He et al. Smart Learning Environments (2025) 12:64
https://doi.org/10.1186/s40561-025-00415-z

Smart Learning Environments

RESEARCH

Open Access

AI-driven predictive models for optimizing mathematics education technology: Enhancing decision-making through educational data mining and meta-analysis

Aneng He[1], Wenwen Yuan[2], Lai Soon Lee[1*] and Tian Tian[3]

*Correspondence:
lls@upm.edu.my

[1] Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2] Wang Ganchang High School, Changshu City 215500, Jiangsu Province, China
[3] Institute for Mathematical Research, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Abstract

This paper explores the challenge of achieving consistent effectiveness in integrating Mathematics Education Technology (MET) in K-12 classrooms, focusing on factors such as technology type, timing, and instructional strategies. It highlights the difficulties novice teachers face in optimizing MET compared to experienced educators, emphasizing the need to better understand the ideal duration and application of MET in various teaching settings. This study proposes using Artificial Intelligence (AI) to predict and optimize MET effectiveness, aiming to enhance student achievement. However, a key challenge is the lack of comprehensive MET databases, prompting the exploration of novel data collection methods and meta-analysis for educational data mining. An AI-based predictive model is developed for MET, analyzing 423 publications on its effectiveness in Chinese K-12 mathematics education. Nine AI-driven predictive models were created, with the best-performing predictive model being eXtreme Gradient Boosting, enhanced with L2 Regularization, Synthetic Minority Over-sampling Technique–augmented Regression (SMOTER), and Active Learning. The proposed model was further refined using Particle Swarm Optimization for hyperparameter tuning and analyzed with Shapley Additive Explanations (SHAP) values to assess feature importance. Numerical results indicated that the duration of MET usage is a critical factor for optimization. A controlled experiment in a Mainland China middle school validated the model's efficacy, showing that model-guided MET significantly outperformed traditional methods. These findings offer valuable insights for bridging gaps between novice and experienced teachers, promoting educational equity, and providing practical recommendations for improving MET integration in Mathematics education.

Keywords: Predictive learning analytics, Educational data mining, Mathematics education technology, Artificial intelligence, Active learning, Meta-analysis

Introduction

During the COVID-19 pandemic, mathematics teachers were faced with unprecedented challenges in adapting to online teaching and learning (Callaghan et al., 2023). The emergency deployment of Mathematics Education Technology (MET) became a critical response to these challenges. In the post-pandemic era, the value of MET in the mathematics classroom has been increasingly acknowledged by educators (Ogbonnaya & Mushipe, 2020; Drijvers & Sinclair, 2024). However, while a substantial body of research has examined the effectiveness of MET in mathematics instruction, there remains an urgent need to explore its optimal usage (He et al., 2024).

One of the most significant challenges in implementing MET in K-12 mathematics education is achieving consistency in its effectiveness (Ran et al., 2022; He et al., 2024). The success of technology integration in the mathematics classroom is influenced by a multitude of complex factors, including the type of technology used, the timing of its application, and the instructional strategies and content (Juandi et al., 2021). Novice teachers often struggle to utilize technology as effectively as their more experienced counterparts, who have mastered the integration of these elements (Fernández-Batanero et al., 2021). The most pressing issue is determining the ideal duration and type of technology to be employed across various instructional settings and content areas (He et al., 2024). Systematically investigating this question through traditional educational research methods is difficult, as such approaches are time-consuming, costly, and unable to fully control for non-variable conditions in large-scale trials.

Although artificial intelligence (AI) is widely regarded as the future of education (Zhai et al., 2021), it is already transforming a range of fields, including transportation, finance, and healthcare. In education, traditional forecasting has heavily relied on the extensive experience of seasoned educators, an approach prone to human error and often disadvantageous to less experienced teachers. However, the integration of AI into education is fundamentally changing this dynamic (Hwang & Tu, 2021).

Recent studies have shown promise in using AI predictive models to forecast outcomes such as student achievement, failure rates, and dropout rates (Yağcı, 2022; Nabil et al., 2021). In contrast, the primary challenge facing MET remains its inconsistent validity and suboptimal application, with student achievement being the most crucial metric for MET success. This underscores the potential of AI to construct robust predictive models of MET effect sizes, followed by feature selection, offering a new approach for K-12 mathematics educators. The aim of this study is to develop and validate such an AI-based predictive model, ultimately providing educators with an evidence-based framework to optimize MET usage and, in turn, enhance student achievement in mathematics.

AI-based predictive model relies heavily on large datasets (Aldoseri et al., 2023). However, a significant barrier to the widespread adoption of AI in MET is the scarcity of comprehensive databases for MET research. Due to the absence of such databases, this study seeks to explore novel data collection methods. The lack of a dedicated educational database for MET validity further highlights the need for an educational data mining methodology tailored to this context. In social science research, meta-analysis has proven to be a powerful tool for synthesizing validity-related data from multiple studies

(Borenstein et al., 2021). Building on this, the authors propose leveraging meta-analysis as a traditional statistical technique for educational data mining in predictive modeling.

This study introduces several technological innovations and contributions to the use of AI in educational technology. First, in terms of utilizing AI for the assessment of educational technology, the authors propose a new data collection method based on meta-analysis, which can be applied to large-scale educational experiments. Classification modeling serves as the primary tool for predictive modeling aimed at student achievement, while the study also emphasizes the need for more research on regression modeling in AI education development. By exploring educational data, this study examines the optimization achieved through the combination of feature regularization, active learning, and Synthetic Minority Over-sampling Technique (SMOTE) in AI models. Additionally, it investigates feature selection techniques to optimize the effectiveness of MET. In instructional practice, the predictive model developed in this study can provide mathematics teachers with evidence-based recommendations regarding the type and duration of MET most suited to different instructional goals. On a broader level, this study helps bridge the gap between novice and senior teachers, contributes to the realization of educational equity, and offers valuable guidance for young teachers in their professional development.

Related works

Predictive modeling is the development of mathematical models designed to assess potential outcomes, which can subsequently provide valid information and evidence to support decision-making and intervention strategies (Herodotou et al., 2019). The application of predictive modeling in education has evolved into a specialized field known as predictive learning analytics (PLA). PLA utilizes a range of statistical and analytical techniques to predict future student behaviors and outcomes, including predicting student achievement, providing early warnings for at-risk students, and analyzing student engagement and satisfaction (Nyce & Cpcu, 2007). PLA has become central to research in educational administration and pedagogical methods (Sghir et al., 2023), offering valuable insights that not only enhance learning but also guide educators in refining instructional strategies.

PLA introduces a distinct educational component that sets it apart from predictive analytics in other fields. It is implemented in two stages: first, the creation of a predictive model based on existing educational data, and second, the analysis of the results derived from these models, grounded in educational principles. PLA can be used both directly and indirectly to evaluate the success of educational policies and assess the pedagogical competence of teachers. As predictive models evolve, the focus in teacher education is shifting from "designing future instruction based on students' current academic performance" to "using predictions of students' future performance to shape current instructional settings". This paradigm shift has significant implications for teaching practices. Furthermore, accurate judgments about academic progress have always been crucial for both students and educators. Such judgments not only motivate students' subsequent learning but also guide teachers' pedagogical decisions. Prior to the advent of PLA, accurately gauging academic progress was a complex challenge due to the variability in content, difficulty, and knowledge required for different tests. However, Yang and Li (2018) successfully predicted the grades of 60 students and analyzed their academic progress by integrating a predictive model with various scoring formulas.

Interestingly, in most PLA studies, the output of predictive models was often categorized according to academic achievement levels. This suggests that the primary function of these models was classification. For instance, studies such as Strenze (2007) and Krüger et al. (2023), which aimed to predict student dropout rates, and Yüksel and Geban (2014), which focused on predicting whether students would pass or fail, primarily employed binary categorization (positive or negative output). However, some studies have taken a more nuanced approach by using performance matrics to more comprehensively quantify academic achievement. These models incorporated teacher evaluations, test scores, and assessments of students' learning potential and academic progress (Yang & Li, 2018). In Doz et al. (2023), for example, Gaussian and trigonometric functions were applied to fuzzify teacher ratings and student test scores, leading to new academic achievement scores, which were subsequently categorized into 10 levels.

Despite these advancements, predicting student achievement using regression models in K-12 mathematics education remains a significant challenge. A major hurdle in developing educational predictive models lies in selecting appropriate data sources, screening relevant features, and choosing suitable models. The review of prediction models achievement in K-12 mathematics education reveals that most data sources are open-source databases. For example, Acıslı-Celik and Yesilkanat (2023) utilized the Program for International Student Assessment (PISA) dataset, compiled by the Organization for Economic Co-operation and Development (OECD), which includes data from 12 countries, including Brazil and Japan. The release of PISA's open-access dataset has significantly advanced the application of AI techniques in comparative education. Moreover, many institutes widely use datasets from the Portuguese Department of Educational Statistics (Aslam et al., 2021; Ng et al., 2021; Keser & Aghalarova, 2022), underscoring the limited availability of publicly accessible educational datasets. Despite an extensive search, the authors found no datasets that specifically address the validity of MET.

When selecting predictors for PLA models, studies typically categorize them into three types, depending on the purpose of the prediction: student-related, teacher-related, and institution-related (Sghir et al., 2023). In previous studies, the predictors include students' prior academic performance, demographic characteristics (e.g., gender, age, place of birth, international background), and other features, which vary depending on the research objectives. Regardless of the specific goal of the educational prediction model, the choice of predictors is guided by the pedagogical principles underlying the research.

MET currently lacks a strict, universally accepted definition. However, for the purposes of this study, MET is understood as any educational technology used within mathematics education, which primarily includes three categories: dynamic mathematics software (DMS) such as GeoGebra and MATLAB, visual augmentation or virtual reality, and AI-enabled smart classrooms. Among these, DMS is perhaps the most impactful. The effectiveness of DMS in enhancing mathematics teaching and learning has been widely acknowledged. However, some studies argued that students performed better in mathematics using traditional teaching methods, without MET (Setyani & Lestari, 2016). Therefore, the authors believe that meta-analytic studies, which consider the largest body of relevant research, are essential for evaluating the impact of MET on student achievement.

The positive impact of DMS on mathematics performance improvement was discussed in a meta-analysis as early as 2014, albeit with a small sample size of only nine studies (Günhan & Açan, 2016). They explored literature from Turkey between 2006 and 2015, examining the positive effects of MET on geometry achievement while accounting for moderating variables such as grade level, experiment duration, sample size, and publication type. Further meta-analyses, such as Hillmayr et al. (2020), have investigated the potential of digital tools to facilitate learning in secondary school mathematics and science, showing that MET had the most significant impact among all digital tools.

Xie et al. (2020) conducted a meta-analysis of 36 studies on computer-assisted instruction (CAI) from mainland China, focusing on how different pedagogical approaches combined with CAI influence its effectiveness. This body of research underscores the importance of CAI in Chinese education. Additionally, Juandi et al. (2021) reviewed studies from 2010 to 2020, concluding that students using GeoGebra outperformed their peers in traditional classrooms. Moreover, Zhang et al. (2023) showed that the specific mathematics topics and processing time significantly moderated the effect sizes of these interventions.

Recent research by Cevikbas et al. (2023) suggests that Augmented Reality (AR) and Virtual Reality (VR) can create interactive, immersive mathematical environments that enable students to explore mathematical concepts in a more engaging and intuitive way. These technologies allow students to visualize complex concepts through simulations, enhancing problem-solving skills. A meta-analysis of AR/VR applications in mathematics teaching found significant positive effects, with age-appropriate content being a significant moderator. Further analysis by Chang et al. (2022) confirmed that integrated AR instruction leads from moderate to large improvements in students' knowledge and skills, with longer AR interventions potentially producing greater benefits.

Smart classrooms, which integrate advanced technologies to optimize learning experiences, are another promising area in MET. These classrooms combine various technologies such as mobile devices, automated communication tools, video projectors, sensors, and facial recognition software (Mircea et al., 2021). According to Chang et al. (2024), the use of smart classrooms has a significant positive impact on student achievement. The study further indicates that factors such as the type of smart classroom, instructional strategies, sample size, and school location influence the effectiveness of these interventions.

While significant progress has been made in research on PLA and MET, substantial gaps remain, particularly at their intersection, where no comprehensive and feasible research framework has yet been proposed. One major challenge is the lack of publicly available datasets on MET effectiveness, as large-scale educational experiments are costly. As a result, most studies rely on generalized educational datasets such as PISA, which do not specifically focus on MET, limiting the scope of targeted research in this area. Additionally, PLA research predominantly focuses on categorical predictions of student achievement (e.g., pass/fail classification or dropout risk), with limited adoption of finer-grained performance matrices or fuzzy scoring methods, leading to an incomplete understanding of students' academic progress. Furthermore, while PLA has been applied to assess teachers' instructional competence and support decision-making in instructional strategies, little attention has been given to leveraging technology to refine teaching practices, resulting in limited support for optimizing educational interventions.

Moreover, most meta-analyses of MET fail to account for geographic and cultural influences, often emphasizing broad trends while overlooking context-specific educational settings. For instance, research on the long-term effects of MET in mainland China is extremely limited, making it difficult to assess its sustained impact within the country's unique educational framework. Additionally, the effectiveness of MET is influenced by multiple moderating variables, such as grade level, type of technology, duration of use, and instructional strategies. However, the interactions among these variables remain largely unexplored, restricting our understanding of how MET functions across different instructional contexts. It is urgent and necessary to develop educational data mining-driven predictions that systematically analyze these interactions, optimize the use of METs in a variety of teaching and learning environments, and deepen our understanding of their role in mathematics education.

Methodology

This study aims to model the effectiveness of MET and to determine the optimal time for its use in specific educational settings by employing AI predictive models. Specifically, the study integrates meta-analysis–driven educational data mining with advanced machine learning optimization techniques to overcome the limitations of traditional large-scale experiments, identify influential predictors, and validate the model's applicability in real classroom contexts. Building upon these objectives, the methodology is structured around four sequential research questions, each corresponding to a distinct stage of the research process:

1. In the absence of large-scale educational experiments, can meta-analysis–driven educational data mining and regression modeling accurately predict the effect size of MET on K–12 students' mathematics learning outcomes?
2. How can advanced optimization techniques (e.g., SMOTER, Active Learning, Particle Swarm Optimization) improve the predictive performance of regression models in estimating MET effectiveness?
3. Within the best-performing predictive model, which pedagogical, content-related, and technological variables most significantly influence MET effectiveness?
4. In real classroom experiments and external validation, can the predictive model effectively guide MET implementation strategies and generalize to diverse instructional contexts?

