A Hybrid Neuro-Fuzzy and Machine Learning Model for Student Performance Prediction in Online Mathematics Education

Beyza Esin Özseven[1], Turgut Özseven[2]
Republic of Turkey, Ministry of National Education, Tokat, Turkiye[1]
Department of Computer Engineering, Tokat Gaziosmanpasa University, Tokat, Turkiye[2]

***Abstract*—Accurate and early prediction of student success in online mathematics education is critical for improving learning processes and developing personalized instruction strategies. However, students' problem-solving behaviors, interaction levels, and learning speeds in online learning environments inherently contain uncertainty, limiting the effectiveness of traditional assessment and singular machine learning approaches. This study proposes a hybrid neuro-fuzzy and machine learning-based student success prediction framework that combines the ability of fuzzy logic to represent uncertainty with the strong generalization and adaptive learning capabilities of machine learning methods. In the proposed approach, student success trends are primarily modeled using ANFIS, Random Forest, and XGBoost models, employing raw and derived features on the ASSISTments dataset. The predictions from these models are treated as continuous success representations reflecting uncertainty in students' learning behaviors and are used as input to a hybrid classification structure to make binary success/failure decisions. Thus, the ANFIS model is positioned as an uncertainty-aware and interpretable context generator, while the Random Forest and XGBoost models provide discriminative classification power. Experimental results demonstrate that both ANFIS and Random Forest models exhibit high individual performance; however, combining their complementary features within a hybrid structure significantly increases prediction stability and generalization. Unlike the limitations of 'black box' models in the literature, the interpretability provided by ANFIS, through its linguistic rules and membership functions, enables the proposed approach to generate pedagogically transparent and actionable implications. The findings reveal that this hybrid framework, which integrates uncertainty management with high prediction accuracy, offers a powerful decision-support mechanism for early identification of at-risk students and for developing personalized intervention strategies in online mathematics education.**

***Keywords*—ANFIS; ensemble learning; fuzzy logic; hybrid models; learning analytics**

I. Introduction

The development of online learning platforms has revolutionized educational analytics, enabling the massive collection and analysis of student behavior [1], [2], [3], [4]. These platforms have become digital ecosystems that enable monitoring students' interactions with the learning process, not only in terms of outcomes but also in terms of the process itself. These digital ecosystems record each student's interactions with the system in the form of time-stamped behavioral logs, such as clickstreams, resource viewing times, and forum participation [1, 5, 6]. This type of data reveals not only what students have learned, but also how and under what conditions they have learned. In particular, intelligent tutoring systems such as ASSISTments [7] reflect students' cognitive states during mathematical problem-solving through multidimensional attributes, such as correct/incorrect answer rates, hint requests, and the number of attempts required to solve the problem [1, 6, 8]. In this respect, the ASSISTments dataset provides a rich, high-resolution dataset for predicting student achievement at an early stage.

Mathematics education, by its very nature, exhibits a high degree of logical interdependence and a cumulative structure. A failure to fully understand a fundamental concept increases the student's risk of subsequent failure [9]. Therefore, predicting student achievement early, before end-of-term assessments are completed, is a critical requirement for educators to intervene promptly.

The literature shows that early warning systems significantly increase student retention and overall success through personalized learning paths, targeted feedback, and supportive content [1, 5, 10 - 14]. However, the effectiveness of these systems depends on the extent to which the forecasting models used can represent uncertainty and individual differences.

Online student behavior exhibits ambiguity, variability, and fuzzy patterns stemming from the nature of human reasoning [15 - 17]. Students' moderate level of engagement or cognitive stability in the problem-solving process cannot be adequately represented by traditional models built on sharp boundaries and linear assumptions. This leads to limitations not only in classical statistical methods but also in machine learning models that prioritize accuracy.

Although artificial neural networks and tree-based learning algorithms are successful at capturing complex patterns, many of these models are criticized for being "black boxes" in terms of the interpretability of decision-making processes [18 - 19]. In the field of education, producing pedagogically meaningful and explainable outcomes is a fundamental requirement, alongside the model's accuracy.

In this context, hybrid neurofuzzy approaches, which combine the uncertainty modeling capabilities of fuzzy logic with the data-learning capacity of machine learning, are among the most promising solutions for predicting student achievement [11, 20]. In particular, the Adaptive-Network-based Fuzzy

Inference System (ANFIS) offers educators a significant advantage by providing high predictive performance through an interpretable structure based on IF-THEN rules [4, 11, 21, 22].

This study proposes a hybrid neurofuzzy and machine learning-based model for predicting student success in online mathematics education using the ASSISTments dataset. Extensive experiments were conducted using ANFIS, Random Forest (RF), and XGBoost models with raw and derived student features. The resulting continuous success representations were integrated into a binary classification structure to determine the risk of student failure. The main contribution of this study is to effectively model uncertainty in online mathematics education, to provide a balanced trade-off between interpretability and predictive performance, and to demonstrate the applicability of the proposed hybrid approach for early warning systems.

The remaining sections of the study are structured as follows: The second section summarizes the relevant literature and research gaps. The third section details the ASSISTments dataset, data preprocessing, and feature engineering. The fourth section presents the proposed hybrid ANFIS-RF-XGBoost architecture and learning processes. The fifth section presents the experimental findings with a comprehensive analysis, and the sixth section discusses these findings in light of the literature, outlining the study's limitations and suggesting future research.

II. Related Works

Predicting student academic performance has witnessed a paradigm shift in the fields of educational data mining and learning analytics over the last decade, moving from statistical models to hybrid intelligent systems [23]. Studies in the literature can be categorized primarily by model architecture, dataset diversity, and interpretability of results.

A. Traditional Machine Learning and Ensemble Models

Early educational data mining studies generally focused on singular algorithms such as decision trees, support vector machines, and logistic regression [1, 3]. However, recent research has shown that ensemble learning methods provide higher accuracy than single classifiers [5, 24]. For example, the success of the XGBoost algorithm in capturing complex, nonlinear relationships in structured training data has been highlighted in numerous studies [4, 25]. Similarly, the RF model has been reported to exhibit robust performance, particularly in addressing feature selection and inter-class imbalance issues [26 - 28]. Ponnan et al. (2025) achieved a test accuracy of 98.06% in student assessment using AdaBoost and RF models [11].

B. ANFIS Models

ANFIS models are gaining prominence for modeling uncertainties in educational processes and for handling fuzzy data based on human reasoning [22]. ANFIS is described as a universal estimator that combines the learning ability of artificial neural networks with the uncertainty management capacity of fuzzy logic [6, 29 - 33]. Taylan and Karagozoglu (2009) stated that the ANFIS model yields results as robust as statistical methods and offers a more natural language for interpreting student outcomes [34]. Hybrid models developed in recent years, such as FCM-PSO-ANFIS, have reduced error rates by fine-tuning parameters using metaheuristic optimization algorithms [35].

C. Deep Learning and Time Series Analysis

The timestamped nature of student interaction data has encouraged the use of RNN and LSTM models [1, 36]. Deep learning models are outperforming traditional methods in discovering hidden patterns and temporal dependencies in large-scale datasets [1, 36]. The combination of a CNN and a BiGRU has achieved up to 97% accuracy in predicting student achievement by providing both local feature extraction and bidirectional temporal modeling [36]. Furthermore, transformer architectures have increased predictive granularity by converting student behaviors into sequential feature vectors through attention mechanisms [1, 8].

D. Explainable Artificial Intelligence (XAI) and Educational Trust

The "black box" nature of models makes it difficult for educators to trust these predictions. In this context, methods such as SHAP and LIME have become widespread to make model decisions more transparent [4, 18, 27]. XAI techniques enable the generation of personalized feedback by revealing which factors influence the prediction and in which direction. In particular, the linguistic rules provided by neuro-fuzzy systems enhance the model's direct interpretability, thereby strengthening educational decision-support processes [18, 22].

E. Indicators Specific to Mathematics Education

Specific research on mathematics education shows that behavioral indicators, such as revisiting resources and resource intensity, are strongly correlated with success [6]. Bakar et al. (2022) suggested incorporating neuroscientific mechanisms such as attention, emotion, and productivity into the AGES model to predict mathematical achievement [15]. Such studies argue that mathematical learning ability can be more accurately predicted not only by academic scores but also by integrating cognitive and psychological processes [37].

F. Research Gaps

A review of the existing literature reveals that the vast majority of studies on the ASSISTments dataset focus either on classical classification-based machine learning algorithms or on information-tracking-based models. These studies mainly address student achievement through sharp class labels, remaining limited in their ability to model uncertainty, incremental progress, and intermediate achievement levels that arise in the student's learning process. In contrast, fuzzy logic-based and neuro-fuzzy approaches have been addressed in only a few studies within the context of the ASSISTments dataset and have not been systematically evaluated.

Although student behavior in online mathematics education inherently involves both continuity and uncertainty, the literature lacks two-stage hybrid approaches that first model this uncertainty as a regression-based achievement score and then proceed to a classification phase for decision support. This highlights the need for a modeling framework that is both pedagogically meaningful and methodologically consistent.

The main gaps in the literature that this study aims to address are summarized below:

- Limited Use of Neuro-Fuzzy Regression Models on the ASSISTments Dataset: The vast majority of studies on the ASSISTments dataset have focused directly on classification using XGBoost, deep learning-based RNN/LSTM models, or knowledge-tracking approaches. However, neuro-fuzzy models like ANFIS can model uncertain learning behaviors more flexibly, as they can predict student success as a continuous score or probability.

While the literature shows that ANFIS provides high accuracy, especially in small and medium-sized datasets, its use as a regression-based predictor in large-scale datasets with time-ordered interaction logs, such as ASSISTments, has been largely neglected. This underscores the necessity of a systematic evaluation of hybrid regression-based neuro-fuzzy models on ASSISTments.

- Treating student achievement solely as a classification problem: A significant portion of current research addresses student performance only as a binary classification problem, ignoring the gradual and uncertain nature of the learning process. However, in online learning environments, student achievement shows gradual increases or decreases over time rather than sudden, sharp transitions.

The literature shows that hybrid models, which first model achievement as a continuous regression outcome and then move it to the classification stage for pedagogical decision support, have not been sufficiently examined on the ASSISTments dataset. This gap necessitates an evaluation of hybrid regression-classification approaches within the context of educational analytics.

