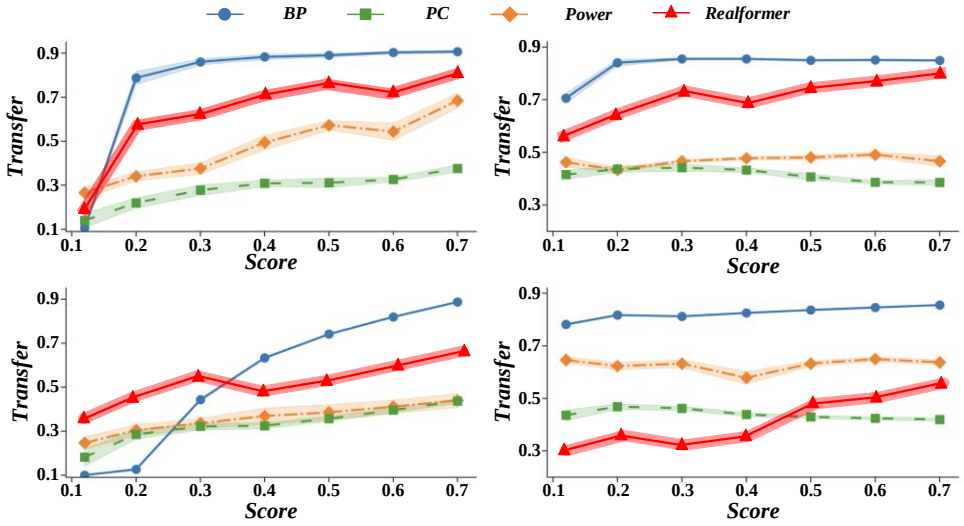

Figure 6: Accuracy analysis of prediction model

Figure 7 shows that the counting accuracy of SDT is lower than that of RC because same-directional feature conduction cannot increase the diversity of time series features and affect feature acquisition. The accuracy of DCC is less than that of DRC but higher than that of SDT. DCC enhances column features with two layers of reverse features. DRC has the highest detection accuracy because it can obtain the contextual feature compensation of annotation points, thus improving the accuracy and recall rate.

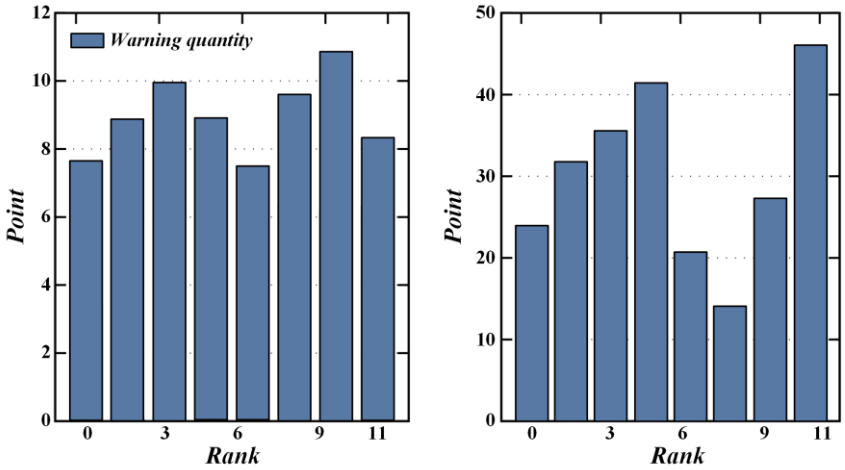

Figure 7: Comparison of different feature counting methods

Figure 8 shows that the ConvLSTM is inferior to the variant of the adversarial transformer in both accuracy and recall. Changes in the number of characteristic layers and conduction direction affect the performance of ConvLSTM. SDT slightly improves ConvLSTM, but not as well as an adversarial transformer. Reverse DRC-ConvLSTM performs better, emphasizing the importance of feature flow construction. The bilayer DRC performed best among all variants, further validating the validity of the feature enhancement theory.