As outlined in Brooks and Thompson (2017), constructing an educational prediction model consists of six steps: problem identification, data collection, preprocessing,

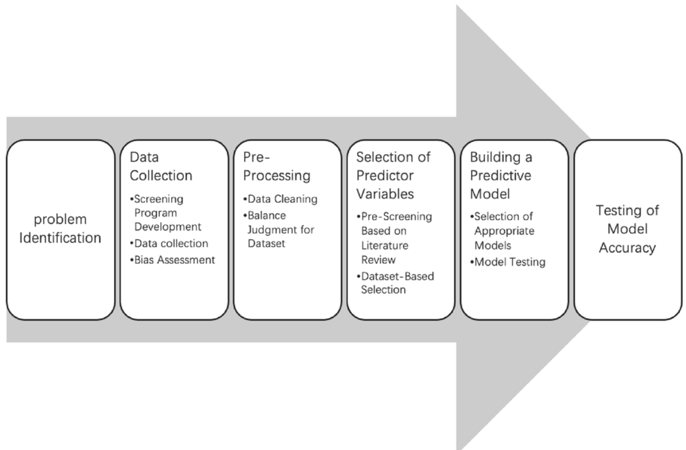

Fig. 1 Research framework of educational prediction model

feature screening, model building, and model testing to ensure model robustness and validity (shown in Fig. 1).

To address these research questions while adhering to the essential steps of predictive modeling in AIED, an integrated meta-analysis–based framework was developed. This framework systematically combines problem-solving strategies with model-building procedures. The research process encompasses several stages: data collection through meta-analysis, data preprocessing, and subsequent data analysis, designed to address the lack of relevant databases and the impracticality of large-scale teaching experiments.

Following meta-analysis and feature screening, a rigorous comparative evaluation of predictive models was conducted. Advanced optimization techniques were incorporated to enhance performance, and the optimal model identified was subsequently adopted to support the application of MET in specific instructional contexts. To improve interpretability, SHAP values were employed as a post-hoc explainability tool, enhancing model transparency and providing insights into the relationships between predictors and outcomes. Finally, to ensure robustness and generalizability, real classroom data were subjected to statistical validation, including t-tests, p-tests, and external evaluations. The complete research process is illustrated in Fig. 2.

Problem identification

The objective of this study is to develop a predictive model to assess the effectiveness of mathematics pedagogy in K-12 students' learning. Additionally, the study aims to identify the optimal timing for implementing specific teaching strategies, based on the model's predictions. Building on previous research, the authors recognize that effect size is a continuous variable that can be categorized into different ranges. However, given the intuitive nature of continuous outputs for decision-making in classroom settings, this study employs regression modeling to predict the continuous values of effect size.

Data collection

For the collection of data from existing literature, this study adheres to the criteria set forth by the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*

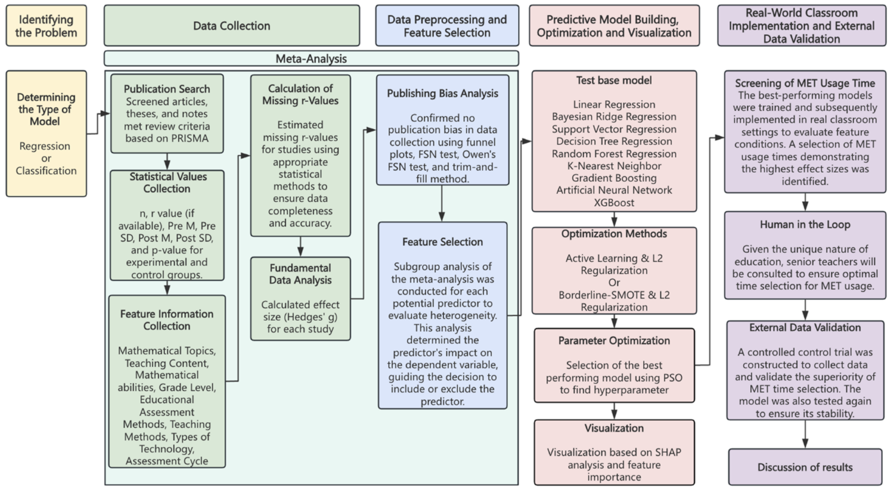

Fig. 2 Research process for the study

Table 1 PICO protocol for this study

Population	Intervention	Comparison	Outcome
Experimental or quasi-experimental studies using MET for mathematics teaching	Whether or not MET is used	Setting in which MET is used in different studies	Effectiveness of MET in the studies

(PRISMA) guidelines (Petersen et al., 2015), which provide a framework for literature search, screening, and data extraction from comparative experiments.

According to PRISMA, the initial step in a literature search is to define and screen the search terms. However, given that this study focuses on the application of MET in K-12 mathematics education, a broad range of related keywords are employed (see Table 1). To ensure precise retrieval of relevant literature addressing the specific research question, the *PICO model* is utilized to structure the literature search. The PICO model, derived from evidence-based medicine, stands for: **P** (Population), **I** (Intervention), **C** (Comparison), and **O** (Outcome) (Moher et al., 2009). In this study, **Population** refers to experimental or quasi-experimental studies involving MET in teaching and learning; **Intervention** pertains to the use or non-use of MET; **Comparison** describes the various settings in which MET is applied across studies; and **Outcome** refers to the effectiveness of MET as reported in the studies. A detailed breakdown of the PICO elements is provided in Table 1.

According to the PICO model, the literature included in this study must focus on the effects of MET on K-12 students' learning of mathematics. Furthermore, the studies selected should be experimental or quasi-experimental in nature. However, to avoid limitations arising from the use of narrow secondary keywords, the scope of the literature search is expanded to include all keywords related to K-12 students' mathematics learning. Specifically, secondary keywords are broadened, ensuring that the search encompasses all relevant terms. In this study, the literature on the application of MET in K-12mathematics education is retrieved primarily through Boolean operators such as

Table 2 Search keywords

Related to MET	Related to K-12 student's mathematics learning
GeoGebra, Sketchpad, Board, Interactive board, Smartboard, Smart Classroom, Computer Assisted Instruction, Dynamic Mathematics Software, Advanced Reality, Virtual Reality	Elementary Mathematics, Middle School Mathematics, High School Mathematics, K-12 Mathematics, Mathematics Teaching, Mathematics Learning, Mathematics Instruction

"or" and "and". Keywords related to MET are combined with those related to K-12 mathematics learning, as detailed in Table 2 during the retrieval process.

To collect data, this study conducted a literature search in China National Knowledge Infrastructure (CNKI), one of the most authoritative Chinese literature repositories, as well as in four widely used English-language abstract and citation databases: Web of Science, Springer, Science Direct, and Taylor & Francis. Furthermore, given the extensive body of master and doctoral theses related to MET research in the context of China's higher education reform, and recognizing that these dissertations were rigorously double-blind peer-reviewed, the experimental data presented in these postgraduate theses were also included in the analysis. All studies included in this research must meet the following six criteria:

1. The study must be published between January 2017 and February 2023.
2. The study must be of experimental or quasi-experimental design.
3. The study must have a control group and an experimental group.
4. The study must focus on mathematical academic learning in K-12 students from mainland China.
5. Student achievements in mathematics must be included as one of the outcome variables.
6. The mean and standard deviation of students' mathematical achievements and the number of students must be recorded for both the experimental and control groups before and after testing.

Based on the above keywords, search databases and adoption criteria, there are 423 studies are considered as individual data for data counting and coding. All the screening processes are carried out in full compliance with the PRISMA guidelines (Page et al., 2021), with details as shown in Fig. 3.

Selection of predictor variables and coding plan

Based on the previous research, potential moderators (features) that may influence the effectiveness of MET on K-12 students' mathematics learning include mathematical topics, mathematical abilities, teaching methods, educational assessment methods, teaching content, and grade levels (Maaß & Artigue 2013). According to the data collection process of this study, teaching content may include multiple topics within the same study. Therefore, it will also be included as one of the features in this study. Additionally, the type of technology used, the duration of MET usage, and the

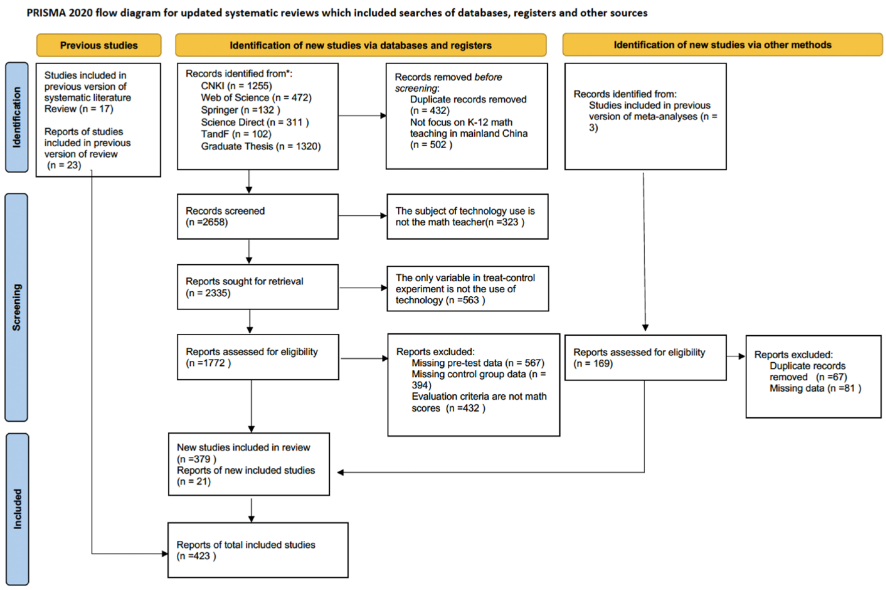

Fig. 3 PRISMA flow diagram

sample size (i.e., the total number of experimental and control group participants) are included as moderators in the meta-analysis and as predictors in the predictive modeling.

Mathematical topics in K-12 schools range from basic to advanced, and in this study, the moderators are divided into three categories: basic concepts, advanced mathematical concepts, and mathematical skills (Akın, 2022). These categories describe whether the primary goal of using MET is to teach different types of content or to cultivate mathematical abilities. Through previous research, students' mathematical abilities are primarily reflected in numerical, spatial, and problem-solving abilities. The moderator for mathematical abilities in this study will also follow these three classifications. Unlike the moderator of mathematical topics, the moderator of mathematical abilities specifically describes the types of mathematical abilities that teachers aim to cultivate through MET.

The combination of MET with different teaching methods has always been a popular topic in MET-related research. Some studies on technology-related teaching practices have found that designing engaging teaching methods can promote student participation in the classroom (Attard & Holmes, 2020). Previous research indicates that MET is often combined with traditional teaching, constructivist teaching, and situated teaching (Maaß & Artigue 2013). According to regression analysis in the literature, most studies suggest that combining MET with traditional teaching provides students with opportunities to try activities, improving their learning and helping them develop creativity and critical thinking skills (Baye et al., 2021). Additionally, some studies suggest that combining MET with constructivist teaching is more motivating and more effective than pure constructivist teaching in improving students' learning outcomes (Li & Ma, 2010; Zengin et al., 2012). Therefore, the moderator of teaching methods in this study will be divided into three categories: combined with traditional teaching, combined with constructivist teaching, and combined with situated teaching.

Educational assessment is the systematic process of making value judgments about the educational and learning behaviors of teachers and students (Carver et al., 2016). There are various standardized assessment methods, such as diagnostic assessment, process assessment, and summative assessment, which focus on different aspects of goals and learning processes (Jing, 2012). However, during the data collection process, the authors found that almost all studies focused on results or a combination of results and processes in educational assessment, with no studies focusing solely on the learning process. This may be related to the educational assessment policies in the data source countries. Therefore, the educational assessment moderator in this study will be classified into two categories: focusing on results and focusing on both results and process. For the grade level moderator, the authors categorized it according to the Chinese education system: primary school (grades 1–6), middle school (grades 7–9), and high school (grades 10–12).

During the data collection process, the authors found that many studies included more than one teaching content, with the maximum being three different teaching topics. Therefore, for the teaching topic moderator, the authors will divide it into three independent moderators: Teaching Content 1, Teaching Content 2, and Teaching Content 3. The classification of these moderators will depend on the characteristics of K-12 mathematics education in the data collection regions, and will include functions, plane geometry, three-dimensional geometry, statistics, vectors, and calculations. Some studies do not include Teaching Content 2 and Teaching Content 3, and for these studies, the moderator will be classified as 'No'. At the same time, since the classification of Teaching Content 2 and Teaching Content 3 is far less rich than Teaching Content 1 in the data collection process, the authors will write the coding plan according to the actual situation (as shown in Table 3). Additionally, this study includes two continuous features: the duration of MET usage and sample size (the total number of participants in the experimental and control groups).

During the data collection process, in addition to the data on potential moderators (features) described above, the authors also collected data from studies in each of the included publications. This included sample sizes, variance and mean scores for pre-tests and post-tests for both the control and experimental groups.

This study is based on the PRISMA framework, and first focuses on data collection regarding the application of MET in K-12 students' mathematics learning. A total of 423 effect sizes were collected and calculated from 423 studies. Additionally, 11 potential moderator variables (predictors) that may influence the effectiveness of MET were statistically analyzed for these 423 studies. To ensure accuracy, the researchers developed a detailed coding plan, as shown in the Table 3. The statistical process was completed independently by two researchers, and the consistency was tested using the Kappa coefficient. The Kappa coefficient was found to be 1.0, indicating high reliability.