- The lack of a clear and in-model approach to addressing uncertainty: In online learning environments, students' interaction speed, problem-solving strategies, and participation levels are inherently fuzzy and variable. However, most existing models treat these behaviors as discrete numerical values, failing to adequately represent situations that require human reasoning, such as a student being "moderately successful" or exhibiting "inconsistent learning behavior."

Although membership functions and linguistic variables in fuzzy logic allow direct modeling of this uncertainty, the use of this approach within hybrid regression and classification frameworks is quite limited in the literature. In particular, the integration of ANFIS as an uncertainty-aware predictor into hybrid structures remains a significant gap in the literature.

- Failure to address pedagogical interpretability and high predictive power together: While powerful ensemble learning methods like RF and XGBoost provide high accuracy, they offer limited transparency for pedagogical decision support systems due to the inability to interpret the internal structure of decision mechanisms directly. Although the literature indicates that XAI methods such as SHAP and LIME address this problem post-prediction, these approaches cannot directly reflect the model's internal decision logic.

In contrast, ANFIS's IF-THEN rule-based structure offers interpretable insights into the learning process. However, studies in the literature that combine this interpretable structure with powerful models such as RF and XGBoost in a hybrid framework are insufficient for the ASSISTments dataset.

- Lack of holistic evaluation of hybrid regression and hybrid classifiers: Existing studies generally focus either solely on regression or solely on classification problems. However, in online mathematics education, student performance needs to be considered both as a continuous indicator of success and as a decision-based classification output.

In the literature, studies that combine ANFIS-based regression outputs with models such as RF and XGBoost to evaluate them within a two-stage hybrid regression-classification architecture are quite limited. This gap necessitates a systematic and comparative examination of hybrid models on the ASSISTments dataset.

- Insufficient modeling of cognitive processes specific to online mathematics education: Mathematics education differs from other disciplines due to its cumulative knowledge structure and high logical dependence. However, most existing models do not address indicators specific to the mathematical problem-solving process, such as hint use, post-error behaviors, cognitive persistence, and learning stability, within a hybrid modeling framework.

This clearly highlights the need for holistic hybrid models that address cognitive and behavioral uncertainties specific to mathematics education at both the regression and classification levels.

In summary, this study addresses complex student interactions in the ASSISTments dataset by combining ANFIS's uncertainty-aware, interpretable regression structure with the highly generalizable capabilities of RF and XGBoost, aiming to fill the gaps identified in the literature from both a hybrid regression and a hybrid classifier perspective. In this respect, the study offers a unique, systematic, and applicable contribution from both methodological and pedagogical standpoints.

III. Materials and Methods

A. Dataset Description

The dataset used in this study was obtained from ASSISTments, an online mathematics learning platform. ASSISTments is an intelligent tutoring system designed to support the problem-solving processes of middle and high school students while collecting detailed, timestamped data on student interactions [38]. The platform integrates teaching and assessment processes by offering students instant feedback, tips, and validation [7, 39, 40].

The ASSISTments dataset records student interactions with math problems in an event-based manner and provides multidimensional behavioral insights for each student-problem interaction [20, 41]. This dataset is widely used, particularly for

student performance prediction, knowledge monitoring, and educational data mining [42]. Furthermore, the fact that ASSISTments data has been used in large-scale data mining competitions such as the KDD Cup 2010 demonstrates that the dataset is an accepted and reliable source within the research community. A sample cross-section of the ASSISTments dataset is given in Table I.

TABLE I. ILLUSTRATIVE SAMPLE FROM THE ASSISTMENTS DATASET

assignment_id	user_id	problem_id	correct	attempt_count	…	skill_id	hint_count	overlap_time
263599	14	93383	0	1		2	2	41131
263599	14	93407	1	1		2	0	29123
263599	14	93400	0	1		2	2	19905
…								

Personality-based features were not used as model inputs; they were evaluated only during student-based data splitting to prevent data leakage. Behavioral features were obtained directly from raw data, while skill-based features were derived to reflect the student's learning progress over time.

The "attempt_count" and "hint_count" attributes reflect students' problem-solving strategies, while "ms_first_response" provides temporal information about engagement and cognitive effort. The "correct" attribute indicates whether the student successfully solved the problem in a given attempt. These attributes allow for the analysis not only of students' outcome-oriented success but also their problem-solving strategies, learning behaviors, and cognitive effort. In particular, "hint_count", "attempt_count", and "ms_first_response" offer significant indicators of students' learning levels and uncertainties in their problem-solving processes [43]. Traditional binary assessment methods are insufficient for modeling these unclear boundaries between learning situations. Managing this uncertainty and predicting student achievement with more humane reasoning constitutes the main motivation for using ANFIS in this study.

B. Data Preprocessing

Educational data obtained from online learning environments inherently contain incomplete records, noisy behavior, and uneven classroom distributions. Therefore, data preprocessing and feature extraction stages are critical for model performance in student performance prediction studies [44]. In this study, a comprehensive data preprocessing and feature engineering process was applied to improve the effectiveness of the proposed hybrid neuro-fuzzy and machine learning model on the ASSISTments dataset.

In the first stage, missing or inconsistent records in the dataset were analyzed. Records lacking "user_id", "assignment_id", and "correct" information were removed from the dataset.

Because the ASSISTments dataset has an event-based structure, the number of interactions among students varies considerably. Some students solve only a few problems, while others have thousands of interactions. This is consistent with the long-tailed distribution frequently reported in the literature [43]. To prevent the model from being dominated by overactive students, student-based (user_id) data splitting and limiting interaction counts were implemented. Furthermore, due to significant variability in student interaction counts within the dataset, students with fewer than 5 interactions were removed, and a maximum interaction limit of 1000 was set for overactive students.

When the "attempt_count" values were analyzed, it was observed that the vast majority were concentrated between 1 and 10. Therefore, in fuzzy category definitions, the maximum number of attempts was set to 10 [7, 43]. Since higher values rarely occur, a threshold has been set at the upper limit, ensuring stable representation and model stability.

The "ms_first_response" property can have excessively large values due to factors such as browser open times and prolonged inactivity. To prevent such values from adding noise to the learning process, the "ms_first_response" property is limited to 600 seconds. Without this limitation, the sensitivity of the fuzzy membership functions decreases, extreme values dominate the model, and the model learns biased rules such as "slow learner = fail." This approach is consistent with clipping strategies proposed in the literature for time-based learning variables [43, 44].

"opportunity_count" is a pedagogically significant attribute representing how many times a student has encountered a particular skill. Therefore, it has not been directly filtered but has been included in the model to reflect the learning process.

Pre-cleaning analyses show that the distribution of student interactions is severely imbalanced, with a small number of students dominating the dataset. After applying cleaning and preprocessing criteria, the number of interactions per student showed a more balanced distribution, with extreme values eliminated. Furthermore, normalization was performed to improve the stability of the fuzzy logic-based model's membership functions and to ensure a fair comparison across machine learning algorithms [45]. Min-max normalization was applied to "attempt_count" and "hint_count", robust normalization to "ms_response_time" and "overlap_time", and log+min-max normalization to "opportunity_count". This helped train the proposed hybrid model more stably and generalizable.

C. Feature Engineering

In this study, to more accurately model students' online mathematics learning behaviors, the features obtained from the raw dataset were normalized, and behavioral and performance-based derived features commonly used in the literature were created.

The raw features in the dataset capture students' core interactions during the problem-solving process. While these

raw features directly reflect students' problem-solving behaviors, their direct use as model input is limited by scale differences and individual variation. Instead of using raw features directly, derived features were created to better represent students' performance and behavioral tendencies during the learning process. These features were designed to reflect the student's cognitive effort and learning status in the problem-solving process [43]. In this context, the following features have been derived: Prior Correct Ratio (PCR), Prior Attempt Mean (PAM), Prior Hint Mean (PHM), Recent Performance Trend (RPT), Skill Difficulty Index (SDI), and Student Experience Level (SEL).

PCR is a ratio-based trait that reflects a student's level of success in previous interactions with a specific problem or skill [6, 46]. The PCR value is in the range of [0,1] and does not require normalization.

$$\mathrm{PCR}_{u,s} = \frac{\sum_{i=1}^{n} Correct_{u,s}^{(i)}}{n} \quad (1)$$

Here, u is a student, s is a problem or a skill, and n is the number of previous attempts.

PAM is a behavioral indicator that represents the average number of attempts a student makes for problems they have encountered in the past. This characteristic reflects the student's mathematical cognitive persistence and learning speed [6], [46], [47]. Since the PAM value is not in the [0,1] range, pre-model min-max normalization should be applied. A high PAM value indicates that the student generally needs more trials.

$$\mathrm{PAM_u} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{attempt\_count_u^{(i)}} \quad (2)$$

Here, u is a student, and n is the number of problems previously solved.

PHM represents the average number of hints a student uses across all problem-solving steps before the current interaction. This characteristic is a historical indicator of a student's help-seeking behavior. Hint usage carries a vague signal about how well a student understands the subject, and overuse is associated with an increased risk of failure [38].

$$\mathrm{PHM}_i = \frac{1}{N_i}\sum_{j=1}^{i-1} hint\_coun\,t_j \quad (3)$$

Here, i is the student's time-ordered current interaction index, $N_i = i - 1$ is the total number of interactions before the current interaction, $hint\_count_j$ is the number of hints the student used in the j[th] interaction.

RPT captures short-term learning dynamics, information fluctuations, and learning trends, which are particularly important in adaptive learning environments. Given the cumulative nature of mathematics education, short-term dips are considered a signal for early intervention [8], [48].

For the last k interactions (3-5):

$$\mathrm{RPT}_i = \frac{1}{k}\sum_{j=i-k}^{i-1} correc\,t_j \quad (4)$$

Here, i is the student's i^{th} problem-solving step.

SDI is the overall difficulty level of a given mathematical skill across all students. This metric allows the model to make fair and balanced assessments by normalizing individual performance with aggregate data [49], [50], [51].

$$\mathrm{SDI}_s = 1 - \frac{1}{M}\sum_{j=1}^{M} Correc\,t_s^{(j)} \quad (5)$$

Here, s is the skill, and M is the total number of trials for that skill.

SDI is calculated at the skill level, independent of the problem, and provides a common difficulty indicator for all students (SDI → 0: Easy skill, SDI → 1: Difficult skill).