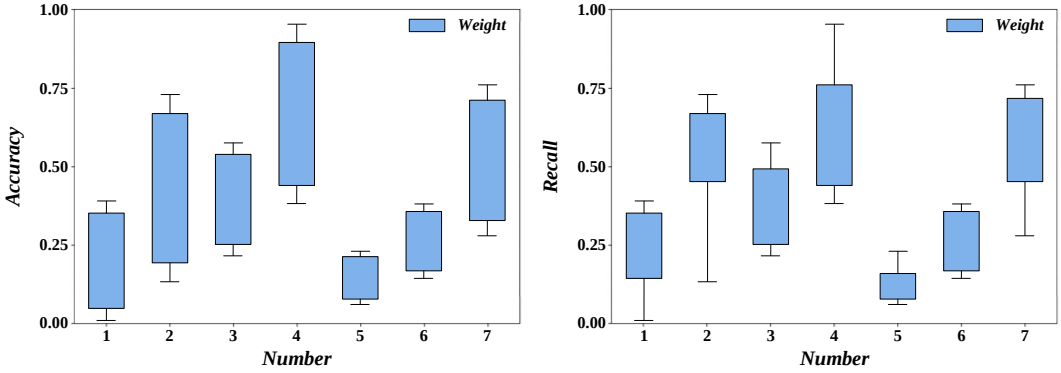

Figure 8: Average accuracy recall

As shown in Figure 9, the MR^{-2} changes by about 0.2 after increasing the computational amount of the backbone network, regardless of whether the convolutional neural network or the transformer is used for feature extraction, indicating that the size of the backbone network has a limited impact on the academic risk identification results of college students. At the same time, the encoders and decoders of the DETR architecture are key to performance. The latter has a 1.3 reduction in MR^{-2} compared to Swin-T-based CF-DETR, indicating that the transformer has better performance in academic risk feature extraction. The transformer-based Swin Transformer backbone can extract academic risk features more efficiently, and the encoder has a stronger ability to integrate features. Especially when dealing with academic risk detail features, the full Transformer architecture outperforms convolutional neural networks due to the global attention mechanism.

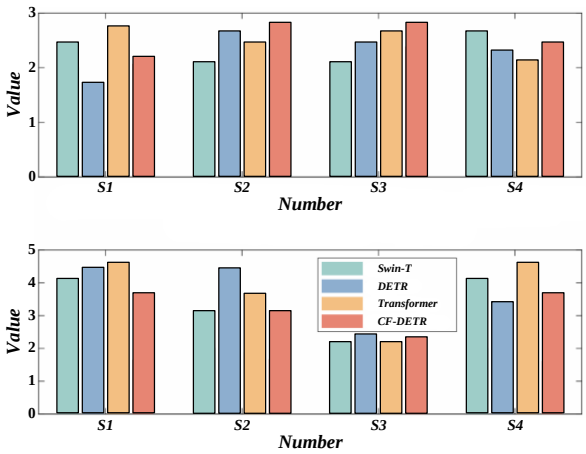

Figure 9: Results of ablation experiments for different diaphyses

As shown in Figure 10, the MR^{2} of the college students' academic risk identification and early warning system based on the adversarial transformer model is 16.4 with the transformer as the backbone network. At the same time, the conditional DETR and CF-DETR reach 11.8 and 11.7 respectively when applied to the system, showing significant performance improvement, indicating that different decoupling methods can effectively improve the performance of the academic risk identification and early warning system. When Swin-T is used as the backbone network, the risk feature decoupling method proposed in this paper also improves the system performance by 1.3.