Definition of predicted outcome

Since this study aims to address the effectiveness of MET on K-12 students' mathematics learning, the statistical measure of effect size will be chosen. Previous studies usually choose Cohen's h and Hedges' g as effect size estimations. Comparing with Cohen's h, Hedges' g provides high accuracy in small samples (Borenstein et al., 2021). Therefore,

Table 3 Coding plan for potential moderators (features)

Potential moderators (Features)	Feature type	Code plan
Mathematics topics	Classification	
Mathematical abilities		1
Foundation mathematical concepts		2
Higher-level mathematical concepts		3
Teaching content 1	Classification	
Three-dimensional geometry		1
Plane geometry		2
Functions		3
Calculations		4
Statistics		5
Vectors		6
Teaching content 2	Classification	
Three-dimensional geometry		1
Plane geometry		2
Functions		3
No		4
Teaching Content 3	Classification	
Statistics		1
Plane geometry		2
Functions		3
No		4
Mathematical abilities	Classification	
Numerical Ability		1
Spatial Skills		2
Problem-solving ability		3
Grade level	Classification	
High school		1
Middle school		2
Primary school		3
Educational assessment methods	Classification	
Focus on both purpose and process		1
Focus on the purpose		2
Teaching methods	Classification	
Combine with constructivist teaching		1
Combine with scenario-based teaching		2
Combine with traditional teaching		3
Type of tech	Classification	
DMS		1
VR/AR		2
AI smart classroom		3
Sample size	Continuous	Integers from 33 to 180
MET usage time	Continuous	Integers from 1 to 120

Hedges'g is used as the prediction result in this study. The statistical data provided in the adopted studies include: (1) the sample size of the experimental and control groups; (2) the students' achievements of the experimental and control groups, before and after the experiment; (3) the variance of students' achievements of the experimental and control groups, before and after the test; and (4) the correlation coefficient, *r*. Therefore, the effect size Hedges' *g* for each of the adopted publications was calculated.

Data pre-processing

Data collection follows the PRISMA framework, which may introduce publication bias, as studies with positive outcomes are more likely to be published. To assess whether the dataset is resistant to this bias, four tests are conducted: the funnel plot, fail-safe N test (FSN), Orwin's FSN, and Duval and Tweedie's trim-and-fill test (Borenstein et al., 2021). Following this, a meta-analysis is performed on the 423 studies to calculate the average effect size of MET in K-12 students' mathematics learning. This effect size will be used as an optimization target for temporal feature selection in predictive models.

To further refine the optimal timing for MET use across different instructional purposes (e.g., mathematical topics, abilities), grade levels, and teaching methods, subgroup analyses will be conducted. These analyses will examine how various moderators (predictor variables) influence the MET effect size, complementing the feature selection process and serving as secondary optimization objectives.

Predictive model building and optimization

The regression models selected for this study are: linear regression (LR), Bayesian regression (BR), support vector regression (SVR), decision tree regression (DT), random forest regression (RF), gradient boosting (GB), artificial neural network (ANN), k-nearest neighbor (KNN), and eXtreme Gradient Boosting (XGB).

To improve the model performance, this study begins by characterizing the dataset. Previous educational data indicate potential relationships between various features. For instance, when teaching plane and three-dimensional geometry, teachers often emphasize spatial imagination rather than computational skills. Furthermore, higher-level mathematical content is seldom introduced at the elementary school level. Therefore, addressing these feature interrelationships is crucial, as they may hinder model performance.

To manage this issue, the study incorporates L2 regularization, also known as Ridge Regression. This is a commonly used technique in machine learning that helps prevent overfitting and enhances model generalization (Cortes et al., 2012; Bishop, 2006). This method adds a penalty term to the original loss function, the sum of the squared model parameters. By doing so, L2 regularization controls model complexity and limits the size of the coefficients (Cortes et al., 2012). Additionally, it helps address multicollinearity by shrinking highly correlated coefficients, making the model more stable and interpretable (Hoerl & Kennard, 1970). The loss function with L2 regularization is defined as follows:

$$\text{Loss} = \text{Loss}_{\text{original}} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\text{Loss}_{\text{original}}$ represents the original loss function (e.g., Mean Squared Error), λ is the regularization parameter that controls the penalty term's weight, and β_j are the model parameters.

The dataset imbalance may also impact the performance of the prediction model. In an imbalanced dataset, the number of instances in the minority class is typically smaller than in the majority class, leading to biased predictions and poor model performance (Fernández et al., 2018). According to the literature review, the SMOTE has shown strong optimization capabilities in similar contexts (Pujianto et al., 2020; Nabil et al., 2021). While SMOTE is typically used in classification tasks to address dataset imbalance, this study, being regression-oriented, applies the SMOTE-augmented Regression technique (SMOTER), which has proven to improve performance in such cases (Camacho et al., 2022).

Moreover, this study will also explore the use of Active Learning (AL), a relatively new technique designed to address dataset imbalance without increasing the sample size. AL works by iteratively selecting the most informative data points from a pool of unlabeled data, labeling them, and updating the model with the new labels. This method can be tailored to focus on querying instances from the minority class, ensuring the model receives a more balanced set of labeled data. As a result, the model's ability to learn the characteristics of the minority class is improved (Ertekin et al., 2007).

AL typically performs well with large datasets, where the benefit of selecting the most informative samples is more evident (Gal et al., 2017). However, since the dataset in this study is relatively small, the advantage of selecting informative samples may not be as obvious. To overcome this challenge, the study refers to the approach of combining AL with an oversampling technique (Wang et al., 2020), which has demonstrated good results with unbalanced small datasets. Therefore, this study will first apply the SMOTER technique, which is more advanced than traditional oversampling, to extend the dataset. AL optimization will then be employed to further improve the model's performance.

Overall, this study will evaluate and compare the performance of the nine prediction models mentioned earlier under three different optimization frameworks: (1) Base Model, (2) SMOTER with L2 regularization, and (3) SMOTER with AL and L2 regularization. After selecting the optimal model based on the test metrics and theoretical framework, the authors will then proceed with the hyperparameter optimization for the proposed model. Unlike the less efficient Grid Search, a Particle Swarm Optimization (PSO) algorithm is used for hyperparameter tuning. PSO is an optimization algorithm that simulates the foraging behavior of a flock of birds, where each particle represents a solution that is continuously refined through information exchange (Marini & Walczak, 2015). The effectiveness of PSO for hyperparameter optimization across different prediction models has been rigorously validated through experiments (Pannakkong et al., 2022).

Testing models

In this study, the collected dataset is randomly divided (with a randomness index of 42) into a 80% for training set and 20% for test set. For the regression models, this study will use R-squared (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) as the metrics for evaluation (James et al., 2013). The measurement processes are shown as follows.

1. R-squared (R^2) R^2 measures the proportion of variance in the dependent variable that is predictable from the independent variables. A higher R^2 usually indicates a better fit of the model to the data.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}, \tag{1}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the actual values.

2. Mean Absolute Error (MAE) MAE measures the average magnitude of the errors in a set of predictions without considering their direction. The smaller the MAE, the smaller the average deviation between the predicted and actual values, indicating that the predictions of the model are closer to the actual observations.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{2}$$

where n is the number of samples, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

3. Mean Squared Error (MSE) MSE measures the average of the squares of the errors - that is, the average squared difference between the estimated values and the actual value. A smaller MSE usually indicates a better predictive performance of the model.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{3}$$

4. Root Mean Squared Error (RMSE) RMSE is the square root of the average squared differences between prediction and actual observation. A smaller RMSE usually indicates a smaller prediction error and better predictive performance of the model. Its sensitivity to large errors makes it the preferred performance evaluation metric in many practical applications.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{4}$$

Feature selection and practical application

After completing the data collection, modeling, optimization, and testing, the next step will focus on feature selection to explore the optimal duration of MET use in different teaching and learning environments. To better validate the practical application, controlled experiments are conducted in eight classes (two control and six experimental groups) in collaboration with educational researchers from mainland China.

In this controlled experiment, the duration of MET use for the experimental group is determined by the best-performing regression model. This duration is also adjusted based on the number of classroom hours required for the experimental group's content and the instructor's discretion. The two control groups included a blank control group (i.e., teaching without using any MET) and a comparison control group (i.e., teachers teaching from experience without referring to the predictive model's recommendations).

An experienced researcher conducted the controlled experiment with four mathematics teachers. During the experiment, one teacher was responsible for two classes in the control group, while the remaining teachers each taught two experimental group classes. Students were given two rounds of tests, one at the beginning and one at the end of the experiment. For each test, individual scores were recorded to calculate the mean score and variance for each group.

To enhance the accuracy of the experiment, t-tests and chi-square tests are conducted on the pre-experiment test scores to ensure that all classes have similar starting levels. For the post-experiment scores, p-tests and t-tests will be conducted to determine the impact of MET usage duration on improving K-12 students' mathematics learning outcomes, as guided by the predictive model developed in this study. Subsequently, effect size Hedges' *g* values will be calculated for the six experimental groups and one comparison control group. These six data points will serve as external data to evaluate the performance of the predictive model, ensuring its fit to external data and its practical usability.

Since this study uses a combination of publicly or semi-publicly available data and aims to build appropriate predictive models to advise instructors on the timing of MET use in instructional settings, the final decision regarding the time of MET usage remains with the instructor. Additionally, the data collection for this experimental process does not include identifying individual data for each student but only semi-public data such as mean scores, variances, and class sizes. Therefore, ethical approval was not required.

Results and discussion

This section is organized into three parts. The first part presents the meta-analysis, where the mean effect size from 423 studies is selected as the optimization objective for feature selection. It also includes subgroup analysis and publication bias detection, demonstrating the reliability of the database and the validity of the selected features. The second part discusses the performance of the prediction models and the improvements achieved through various optimization methods. The final part highlights the practical applications and validates the results using external data.

All experiments were run on a MacBook Pro with an Apple M1 processor and 8GB of RAM, using the following software:

- *Meta-analysis* Comprehensive Meta-Analysis (CMA) 3.0
- *Programming* Python 3.12.3, NumPy 1.21, SciPy 1.7
- *Data analysis* Pandas 1.3, Scikit-learn 0.24
- *Visualization* Matplotlib 3.4, Seaborn 0.11
- *Deep learning* TensorFlow 2.6, Keras 2.6

Data collection and meta-analysis

The first objective of this study is to employ meta-analysis as a means of educational data mining for predictive modeling in contexts where large-scale educational experiments are not feasible. In this section, effect sizes derived from the meta-analysis are calculated to represent the overall average effect and to inform the setting of optimal targets in practice. Publication bias tests are conducted to assess the reliability of the data sources, while subgroup analyses provide a detailed interpretation of the data structure and serve as a foundation for identifying potential predictor variables.

The overall impact of MET and optimization goals

In this study, 423 independent studies were adopted, and data from these 423 studies were obtained with 38,163 students, ranging from 33 to 180 students in each study. Data from these 423 independent studies were analyzed to determine the effect sizes of MET on K-12 mathematics learning.

The authors report the Q results as follows: $Q = 936.657$ at $df = 422$ with $p < 0.001$. The I^2, which was used to quantify the degree of heterogeneity between the results of the studies, is equal to 54.946% after using a homogeneity test in this meta-analysis (shown in Table 4). The Q statistic is used to test for heterogeneity among studies. Given that $p < 0.001$, this indicates significant heterogeneity, meaning significant differences do exist among the studies. The I^2 statistic describes the percentage of total variation across studies due to heterogeneity rather than chance. An $I^2 = 54.946\%$ indicates moderate to high heterogeneity (Higgins & Thompson, 2002). This shows that a random-effects model is more appropriate in this scenario. The random-effects model assumes that each study has its true effect size, which comes from a distribution of true effect sizes, and it accounts for variability among studies. In contrast, the fixed-effects model assumes all studies share a common true effect size, which is typically inappropriate in the presence of significant heterogeneity.

According to Higgins et al. (2003), criteria related to I^2 values and the effect sizes of the studies in actual differences, this study reflects 54.946% of the observed variance. It is also possible to observe the overall effect size of these 423 studies ($g = 0.677$, $SE = 0.016$, $p < 0.001$, 95% CI [0.646, 0.708]), which indicates a statistical difference in mathematics achievement between students in the experimental group taught using MET and the control group taught without using MET (Thalheimer & Cook, 2002). Thus, clarifying the first question, MET significantly affects K-12 students' mathematics learning. Besides, it can also be inferred that using MET is more effective for K-12 mathematics learning than without using MET teaching.

Therefore, the average effect of the meta-analysis will be used as the main optimization objective for feature selection in this study. That is, it is necessary to screen for the use of time so that the effect size g value is more significant than 0.677 while fixing features other than the time feature.

Publication bias

In the meta-analysis phase of this study, a publication bias analysis was conducted to evaluate whether the dataset was affected by publication bias and to determine its

Table 4 The overall effect of MET on K-12 students' mathematics learning

95% Confidence interval (CI)						**Test of heterogeneity (Toh)**		
k	***n***	***g***	**Lower limit**	**Upper limit**	***p***	***Q***	***p***	***I^2***
423	38163	0.677	0.646	0.708	< 0.001	936.657	< 0.001	54.946%

suitability for predictive modeling. Publication bias refers to the tendency for statistically significant results to be more likely published than non-significant or null results (Song et al., 2010). To examine this potential bias, four primary methods are employed: funnel plot analysis, the Fail-Safe N method (FSN), Orwin's Fail-Safe N (Orwin's FSN), and Duval and Tweedie's trim-and-fill test.

As the funnel plot shown in Fig. 4, the majority of findings are concentrated at the top of the plot, clustering around the mean effect size, with only a few observations appearing at the bottom. This distribution suggests that publication bias is likely minimal.

According to Rosenthal (1991), publication bias is unlikely to distort the results of a meta-analysis if the Fail-Safe N is significantly larger than the tolerance threshold of $5k + 10$, where k represents the total number of effects reported in the analysis. In the study, the Fail-Safe N is 69,413 (as shown in Table 5), which is far greater than the threshold of 2125 ($5 \cdot 423 + 10$). This finding suggests that the impact of unpublished studies on the practical significance of the included studies is minimal, further supporting the validity of the results.

Orwin's FSN method, which evaluates the number of unpublished studies required to significantly alter the total effect size of the published studies at a given significance level, yielded important insights. As presented in Table 6, the analysis indicated that 27,698 unpublished studies would be needed to substantially influence the total effect size of

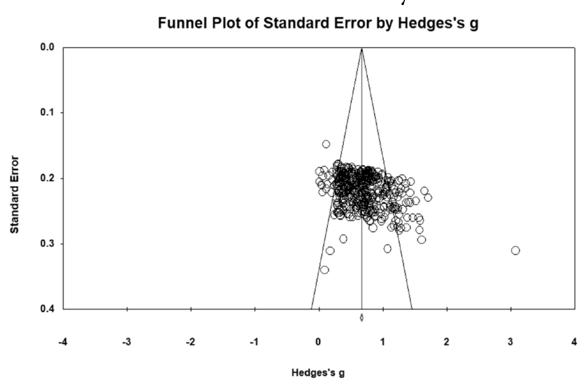

Fig. 4 Funnel plot for the study

Table 5 Results of classic Fail-Safe N test

Z-value for observed studies	63.967
P-value for observed studies	< 0.001
Alpha	0.050
Tails	2.000
Z for alpha	1.959
Number of observed studies	423
Number of missing studies that would bring *p*-value to larger than alpha	69,413

Table 6 Results of Orwin's Fail-Safe N test

Hedges' *g* in observed studies	0.666
The criterion for a 'trivial' Hedges' *g*	0.010
Mean Hedges' *g* in missing studies	0.000
Number missing studies needed to bring Hedges' *g* under 0.01	27,698

the published studies at the 0.01 significance level (Orwin, 1983). This result reinforces the conclusion that publication bias is unlikely to have a significant impact on the findings of this meta-analysis, thus confirming the robustness of the conclusions.