SEL is a composite pedagogical indicator representing a student's overall experience and proficiency level in an online mathematics learning environment. This characteristic translates the student's familiarity with the system and resource-intensive usage habits into the model [15]. SEL is not a variable with an implicit formula. It is defined as follows:

$$\mathrm{SEL}_i = F(\mathrm{PCR}_i, \mathrm{PAM}_i, \mathrm{PHM}_i, \mathrm{OC}_i) \quad (6)$$

Here, i is the student's time-ordered current interaction index, and $F(\cdot)$ is the Fuzzy Inference Mechanism (ANFIS).

The derived features capture both past performance and real-time behavioral patterns, providing a richer representation of student learning dynamics. All features used in the study are listed in Table II.

TABLE II. CHARACTERISTICS USED IN THE STUDY

Feature	Reason
attempt_count (AC)	The solution reflects the intensity of the behavior; in ANFIS, it is meaningful in terms of "Low–Medium–High" memberships.
hint_count (HC)	Help-seeking behavior can be interpreted pedagogically using vague rules.
ms_response_time (RT)	ML models can tolerate high-dimensional correlation; this increases rule complexity in ANFIS.
Prior Correct Ratio (PCR)	The student's past achievements are critical to their learning.
Prior Attempt Mean (PAM)	Long-term effort level: a stable statistical characteristic.
Prior Hint Mean (PHM)	The constant need for help reflects the learning strategy.
Recent Performance Trend (RPT)	The short-term learning tendency is suitable for trend-based rules in ANFIS.
opportunity_count (OC)	The number of times a student encounters a skill is a key context for both statistical and rule-based models.
overlap_time (OT)	Student waiting time for the question
Skill Difficulty Index (SDI)	Content difficulty: pedagogically meaningful within ANFIS guidelines.
Student Experience Level (SEL)	Cumulative learning level: intuitive linguistic variable in ANFIS.

D. Inter-Feature Relationship Analysis

As part of the feature identification process, a correlation analysis was performed to assess relationships among the extracted features. The resulting Pearson correlation matrix, shown in Table III, provides insight into potential feature dependencies before neuro-fuzzy logic and machine learning

modeling. The analysis included both raw and derived numerical features, excluding descriptive variables.

The vast majority of correlation coefficients are at the $|r|<0.40$ level. This indicates that there is no high-level linear dependence between the features and that the risk of multicollinearity is generally low. A moderate positive correlation ($r=0.46$) was observed between RT and OT. This relationship shows that time overlap increases as the student's first response time increases. There is a relatively high positive correlation ($r = 0.58$) between HC and PHM. This indicates that the student's past habit of using hints is reflected in their current problem behavior. PCR shows a strong negative relationship with both PAM ($r = -0.53$) and PHM ($r = -0.62$). This finding reveals that students with high past success tend to try less and use fewer hints. The correlation between the RPT feature and all other features is negligible, suggesting that the variable provides an independent and complementary source of information for the model. The very low correlation of OC with other characteristics suggests that this variable is largely independent of other behavioral traits as a time-based measure.

TABLE III. PEARSON CORRELATION MATRIX FOR THE FEATURE SET

Feature	AC	HC	RT	OT	SDI	SEL	RPT	PCR	PAM	PHM	OC
AC	1.00	0.38	0.04	0.17	0.11	-0.04	-0.02	-0.19	0.27	0.16	0.00
HC	0.38	1.00	0.00	0.10	0.22	-0.09	-0.03	-0.37	0.19	**0.58**	-0.01
RT	0.04	0.00	1.00	**0.46**	0.09	0.03	0.02	0.02	0.02	0.00	0.03
OT	0.17	0.10	**0.46**	1.00	0.07	0.01	0.01	-0.03	0.05	0.04	0.01
SDI	0.11	0.22	0.09	0.07	1.00	0.03	0.01	-0.35	0.21	0.31	-0.06
SEL	-0.04	-0.09	0.03	0.01	0.03	1.00	0.01	0.13	-0.08	-0.11	0.16
RPT	-0.02	-0.03	0.02	0.01	0.01	0.01	1.00	-0.01	0.01	0.01	0.01
PCR	-0.19	-0.37	0.02	-0.03	-0.35	0.13	-0.01	1.00	**-0.53**	**-0.62**	0.08
PAM	0.27	0.19	0.02	0.05	0.21	-0.08	0.01	**-0.53**	1.00	**0.41**	-0.04
PHM	0.16	**0.58**	0.00	0.04	0.31	-0.11	0.01	**-0.62**	**0.41**	1.00	-0.04
OC	0.00	-0.01	0.03	0.01	-0.06	0.16	0.01	0.08	-0.04	-0.04	1.00

Overall, the correlation analysis results indicate that including the selected features in both neurofuzzy and machine learning models is statistically feasible and that there is no significant multicollinearity.

In addition to correlation analysis, feature redundancy was further examined using the Variance Inflation Factor (VIF). VIF quantifies the degree of multicollinearity by measuring how much the variance of a regression coefficient is inflated due to linear dependence among predictors. The VIF analysis was conducted on all numerical raw and derived features extracted from the ASSISTments dataset. Following established guidelines, a VIF threshold of 10 was adopted, where values below this limit indicate acceptable levels of multicollinearity (Table IV).

TABLE IV. VIF TABLE

Feature	VIF	Comment
AC	3.79	Moderate correlation - acceptable.
HC	2.03	Low multi-connectivity
RT	1.96	Low multi-connectivity
OT	1.49	Very low multi-connectivity
SDI	**8.04**	**High – caution advised**
SEL	1.98	Low multi-connectivity
RPT	1.01	Independent variable
PCR	**4.26**	Medium-high
PAM	**5.84**	**High on the border**
PHM	2.78	Middle
OC	2.21	Low

VIF was calculated to quantify the degree of multicollinearity among the independent variables. Generally, VIF values <5 are acceptable; values between 5 and 10 require attention; and values ≥ 10 indicate significant multicollinearity. According to the analysis results, the vast majority of features have VIF<5. This indicates that the features contribute largely independently to the model. The RPT feature, with a VIF$\approx$1, is an almost completely independent source of information and a strong complement to the model. AC, HC, RT, OT, and SEL have low-to-medium VIF values and do not pose a risk in terms of linear dependence. PAM (VIF$\approx$5.84) and especially SDI (VIF$\approx$8.04) exhibit relatively high VIF values. This suggests that these variables share common variance with some other inputs in the model. However, since these values are below the critical threshold of 10, they do not warrant direct exclusion. These results indicate that the feature set is generally well-balanced and does not exhibit significant multicollinearity. When the VIF results are evaluated together with the Pearson correlation matrix, it becomes clear that the relatively high VIF values observed are due to natural overlap.

The SDI feature shows moderate negative correlations with PCR and moderate positive correlations with PHM and PAM. These relationships stem from the fact that problem difficulty is naturally linked to the student's previous success, trial, and hint-use behaviors. Therefore, the high VIF value is not a statistical problem, but a reflection of a pedagogically significant overlap.

There is a moderate-to-high negative correlation between PAM and PCR features. It is expected that a student's experience of more trials is generally associated with a lower success rate. The increase in VIF values is a mathematical consequence of

this behavioral correlation and does not mean the model is meaningless.

There is a significant positive relationship between HC and PHM. The relatively higher VIF values result from the inclusion of both short- and long-term representations of the same concept in the model. This requires attention, especially in linear models, but does not pose a problem for nonlinear models.

Features such as RPT, OT, and RT exhibit low correlation and low VIF values. These variables provide independent and complementary information to the model.

No feature exceeds the critical multicollinearity threshold (VIF≥10). Therefore, all features are preserved, especially in ANFIS and machine learning-based nonlinear models.

E. Problem Formulation

In online learning environments, students' problem-solving performance exhibits a nonlinear, complex structure due to individual learning differences, prior knowledge levels, problem difficulty, and behavioral uncertainties arising during interaction. This uncertainty makes it difficult to predict student success with both high accuracy and pedagogical interpretation. Therefore, hybrid prediction mechanisms that can explicitly model uncertainty in student performance and offer high predictive power are needed in online mathematics education.

The ASSISTments dataset consists of multidimensional observations derived from students' time-ordered interactions, each reflecting the student's cognitive and behavioral state. Each student interaction is defined as follows:

$$\mathrm{x}_i = [x_{i1}, x_{i2}, \ldots, x_{in}] \in \mathbb{R}^n \tag{7}$$

Here, x_i represents raw and statistical features derived from the student's past performance, problem characteristics, and learning process.

In the ASSISTments dataset, the output variable is defined as a binary label containing correct or incorrect response information for each student-problem interaction:

$$y_i \in \{0,1\} \tag{8}$$

However, the Subtractive Clustering-based ANFIS model used in this study inherently produces a continuous-valued output due to its Sugeno-type fuzzy inference mechanism. Therefore, in the first stage of the study, the problem was treated as a hybrid regression problem, and the model output was defined as follows:

$$\hat{y}_i^{(m)} = f^{(m)}(\mathrm{x}_i), m \in \{\text{ANFIS}, RF, \text{XGBoost}\} \tag{9}$$

Here, $\hat{y}_i^{(m)}$ is interpreted as a continuous value corresponding to the student's success score on the given problem or the probability of giving the correct answer.

In this context, ANFIS is an interpretable regression estimator that models ambiguous student behavior using fuzzy rules, whereas RF and XGBoost are robust regression models with high capacity to capture nonlinear patterns. All models were trained and evaluated to predict the same continuous target variable, ensuring methodological consistency.

While regression-based predictions offer rich information about the probability of student success, decision-based outcomes are frequently needed in educational applications. Therefore, in the second phase of the study, the problem was reformulated as a hybrid classification problem. In this phase, using the continuous outputs obtained from the regression models, the final class label was defined as follows:

$$\tilde{y}_i = \begin{cases} 1, & \hat{y}_i \geq \tau \\ 0, & \hat{y}_i < \tau \end{cases} \tag{10}$$

Here, τ represents the decision threshold set on the validation data.

In this approach, ANFIS acts as a pre-predictor, providing fuzzy-logic-based intermediate outputs to the decision-making process, while RF and XGBoost serve as higher-level classifiers that make the final decision based on these outputs. This structure offers a unique hybrid classification framework that combines both uncertainty modeling and high classification accuracy under one roof.