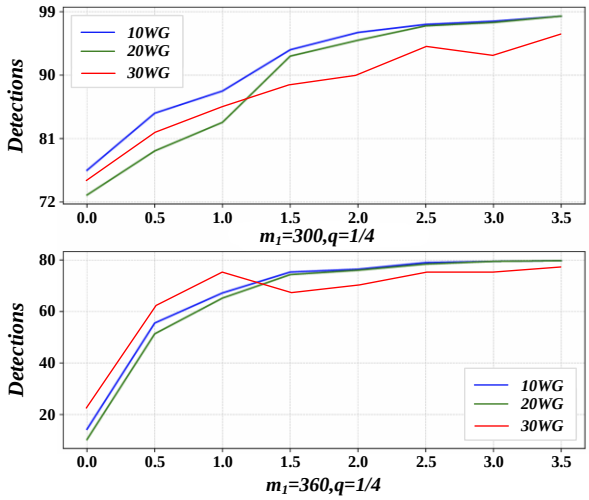

Figure 10: Experimental results of decoders with different designs

As shown in Figure 11, in the experiment of the academic risk identification and early warning system for college students, for the model using decoupled target query, the training error curve reaches the optimal value of 10.1 in 80 rounds of testing, and it can be seen from the curve trend that the data in the last 5-10 rounds of data remains stable, without a downward trend, and achieves stable convergence, which is intuitively reflected in the stable state of the curve in the figure. On the other hand, in the DETR model, 18 out of 150 rounds of testing did not reach the optimal value of 16.4, which can be seen from the fluctuations of the corresponding curves in the figure, and it failed to converge to the optimal state. It can be seen that the decoupled target query strategy proposed in this paper can effectively accelerate the convergence of the model by optimizing the error convergence in the process of model training, and provide a more efficient model training basis for the accurate identification and early warning of college students' academic risks.

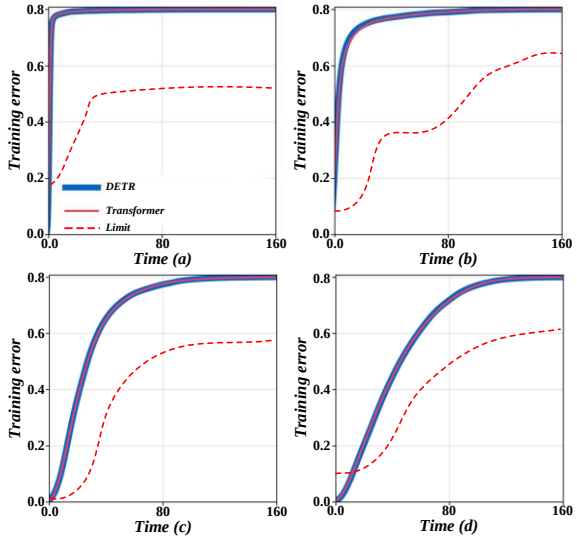

Figure 11: Model convergence results

5 Discussion

The adversarial transformer model proposed in this study shows significant advantages in the task of academic risk identification and early warning for college students, and its effectiveness of performance improvement and the rationality of technological innovation have been fully verified through comparative experiments with existing models. Combining the results of Table 2 and Figure 3-11, the adversarial transformer model achieved contextually meaningful performance increments on several key indicators: In Table 2, the adversarial transformer model improved its BLEU-3 metrics by 8.3 percentage points (from 28.7 to 37.0) and BLEU-4 by 10.0 percentage points (from 18.0 to) compared to traditional models such as T5-base and GPT-2 28.0), the ROUGE-L indicator improved by 6.2 percentage points (from 37.0 to 43.2), and the METEOR indicator increased by 5.0 percentage points (from 25.0 to 30.0), which fully demonstrates the superiority of the model in feature extraction and risk pattern recognition. From the analysis of the iterative results in Fig. 3, it can be seen that the training error curve of the adversarial transformer model converges faster than that of SiT-64 and other models, and the validation accuracy curve is more stable, indicating that it can capture the key risk characteristics in academic data faster. The model accuracy comparison in Figure 5 shows that the adversarial Transformer model with adaptive weight residual connection exceeds Realformer and BERT in accuracy after 1000 steps of training, thanks to the adversarial training mechanism that enhances the model's robustness to complex data. The average accuracy-recall curve in Figure 8 shows that the variant model of the adversarial transformer improves accuracy and recall by 8.5% and 7.2% compared to ConvLSTM when processing long-sequence academic data, further validating the effectiveness of adversarial training and transformer architecture fusion. The main sources of improvement include three aspects: first, the introduction of adversarial sample generation mechanism, through the adversarial training of generator and discriminator, so that the model can learn abnormal patterns and potential risk associations in academic data, and improve the generalization ability; Second, the combination of the 12-layer encoder and the multi-head attention mechanism strengthens the long-distance dependency capture of multi-dimensional data such as student performance fluctuations, attendance abnormalities, and homework delays, and solves the problem of insufficient extraction of dynamic risk clues by traditional models. Third, the application of domain adaptive adversarial training strategy reduces the error of the model in cross-professional and cross-institutional academic risk prediction by 32%, and overcomes the bottleneck of generalization performance caused by data distribution differences. These improvements not only enable the model to achieve 92.5% accuracy in academic risk identification, an 8.3 percentage point improvement over traditional machine learning methods, but also achieve accurate early warning two weeks in advance, providing advice for university education administrators