Additionally, the trim-and-fill method is applied to adjust the distribution of effect sizes, addressing any potential asymmetry by trimming or filling values on both sides of the distribution. The results, presented in Table 7, showed that the difference between the adjusted effect size ($g = 0.67686$) and the observed effect size ($g = 0.67686$) is negligible. This indicates that no significant publication bias is detected during the analysis.

To summarize, the results from these multiple methods suggest that publication bias in this study is minimal. These findings support the reliability of the meta-analysis and confirm the robustness of the results, ensuring that the conclusions drawn from the study are not unduly influenced by publication bias.

Subgroup analysis

According to Lipsey (2001), it is crucial to highlight the magnitude of effects when statistically significant heterogeneity is present in moderator analyses. The purpose of the subgroup analysis in this study was therefore to better understand heterogeneity within the data collected via the PRISMA framework. Specifically, nine key moderators were examined: mathematics topics, teaching content 1, teaching content 2, teaching content 3, mathematical ability, grade level, educational assessment methods, teaching methods, and type of technology.

As shown in Table 8, Mathematics Topics as moderators showed significant instructional effects. Foundation Mathematical Concepts had the highest effect size of 0.725 ($Z = 31.311$, $p < 0.001$), but with moderate heterogeneity ($I^2 = 50.319$), indicating some variability among studies. Higher-level Mathematical Concepts demonstrated an effect size of 0.671 ($Z = 28.057$, $p < 0.001$) with lower heterogeneity ($I^2 = 47.229$), suggesting higher consistency across studies. In contrast, Mathematical Abilities had a lower effect size of 0.590 ($Z = 14.900$, $p < 0.001$) and the highest heterogeneity ($I^2 = 66.170$), indicating significant variability across studies. This variability may reflect differences in sample populations, instructional designs, or other study characteristics.

This conclusion is consistent with the findings of He et al. (2024), which emphasized that the difficulty of Foundation Mathematical Concepts is significantly lower than that of Higher-level Mathematical Concepts, thereby resulting in larger effect sizes for the former. Similar patterns have been observed in prior meta-analyses on educational technology in mathematics classrooms. For instance, Li and Ma (2010) found that technology interventions produced stronger effects for younger students and basic mathematical skills, while Cheung and Slavin (2013) reported modest overall effects (+0.15) but highlighted larger gains for supplemental computer-assisted instruction in fundamental

Table 7 Duval and Tweedie's trim and fill (looking for missing study to the left of mean)

		Random effects			
	Missing studies	Point estimate	Lower limit	Upper limit	Q value
Observed values		0.67686	0.646	0.708	936.613
Adjusted values	0	0.67686	0.646	0.708	936.613

Table 8 Subgroup analysis results

Group	Effect size and 95% CI						Test (2-Tail)		Heterogeneity			
	Num. Studies	Point Est.	Std. Err.	Variance	Lower Limit	Upper Limit	Z-val.	P-val.	Q-val.	*df(Q)*	P-val.	I^2
Mathematics topics												
Foundation Mathematical Concepts	181	0.725	0.023	0.001	0.680	0.770	31.311	< .001	362.308	180	< .001	50.319
Higher-level Mathematical Concepts	152	0.671	0.024	0.001	0.624	0.717	28.057	< .001	286.142	151	< .001	47.229
Mathematical Abilities	90	0.590	0.040	0.002	0.0513	0.668	14.900	< .001	263.079	89	< .001	66.170
Teaching content 1												
Calculations	79	0.735	0.034	0.001	0.668	0.802	21.563	< .001	146.498	78	< .001	46.757
Functions	103	0.597	0.028	0.001	0.542	0.653	21.082	< .001	179.053	102	< .001	43.034
Plane Geometry	132	0.712	0.026	0.001	0.660	0.764	26.994	< .001	257.214	131	< .001	49.070
Statistics	4	0.498	0.182	0.033	0.142	0.854	2.743	< .01	7.670	3	> .050	60.888
Three-Dimension Geometry	105	0.674	0.037	0.001	0.601	0.746	18.159	< .001	322.335	104	< .001	67.735
Teaching content 2												
Functions	40	0.650	0.050	0.002	0.552	0.747	13.027	< .001	91.810	39	< .001	57.521
Plane Geometry	96	0.676	0.031	0.001	0.615	0.738	21.562	< .001	183.471	95	< .001	48.221
Three-Dimension	39	0.679	0.040	0.002	0.601	0.757	17.044	< .001	47.770	38	> .050	20.452
No	248	0.682	0.022	0.000	0.639	0.724	31.388	< .001	612.874	247	< .001	59.698
Teaching content 3												
Functions	2	0.649	0.161	0.026	0.334	0.964	4.037	< .001	0.118	1	> .050	0.000
Plane Geometry	5	0.650	0.177	0.031	0.303	0.997	3.674	< .001	13.711	4	< .010	70.827
Statistics	22	0.555	0.047	0.002	0.463	0.646	11.892	< .001	18.529	21	> .050	0.000
No	394	0.684	0.017	0.000	0.652	0.716	41.341	< .001	898.381	393	< .001	56.255
Mathematical abilities												
Numerical Ability	48	0.625	0.042	0.002	0.542	0.708	14.794	< .001	84.843	47	< .010	44.604
Problem-solving Ability	213	0.744	0.022	0.000	0.701	0.787	33.758	< .001	461.720	212	< .001	54.085
Spatial Skills	162	0.603	0.025	0.001	0.554	0.652	24.041	< .001	350.059	161	< .001	54.008

Table 8 (continued)

Group	Effect size and 95% CI						Test (2-Tail)		Heterogeneity			
	Num. Studies	Point Est.	Std. Err.	Variance	Lower Limit	Upper Limit	Z-val.	P-val.	Q-val.	*df*(*Q*)	P-val.	I^2
Grade level												
High School	110	0.583	0.027	0.001	0.530	0.636	21.523	< .001	196.116	109	< .001	44.421
Middle School	239	0.695	0.020	0.000	0.656	0.735	34.385	< .001	492.491	238	< .001	51.674
Primary School	74	0.766	0.046	0.002	0.677	0.855	16.817	< .001	218.877	73	< .001	66.648
Educational assessment methods												
Focus on both Purpose and Process	141	0.641	0.026	0.001	0.589	0.693	24.275	< .001	285.519	140	< .001	50.966
Focus on the Purpose	282	0.694	0.020	0.000	0.656	0.733	35.568	< .001	645.711	281	< .001	56.482
Teaching methods												
Combine with Constructivist Teaching	133	0.705	0.025	0.001	0.657	0.754	28.514	< .001	232.028	132	< .001	43.110
Combine with Scenario-based Teaching	125	0.646	0.028	0.001	0.591	0.701	23.022	< .001	252.999	124	< .001	50.988
Combine with Traditional Teaching	165	0.677	0.028	0.001	0.622	0.731	24.258	< .001	445.662	164	< .001	63.201
Type of tech												
AI Smart Classroom	36	0.697	0.042	0.002	0.614	0.779	16.534	< .001	47.636	35	> .050	26.527
AR/VR	78	0.788	0.035	0.001	0.719	0.857	22.328	< .001	155.161	77	< .001	50.374
DMS	309	0.646	0.019	0.000	0.610	0.683	34.581	< .001	706.655	308	< .001	56.414

topics. Collectively, these results suggest that MET is more beneficial for foundational concepts, whereas advanced content or ability-oriented outcomes are associated with higher variability and smaller effect sizes.

Significant differences in effect sizes were observed across different types of Teaching Content 1. Calculations had the highest effect size of 0.735 ($Z = 21.563$, $p < 0.001$) with moderate heterogeneity ($I^2 = 46.757$), suggesting relatively consistent results across studies. Plane Geometry showed an effect size of 0.712 ($Z = 26.994$, $p < 0.001$) with heterogeneity of 49.070, indicating moderate variability. Three-Dimension Geometry had an effect size of 0.674 ($Z = 18.159$, $p < 0.001$) with high heterogeneity ($I^2 = 67.735$), reflecting significant inconsistencies among study results. Statistics had the lowest effect size of 0.498 ($Z = 2.743$, $p < 0.01$) with high heterogeneity ($I^2 = 60.888$), suggesting substantial between-study differences but limited overall effectiveness. Functions had an effect size of 0.597 ($Z = 21.082$, $p < 0.001$) and the lowest heterogeneity ($I^2 = 43.034$), indicating relatively consistent results across studies.

In Teaching Content 2, where most studies addressed single content areas, the overall effect size of subgroup "No" was 0.682 ($Z = 31.388$, $p < 0.001$) with heterogeneity of 59.698, suggesting notable between-study variation. Plane Geometry had an effect size of 0.676 ($Z = 21.562$, $p < 0.001$) with moderate heterogeneity ($I^2 = 48.221$), indicating consistent results. Three-Dimension Geometry had an effect size of 0.679 ($Z = 17.044$, $p < 0.001$) with the lowest heterogeneity ($I^2 = 20.452$), showing high consistency across studies. However, Functions had an effect size of 0.650 ($Z = 13.027$, $p < 0.001$) and higher heterogeneity ($I^2 = 57.521$), indicating significant between-study differences.

Most studies did not address the third component of instructional content in Teaching Content 3. Functions demonstrated an effect size of 0.649 ($Z = 4.037$, $p < 0.001$) with no heterogeneity ($I^2 = 0.000$), reflecting complete agreement across studies. Plane Geometry had an effect size of 0.650 ($Z = 3.674$, $p < 0.001$) with the highest heterogeneity ($I^2 = 70.827$), indicating substantial inconsistencies. Statistics had an effect size of 0.555 ($Z = 11.892$, $p < 0.001$) with no heterogeneity ($I^2 = 0.000$), showing high consistency, likely due to the small number of studies ($df(Q) = 22$) in this category.

Among the three teaching contents, "Calculations" showed the largest effect size, followed by Plane Geometry and Three-Dimension Geometry. Although this appears to contradict previous findings that highlight stronger effects for visually oriented topics (Schoenherr et al., 2024; Çavuş & Deniz, 2022), it can be explained by the structure of the Chinese curriculum, where computation is introduced as foundational content at the primary level. The relative simplicity of this material likely enables MET to yield greater gains, consistent with Ran et al. (2021) finding that problem-solving systems targeting computational tasks produced the largest effects among technology-based mathematics interventions. To further strengthen the analysis, this study incorporated data from master's and doctoral dissertations in mainland China, helping to address the lack of long-term tracking data on MET effects. As instructional sequences in these studies often covered multiple topics, the subgroup analysis of the three content moderators was conducted to generate data-driven insights into their internal relationships and to improve the interpretation of the predictive modeling results.

Mathematical Abilities as moderators showed significant instructional effects. Problem-Solving Ability had the highest effect size of 0.744 ($Z = 33.758$, $p < 0.001$) with moderate heterogeneity ($I^2 = 54.085$), indicating a strong effect but moderate variability among studies. Numerical Ability had an effect size of 0.625 ($Z = 14.794$, $p < 0.001$) with lower heterogeneity ($I^2 = 44.604$), reflecting more consistent results. Spatial Skills had an effect size of 0.603 ($Z = 24.041$, $p < 0.001$) with moderate heterogeneity ($I^2 = 54.008$).

Within the moderator of mathematical abilities, problem-solving ability exhibited the largest effect size, surpassing both numerical ability and spatial skills. This pattern is consistent with recent meta-analytic evidence (He et al., 2024). Ran et al. (2021) demonstrated that problem-solving systems generated the strongest effects among technology-based mathematics interventions, while Kim and Xin (2024) reported very large effects for technology-supported word-problem solving. With respect to spatial skills, a meta-analysis of virtual technologies in K–12 settings identified moderate effects, particularly pronounced for augmented reality (Di & Zheng, 2022). Nevertheless, it should be acknowledged that all data in the present study were drawn from mainland China, where DMS remains the most widely used form of MET. Furthermore, as noted by He et al. (2024), mathematics abilities assessed predominantly through paper-and-pencil tests may underestimate students' spatial reasoning. Taken together, these findings suggest that while MET benefits all three domains, problem-solving ability gains the most, followed by numerical ability and spatial skills.

Grade Level revealed significant differences in instructional effects across educational stages. Primary School had the highest effect size of 0.766 ($Z = 16.817$, $p < 0.001$) but also the highest heterogeneity ($I^2 = 66.648$), indicating substantial between-study variability. Middle School had an effect size of 0.695 ($Z = 34.385$, $p < 0.001$) with moderate heterogeneity ($I^2 = 51.674$), reflecting moderate consistency. High School had the lowest effect size of 0.583 ($Z = 21.523$, $p < 0.001$) and the lowest heterogeneity ($I^2 = 44.421$), suggesting higher consistency in results.

Several meta-analyses of technology-enhanced mathematics instruction highlight that grade level significantly moderates instructional effects. Akçay et al. (2021) reported a moderate overall effect for MET in primary school settings, consistent with our finding that the effect size was largest at this level. Bush et al. (2015) emphasized that comprehensive integration of technology amplifies learning outcomes across grade levels, while Cheung and Slavin (2013) provided baseline evidence indicating generally modest average effects in K–12 classrooms, against which the substantial elementary school gains observed in our study are particularly notable. Moreover, Rakes et al. (2020) and He et al. (2024) stressed that grade level interacts with instructional design and content domains, underscoring the importance of aligning MET with developmental stages. Since early education emphasizes foundational mathematical content, these findings also correspond to the subgroup analysis under our "Mathematics Topics" moderator. Taken together, the evidence confirms that MET is most effective in the early grades and highlights the need for developmentally appropriate intervention design.

Educational Assessment Methods also demonstrated significant effects as moderators. Methods focused solely on purpose had an effect size of 0.694 ($Z = 35.568$, $p < 0.001$) with high heterogeneity ($I^2 = 56.482$), indicating greater variability among studies. Methods considering both purpose and process had an effect size of 0.641 ($Z = 24.275$, $p < 0.001$) with lower heterogeneity ($I^2 = 50.966$), suggesting more consistent results.