F. ANFIS Model and Learning Process

In the field of educational data mining, student knowledge levels, learning speeds, and performance changes often involve ambiguous transitions rather than precise boundaries; therefore, ANFIS offers a suitable and powerful framework for modeling such patterns [42].

In the generated ANFIS model, the number of membership functions and rules is automatically determined based on data density using the subtractive clustering method. Sugeno-type output is used. The model's training parameters are defined as Epoch=100, StepSize=0.01, and ErrorGoal=0. In the subtractive clustering method, the cluster radius parameter is a critical hyperparameter that directly determines the complexity of the ANFIS model. Small radius values result in more clusters in the data space and, consequently, a high number of fuzzy rules, while large radius values lead to oversimplification of the model. With a cluster radius of 0.5, an average of 5–8 membership functions was obtained for each input variable.

ANFIS is a model based on the supervised learning paradigm, which requires the initial definition of the target variable during training. In the subtractive clustering method, the fuzzy rule parameters are initially estimated using the least-squares method. At this stage, the prior parameters are assumed to be constant, and the rule firing strengths are calculated for each training instance [34, 45]. The normalized firing strength for each rule is defined as follows:

$$\bar{w}_k = \frac{w_k}{\sum_{l=1}^{K} w_l} \tag{11}$$

Here, w_k represents the firing strength of the kth fuzzy rule.

The model output is calculated as follows, using normalized firing strengths:

$$\hat{y} = \sum_{k=1}^{K} \bar{w}_k \left(p_{k0} + \sum_{j=1}^{n} p_{kj}\, x_j \right) \tag{12}$$

When this expression is rearranged in terms of linear parameters, a classical linear regression problem is obtained. This linear system is solved using the least-squares method,

considering all training examples, and the initial output parameters $\{p_{k0}, p_{k1}, \ldots, p_{kn}\}$ are computed for each rule.

After determining the initial output parameters, the ANFIS model was trained using a hybrid learning algorithm. In this algorithm, during the forward traversal phase, the prior parameters are held constant, and the output parameters are updated using the least-squares method. In the backpropagation phase, the output parameters are treated as constants, and the prior parameters are optimized using gradient descent. This two-stage learning process increases both the accuracy and generalization ability of the model [45].

G. Random Forest and Learning Process

In this study, the RF model was used in regression mode to predict a continuous-valued student achievement score from pairwise-labeled student interactions. RF generates the final prediction by averaging the outputs of multiple decision trees created using bootstrap sampling [52]. This structure offers significant advantages, including the ability to model nonlinear relationships and resilience against overlearning. Each decision tree learns the relationship between the input features and the target variable to produce a continuously valuable output. The RF output is calculated as follows:

$$\hat{y}_i^{RF} = \frac{1}{T}\sum_{t=1}^{T} f_t\,(\mathbf{x}_i) \tag{13}$$

Here, T represents the number of trees, and $f_t(\cdot)$ represents the t[th] decision tree [52].

This structure allows RF to generate predictions in the same target space as the continuous output produced by the ANFIS model, enabling a fair comparison.

H. XGBoost and Learning Process

XGBoost is a regularized decision tree ensemble method based on the gradient boosting principle. The model improves its performance by learning a new tree in each iteration, focusing on minimizing the errors of previous predictions [52]. In this study, the XGBoost method is used as a regression method to estimate continuous student achievement scores. The general form of the XGBoost model can be expressed as follows:

$$\hat{y}_i^{XGB} = \sum_{m=1}^{M} \gamma_m\, h_m(\mathrm{x}_i) \tag{14}$$

Here, $h_m(\cdot)$ represents the m^{th} regression tree, and γ_m represents the learning rate [53].

This structure allows XGBoost to generate predictions in the same target space as the continuous outputs of other models, enabling a fair comparison.

I. Proposed Hybrid Model: ANFIS–RF–XGBoost-Based Student Achievement Prediction Framework

In this study, the proposed architecture comprises two complementary learning stages: regression-based pre-prediction with uncertainty awareness and classification-based final output generation, both focused on decision support. Fig. 1 presents the general flowchart of the hybrid architecture proposed in this study.

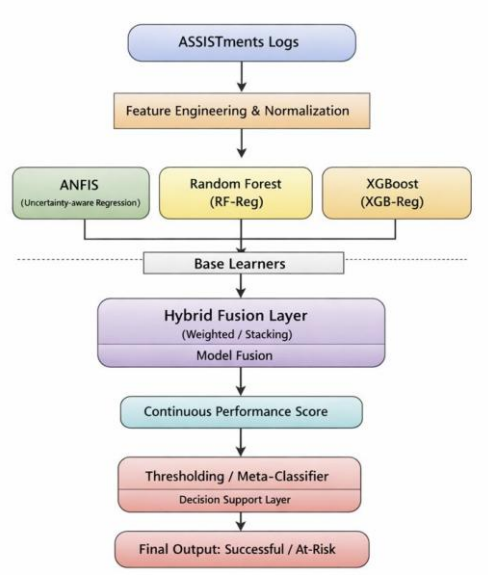

Fig. 1. Proposed hybrid model.

The raw interaction logs in the ASSISTments dataset reflect students' problem-solving behaviors in chronological order. The feature engineering process performed on the dataset is described in detail in Section III-C.

In the Base Learners phase, ANFIS, RF, and XGBoost regression processes are performed. With ANFIS, student behaviors are represented by linguistic, interpretable rules, and a continuously valued achievement score is obtained. In addition to ANFIS, XGBoost models were trained in regression mode to capture nonlinear patterns with high accuracy, and RF models were trained for robustness to noise and stable prediction. These models complement ANFIS by enabling learning of complex feature interactions, robustness to noise, and high generalization performance.

Three independent models are trained using the same input properties x_i

$$\begin{aligned} \hat{y}_i^{ANFIS} &= f_{ANFIS}(\mathrm{x}_i) \\ \hat{y}_i^{RF} &= f_{RF}(\mathrm{x}_i) \\ \hat{y}_i^{XGB} &= f_{XGBoost}(\mathrm{x}_i) \end{aligned} \tag{15}$$

All of these models are structured in regression mode to predict a continuous student achievement score from binary-labeled data.

In the next stage, the regression-based predictions are fed into the fusion layer, which is at the heart of the hybrid architecture. Two different hybridization strategies, weighted ensemble and stacking, were evaluated in this study.

In the weighted ensemble approach, the output of each model is weighted according to its performance on the validation data:

$$\hat{y}_i^{Hybrid} = \alpha\hat{y}_i^{ANFIS} + \beta\hat{y}_i^{RF} + \gamma\hat{y}_i^{XGB} \tag{16}$$

Here, $\alpha + \beta + \gamma = 1, \alpha, \beta, \gamma \geq 0$, and this can be determined based on validation set performance, model reliability, or grid search optimization. Thus, it balances the weaknesses of individual models and increases the stability of the regression output.

In stacking-based hybrid learning, the ANFIS, RF, and XGBoost outputs are treated as meta-features and fed to a second-level learner. This meta-learner is structured as a linear regression classifier for regression and a logistic regression classifier for classification. This stacking structure produces a more powerful prediction by learning complementary information between models.

$$z_i = [\hat{y}_i^{ANFIS}, \hat{y}_i^{RF}, \hat{y}_i^{XGB}] \quad (17)$$

Here, z_i is the meta-input vector.

For the final estimate:

$$\hat{y}_i^{Hybrid} = g(z_i) \quad (18)$$

Here, $g(\cdot)$ is chosen to be either a linear or a low-complexity regression model.

The consistent success scores were transformed into a binary classification problem for decision support in educational applications. This transformation was performed using the following thresholding function:

$$\tilde{y}_i = \begin{cases} 1, & \hat{y}_i^{\text{hybrid}} \geq \tau \\ 0, & \hat{y}_i^{\text{hybrid}} < \tau \end{cases} \quad (19)$$

Here, the threshold value τ is optimized on the validation set.

The proposed hybrid architecture offers the following advantages:

- Uncertainty Awareness: ANFIS explicitly models fuzzy patterns in student behavior.
- High Predictive Power: RF and XGBoost successfully capture complex and nonlinear relationships.
- Interpretability: ANFIS rules allow for pedagogical interpretation of prediction results.
- Flexible Use: The architecture is naturally adaptable to both regression and classification scenarios.

J. Performance Evaluation Metrics

The proposed hybrid architecture treats student achievement as both a continuous performance score (regression) and a binary decision outcome (classification). Therefore, model performance was analyzed using a two-stage evaluation strategy appropriate to the nature of the problem. In the first stage, the accuracy of regression-based predictions was measured, and in the second stage, the classification performance of the resulting hybrid outcomes was evaluated.

This multi-level performance evaluation strategy comprehensively demonstrates not only the prediction accuracy of the proposed hybrid model but also its practical usability for early warning and pedagogical decision-support systems.

1) Regression-based evaluation metrics: The ANFIS, RF, and XGBoost models were structured in regression mode to predict the student's consistent success score on the relevant problem. The metrics used at this stage measure the level of error between the predicted values and the actual labels. For regression-based models, the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2) metrics were used [11, 22, 24].

Furthermore, the continuous predictions generated by regression models are subsequently converted to a binary classification problem for evaluating decision performance. Fixed threshold values are commonly used for this process. However, the literature indicates that such fixed thresholds do not guarantee optimal decision performance, especially when the class distribution is unbalanced and false-positive and false-negative errors have asymmetric costs [54, 55]. To overcome this limitation, an adaptive thresholding method based on the ROC curve was used in this study, rather than a fixed thresholding approach. The adaptive threshold was set to the point on the ROC curve that best balances sensitivity and specificity. For this purpose, the Youden Index (J), widely used in the literature, was adopted [56]. The Youden index is defined as follows:

$$J = TPR - FPR \quad (20)$$

This criterion determines the decision boundary that best distinguishes between classes by selecting a threshold value that maximizes the true positive rate while minimizing the false positive rate.

The ROC-based adaptive thresholding approach allows evaluation not only of the continuous predictive accuracy of models but also of their decision-making success, in a manner sensitive to data distribution. Particularly in online mathematics learning environments, considering the pedagogical consequences of incorrectly classifying a failing student as successful, this approach offers a more realistic and reliable assessment framework [44, 57].

2) Classification-based evaluation metrics: Regression-based hybrid outputs were converted into a binary classification problem using a threshold or a meta-classifier for decision support. Classification performance was evaluated using accuracy, recall, precision, and F1-score [4, 47, 58].