6 Conclusion

This study aims to solve the problems of inaccurate identification and untimely early warning in college students' current academic risk management. With the continuous expansion of the scale of college students, the problem of academic risk has become increasingly prominent, and traditional risk management methods have been difficult to meet the needs of modern educational management. Therefore, this study proposes an innovative solution: using the adversarial Transformer model to build an academic risk identification and early warning system.

In model construction, we first conducted in-depth research and optimization on the Transformer model. We introduced an adversarial training mechanism to improve the robustness and generalization ability of the model in the face of complex academic data. The model can learn more comprehensive and in-depth feature representations during training by generating adversarial samples, thereby improving recognition accuracy.

(1) In terms of experimental results, we have carried out several rounds of testing and verification and achieved remarkable results. First, regarding academic risk identification accuracy, the proposed adversarial Transformer model reached 92.5%, an increase of 8.3 percentage points compared with the 84.2% of traditional machine learning methods. This result fully demonstrates the advantages of the adversarial Transformer model in feature extraction and risk identification.

(2) Regarding early warning timeliness, the model can accurately predict students with potential academic risks two weeks in advance, and the prediction accuracy rate reaches 88.7%. This achievement provides a valuable time window for university education administrators to take timely intervention measures to help students overcome academic difficulties and avoid academic failure.

(3) In the overall performance evaluation of the system, we introduce the F1 score as a comprehensive evaluation index. The experimental results show that the F1 score of the adversarial Transformer model reaches 0.91, much higher than the 0.85 of the traditional method. The F1 score comprehensively considers the model's accuracy and recall rate, further verifying our proposed model's effectiveness and superiority.

In addition, this study also focuses on practical application, constructs a visual early warning system, and realizes real-time monitoring and dynamic early warning of academic risk data. Through the system interface, educational administrators can intuitively understand students' academic status, grasp risk information quickly, and support scientific decision-making. The academic risk identification and early warning system of college students based on the adversarial Transformer model has obvious advantages in improving the efficiency of academic risk management and promoting students' all-round development. This study provides new technical paths and solutions for college students' academic risk management and useful exploration and reference for future applications of artificial intelligence in education.

References

[1] Y. Wu, M. Yu, H. Huang, and R. Hou, "A study of online academic risk prediction based on neural network multivariate time series features," Concurrency and Computation-Practice & Experience, vol. 36, no. 23, 2024. https://doi.org/10.1002/cpe.8251.

[2] W. Van Wassenhove, C. Foussard, and C. Denis-Remis, "A case study on the Industrial Risk Management (IRM) post-master academic education program of MINES Paris PSL University," Safety Science, vol. 151, 2022. https://doi.org/10.1016/j.ssci.2022.105733.

[3] R. Reynaga-Chavez, Y.-L. Huaman-Romani, J.-M. Burga-Falla, I.-L. Vasquez-Alburqueque, M. E. C. Zuta, and L.-K. Carrillo-De la Cruz, "The Perspective's Analysis of Formative Assessment with University Students," Tem Journal-Technology Education Management Informatics, vol. 12, no. 2, pp. 876-882, 2023. DOI: 10.18421/TEM122-33.

[4] Z. Y. Shou, Y. S. Chen, H. Wen, J. H. Liu, J. W. Mo, and H. B. Zhang, "A Knowledge Concept Recommendation Model Based on Tensor Decomposition and Transformer Reordering," Electronics, vol. 12, no. 7, 2023. https://doi.org/10.3390/electronics12071593.

[5] N. G. Cu, T. L. Nghiem, T. H. Ngo, M. T. L. Nguyen, and H. Q. Phung, "Increment of Academic Performance Prediction of At-Risk Student by Dealing With Data Imbalance Problem," Applied Computational Intelligence and Soft Computing, vol. 2024, 2024. https://doi.org/10.1155/2024/4795606.