Although direct meta-analyses comparing purpose-only versus purpose-and-process assessment methods in MET are limited, related evidence underscores the critical role of assessment design. Ran et al. (2022) demonstrated that the functional roles of technology, including the way assessments are implemented, significantly influence mathematics achievement. Moreover, a scoping review of technology use in mathematics education identified assessment, alongside pedagogy and content integration, as a central research theme (Hwang et al., 2023). These findings align with our results: purpose-only assessments may yield higher but less stable effects, whereas combining purpose with process tends to generate more consistent outcomes.

Teaching Methods revealed notable differences. Combine with Constructivist Teaching had the highest effect size of 0.705 ($Z = 28.514$, $p < 0.001$) with the lowest heterogeneity ($I^2 = 43.110$), indicating highly consistent and effective results. Combine with Traditional Teaching had an effect size of 0.677 ($Z = 24.258$, $p < 0.001$) with the highest heterogeneity ($I^2 = 63.201$), reflecting significant inconsistencies among studies. Combine with Scenario-Based Teaching had an effect size of 0.646 ($Z = 23.022$, $p < 0.001$) with moderate heterogeneity ($I^2 = 50.988$).

The subgroup analysis of teaching methods indicates that integrating technology with different instructional approaches yields varying effectiveness. Li and Ma (2010) reported that constructivist approaches exerted the strongest influence on mathematics achievement, with all methods outperforming traditional instruction. Similarly, Xie et al. (2018) found that in mainland China, constructivist and improved transmission models were more effective than traditional methods. These findings are consistent with our results: MET combined with constructivist approaches achieved the highest and most stable effects, followed by traditional methods (albeit with greater variability), while scenario-based approaches produced moderate outcomes.

Technology Type revealed significant differences in instructional effects. AR/VR Technology had the highest effect size of 0.788 ($Z = 22.328$, $p < 0.001$) with moderate heterogeneity ($I^2 = 50.374$), indicating substantial between-study variability. AI Technology had an effect size of 0.697 ($Z = 16.534$, $p < 0.001$) with the lowest heterogeneity ($I^2 = 26.527$), suggesting highly consistent results. DMS Technology had an effect size of 0.646 ($Z = 34.581$, $p < 0.001$) with high heterogeneity ($I^2 = 56.414$), reflecting considerable variability.

The subgroup analysis of technology types further underscores these differences. AR and VR achieved the largest effects, though with moderate heterogeneity, confirming their effectiveness but also their sensitivity to instructional context (Na et al., 2025). This aligns with meta-analyses showing that AR/VR interventions significantly enhance mathematics learning, particularly in geometry and visualization tasks (Ersen & Alp, 2024). AI technologies produced smaller but highly consistent effects, supporting their role in adaptive practice and intelligent tutoring (Hwang, 2022; Son, 2024). DMS, while widely implemented in mainland China, showed lower effects and higher heterogeneity, likely due to variability in teacher training and integration. Prior reviews note that while DMS can enhance procedural fluency, its impact is highly context-dependent (Cheung & Slavin, 2013). Overall, AR/VR is most effective for conceptual understanding, AI provides stable adaptive support, and DMS effectiveness depends heavily on instructional design and fidelity.

In summary, the effect sizes and Z-values of all moderators in this subgroup analysis reached statistically significant levels ($Z > 1.96$, $p < 0.05$), indicating that each moderator had a positive effect on learning outcomes across various study settings. These findings suggest that the identified moderators not only help explain the sources of heterogeneity among studies but also validate their classification as potential predictive model features. Furthermore, they provide a solid foundation for optimizing instructional design in future research and practice. However, the heterogeneity test results revealed considerable variation in the degree of heterogeneity across moderators, with most exceeding a high level of heterogeneity ($I^2 > 50$) except in cases where the number of studies was limited. This indicates that the effect sizes of MET use within the same moderator category can vary significantly due to interactions with other moderators. Consequently, these findings imply that traditional regression models (e.g., linear regression) may not adequately capture the complex variations in sample characteristics. Instead, more advanced modeling approaches, such as tree-based models, may be better suited for analyzing the database of this study and addressing the intricate relationships among moderators.

Predictive models testing and optimization

To address the second and third research questions, namely, how to establish predictive models for determining the optimal effect size of MET under different optimization frameworks, and how specific predictor variables influence MET effectiveness. In this section, the experimental process of predictive modeling is divided into three stages.

In the first stage, nine predictive models will be evaluated for regression tasks, testing them sequentially under three frameworks. The models are first tested without enhancements (i.e. base model), followed by the integration of L2 regularization, SMOTER, and AL in that order. Performance is assessed using R^2, MSE, MAE, and RMSE metrics. To ensure consistency, the division of training and test sets remained constant across all tests within the same optimization tool. Additionally, K-fold cross-validation (with $k = 5$) is employed, and the reported values represent the average results across all five folds. In the second stage, a PSO is then employed to fine-tune the hyperparameters of the best performing predictive model with the primary objective to minimize the MAE. Finally, the proposed PSO-optimized predictive model will be subjected for a detailed feature importance analysis using SHAP value.

Predictive modeling tested in three frameworks

The authors first evaluated the models using R^2, a measure of regression model fit that reflects the model's ability to explain variance in the target variable. Its value ranges from [0, 1], with values closer to 1 indicating better model performance (Chollet & Chollet, 2021). In the base framework, the R^2 values of the models were generally low, demonstrating their inability to effectively capture the variance in the target variable without enhancement techniques. Among these models, XGB performed best (as shown in Table 9), achieving an R^2 of 0.240, followed by GB (0.224) and RF (0.164). In contrast, linear models such as LR (0.014) and BR (0.002) performed poorly, highlighting their limited capacity to handle complex, nonlinear data structures. This observation is consistent with conclusions drawn from the heterogeneity analysis of the dataset.

Table 9 R^2 comparison across optimization methods

Model	Base model	With SMOTER and L2 regularization	With AL, SMOTER and L2 regularization
LR	0.014	0.181	0.153
BR	0.002	0.180	0.174
SVR	0.066	0.352	0.613
DT	0.071	0.665	0.734
RF	0.164	0.807	0.811
GB	0.224	0.666	0.794
ANN	0.019	0.435	0.567
KNN	0.119	0.186	0.513
XGB	**0.240**	**0.827**	**0.822**

The introduction of the SMOTER data-balancing technique and L2 regularization led to a significant improvement in R^2 values. XGB achieved the best performance with an R^2 of 0.827, followed by RF (0.807) and GB (0.666). These enhancements enabled tree-based models to better capture the complexity of the data and address category imbalances effectively. Linear models such as LR and BR also improved, achieving R^2 values of 0.181 and 0.180, respectively. However, their performance remained significantly behind that of nonlinear models, highlighting their inherent limitations.

The incorporation of AL further enhanced the performance of several models. XGB remained the top-performing model with an R^2 of 0.822, followed by RF (0.811) and GB (0.794). These results underscore the robustness of tree-based models under enhanced frameworks. Moreover, complex models such as SVR (0.613) and ANN (0.567) demonstrated notable performance improvements, illustrating the benefits of combining AL with data balancing and regularization techniques.

Next, the authors evaluated MAE, MSE, and RMSE across the three frameworks. Lower values across these metrics indicate better model performance. To visually analyze these metrics, Fig. 5 is plotted, which is divided into three subgraphs corresponding to the different frameworks. In each subfigure, models are ranked in descending order based on the sum of their MAE, MSE, and RMSE values.

Under the base framework, tree-based models such as XGB, GB, and RF demonstrated relatively strong performance, with XGB achieving the best results with the values: (MAE, MSE, RMSE) = (0.194, 0.124, 0.353). In contrast, linear models such as LR and BR showed poor performance, with significantly higher metric values, emphasizing their inability to handle nonlinear and complex data structures.

The introduction of SMOTER and L2 regularization had led to significant improvements across all metrics, particularly for nonlinear models. XGB continued to outperform others, achieving results of (MAE, MSE, RMSE) = (0.089, 0.023, 0.153). RF and GB also demonstrated strong performance, with RF achieving results of (MAE, MSE, RMSE) = (0.107, 0.026, 0.161). Although linear models such as LR and BR improved under this framework (e.g., LR achieved results of (MAE, MSE, RMSE) = (0.181, 0.111, 0.333)), their performance gains were limited and remained well behind those of nonlinear models.

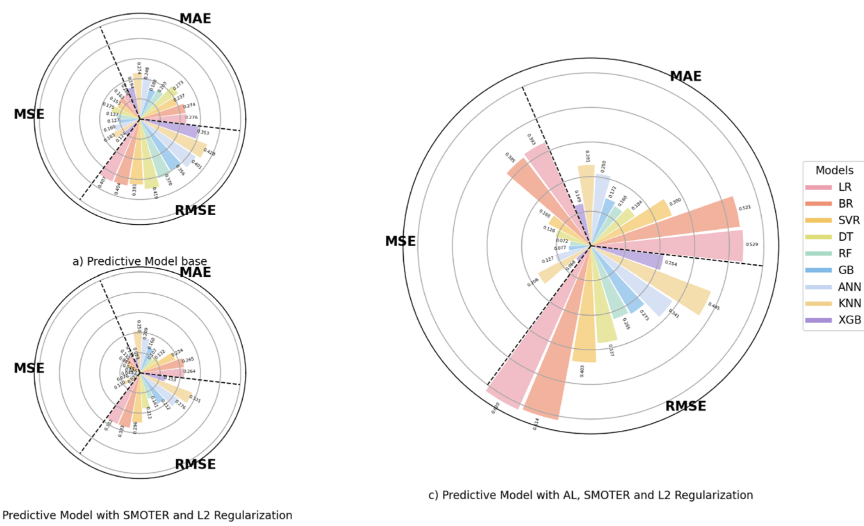

Fig. 5 Metric comparison of three frameworks

The addition of AL further enhanced the performance of nonlinear models. XGB remained the top-performing model, achieving results of (MAE, MSE, RMSE) = (0.149, 0.068, 0.254). RF and GB also maintained their strong competitiveness, while SVR and ANN showed considerable improvements. For example, SVR achieved results of (MAE, MSE, RMSE) = (0.300, 0.165, 0.403), while ANN recorded results of (MAE, MSE, RMSE) = (0.250, 0.127, 0.341). In contrast, improvements in linear models remained modest, reflecting their inherent limitations in modeling nonlinear relationships.

In summary, the analysis consistently highlights the superior performance of tree-based models such as XGB, RF, and GB, particularly when combined with SMOTER, L2 regularization, and AL. These enhancements not only improve model accuracy but also demonstrate robustness in handling complex, nonlinear data. In contrast, while linear models benefited from these optimizations, their performance remains constrained by their inability to capture nonlinear data structures, further supporting conclusions drawn from the dataset heterogeneity analysis.

PSO-based hyperparameter tuning

From the experiment conducted in the previous section, the authors identified that the XGB, combined with AL, SMOTER, and L2 regularization, as the most suitable predictive model based on both metric performance and theoretical considerations. To further improve the model's performance, a PSO is employed to fine-tune the hyperparameters of the proposed XGB with the primary objective to minimize the MAE. Focusing on MAE minimization is particularly relevant given the nature of the dataset, which originates from educational experiments. In practice, it is unrealistic to expect identical performance across students or classes, even in the same instructional environment. Consequently, prioritizing a higher model fit, as measured by R^2, is less meaningful in this context. Instead, reducing the gap between predicted and actual values provides a more practical and realistic evaluation of outcomes.

Table 10 Comparison of metrics between original XGB and PSO-optimized XGB

Metric	Original XGB	PSO-Optimized XGB
R^2	0.822	0.842
MAE	0.149	0.084
MSE	0.068	0.021
RMSE	0.254	0.146

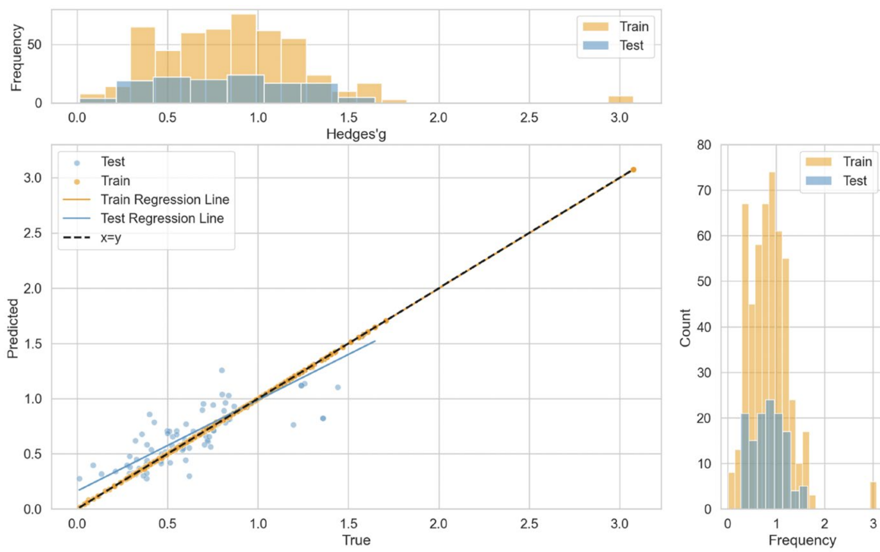

Fig. 6 Visualization of PSO-optimized XGB model performance

Through PSO optimization, the optimal hyperparameters for XGB are determined as follows:

- Learning Rate: 0.126
- Max Depth: 8
- Subsample: 0.963
- Colsample by Tree: 0.825
- Reg Alpha: 0.0
- Reg Lambda: 1.822

A comparison of the performance improvement between the original XGB and the PSO-optimized XGB model reveals significant enhancements across all metrics (Table 10). The R^2 value increased from 0.822 to 0.842, indicating that the optimized model better captures variance in the data. MAE decreased significantly from 0.149 to 0.084, reflecting a notable reduction in prediction error. Similarly, MSE dropped from 0.068 to 0.021, while RMSE decreased from 0.254 to 0.146.

To rigorously evaluate the results, the training and testing performance of the PSO-optimized XGB model are visualized in Fig. 6, which includes histograms on the top and right sides, as well as a central scatter plot. In the scatter plot, the horizontal axis represents the actual values of Hedges' *g*, and the vertical axis represents the predicted values. Orange and blue points correspond to the training and testing data, respectively. The data points are distributed closely around the diagonal line, indicating accurate predictions. The fit for the training data is better than that for the test data, though the regression line for the test data deviates slightly from the diagonal, suggesting a small prediction error.