IV. Experimental Results

A. Experimental Setup

This section describes the experimental setup used to objectively and reproducibly evaluate the performance of the proposed hybrid neuro-fuzzy and machine-learning-based architecture. The experiments were designed within a multi-stage framework aiming to predict student achievement levels in online mathematics learning environments in both continuous (regression) and binary (classification) forms. In this context, the comparative performance of single learners and hybrid regression and hybrid classifier models was analyzed on raw and derived feature sets using the ASSISTments dataset. The entire experimental process was structured with the principles of fair comparison and data leakage prevention in mind.

B. Dataset and Splitting Strategy

In the experimental analyses, a student-based splitting strategy was adopted to prevent data leakage and to realistically evaluate the models' generalization ability. In this approach, the interactions of the same student were included in only one subset (training, validation, or testing). Thus, by measuring the model's performance on students it has not previously encountered, an evaluation closer to practical use cases was provided. The dataset was divided into three subsets: 60% for training, 20% for validation, and 20% for testing. The training set was used to learn the model parameters, the validation set for hyperparameter adjustments and determining hybrid model weights, and the testing set for the final performance evaluation. In the classification phase, the distribution of success and failure labels was analyzed, focusing on recall and F1-score to mitigate class imbalance and prevent skewed performance metrics. Thanks to this splitting strategy, the reliability of the proposed hybrid regression and hybrid classifier models in terms of both prediction accuracy and early warning systems was evaluated consistently and fairly.

C. Experimental Scenarios

To comprehensively evaluate the effectiveness of the proposed architecture, numerous experimental scenarios were designed that incorporated different model types, feature sets, and learning paradigms. These experimental scenarios aim to analyze the performance of individual models and clearly demonstrate the contribution of hybrid structures. In this context, the experiments were grouped into three main categories: individual regression models, hybrid regression models, and hybrid classification models.

In individual regression models, student achievement was treated as a continuous performance score, and ANFIS, RF, and XGBoost models were trained in regression mode. The goal was to compare the fundamental predictive capabilities of different learning paradigms on the ASSISTments dataset. In this context, analysis was performed with three regression models and three feature sets (raw, derived, and hybrid). These scenarios enable analysis of the impact of feature engineering and model structure on regression performance.

Hybrid regression scenarios aim to achieve more stable, accurate predictions by combining outputs from individual models. In this stage, ANFIS, RF, and XGBoost regression outputs were used as inputs to hybrid coupling mechanisms. These experiments were designed to evaluate the performance change and generalization status of the hybrid approach compared to single models.

In the hybrid classification scenario, the continuous success scores from the hybrid regression models were converted into binary classification labels for decision support. In this scenario, students were divided into two classes. This scenario aims to evaluate the applicability of the proposed architecture for early warning systems and its potential for pedagogical decision support.

D. Comparison of Individual Models

In this section, the performance of the individual regression models that constitute the core components of the proposed hybrid architecture is comprehensively analyzed. Student achievement is treated as a continuous output variable, and the ANFIS, RF, and XGBoost models are evaluated within the regression framework. Each model was trained and tested separately using raw features, derived features, and hybrid feature sets to examine the effect of feature representations on prediction performance. The test results obtained after nine analyses are given in Table V.

TABLE V. PERFORMANCE OF SINGLE REGRESSION MODELS

	Regression				Decision				
Model	**Features**	**RMSE**	**MAE**	**R^2**	**Threshold**	**Accuracy**	**Precision**	**Recall**	**F1**
ANFIS	Raw	0.2633	0.1409	0.6699	0.4825	0.922063	0.902544	0.9962	0.9471
	Derived	0.4159	0.3461	0.1765	0.7169	0.681849	0.828835	0.6874	0.7515
	Hybrid	0.2545	0.1411	0.6917	0.6971	0.916655	0.918273	0.9670	0.9420
RF	Raw	0.2561	0.1352	0.6878	0.8174	0.916943	0.910889	0.9769	0.9427
	Derived	0.4082	0.3301	0.2065	0.6790	0.714277	0.821494	0.7561	0.7874
	Hybrid	0.2369	0.1146	0.7327	0.7135	0.912187	0.936491	0.9382	0.9373
XGBoost	Raw	0.2597	0.1386	0.6790	0.8058	0.918995	0.906226	0.9863	0.9446
	Derived	0.4109	0.3399	0.1961	0.6966	0.704823	0.824742	0.7343	0.7769
	Hybrid	0.2434	0.1261	0.7179	0.6899	0.914872	0.927508	0.9528	0.9400

The table's results point to three clear and consistent findings. There is a significant performance decrease in all models when derived features are used alone. The hybrid feature set (Raw + Derived) is the most balanced and robust structure for all models. RF and XGBoost perform better in regression than ANFIS; however, ANFIS remains competitive in decision-making accuracy.

The lowest RMSE was obtained with the RF-Hybrid (0.2369), and the lowest MAE was obtained with the RF-Hybrid (0.1146). This result shows that RF is the most successful model in the hybrid feature space, with both lower mean error and lower extreme sensitivity. ANFIS-Hybrid (RMSE = 0.2545) and XGB-Hybrid (RMSE = 0.2434) are robust; however, RF-Hybrid is significantly ahead.

The highest R^2 values were obtained in the RF-Hybrid (0.7327) model, followed by the XGBoost-Hybrid (0.7179) and ANFIS-Hybrid (0.6917) models. This clearly shows that hybrid features enable the model to better explain the variance in the output variable. In particular, the fact that derived features alone dramatically reduce R^2 (≈0.17–0.20) indicates that they do not replace raw information but rather play a complementary role.

The highest accuracy was obtained with the ANFIS-Raw (0.9221) model, followed by the XGBoost-Raw (0.9190) and RF-Raw (0.9169) models. It is worth noting that raw features yield clearer classification signals when using decision thresholds.

The highest recall value was obtained in the ANFIS – Raw (0.9962) model, and the most balanced precision-recall value was obtained in the RF – Hybrid (0.9365 / 0.9382) model. The ANFIS-Raw configuration exhibits a recall-oriented character, while the RF-Hybrid model shows a balance-oriented character.

The highest F1 value was obtained in the ANFIS – Raw (0.9471) model, followed by the closely related XGBoost – Raw (0.9446) and RF – Hybrid (0.9373) models. This indicates that raw features are highly effective for classification, especially with ANFIS and XGB.

The findings show that using derived features alone reduces both regression and decision success, whereas the hybrid structure, which combines raw and derived features, provides significant and consistent improvements, particularly in regression performance. In terms of classification, raw features were observed to contain more distinct discriminative information, especially in the ANFIS and XGBoost models.

E. Hybrid Regression Results

1) Weighted ensemble performance: Ensemble learning approaches aim to reduce the generalization errors of individual models by combining their outputs, leveraging models with different assumptions and learning mechanisms [59], [60]. Weighted ensemble methods, in particular, offer a more balanced and effective estimation mechanism by accounting for each sub-model's validation performance, whereas simple averaging approaches assume all models contribute equally [60]. In this study, the uncertainty modeling and interpretability advantages of ANFIS are combined with the high generalization capabilities of RF and XGBoost models, and the contributions of the sub-models to the ensemble output are adaptively determined based on their validation-set error. Optimizing the weights inversely proportional to the RMSE values calculated on the validation data ensures that models with lower prediction errors have a greater impact on the final decision. This approach is consistent with the performance-based weighting strategies proposed in the literature [60], [61] and aims to contribute to more reliable prediction of student success in online mathematics education, in terms of both accuracy and stability. The most successful validation results, including RMSE-based weighted ensemble analysis and single regression, are shown in Table VI.

TABLE VI. WEIGHTED HYBRID VS. SINGLE MODELS

Model	RMSE	MAE	R^2	Threshold	Accuracy	Precision	Recall	F1
RF+Hybrid	0.2369	0.1146	0.7327	0.7135	0.912187	0.936491	0.9382	0.9373
Weighted Hybrid	0.2414	0.1255	0.7226	0.7014	0.916732	0.928630	0.9544	0.9413

Table VI evaluates the regression-based prediction performance and the classification performance derived from these predictions. The RF+Hybrid model has lower RMSE (0.2369) and MAE (0.1146) than the weighted hybrid model. However, its R^2 (0.7327) value is higher. These findings indicate that the RF-based hybrid approach has higher explanatory power and a lower error rate in predicting continuous success scores. Therefore, the RF+Hybrid model is more advantageous in terms of pure regression accuracy.

In contrast, when classification performance is examined, the weighted hybrid model offers more balanced and robust results. Although the RMSE and MAE values of the weighted hybrid model are slightly higher than those of the RF+Hybrid, it is superior in terms of accuracy (0.9167), recall (0.9544), and, especially, the F1 score (0.9413). This increase in recall value is critical, particularly in avoiding missing the "successful student" class, and yields a more pedagogically meaningful result in the context of educational analytics.

When the threshold values are examined, it is seen that the RF+Hybrid model uses a higher threshold (0.7135), whereas the weighted hybrid model uses a lower one (0.7014). This indicates that the weighted hybrid model creates a more comprehensive decision threshold, thereby increasing true positives and positively impacting Recall and F1 scores.

In conclusion, while the RF+Hybrid model provides lower error and higher explanatory power for continuous success prediction, the weighted hybrid model offers a more appropriate, balanced performance in classification-based decision-support scenarios—especially in online learning environments that require early risk detection and intervention. This finding clearly demonstrates that, in hybrid learning analytics studies that consider both regression and classification objectives, model selection should be guided by the problem context.

2) Stacking-based hybrid performance: Stacking (stacked generalization) is an advanced ensemble learning approach that combines the outputs of multiple base learners via a higher-level meta-learner and is particularly notable for its ability to effectively integrate the complementary strengths of different model types [62, 63]. In this approach, instead of using the predictions generated by the baseline models directly as the final decision, they are reweighted by a model learned in the second stage, thus incorporating nonlinear relationships between model outputs into the learning process [60]. In this

study, the outputs of ANFIS, RF, and XGBoost models were combined within a stacking framework. A linear regression model, which is simple and limits overfitting, was chosen as the meta-learner, and the final student success prediction was made using the continuous success scores from the baseline models. This structure is consistent with studies in the literature showing that heterogeneous ensemble models offer higher prediction accuracy compared to singular and weighted ensemble approaches [61, 63] and aims to more effectively model complex and uncertain student interactions in online mathematics education. The results of the hybrid regression analysis and the most successful singular regression results are given in Table VII.