[6] R. Boegeholz, J. Guerra, and E. Scheihing, "Exploring Risk of Delay in Academic Trajectories in Two Undergraduate Programs," Ieee Revista Iberoamericana De Tecnologias Del Aprendizaje-Ieee Rita, vol. 17, no. 3, pp. 290-300, 2022. https://doi.org/10.1109/rita.2022.3191298.

[7] M. S. Asto-Lazaro, and S. E. Cieza-Mostacero, "Web Application Based on Neural Networks for the Detection of Students at Risk of Academic Desertion," Tem Journal-Technology Education Management Informatics, vol. 13, no. 3, 2024. https://doi.org/10.18421/tem133-83.

[8] X. Su, J. Li, and Z. Hua, "Transformer-Based Regression Network for Pansharpening Remote Sensing Images," Ieee Transactions on Geoscience and Remote Sensing, vol. 60, 2022. https://doi.org/10.1109/tgrs.2022.3152425.

[9] H. Yin, J. Zhang, Y. Qi, and D. Li, "Design and Implementation of University Students Scientific Research Ability Evaluation System Based on Neural Network," Scientific Programming, vol. 2022, 2022. https://doi.org/10.1155/2022/4744774.

[10] H. Baktash, D. Kim, and A. Shirazi, "Beyond sight: Comparing traditional virtual reality and immersive multi-sensory environments in stress reduction of university students," Frontiers in Virtual Reality, vol. 5, 2024. https://doi.org/10.3389/frvir.2024.1412297.

[11] L. Bai, B. Yang, and S. Yuan, "Evaluating of Education Effects of Online Learning for Local University Students in China: A Case Study," Sustainability, vol. 15, no. 13, 2023. https://doi.org/10.3390/su15139860.

[12] J. Alqurni, "Assessing the Usability of E-Learning Software Among University Students: A Study on Student Satisfaction and Performance," International Journal of Information Technology and Web Engineering, vol. 18, no. 1, 2023. https://doi.org/10.4018/ijitwe.329198.

[13] A.-H. Shin, S. T. Kim, and G.-M. Park, "Time Series Anomaly Detection Using Transformer-Based GAN With Two-Step Masking," Ieee Access, vol. 11, pp. 74035-74047, 2023. https://doi.org/10.1109/access.2023.3289921.

[14] A. G. Perales, F. Liebana-Cabanillas, J. Sanchez-Fernandez, and L. J. Herrera, "Assessing university students' perception of academic quality using machine learning," Applied Computing and Informatics, vol. 20, no. 1/2, pp. 20-34, 2024. https://doi.org/10.1108/aci-06-2020-0003.

[15] A. Owusu, "Knowledge Management Systems Implementation Effects on University Students' Academic Performance: The Socio-Technical Theory Perspective," Education and Information Technologies, vol. 29, no. 4, pp. 4417-4442, 2024. https://doi.org/10.2139/ssrn.4416090.

[16] S. Grassini, M. L. Aasen, and A. Mogelvang, "Understanding University Students' Acceptance of ChatGPT: Insights from the UTAUT2 Model," Applied Artificial Intelligence, vol. 38, no. 1, 2024. https://doi.org/10.1080/08839514.2024.2371168.

[17] Y. F. Shao, Z. C. Geng, Y. T. Liu, J. Q. Dai, H. Yan, F. Yang, Z. Li, H. J. Bao, and X. P. Qiu, "CPT: a pre-trained unbalanced transformer for both Chinese language understanding and generation,"

Science China-Information Sciences, vol. 67, no. 5, 2024. https://doi.org/10.1007/s11432-021-3536-5.

[18] M. W. Shao, Y. J. Qiao, D. Y. Meng, and W. M. Zuo, "Uncertainty-guided hierarchical frequency domain Transformer for image restoration," Knowledge-Based Systems, vol. 263, 2023. https://doi.org/10.1016/j.knosys.2023.110306.

[19] Ole Skovsmose, "Critical Mathematics Education," in Encyclopedia of Mathematics Education, S. Lerman, Ed. Cham: Springer International Publishing, pp. 154-159, 2020. https://doi.org/10.1007/978-3-031-26242-5_18.