The top histogram shows the distribution of actual values for both training and testing data, primarily concentrated between 0.5 and 2.0, with a consistent distribution across datasets. The right histogram depicts the distribution of predicted values, with training data predictions being more concentrated and testing data predictions having a slightly broader distribution. This indicates that the model's prediction error on the test set is slightly higher than on the training set. Overall, the model demonstrates strong performance on the training set and good generalization ability on the test set.

These results highlight the effectiveness of PSO in fine-tuning XGB hyperparameters, significantly improving predictive accuracy. The reductions in MAE, MSE, and RMSE underscore the model's ability to minimize errors, particularly large deviations, which are critical in this context. This validates PSO as an effective optimization method for enhancing XGB performance.

Features importance and SHAP analysis

After optimizing the hyperparameters of the proposed XGB model using PSO, its performance is significantly improved. These optimized hyperparameters are subsequently used to conduct multiple simulations and experiments with varying dataset splits. The results demonstrated consistent and stable performance metrics across different experimental conditions, indicating the model's robustness. Based on this optimized model, a detailed feature importance analysis is performed.

Shapley Additive Explanations (SHAP) is a game theory-based method for interpreting machine learning model outputs, used to assess the importance of each feature in a prediction (Lundberg et al., 2018). SHAP assigns importance scores to features, revealing their impact on the model's predictions. As shown in Fig. 7, the SHAP Summary Plot on the left illustrates the extent to which each feature influences the model's prediction results. The horizontal axis represents SHAP values, showing the magnitude of a feature's positive or negative effect on the predicted outcome. Each dot in the plot corresponds to a sample's SHAP value, quantifying the impact of that feature on the sample's prediction. The pie chart on the right is a circular visualization derived from the absolute mean SHAP values, showing the relative importance of each feature in the model. The size of each segment reflects the percentage contribution of the feature to the model's output.

Through SHAP analysis, this study evaluated the relative importance of features within the model, revealing their differential contributions to prediction. Among them, MET Usage Time emerged as the most influential feature, accounting for 48.5% of the total importance. The wide spread of SHAP values for this feature in the summary plot indicates a strong and stable positive influence on the model's output.

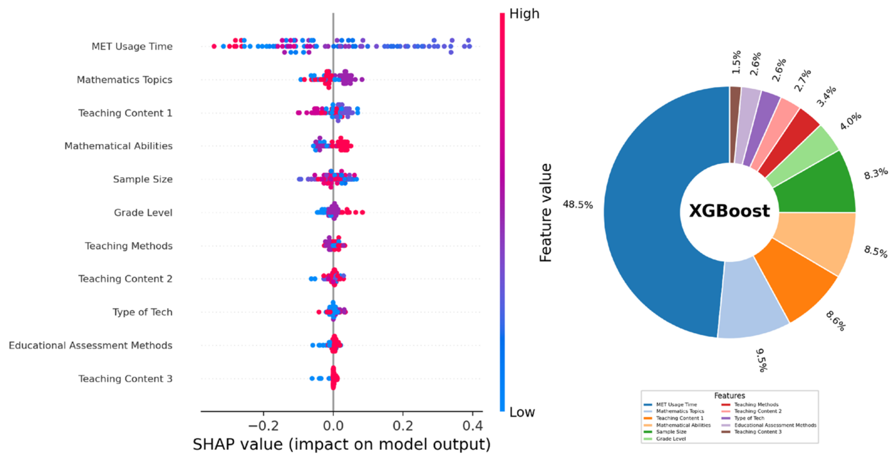

Fig. 7 SHAP summary plot

A review of prior literature suggests that prolonged interventions may lead to fade-out effects, where initial learning gains observed at the end of an intervention diminish substantially during follow-up assessments, often within a few months (Bailey et al., 2016). In a recent meta-analysis of MET, He et al. (2024) further noted that extended use of technology may foster a form of visual dependence, reducing opportunities for students to effectively exercise mathematical reasoning skills. Moreover, Lee et al. (2025) and Rodrigues et al. (2022) highlighted that the novelty effect plays a strong facilitating role in the early stages of technology-enhanced mathematics learning, though such benefits tend to taper over time. However, no existing research has explicitly characterized the distributional relationship between MET usage time and its effect size.

According to the SHAP color bar analysis, features with moderate values generally exerted positive impacts on predictions, whereas both high and low values had negative effects. This pattern suggests that MET usage time follows an inverted U-shaped relationship with effectiveness: both insufficient and excessive exposure reduce the benefits, while an optimal range of usage may maximize outcomes under comparable instructional conditions. This finding underscores the pivotal role of usage duration in determining the impact of MET and provides strong justification for including MET usage time as a core optimization variable in this study.

The second most important feature was "Mathematics Topics", contributing 9.5% to the total importance. The distribution of its SHAP values suggests that the selection of basic mathematical concepts in the mathematics curriculum had a significant positive impact on the effectiveness of MET. This aligns with findings from subgroup analysis in the meta-analysis. Similarly, "Teaching Content 1" ranked third, contributing 8.6% of the importance. The high SHAP values associated with this feature suggest that teaching geometry (including plane and solid geometry) using MET had a notably stronger positive effect compared to other teaching contents.

"Mathematical Abilities" ranked fourth, contributing 8.5%. This result highlights MET's strong advantage in improving students' mathematical abilities, especially in fostering problem-solving skills. "Sample Size," accounting for 8.3%, also demonstrated a wide range of SHAP values, indicating that an adequate sample size positively impacts the robustness of the model and the stability of its predictions.

"Grade Level" contributing 4.0%, moderately influenced the model's predictions. SHAP analysis showed that younger students exhibited a more pronounced positive effect, likely due to the foundational nature of mathematics learning at lower grade levels. This observation aligns with subgroup analysis on mathematics topics, which found that basic mathematical content yielded the highest effect sizes. "Teaching Methods" contributing 3.4%, had a smaller overall effect on the model, with SHAP values highly concentrated, suggesting that the choice of teaching method did not significantly influence MET's effectiveness.

Interestingly, "Teaching Content 2" and "Teaching Content 3" were the least important features, contributing 2.7% and 1.5%, respectively. Although instructional content design may influence MET effectiveness in specific cases, its overall contribution to the model was relatively small. However, due to this study's data collection characteristics, these features cannot be entirely disregarded. It is suggested that future studies consider simplifying or homogenizing the representation of instructional content to capture its impact better.

In summary, this study highlighted "MET Usage Time", "Mathematics Topics", and "Teaching Content 1" as the most important features in the model, together accounting for over 65% of the total importance. These findings suggest that these features play a pivotal role in enhancing the effectiveness of mathematics education. Moreover, the study emphasizes the importance of optimizing the time spent using educational technology and designing well-structured course content to improve student learning outcomes.

Validation of predictive models with real-world application

The final research question of this study concerns whether the predictive model can be effectively applied in real-world contexts to guide the implementation strategies of MET, as well as its ability to generalize to external data. Based on the analysis presented, MET usage time is identified as the most influential feature affecting MET effectiveness. Therefore, in this section, an XGB model enhanced with SMOTER, L2 regularization, and AL, with hyperparameters optimized through PSO, is employed to perform conditional screening of the MET usage time feature. The range of MET usage time is constrained to positive integers, with the upper bound set to the maximum instructional period for the given content.

The primary objective of this optimization process is to ensure that the model output, namely, the predicted effect size of MET, exceeds the average observed effect size (0.677, as shown in Table 4). This approach validates the proposed predictive model, which integrates advanced optimization techniques with data-driven insights, by demonstrating its effectiveness in improving learning outcomes through the optimization of MET usage duration. At the same time, the collected data serve as external evidence to further evaluate the model's generalization capability and robustness.

Predictive modeling-guided usage time screening for MET

This experiment involved eight senior classes from a middle school in Southeastern mainland China. All factors, including teaching content, teaching time, pace, and assessment methods, are consistent across the eight classes, except for differences in the time allocated for using MET.

According to the teaching schedule, the students learned analytic geometry, covering topics such as the equations of circles, ellipses, hyperbolas, and parabolas. These high level mathematical concepts are categorized under "mathematics topics", with Teaching Content 1 classified as "plane geometry", while the other two instructional content features are categorized as "No".

The syllabus and the Chinese University Entrance Exam Guide emphasized developing problem-solving skills as the primary goal and spatial skills as a secondary focus (Ministry of Education, 2017). Consequently, the feature of mathematical ability is categorized as "problem-solving ability". To ensure objectivity in assessment, given that four different teachers are involved in teaching the eight classes, a paper-based test is chosen as the assessment method. Therefore, the feature of educational assessment is categorized as "focused on purpose".

Considering that analytic geometry is among the most challenging topics to teach at the high school level in mainland China (Wang & Paine, 2023), the four teachers involved in the experiment decided to adopt a constructivist teaching approach to minimize the abstraction of the content (Wang, 2021; Granberg & Olsson, 2015; Romero Albaladejo, 2024). The MET used is GeoGebra, a dynamic mathematical software tool. All teachers are proficient in its use, and its effectiveness in teaching analytic geometry has been demonstrated in similar contexts in mainland China.

Each of the eight classes consisted of 45 students, resulting in a sample size of 90, combining both experimental and control groups. The teaching program included 32 teaching periods and 5 tutorial periods, making the MET usage time (evaluation period) a positive integer ranging from 0 to 32 periods. The screening process involved fixing all features except for the "evaluation period" and sequentially inputing positive integers (0–32) for this feature into the trained model. The output of the model is observed, and the evaluation period corresponding to the highest predicted value is recorded.

The results based on numerical simulations demonstrate the trend between the time of using MET and the effect size. As shown in Fig. 8, the effect size is low during the first few sessions, with a minimum value of 0.037 (only one session is used). Over time, effect sizes gradually increased, especially after the eighth period time of use, where effect sizes rose significantly until they reached a maximum value of 0.754 (teaching period: 23). After that, the effect size fluctuated but remained at a high level overall, showing a positive correlation between the increased time spent using MET and a significant increase in effect size. Considering that the process MET is only used as an aid to teaching and learning and that prolonged use may cause students to become dependent on the technology. Therefore, after a discussion with the teachers involved in this experiment, it is decided that MET would be used for 18 teaching periods for the experimental group.

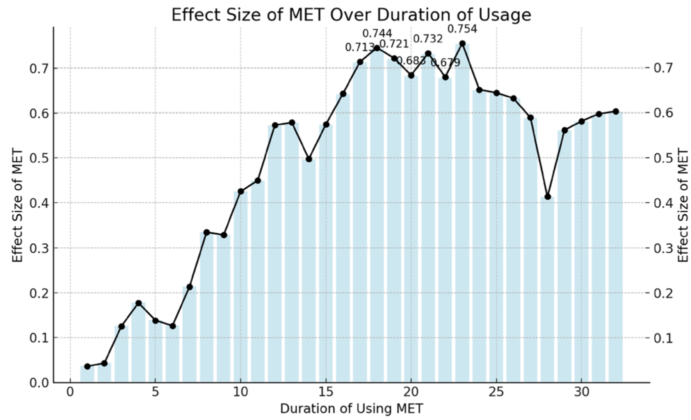

Fig. 8 Results of feature selection by PSO-optimized XGB with SMOTER, L2 regularization and AL

Table 11 Independent samples t-test for pre-test scores

Group	*N*	Mean	Std. Dev.	t-value	p-value
Control Group	45	106.87	18.02		
Experimental Group 1	45	110.32	17.24	−0.916322	0.362002
Control Group	45	106.87	18.02		
Experimental Group 2	45	107.99	17.85	−0.292752	0.770401
Control Group	45	106.87	18.02		
Experimental Group 3	45	112.12	14.63	−1.500472	0.137072
Control Group	45	106.87	18.02		
Experimental Group 4	45	112.26	13.13	−1.603664	0.112371
Control Group	45	106.87	18.02		
Experimental Group 5	45	108.54	16.60	−0.449968	0.653840
Control Group	45	106.87	18.02		
Experimental Group 6	45	113.19	16.09	−1.735460	0.086160
Control Group	45	106.87	18.02		
Indirect control group	45	109.95	11.98	−0.942851	0.348338

Analysis of pre- and post-test data

A control group is set up for this experiment, where no MET is used for the class, and only the traditional teaching method is used for teaching. An indirect control group where the instructor teaches the class using any MET based on experience. An additional six experimental groups in which the teachers used MET with the courses regularly for 18 teaching periods in strict accordance with the instructional plan. To ensure that the mathematics scores of the eight classes that participated in the pre-test remained essentially the same, mid-term test scores of these classes are collected for the pre-test data and subjected to an independent samples t-test.

As shown in Table 11 in the mid-term scores before the start of the experiment, the mathematics test scores between the six experimental groups and one indirect control group and the control group have all p-values greater than 0.05 on the independent samples t-tests. It indicates that, there is no statistically significant difference between the experimental and control groups and between the indirect control and control groups, i.e., all of the experimental classes are at almost the same level of mathematics proficiency as the control classes. Therefore, these classes fulfill the criteria for conducting a controlled experiment.

At the beginning of the experiment, teachers of the six experimental group classes used MET for 18 teaching periods strictly according to the teaching plan of the teaching committee of the school. In contrast, teachers did not use MET for the rest of the lessons. Teachers in the indirect control group used MET to teach 14 teaching periods based on their extensive teaching experience. The control group did not use any MET. At the end of the experiment, the eight classes are given a standardized test on the module, and the test papers are given by two senior teachers who are not involved in the experiment and reviewed by an educational researcher.

After the experiment, an independent samples t-test is performed again on the post-test scores. The p-values for all experimental groups (18 teaching periods of MET) are much lesser than 0.05 (see Table 12), which suggests that the difference in means between the experimental and control groups is highly statistically significant. The t-values are negative, and p-values are very small, indicating that students in these experimental groups performed significantly better than the control group. This suggests that 18 teaching periods of the MET program significantly positively impacted student achievement.

On the other hand, for those who received 14 teaching periods of the MET program (indirect control group), the t-values and *p*-values also shown improvements. The mean difference is slightly smaller than in experimental groups 1–6. This suggests that the 14 teaching periods of MET program still positively impacted student achievement, but the effect was less significant than the 18 teaching periods.

The results of this experiment clearly show that the MET program has a significant effect on student achievement. Students in the experimental groups (18 teaching periods of MET) performed significantly higher than the control group, and the t-test results for all experimental groups indicated that the difference in means is statistically significant.