TABLE VII. HYBRID REGRESSION AND SINGLE-DOMAIN REGRESSION RESULTS

Model	RMSE	MAE	R^2	Threshold	Accuracy	Precision	Recall	F1
RF+Hybrid	0.2369	0.1146	0.7327	0.7135	0.912187	0.936491	0.9382	0.9373
Weighted Hybrid	0.2414	0.1255	0.7226	0.7014	0.916732	0.928630	0.9544	0.9413
Stacking Hybrid	0.2503	0.1280	0.7017	0.5801	0.916157	0.929559	0.9524	0.9408

The results show that different hybridization strategies achieve a significant balance between minimizing error and maintaining decision sensitivity. In terms of regression performance, the RF+Hybrid model exhibits the lowest RMSE (0.2369) and MAE (0.1146), while also offering the highest R^2 (0.7327). This indicates that the RF-based hybrid structure produces more accurate continuous predictions of student achievement scores by effectively modeling nonlinear relationships. Therefore, the RF+Hybrid approach stands out in scenarios where pure regression accuracy and error minimization are prioritized. In contrast, while weighted and stacking hybrid models exhibit relatively higher regression errors, they offer more balanced and competitive classification performance. Specifically, the weighted hybrid model achieves the highest recall (0.9544) and F1 score (0.9413), demonstrating a more sensitive structure for accurately identifying at-risk students. Similarly, the stacking hybrid model, despite using a lower decision threshold (0.5801), demonstrated comparable recall (0.9524) and F1 (0.9408) scores. This finding shows that the meta-learning-based stacking approach can leverage the complementary information content of different regressors.

When examining the decision thresholds, it is observed that the RF+Hybrid model adopts a more conservative classification strategy with a higher threshold (0.7135), whereas the weighted hybrid and, especially, the stacking hybrid models increase the true positive rate by operating with lower thresholds. This preference directly affected the precision-recall trade-off, resulting in higher recall and F1 scores.

Overall, the RF+Hybrid model offers the strongest performance in continuous success prediction, while the weighted and stacking hybrid models are better suited for classification-based decision-support applications. These results clearly demonstrate that in online education analytics, rather than a single "best" model, different hybrid regression strategies should be preferred depending on the problem objectives (error minimization vs. early risk detection).

F. Meta-Classification Results

This section presents the results obtained by combining continuous success scores derived from regression models within weighted ensemble and stacking-based hybrid classification structures. The analyses show that hybrid approaches, while not significantly improving classification accuracy over single models, produce more stable and consistent results across different experimental scenarios and feature sets. This supports the usability of hybrid architectures as reliable early warning systems in online mathematics education.

In the proposed meta-classification structure, raw or derived features were not directly fed to the classifier; instead, continuous success scores obtained from ANFIS, RF, and XGBoost regression models were used as meta-features. Thus, different learning biases of the base models were integrated at the meta-level to arrive at the final decision. Classification was performed using logistic regression. After analysis, the accuracy was 0.916560, the precision was 0.929140, the recall was 0.9535, and the F1 score was 0.9412.

When the experimental results are examined, it is generally observed that the individual models and the proposed hybrid structures exhibit similar performance on metrics such as accuracy and F1-score. This proves that the pedagogical variables (PCR, PAM, RPT, etc.) derived from raw data during the feature engineering phase have a very strong signal capacity in representing the fundamental patterns in student behavior. Therefore, the similarity in performance between the models is considered a saturation point resulting from the high quality of the distinguishing features in the dataset.

G. Interpretability Analysis (ANFIS)

This section outlines the interpretability advantages of the ANFIS model within the proposed hybrid architecture. In online mathematics learning environments, high prediction accuracy is as critical as the pedagogical interpretability of these predictions. Unlike black-box machine learning models, ANFIS's Sugeno-type fuzzy inference mechanism allows for a clear examination of the underlying logical relationships in predicting success by representing student interactions through "If-Then" rules. In this context, the ANFIS model is considered not only a predictor within the hybrid structure but also an explanatory component that transforms the uncertainty in student behavior into a pedagogically interpretable form. Three rules have been established in the educational process using the hybrid feature set and the ANFIS model. These are:

Rule 1: IF AC is low [MF 1] AND HC is low [MF 1] AND RT is low [MF 1] AND OT is low [MF 1] AND SDI is low [MF 1] AND SEL is low [MF 1] AND RPT is high [MF 1] AND PCR is low [MF 1] AND PAM is low [MF 1] AND PHM is low [MF 1] AND OC is low [MF 1] THEN output = -24.596*x -6.812*x

-0.123*x -12.608*x -0.436*x +0.764*x -0.033*x -0.226*x +0.689*x +0.474*x +0.309*x +2.747

Rule 2: IF AC is low [MF 2] AND HC is low [MF 2] AND RT is low [MF 2] AND OT is low [MF 2] AND SDI is low [MF 2] AND SEL is low [MF 2] AND RPT is high [MF 2] AND PCR is low [MF 2] AND PAM is low [MF 2] AND PHM is low [MF 2] AND OC is low [MF 2] THEN output = -0.029*x -0.058*x +0.006*x -0.059*x -0.104*x -0.001*x +0.006*x +0.050*x +0.002*x +0.009*x -0.009*x +0.054

Rule 3: IF AC is low [MF 3] AND HC is low [MF 3] AND RT is low [MF 3] AND OT is low [MF 3] AND SDI is low [MF 3] AND SEL is low [MF 3] AND RPT is high [MF 3] AND PCR is low [MF 3] AND PAM is low [MF 3] AND PHM is low [MF 3] AND OC is low [MF 3] THEN output = -4.253*x -1.033*x -0.507*x +8.283*x -0.153*x +0.201*x -0.056*x +0.462*x +0.033*x -0.106*x +0.088*x -0.425

The Takagi-Sugeno-Kang (TSK) type fuzzy rules obtained in the study reflect the system's behavioral tendencies towards a specific student profile. The antecedents of the three rules examined are identical: they describe a group of students with low experience and low past achievement, but whose current performance is on the rise.

The linear regression coefficients in the consequences of the rules determine the weight of the variables on the outcome.

- Rule 1 (High-Precision Estimation): This rule has the highest coefficient ($w_1 = -24.596$). Academically, this indicates that the model attempts to capture "extreme values" in this specific student group. It is the component of the system that gives the harshest punitive response to the increase in the number of erroneous trials.
- Rule 2 (Stable Estimation): The proximity of the coefficients to zero ($(w_i \approx 0)$) proves that this rule acts as an "anchor" in the system. It stabilizes the output by filtering noise in the variables and prevents overfitting.
- Rule 3 (Balanced Effect): Rule 3 establishes a balance between positive and negative weights. Specifically, the +8.283 coefficient on the fourth variable indicates that a particular parameter (e.g., dwell time or response speed) is considered strong evidence in favor of "learning" under this rule.

When the rules are examined as a whole, it is seen that the factors "Low Experience" and "Low Past Success" are treated as negative starting points (biases) by the model. However, the "RPT is high" condition causes the model to categorize the student as a "recovering student".

The fact that these three rules share the same antecedent confirms that the system uses a fuzzy clustering method and that the final prediction is produced by taking the weighted average of these rules (defuzzification). This shows that the model can make more flexible predictions by combining multiple local models with different gradients rather than a single linear model. The model's membership functions are determined by the model and shown in Fig. 2.

The membership functions used in the system are generally Gaussian or Generalized Bell-type curves. This choice ensures that the transitions between input values are smooth and continuous, rather than sharp, allowing the model to more accurately capture non-linear student behavior. As shown in Fig. 2, features such as HC and PAM show asymmetric intensity distributions at low values, whereas cognitive load indicators such as RT and SDI exhibit normal distributions. This configuration allows the system to respond aggressively to student errors while calculating success trends with a more balanced weighted average.

In the AC, HC, PAM, and PHM graphs, the curves are leaning to the left (towards zero). The model considers the amount of student use of systemic assistance tools as an indicator of "independence." The rapid damping of the curves after a scale of 0.4 indicates that the system labels a small number of trials as "positive/low" and that, when this threshold is exceeded, the student quickly moves into the "struggling student" category. The fact that the OC graph also peaks at similarly low values indicates that the initial responses to newly encountered tasks are weighted.

In the PCR and SEL graphs, membership levels span a wider range. These variables determine the model's "long-term student profile." In the PCR graph, high membership values in the 0.6-0.8 range indicate that the system evaluates students with moderate-to-high achievement levels within a similar confidence interval. In the SEL graph, the distinct separation of different clusters (blue, orange, and yellow curves) suggests that the model triggers different rule sets for novice and expert students.

The RT and SDI graphs exhibit a complete bell-shaped curve. These parameters represent the student's cognitive state. The RT graph peaking around 0.7-0.8 indicates that the system considers responses that follow a specific thought process "ideal," rather than those that are too fast (predictive) or too slow (disconnected). The SDI, which peaks in the 0.3-0.4 range, indicates that the system measures performance more stably on medium-easy tasks.

The RPT function is perfectly symmetrical and centered at 0. These variables measure the model's "learning momentum." Centering the momentum at zero indicates that the model has an unbiased starting point for the student's direction (progress or regression), and even a slight deviation in the trend immediately affects the output (via Rule 1-3 coefficients).

During the fuzzification of the 11 model input parameters, the variances of the Gaussian membership functions were optimized based on the educational significance of each variable.