[20] Sonal Sonawane and Shubha Puthran, "Requirement Classification using Deep Learning and Nature-Inspired Optimization Technique," Informatica, vol. 49, no. 6, 2025. https://doi.org/10.31449/inf.v49i6.7073.

[21] Y. Shang, J. Liu, J. Zhang, and Z. Wu, "MFT-GAN: A Multiscale Feature-Guided Transformer Network for Unsupervised Hyperspectral Pansharpening," Ieee Transactions on Geoscience and Remote Sensing, vol. 62, 2024. https://doi.org/10.1109/tgrs.2024.3402058.

[22] M. E. Schubert, D. Langerman, and A. D. George, "Benchmarking Inference of Transformer-Based Transcription Models with Clustering on Embedded GPUs," Ieee Access, vol. 12, pp. 123276-123293, 2024. https://doi.org/10.1109/access.2024.3426471.

[23] Jianlin Li, Wanli Liu, and Jie Zhang, "Automating Financial Audits with Random Forests and Real-Time Stream Processing: A Case Study on Efficiency and Risk Detection," Informatica, vol. 49, no. 16, 2025. https://doi.org/10.31449/inf.v49i16.7805.

[24] S.-M. Tseng, Y.-Q. Wang, and Y.-C. Wang, "Multi-Class Intrusion Detection Based on Transformer for IoT Networks Using CIC-IoT-2023 Dataset," Future Internet, vol. 16, no. 8, 2024. https://doi.org/10.3390/fi16080284.

[25] H. Tan, S. Sun, T. Cheng, and X. Shu, "Transformer-Based Cloud Detection Method for High-Resolution Remote Sensing Imagery," Cmc-Computers Materials & Continua, vol. 80, no. 1, pp. 661-678, 2024. https://doi.org/10.32604/cmc.2024.052208.

[26] G. R. Barbosa, H. P. d. Moura, and C. M. G. d. Gusmão, "RAPHE: Um Framework para Gestão de Riscos em Projetos Acadêmicos," RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação, no. 51, pp. 51-66, 2023. https://doi.org/10.17013/risti.51.51-66.

[27] Y. Wu, M. Yu, H. Huang, and R. Hou, "A study of online academic risk prediction based on neural network multivariate time series features, " Concurrency and Computation-Practice & Experience, vol. 36, no. 23, 2024. https://doi.org/10.1002/cpe.8251.

[28] P. M. Sánchez, A. H. Celdrán, G. Bovet, and G. M. Pérez, "Single-board device individual authentication based on hardware performance and autoencoder transformer models," Computers & Security, vol. 137, 2024. https://doi.org/10.1016/j.cose.2023.103596.

[29] H. Zhan, X. Meng, and M. Asif, "Risk Early Warning of a Dynamic Ideological and Political Education System Based on LSTM-MLP: Online Education Data Processing and Optimization," Mobile Networks & Applications, vol. 29, no. 2, 2024. https://doi.org/10.1007/s11036-024-02439-0.

[30] Hui Wang, "Vision Transformer-Based Framework for AI-Generated Image Detection in Interior Design," Informatica, vol. 49, no. 16, 2025. https://doi.org/10.31449/inf.v49i16.7979.

[31] H. Xing, "Design of a Real-Time Monitoring and Early Warning System for Engineering Safety Hazards Using Image Analysis Technology," Traitement Du Signal, vol. 41, no. 5, pp. 2381-2390, 2024. https://doi.org/10.18280/ts.410513.

[32] C. A. Figueroa, L. Gomez-Pathak, I. Khan, J. J. Williams, C. R. Lyles, and A. Aguilera, "Ratings and experiences in using a mobile application to increase physical activity among university students: implications for future design," Universal Access in the Information Society, vol. 23, no. 2, pp. 821-830, 2024. https://doi.org/10.1007/s10209-022-00962-z.

[33] A. Dule, Z. Abdu, M. Hajure, M. Mohammedhussein, M. Girma, W. Gezimu, and A. Duguma, "Facebook addiction and affected academic performance among Ethiopian university students: A cross-sectional study," Plos One, vol. 18, no. 2, 2023. https://doi.org/10.1371/journal.pone.0280306.