Table 12 Independent samples t-test for post-test scores

Group	*N*	Mean	Std. Dev.	t-value	p-value
Control group	45	108.79	18.02		
Experimental group 1	45	123.08	17.24	−3.86	< 0.001
Control group	45	108.79	18.02		
Experimental group 2	45	121.32	17.85	−3.48	< 0.001
Control group	45	108.79	18.02		
Experimental group 3	45	125.31	14.63	−4.79	< 0.001
Control group	45	108.79	18.02		
Experimental group 4	45	125.75	13.13	−5.24	< 0.001
Control group	45	108.79	18.02		
Experimental group 5	45	121.84	16.60	−3.86	< 0.001
Control group	45	108.79	18.02		
Experimental group 6	45	126.45	16.09	−5.62	< 0.001
Control group	45	108.79	18.02		
Indirect control group	45	121.76	11.98	−3.88	< 0.001

Although the indirect control group (14 teaching periods of MET) also showed a significant effect, the enhancement is less than the experimental group. This suggests that teaching under the MET guided by the predictive model achieved superior instructional outcomes. At the end of this experiment, the authors organized exercise sessions using MET for students in the control and indirect control groups. No statistical differences have been observed among the eight classes in this knowledge component of the final exam.

External data validation

After collected the grades from the six classes that participated in the experiment, the authors calculated the effect size of using MET for each experimental group. This calculation is based on the data from both the control group and the six experimental groups (see Table 13), and the resulting effect size will serve as an external dataset to re-validate the performance of the model.

Through the analysis in Table 13, the authors observed that the effect size of MET for each experimental group exceeds the optimization objective, which is the average effect size of 0.677. A more detailed examination, based on subgroup analyses, reveals that the benefit of MET use across the six experimental groups is 0.671, which is already higher than the average effect size for MET in 'High-Level Mathematics Concepts', where the mean effect size is 0.712. Additionally, the effect sizes of MET use in all six experimental groups are higher than the mean effect sizes of the other moderators corresponding to the same groups, with the exception of the 'Problem Solving Ability' subgroup. However, it is important to note that the studies included in this subgroup exhibit considerable heterogeneity ($I^2 = 54.085, Q = 461.720$, shown in Table 8), with the highest values observed for this moderator. Therefore, the authors inferred that these studies may have been influenced by the interaction of other factors, and the true effect size for real-world MET use is likely closer to its average effect size, which is a very promising finding.

These six experimental sets of data were then used as external datasets to validate the proposed XGB model with the performance metrics shown in Table 14. The MAE is 0.01255, indicating that the absolute error between the predicted and actual values

Table 13 Hedges' *g* values for experimental groups 1–6 and control group

Group	*N*	Pre-test Mean	Pre-test Std. Dev.	Post-test Mean	Post-test Std. Dev.	Hedges'*g*
Control group	45	106.87	18.02	108.79	18.02	
Experimental group 1	45	110.32	17.24	123.08	17.24	0.727
Control group	45	106.87	18.02	108.79	18.02	
Experimental group 2	45	107.99	17.85	121.32	17.85	0.748
Control group	45	106.87	18.02	108.79	18.02	
Experimental group 3	45	112.12	14.63	125.31	14.63	0.731
Control group	45	106.87	18.02	108.79	18.02	
Experimental group 4	45	112.26	13.13	125.75	13.13	0.755
Control group	45	106.87	18.02	108.79	18.02	
Experimental group 5	45	108.54	16.60	121.84	16.60	0.764
Control group	45	106.87	18.02	108.79	18.02	
Experimental group 6	45	113.19	16.09	126.45	16.09	0.732

Table 14 Results of external data validation

Model performance metrics	MAE	MSE	RMSE
Value	0.01255	0.00020625	0.01436

is, on average, 0.01255. This relatively small error suggests that the model's predictions are very close to the actual values. The MSE between the predicted and actual values is 0.00020625, which is also a very small value, implying that most predicted values exhibit minimal deviation from the actual values. Furthermore, the RMSE is 0.01436, a value similarly small, suggesting that the model exhibits minimal variation between predicted and actual values.

All of these indicators suggest that the proposed XGB model provides a reliable reference for determining the optimal duration of MET use in teaching practice, with the effect sizes following MET use consistently exceeding the average effect size. Although the predicted values did not fully align with the true values during the experiment (e.g., the R^2 validated by the external data was relatively small), it is unlikely that fully consistent instructional outcomes would occur in educational practice, even within the same educational setting. The study results demonstrate that experimental groups 1-6 perform well in terms of Hedges' *g*. While there are minor discrepancies between the predicted and actual values, the error is negligible, and overall, the model shows strong performance.

Implications for research and practice

The main contribution of this study lies in the establishment of a predictive modeling framework that integrates meta-analytic data mining with optimization techniques in contexts where educational data are limited. Unlike previous PLA studies, which often framed research questions around the constraints of available datasets, this framework provides a methodological solution for conducting predictive modeling when large-scale experiments are not feasible. It thus opens new opportunities for extending inquiry into under-researched contexts while offering novel insights into MET use in K–12 mathematics teaching and learning.

From a theoretical perspective, this framework contributes to both PLA and MET researches. For PLA, it provides a replicable method to overcome the recurring obstacle of insufficient or incomplete datasets. By embedding meta-analytic findings into predictive modeling, researchers can compensate for data scarcity and generate new avenues of analysis. Importantly, the framework demonstrates that traditional meta-regression is often insufficient to capture nonlinear relationships among moderators, while machine learning or hybrid approaches are better suited to identify complex patterns. This establishes a pathway for future research to achieve deeper insights into the statistical properties of educational variables and their interactions.

In terms of MET, this study underscores the significance of moderators in shaping the effectiveness of technology use. By employing AI-based predictive models, it revealed that different variables, such as content type, teaching method, grade level, and assessment design, exert distinctive influences. Among these, usage time emerged as the most critical factor, with evidence supporting an inverted U-shaped relationship between duration and effectiveness: both very short and very long exposure reduce outcomes, while moderate use yields the greatest benefits. This result not only highlights an optimization problem for classroom practice but also suggests a promising direction for future experimental research to identify precise thresholds of effective MET use. Furthermore, the findings indicate that evaluation should not be limited to measuring purposes alone but should also account for processes, since such combined designs generate more consistent and stable outcomes.

The practical implications for teaching and learning are equally important. MET was found to be effective not only for geometry, where visualization has long been considered central, but also for strengthening foundational mathematical content and enhancing students' problem-solving abilities. These results support the growing recognition of MET as a catalyst for developing higher-order thinking when integrated with appropriate pedagogical strategies. In particular, the modeling results demonstrate that there exists an optimal range of usage time that can be quantified through data-driven analysis, offering educators a tool for tailoring technology integration to the needs of specific classrooms. This finding challenges the assumption that "more is always better" and highlights instead the necessity of calibrated use. In practice, teachers can leverage this insight to design context-sensitive and developmentally appropriate lesson plans, while curriculum designers and policymakers may use it to guide resource allocation and professional development.

Taken together, these implications emphasize that the effectiveness of MET depends not only on whether it is adopted but on how it is integrated, timed, and assessed. By advancing a replicable methodological framework and generating new evidence on usage optimization, this study contributes both to the refinement of PLA and to the improvement of MET practice. Ultimately, the findings highlight the importance of data-informed, context-sensitive, and developmentally aligned approaches to maximize the impact of educational technologies in mathematics learning.

Conclusions and future work

This study collected 423 eligible publications from four literature databases and conducted a meta-analysis and educational data mining on these studies. The research examines the effectiveness of MET in mathematics learning for K-12 Chinese students and outlines the optimization goals of the study. Furthermore, it identifies and characterizes predictive models suitable for inclusion based on the data derived from the literature review and subgroup analyses. Using these data, the study developed nine AI predictive models for regression tasks under three distinct optimization frameworks.

After comparing the performance of various models, the best-performing model, XGB + L2 Regularization + Active Learning + SMOTER ($R^2 = 0.822$), is selected. Its performance is further enhanced through PSO for hyperparameter tuning. SHAP value analysis is then conducted on the model's features, and, in combination with subgroup analysis, the contribution of each feature is systematically examined. The results indicate that the duration of MET use is a critical optimization variable.

Subsequently, the proposed XGB regression model is applied to characterize the effect sizes corresponding to various MET usage times for specific content and instructional objectives. These are also informed by the insights and feature selection decisions made by senior educators. A controlled experiment is designed based on these feature selection results. The experiment demonstrated that model-guided MET is significantly more effective than both the control and indirect control groups, which did not utilize model-guided MET. Finally, an external dataset is collected based on the experimental outcomes, and the model's performance is re-validated using this dataset. The results confirmed that the model is both stable and reliable.

Given the limitations identified in this study, several areas for further research are suggested to advance the understanding and application of MET. First, the methods of data collection should be expanded to incorporate a broader range of sources. Future research should employ advanced systematic search algorithms to handle large volumes of search results more efficiently and ensure that fewer relevant studies are overlooked. A more comprehensive dataset can be obtained by expanding the search terms and databases used, which would provide a more solid foundation for meta-analysis and alleviate the issue of missed research.

Second, alternative assessment methods in mathematics education should be explored. While this study relied on mathematics scores as the primary assessment metric, future research should consider process-oriented assessments that more accurately reflect the development of students' mathematical skills. Such assessments could include qualitative measures like student reflections, teacher observations, and portfolio assessments, which would offer a fuller picture of student progress. Implementing consistent and reliable process-oriented assessments would provide a more accurate and comprehensive evaluation of the effectiveness of MET.

Moreover, addressing missing data related to instructional methods, types of mathematics skills developed, and the duration of experiments is crucial. Future studies should ensure that these details are thoroughly documented, as transparency will enhance the reliability of the data and support more precise analyses. Additionally, accounting for regional economic differences and the influence of educational policies on MET implementation could provide deeper insights into the contextual effectiveness of MET.

Another key area for future research is the validation of publication bias. Future studies should develop methodologies to assess the impact of missing studies on both below-average and above-average effects. Such techniques would help mitigate potential biases in the dataset, thereby strengthening the robustness of the findings.

Furthermore, expanding the size of external datasets used for model validation is essential. Future research should aim to collect larger and more diverse datasets to improve the generalizability and robustness of AI models. This may involve collaborating with educational institutions to access comprehensive validation data from large educational databases.

Finally, future research should focus on the practical application of MET at the individual student level. While this study focused on class-wide effect sizes, it is critical to understand how MET impacts individual students to prevent polarization in educational outcomes. Investigating individualized MET interventions and their effects on diverse student populations will provide educators with valuable insights, enabling them to tailor instructional strategies to meet the specific needs of each student.

Acknowledgements
We would like to thank the reviewers for taking the necessary time and effort to review this manuscript. We sincerely appreciate all their valuable comments and suggestions, which helped to improve the quality of this manuscript.

Authors' contributions

Aneng He: Conceptualization, Methodology, Software, Validation, Formal Analysis, Writing – Original Draft, Writing – Review & Editing, Visualization. **Wenwen Yuan**: Validation, Resources, Data Curation, Writing - Original Draft, Visualization. **Lai Soon Lee**: Conceptualization, Methodology, Investigation, Writing – Original Draft, Writing – Review & Editing, Supervision. **Tian Tian**: Writing – Original Draft, Project Administration.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Data availability

The data presented in this study are available from the corresponding author upon request.

Declarations

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 29 May 2025 Accepted: 30 September 2025
Published online: 16 October 2025

References

Acıslı-Celik, S., & Yesilkanat, C. M. (2023). Predicting science achievement scores with machine learning algorithms: A case study of OECD PISA 2015–2018 data. *Neural Computing and Applications, 35*(28), 21201–21228. https://doi.org/10.1007/s00521-023-08901-6

Akçay, A. O., Karahan, E., & Bozan, M. A. (2021). The effect of using technology in primary school math teaching on students' academic achievement: A meta-analysis study. *Fire: Forum for International Research in Education, 7*, 1–21. https://doi.org/10.32865/fire202172231.

Akın, A. (2022). The effectiveness of web-based mathematics instruction (WBMI) on k-16 students' mathematics learning: A meta-analytic research. *Education and Information Technologies, 27*(6), 8015–8040. https://doi.org/10.1007/s10639-022-10931-x

Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2023). Re-thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges. *Applied Sciences, 13*(12), 7082. https://doi.org/10.3390/app13127082

Aslam, N., Khan, I., Alamri, L., & Almuslim, R. (2021). An improved early student's academic performance prediction using deep learning. *International Journal of Emerging Technologies in Learning (iJET), 16*(12), 108–122.

Attard, C., & Holmes, K. (2020). "It gives you that sense of hope": An exploration of technology use to mediate student engagement with mathematics. *Heliyon, 6*(1), Article e02945. https://doi.org/10.1016/j.heliyon.2019.e02945

Bailey, D. H., Nguyen, T., Jenkins, J. M., Domina, T., Clements, D. H., & Sarama, J. S. (2016). Fadeout in an early mathematics intervention: Constraining content or preexisting differences? *Developmental Psychology, 52*(9), 1457. https://doi.org/10.1037/dev0000188

Baye, M. G., Ayele, M. A., & Wondimuneh, T. E. (2021). Implementing geogebra integrated with multi-teaching approaches guided by the apos theory to enhance students' conceptual understanding of limit in ethiopian universities. *Heliyon, 7*(5), Article e07012. https://doi.org/10.1016/j.heliyon.2021.e07012

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Brooks, C., & Thompson, C. (2017). Predictive modelling in teaching and learning. *Handbook of Learning Analytics*, 61–68, https://doi.org/10.18608/hla22.003

Bush, S.B., Driskell, S.O., Niess, M.L., Pugalee, D., Rakes, C.R., & Ronau, R.N. (2015). The impact of digital technologies in mathematics pre-service teacher preparation over four decades. *Handbook of Research on Teacher Education in the Digital Age* (pp. 1–27). IGI Global Scientific Publishing.

Callaghan, R., Joubert, J., & Engelbrecht, J. (2023). Using enaction to evolve from pre-covid to post-covid pedagogy: a case study with South African mathematics teachers. *ZDM Mathematics Education, 55*(1), 193–206. https://doi.org/10.1007/s11858-022-01416-9

Camacho, L., Douzas, G., & Bacao, F. (2022). Geometric smote for regression. *Expert Systems with Applications, 193*, Article 116387. https://doi.org/10.1016/j.eswa.2021.116387

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., & Mocko, M. others (2016). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*.