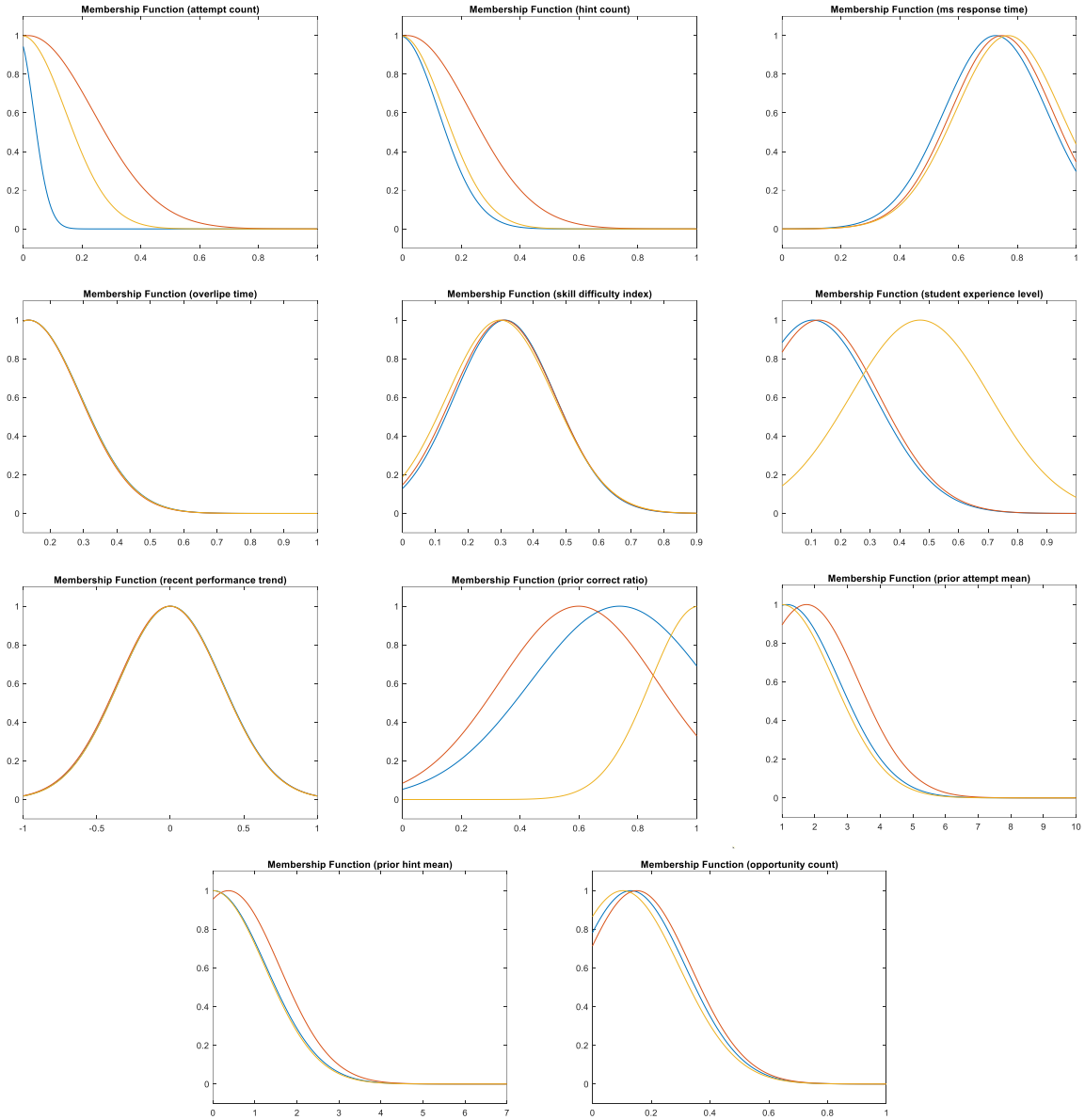

Fig. 2. Membership functions created with ANFIS and hybrid features ("AC", "HC", "RT", "OT", "SDI", "SEL", "RPT", "PCR", "PAM", "PHM", and "OC", respectively).

V. Discussion of Results

The findings of this study reveal, from both methodological and pedagogical perspectives, the advantages that hybrid approaches offer in predicting student achievement in online mathematics education compared to singular models and traditional assessment methods.

A. Hybrid Model Performance and Predictive Stability

Experimental results show that hybrid structures combining RF, XGBoost, and ANFIS models exhibit more stable performance in predicting continuous success scores (regression) and identifying at-risk students (classification). Specifically, the RF+Hybrid model, achieving the lowest RMSE (0.2369) and the highest R^2 (0.7327), confirms the power of tree-based ensemble learning algorithms in modeling complex, nonlinear student behavior. However, the weighted hybrid model, which achieves the highest F1-score (0.9413) and recall value (0.9544) in the classification phase, demonstrates that combining the models with weights minimizes erroneous decisions and produces more reliable early warning signals.

The numerical similarity in the classification performance of the models necessitates considering not only 'total accuracy' but also strategic metrics such as 'prediction stability' and 'error type cost' when selecting a model. While single-domain models (e.g., ANFIS-Raw) achieve high accuracy in some scenarios, the weighted hybrid model's high recall (0.9544) and F1 score (0.9413) during the classification phase provide a more reliable decision boundary for identifying at-risk students. This finding confirms that the hybrid architecture minimizes asymmetric error costs (mistaking failing students for passing students),

generating a more reliable early warning signal in the context of educational analytics.

B. The Role of Feature Engineering and Information Complementarity

Feature analysis results indicate that combining raw data and derived features is critical for model performance. The dramatic decrease in R^2 values (≈0.17–0.20) in scenarios where derived features are used alone indicates that these features cannot replace raw interaction information and should be included in the model as "context providers". In particular, the negligible correlation of the RPT feature with other features indicates that it provides independent, high-value "learning momentum" information for the model.

C. Uncertainty Management and Pedagogical Interpretability

The IF-THEN rules and membership functions provided by the ANFIS model have made the logical process underlying predictions transparent, unlike those of machine learning models, which are often described as "black boxes."

- Help-Seeking Behavior: The concentration of membership functions for HC and AC features at low values indicates that the model positively weights the use of a small number of hints as "learning independence" and quickly moves the learner to the "risky" category when a certain threshold is exceeded.
- Cognitive Load: The bell-curve form of RT documents that the system considers responses that come after a reasonable thought process as "ideal," rather than responses that are too fast (predictive) or too slow (disconnected).

D. Practical Applicability and Decision Support Potential

The study's two-stage (regression and classification) structure provides educators with both a continuous "probability of success" and a clear "successful/at-risk" label. The use of ROC-based adaptive thresholding created a more reliable decision boundary, particularly by reducing the pedagogical risks of false negatives (counting a failing student as successful). This suggests that the proposed model can be integrated into an early warning system for real-time monitoring on online mathematics platforms.

VI. Conclusion and Future Work

This study presents a unique hybrid framework combining fuzzy logic and machine learning techniques to predict student performance in online mathematics education. The main findings of the research are as follows:

- Superiority of the Hybrid Structure: Compared to single models, combining ANFIS's uncertainty modeling with the high predictive power of Random Forest and XGBoost yielded more balanced, reliable results.
- Prediction Accuracy: While the Random Forest (RF)-based hybrid model yielded the lowest regression error (RMSE: 0.2369), the Weighted Hybrid model excelled in classification accuracy (Accuracy: 0.9167) and risk detection sensitivity (Recall: 0.9544).
- Feature Interaction: Supporting raw interaction data with derived pedagogical features significantly increased the model's explanatory power (R^2).
- Explainability: The linguistic rules generated by ANFIS, unlike "black box" models, have enabled understanding of the cognitive and behavioral reasons behind the risk of failure (such as overuse of hints or temporal disconnections).

Research results have shown that different modeling paradigms (neuro-fuzzy and tree-based ensemble learning) achieve similar accuracy levels. However, the high predictive power offered by Random Forest and XGBoost models, supported by ANFIS's pedagogically interpretable IF-THEN rules, distinguishes the proposed hybrid framework from 'black box' models, making it unique. This not only provides high predictive performance but also enhances the pedagogical value of the study by enabling an understanding of the cognitive reasons behind the risk of failure (such as overuse of hints or temporal disconnections).

In the future, this model is planned to be tested with datasets from different disciplines and integrated into a real-time educational interface to provide teachers with instant intervention reports.

VII. Pedagogical Recommendations

Based on the findings from the model's membership functions and rule analysis (ANFIS), the following recommendations have been developed for online mathematics platforms:

- Balanced Use of Hints: Since excessive hinting (HC) is perceived by the model as a direct risk signal, the system should direct students to additional explanatory videos after a certain point instead of overwhelming them with hints.
- Encouragement of Ideal Thinking Time: Since very fast responses are often considered guessing (gaming the system), a minimum "waiting time" should be encouraged for students to read and analyze the question.
- Cumulative Follow-up: Due to the cumulative nature of mathematics education, the system should automatically define micro-tasks to "remind" students of past topics for those with low past achievement rates (PCR).
- Feedback Focused on Learning Momentum: Motivational messages should be delivered focusing not only on the result but also on the student's rising performance trend (RPT).

VIII. Limitations and Threats to Validity

The following limitations should be considered when evaluating the results of the study:

- Dataset Coverage: Analyses were performed only for the mathematics domain using the ASSISTments dataset (2009-2010); additional validations are needed to generalize the results to other subject areas (language, science, etc.).

- External Factors: Off-platform data, such as students' off-platform study habits, physical environment, or socio-economic status, could not be included in the model.
- Interaction Limitations: To maintain the stability of the model, students with very few (below 5) or many (above 1000) interactions were excluded from the analysis, as this may affect performance in extreme profiles.
- Computational Cost: The hybrid structure combining ANFIS and ensemble models may require high computational power in systems with very large-scale and real-time applications.

REFERENCES

[1] Z. Shou, M. Xie, J. Mo, and H. Zhang, "Predicting Student Performance in Online Learning: A Multidimensional Time-Series Data Analysis Approach," *Appl. Sci.*, vol. 14, no. 6, p. 2522, 2024.

[2] J. Stojanović *et al.*, "Application of distance learning in mathematics through adaptive neuro-fuzzy learning method," *Comput. Electr. Eng.*, vol. 93, p. 107270, 2021.

[3] W. Zhang, X. Huang, S. Wang, J. Shu, H. Liu, and H. Chen, "Student performance prediction via online learning behavior analytics," in *2017 International Symposium on Educational Technology (ISET)*, IEEE, 2017, pp. 153–157. Accessed: Jan. 24, 2026.

[4] A. Prasanth, "TabNet–XGBoost Hybrid Model for Student Performance Prediction and Customized Feedback.," *Int. J. Adv. Comput. Sci. Appl.*, vol. 16, no. 10, 2025.

[5] E. Evangelista, "A hybrid machine learning framework for predicting students' performance in virtual learning environment," *Int. J. Emerg. Technol. Learn. IJET*, vol. 16, no. 24, pp. 255–272, 2021.

[6] J. Wang and Y. Yu, "Machine learning approach to student performance prediction of online learning," *PloS One*, vol. 20, no. 1, p. e0299018, 2025.