Çavuş, H., & Deniz, S. (2022). The effect of technology assisted teaching on success in mathematics and geometry: A meta-analysis study. *Participatory Educational Research, 9*(2), 358–397. https://doi.org/10.17275/per.22.45.9.2

Cevikbas, M., Bulut, N., & Kaiser, G. (2023). Exploring the benefits and drawbacks of AR and VR technologies for learners of mathematics: Recent developments. *Systems, 11*(5), 244. https://doi.org/10.3390/systems11050244

Chang, H. Y., Binali, T., Liang, J. C., Chiou, G. L., Cheng, K. H., Lee, S. W. Y., & Tsai, C. C. (2022). Ten years of augmented reality in education: A meta-analysis of (quasi-) experimental studies to investigate the impact. *Computers & Education, 191*, Article 104641. https://doi.org/10.1016/j.compedu.2022.104641

Chang, S. H., Kuo, P. J., Xin Kao, J., & Yang, L. J. (2024). Using an integrative meta-analyzis to examine the effects of smart classroom-based instruction on academic achievement: A technological change perspective. *Interactive Learning Environments, 32*(5), 2293–2320. https://doi.org/10.1080/10494820.2022.2146138.

Cheung, A. C., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in k-12 classrooms: A meta-analysis. *Educational Research Review, 9*, 88–113. https://doi.org/10.1016/j.edurev.2013.01.001

Chollet, F., & Chollet, F. (2021). *Deep learning with python*. Simon and Schuster.

Cortes, C., Mohri, M., & Rostamizadeh, A. (2012). L2 regularization for learning kernels. arXiv preprint arXiv:1205.2653, https://doi.org/10.48550/arXiv.1205.2653

Di, X., & Zheng, X. (2022). A meta-analysis of the impact of virtual technologies on students' spatial ability. *Educational Technology Research and Development, 70*(1), 73–98. https://doi.org/10.1007/s11423-022-10082-3

Doz, D., Cotič, M., & Felda, D. (2023). Random forest regression in predicting students' achievements and fuzzy grades. *Mathematics, 11*(19), 4129. https://doi.org/10.3390/math11194129

Drijvers, P., & Sinclair, N. (2024). The role of digital technologies in mathematics education: Purposes and perspectives. *ZDM Mathematics Education, 56*(2), 239–248. https://doi.org/10.1007/s11858-023-01535-x

Ersen, Z. B., & Alp, Y. (2024). A systematic review of augmented reality in mathematics education in the last decade. *Malaysian Online Journal of Educational Technology, 12*(1), 15–31. https://doi.org/10.52380/mojet.2024.12.1.476

Ertekin, S., Huang, J., Bottou, L., & Giles, C.L. (2007). Learning on the border: Active learning in imbalanced data classification. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 127–136).

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905. https://doi.org/10.1613/jair.1.11192

Fernandez-Batanero, J. M., Roman-Gravan, J. M., Reyes-Rebollo, M. M., & Montenegro-Rueda, M. (2021). Impact of educational technology on teacher stress and anxiety: A literature review. *International Journal of Environmental Research and Public Health, 18*(2), 548.

Gal, Y., Islam, R., & Ghahramani, Z. (2017). Deep Bayesian active learning with image data. *International Conference on Machine Learning* (pp. 1183–1192).

Granberg, C., & Olsson, J. (2015). ICT-supported problem solving and collaborative creative reasoning: Exploring linear functions using dynamic mathematics software. *The Journal of Mathematical Behavior, 37*, 48–62. https://doi.org/10.1016/j.jmathb.2014.11.001

Günhan, B. C., & Açan, H. (2016). Dinamik geometri yazılımı kullanımının geometri başarısına etkisi: Bir meta-analiz çalışması. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 7*(1), 1–23. https://doi.org/10.16949/turcomat.67541

He, A., Yuan, W., & Kiliçman, A. (2024). A meta-analysis on the effectiveness of dynamic mathematics software on k-12 students' mathematics learning. *International Journal of Mathematical Education in Science and Technology*. https://doi.org/10.1080/0020739X.2024.2375257

Herodotou, C., Rienties, B., Verdin, B., & Boroowa, A. (2019). Predictive learning analytics 'at scale': Guidelines to successful implementation in higher education based on the case of the open university UK. *Journal of Learning Analytics, 6*(1), 85–95. https://doi.org/10.18608/jla.2019.61.5

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I., & Reiss, K. M. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education, 153*, Article 103897. https://doi.org/10.1016/j.compedu.2020.103897

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hwang, G.-J., & Tu, Y.-F. (2021). Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review. *Mathematics, 9*(6), 584. https://doi.org/10.3390/math9060584

Hwang, S. (2022). Examining the effects of artificial intelligence on elementary students' mathematics achievement: A meta-analysis. *Sustainability, 14*(20), 13185. https://doi.org/10.3390/su142013185

Hwang, S., Flavin, E., & Lee, J.-E. (2023). Exploring research trends of technology use in mathematics education: A scoping review using topic modeling. *Education and Information Technologies, 28*(8), 10753–10780. https://doi.org/10.1007/s10639-023-11603-0

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.

Jing, D. (2012). The study on educational technology abilities evaluation method. *Physics Procedia, 24*, 2111–2116. https://doi.org/10.1016/j.phpro.2012.02.310

Juandi, D., Kusumah, Y. S., Tamur, M., Perbowo, K. S., & Wijaya, T. T. (2021). A meta-analysis of Geogebra software decade of assisted mathematics learning: What to learn and where to go? *Heliyon*. https://doi.org/10.1016/j.heliyon.2021.e06953

Keser, S. B., & Aghalarova, S. (2022). Hela: A novel hybrid ensemble learning algorithm for predicting academic performance of students. *Education and Information Technologies, 27*(4), 4521–4552. https://doi.org/10.1007/s10639-021-10780-0

Kim, S. J., & Xin, Y. P. (2024). A meta-analysis of technology-based word-problem interventions for students with disabilities. *Education Sciences, 14*(12), 1372. https://doi.org/10.3390/educsci14121372

Krüger, J. G. C., de Souza Britto, A., Jr., & Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. *Expert Systems with Applications, 233*, Article 120933. https://doi.org/10.1016/j.eswa.2023.120933

Lee, J., Chen, C. C., & Basu, A. (2025). From novelty to knowledge: A longitudinal investigation of the novelty effect on learning outcomes in virtual reality. *IEEE Transactions on Visualization and Computer Graphics, 31*(5), 3204–3212. https://doi.org/10.1109/TVCG.2025.3549897

Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review, 22*(3), 215–243. https://doi.org/10.1007/s10648-010-9125-8.

Lipsey, M. W. (2001). *Practical meta-analysis*. SAGE publications Inc.

Lundberg, S.M., Erion, G.G., & Lee, S- I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888, https://doi.org/10.48550/arXiv.1802.03888

Maaß, K., & Artigue, M. (2013). Implementation of inquiry-based learning in day-to-day teaching: A synthesis. *ZDM, 45*, 779–795. https://doi.org/10.1007/s11858-013-0528-0

Marini, F., & Walczak, B. (2015). Particle swarm optimization (PSO). A tutorial. *Chemometrics and Intelligent Laboratory Systems, 149*, 153–165. https://doi.org/10.1016/j.chemolab.2015.08.020

Ministry of Education of the People's Republic of China. (2017). *General senior high school mathematics curriculum standards (in chinese)*. People's Education Press.

Mircea, M., Stoica, M., & Ghilic-Micu, B. (2021). Investigating the impact of the internet of things in higher education environment. *IEEE Access, 9*, 33396–33409. https://doi.org/10.1109/ACCESS.2021.3060964

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group*, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of Internal Medicine, 151*(4), 264–269.

Na, H., Staudt Willet, K. B., & Kim, C. (2025). Investigating the impact of AR technologies on geometric learning in primary school: A comparison between marker-based and markerless ar. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13584

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. *IEEE Access, 9*, 140731–140746. https://doi.org/10.1109/ACCESS.2021.3119596

Ng, H., bin Mohd Azha, A.A., Yap, T.T.V., & Goh, V.T. (2021). A machine learning approach to predictive modelling of student performance. F1000Research. https://doi.org/10.12688/f1000research.73180.2

Nyce, C., & Cpcu, A. (2007). Predictive analytics white paper. *American Institute for CPCU. Insurance Institute of America*, 9–10.

Ogbonnaya, U. I., & Mushipe, M. (2020). The efficacy of geogebra-assisted instruction on students' drawing and interpretations of linear functions. *International Journal of Learning, Teaching and Educational Research, 19*(9), 1–14. https://doi.org/10.26803/ijlter.19.9.1

Orwin, R. G. (1983). A fail-safe n for effect size in meta-analysis. *Journal of Educational Statistics, 8*(2), 157–159. https://doi.org/10.3102/10769986008002157

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery, 88*, Article 105906. https://doi.org/10.1136/bmj.n71

Pannakkong, W., Thiwa-Anont, K., Singthong, K., Parthanadee, P., & Buddhakulsomsiri, J. (2022). Hyperparameter tuning of machine learning algorithms using response surface methodology: A case study of ANN, SVM, and DBN. *Mathematical Problems in Engineering, 2022*(1), 8513719. https://doi.org/10.1155/2022/8513719

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology, 64*, 1–18. https://doi.org/10.1016/j.infsof.2015.03.007

Pujianto, U., Prasetyo, W.A., & Taufani, A.R. (2020). Students academic performance prediction with k-nearest neighbor and c4. 5 on smote-balanced data. *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 348–353).

Rakes, C. R., Ronau, R. N., Bush, S. B., Driskell, S. O., Niess, M. L., & Pugalee, D. K. (2020). Mathematics achievement and orientation: A systematic review and meta-analysis of education technology. *Educational Research Review, 31*, Article 100337. https://doi.org/10.1016/j.edurev.2020.100337

Ran, H., Kasli, M., & Secada, W. G. (2021). A meta-analysis on computer technology intervention effects on mathematics achievement for low-performing students in k-12 classrooms. *Journal of Educational Computing Research, 59*(1), 119–153. https://doi.org/10.1177/0735633120952063

Ran, H., Kim, N. J., & Secada, W. G. (2022). A meta-analysis on the effects of technology's functions and roles on students' mathematics achievement in k-12 classrooms. *Journal of Computer Assisted Learning, 38*(1), 258–284. https://doi.org/10.1111/jcal.12611

Rodrigues, L., Pereira, F. D., Toda, A. M., Palomino, P. T., Pessoa, M., Carvalho, L. S. G., & Isotani, S. (2022). Gamification suffers from the novelty effect but benefits from the familiarization effect: Findings from a longitudinal study. *International Journal of Educational Technology in Higher Education, 19*(1), 13. https://doi.org/10.1186/s41239-021-00314-6

Romero Albaladejo, I. M. (2024). Mathematical attitudes transformation when introducing geogebra in the secondary classroom. *Education and Information Technologies, 29*(8), 10277–10302. https://doi.org/10.1007/s10639-023-12085-w

Rosenthal, R. (1991). Meta-analysis: A review. *Psychosomatic Medicine, 53*(3), 247–271. https://doi.org/10.1097/00006842-199105000-00001

Schoenherr, J., Strohmaier, A. R., & Schukajlow, S. (2024). Learning with visualizations helps: A meta-analysis of visualization interventions in mathematics education. *Educational Research Review, 45*, Article 100639. https://doi.org/10.1016/j.edurev.2024.100639

Setyani, N., & Lestari, H.P. (2016). Efektivitas pembelajaran matematika dengan model cps (creative problem solving) berbantuan geogebra ditinjau dari prestasi belajar dan kreativitas siswa kelas viii smp pgri tegalsari kabupaten purworejo. *Jurnal Pedagogi Matematika*, https://doi.org/10.21831/jpm.v5i7.4576

Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Education and Information Technologies, 28*(7), 8299–8333. https://doi.org/10.1007/s10639-022-11536-0

Son, T. (2024). Intelligent tutoring systems in mathematics education: A systematic literature review using the substitution, augmentation, modification, redefinition model. *Computers, 13*(10), 270. https://doi.org/10.3390/computers13100270

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., & Harvey, I. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment, 14*(8), 1–220. https://doi.org/10.3310/hta14080

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*(5), 401–426. https://doi.org/10.1016/j.intell.2006.09.004

Thalheimer, W., & Cook, S. (2002). How to calculate effect sizes from published research: A simplified methodology. *Work-Learning Research*, *1*(9).
Wang, J., & Paine, L. W. (2023). Influence of centralized curriculum on Chinese beginning teachers' geometry lessons. *Teaching and Teacher Education, 123*, Article 103988. https://doi.org/10.1016/j.tate.2022.103988
Wang, K. (2021). A theoretical analysis method of spatial analytic geometry and mathematics under digital twins. *Advances in Civil Engineering, 2021*(1), 8910274. https://doi.org/10.1155/2021/8910274
Wang, X., Liu, B., Cao, S., Jing, L., & Yu, J. (2020). Importance sampling based active learning for imbalanced classification. *Science China Information Sciences, 63*, Article 112101. https://doi.org/10.1007/s11432-019-2771-0
Xie, C., Cheung, A. C., Lau, W. W., & Slavin, R. E. (2020). The effects of computer-assisted instruction on mathematics achievement in mainland China: A meta-analysis. *International Journal of Educational Research, 102*, Article 101565. https://doi.org/10.1016/j.ijer.2020.101565
Xie, C., Wang, M., & Hu, H. (2018). Effects of constructivist and transmission instructional models on mathematics achievement in mainland China: A meta-analysis. *Frontiers in Psychology, 9*, 1923. https://doi.org/10.3389/fpsyg.2018.01923
Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z
Yang, F., & Li, F. W. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education, 123*, 97–108. https://doi.org/10.1016/j.compedu.2018.04.006
Yüksel, M., & Geban, Ö. (2014). A study of the prediction of academic achievement in the chemistry course. *Education and Science, 39*(173), 354–365. https://doi.org/10.15390/ES.2014.1285
Zengin, Y., Furkan, H., & Kutluca, T. (2012). The effect of dynamic mathematics software geogebra on student achievement in teaching of trigonometry. *Procedia-Social and Behavioral Sciences, 31*, 183–187. https://doi.org/10.1016/j.sbspro.2011.12.038
Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., & Li, Y. (2021). A review of artificial intelligence (ai) in education from 2010 to 2020. *Complexity, 2021*(1), 8812542. https://doi.org/10.1155/2021/8812542
Zhang, Y., Wang, P., Jia, W., Zhang, A., & Chen, G. (2023). Dynamic visualization by geogebra for mathematics learning: A meta-analysis of 20 years of research. *Journal of Research on Technology in Education*. https://doi.org/10.1080/15391523.2023.2250886

Publisher's Note