[7] M. Feng, N. Heffernan, and K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses," *User Model. User-Adapt. Interact.*, vol. 19, no. 3, pp. 243–266, 2009.

[8] X. Zhang, Y. Zhang, A. L. Chen, M. Yu, and L. Zhang, "Optimizing multi label student performance prediction with GNN-TINet: A contextual multidimensional deep learning framework," *PloS One*, vol. 20, no. 1, p. e0314823, 2025.

[9] Z. R. Khudhair, M. Raheema, and J. Salman, "Evaluation of the Use of ANFIS in Predicting Student's," *Kerbala J. Eng. Sci.*, vol. 3, no. 3, pp. 101–113, Sep. 2023, doi: 10.63463/kjes1088.

[10] F. Widyahastuti and V. U. Tjhin, "Student Performance Prediction using Online Behavior Discussion Forum with Data Mining Techniques:," in *Proceedings of the Borneo International Conference on Education and Social Sciences*, Banjarmasin, Indonesia: SCITEPRESS - Science and Technology Publications, 2018, pp. 90–95. doi: 10.5220/0009017000900095.

[11] V. Ponnan, A. Abdulla, H. Muthukumaran, and V. Natarajan, "Data-Driven Hybrid Fuzzy-ML Model for Comprehensive Student Assessment," in *Understanding Uncertainty: Modern Approaches in Fuzzy System and Applications*, 2025, pp. 100–123.

[12] "Predicting Student Academic Success Using Deep Learning: A Multi-Factor Approach to Performance Prediction," *J. Logist. Inform. Serv. Sci.*, Mar. 2025, doi: 10.33168/JLISS.2025.0114.

[13] M. Nazir, A. Noraziah, M. Rahmah, M. Fakherldin, and A. Khawaji, "Transforming Education with Deep Learning: A Systematic Review on Predicting Student Performance and Critical Challenges," *Fusion Pract. Appl.*, no. Issue 2, pp. 79–99, Jan. 2025, doi: 10.54216/FPA.180207.

[14] A. E. Villegas-Espinoza and J. I. Necochea-Chamorro, "Using Deep Learning in Student Performance Prediction: A Systematic Review," *TEM J.*, vol. 14, no. 3, p. 2472, 2025.

[15] M. A. A. Bakar, A. T. Ab Ghani, M. L. Abdullah, N. Ismail, and S. Ab Aziz, "Adaptive Neuro-Fuzzy Inference System (ANFIS) Formulation to Predict Students' Neuroscience Mechanistic: A Concept of an Intelligent Model to Enhance Mathematics Learning Ability.," *TEM J.*, vol. 11, no. 4, 2022.

[16] I. Sapuguh, N. Ahlina, A. Wahyudi, B. Setyawan, and A. S. Rosalinda, "Development of fuzzy logic based student performance prediction system," *J. Tek. Inform. CIT Medicom*, vol. 15, no. 6, pp. 284–290, 2024.

[17] A. Al-Hmouz, J. Shen, R. Al-Hmouz, and J. Yan, "Modeling and simulation of an adaptive neuro-fuzzy inference system (ANFIS) for mobile learning," *IEEE Trans. Learn. Technol.*, vol. 5, no. 3, pp. 226–237, 2011.

[18] E. B. George, "Explainable AI Methods for Predicting Student Grades and Improving Academic Success," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 23s, pp. 117–126, Mar. 2025, doi: 10.52783/jisem.v10i23s.3680.

[19] R. Chimatapu, H. Hagras, A. Starkey, and G. Owusu, "Explainable AI and Fuzzy Logic Systems," in *Theory and Practice of Natural Computing*, vol. 11324, D. Fagan, C. Martín-Vide, M. O'Neill, and M. A. Vega-Rodríguez, Eds., in Lecture Notes in Computer Science, vol. 11324. , Cham: Springer International Publishing, 2018, pp. 3–20. doi: 10.1007/978-3-030-04070-3_1.

[20] A. Altaher and O. BaRukab, "Prediction of student's academic performance based on adaptive neuro-fuzzy inference," *Int. J. Comput. Sci. Netw. Secur. IJCSNS*, vol. 17, no. 1, p. 165, 2017.

[21] T. O. Soyoye, T. Chen, and R. Hill, "Predicting Academic Performance of University Students Using Adaptive Neuro Fuzzy Inference System (ANFIS)- Subtractive Clustering Algorithm (ANFIS-SC): A Case Study in the UK," in *Advances in Computational Intelligence Systems*, vol. 1462, H. Zheng, D. Glass, M. Mulvenna, J. Liu, and H. Wang, Eds., in Advances in Intelligent Systems and Computing, vol. 1462. , Cham: Springer Nature Switzerland, 2024, pp. 315–333. doi: 10.1007/978-3-031-78857-4_24.

[22] D. Yang and M. S. A. Malik, "Design of performance evaluation method for higher education reform based on adaptive fuzzy algorithm," *PeerJ Comput. Sci.*, vol. 11, p. e3090, 2025.

[23] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Appl. Sci.*, vol. 11, no. 1, p. 237, 2020.

[24] W. Ahmed, M. A. Wani, P. Plawiak, S. Meshoul, A. Mahmoud, and M. Hammad, "Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions," *Sci. Rep.*, vol. 15, no. 1, p. 26879, 2025.

[25] C. S. Sanaboina and D. R. Sri, "Comparative Analysis of Hybrid Machine Learning Models for Predicting Student Performance with Data Balancing Via SMOTE and GAN," *Int. J. Eng. Res. Technol.*, vol. 14, no. 9, Sep. 2025, doi: 10.5281/zenodo.18096036.

[26] S. Sayadi and M. K. Sayadi, "A hybrid of random forest and SVM for predicting student performance," *Int. J. Ind. Eng. Manag. Sci.*, vol. 9, no. 1, pp. 44–51, 2022.

[27] C. Li and Z. Cao, "Deep learning-based AI model for predicting academic success and engagement among physical higher education students," *Sci. Rep.*, 2025.

[28] R. Hemidov, "Optimizing Student Performance Prediction Employing Hybrid Random Forest Models Boosted by Nature-Inspired Algorithms," *J. Artif. Intell. Syst. Model.*, vol. 3, no. 03, pp. 55–70, 2025.

[29] J. A. Goguen, "LA Zadeh. Fuzzy sets. Information and control," *J. Symb. Log.*, vol. 38, no. 4, pp. 656–657, 1973.

[30] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "PREDICTING STUDENTS' PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES," *Appl. Artif. Intell.*, vol. 18, no. 5, pp. 411–426, May 2004, doi: 10.1080/08839510490442058.

[31] T. Bannet, S. Kumar, and K. Chidananda, "Machine Learning techniques for student performance prediction," *J. Emerg. Technol. Innov. Res.*, vol. 8, no. 10, pp. c616–c620, 2021.

[32] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," *Comput. Educ.*, vol. 61, pp. 133–145, 2013.

[33] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.

[34] O. Taylan and B. Karagözoğlu, "An adaptive neuro-fuzzy model for prediction of student's academic performance," *Comput. Ind. Eng.*, vol. 57, no. 3, pp. 732–741, 2009.

[35] V. O. Eguavoen and E. Nwelih, "A Hybrid FCM-PSO-ANFIS Model for Predicting Student Academic Performance," *J. Sarj. Tek. Inform.*, vol. 12, no. 3, pp. 91–98, 2024.

[36] K. O. Adefemi, M. B. Mutanga, and V. Jugoo, "Hybrid Deep Learning Models for Predicting Student Academic Performance," *Math. Comput. Appl.*, vol. 30, no. 3, p. 59, 2025.

[37] S. Göktepe Körpeoğlu, A. Filiz, and S. Göktepe Yıldız, "AI-driven predictions of mathematical problem-solving beliefs: Fuzzy logic, adaptive neuro-fuzzy inference systems, and artificial neural networks," *Appl. Sci.*, vol. 15, no. 2, p. 494, 2025.

[38] M. Zerkouk, M. Mihoubi, and B. Chikhaoui, "A Comprehensive Review of AI-based Intelligent Tutoring Systems: Applications and Challenges," Jul. 25, 2025, *arXiv*: arXiv:2507.18882. doi: 10.48550/arXiv.2507.18882.

[39] A. Kobsa, "User modeling and user-adapted interaction," in *Conference companion on Human factors in computing systems - CHI '94*, Boston, Massachusetts, United States: ACM Press, 1994, pp. 415–416. doi: 10.1145/259963.260532.

[40] N. T. Heffernan and C. L. Heffernan, "The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching," *Int. J. Artif. Intell. Educ.*, vol. 24, no. 4, pp. 470–497, Dec. 2014, doi: 10.1007/s40593-014-0024-x.

[41] M. Gray, "Predicting Student Performance Using Discussion Forums' Participation Data," Master's Thesis, Duke University, 2024. Accessed: Jan. 24, 2026.

[42] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 40, no. 6, pp. 601–618, 2010.

[43] Z. A. Pardos and N. T. Heffernan, "Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing," in *User Modeling, Adaptation, and Personalization*, vol. 6075, P. De Bra, A. Kobsa, and D. Chin, Eds., in Lecture Notes in Computer Science, vol. 6075. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 255–266. doi: 10.1007/978-3-642-13470-8_24.

[44] R. S. Baker, T. Martin, and L. M. Rossi, "Educational Data Mining and Learning Analytics," in *The Handbook of Cognition and Assessment*, A. A. Rupp and J. P. Leighton, Eds., Hoboken, NJ, USA: John Wiley & Sons, Inc., 2016, pp. 379–396. doi: 10.1002/9781118956588.ch16.

[45] J.-S. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Trans. Syst. Man Cybern.*, vol. 23, no. 3, pp. 665–685, 1993.

[46] J. Ferrandiz, D. Fonseca, and A. Banawi, "Mixed Method Assessment for BIM Implementation in the AEC Curriculum," in *Learning and Collaboration Technologies*, P. Zaphiris and A. Ioannou, Eds., Cham: Springer International Publishing, 2016, pp. 213–222. doi: 10.1007/978-3-319-39483-1_20.

[47] Z. Ibrahim and D. Rusli, "Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression," in *21st Annual SAS Malaysia Forum, 5th September*, Kuala Lumpur, 2007.

[48] D. Duan and S. Zhang, "Hybrid Model of Fuzzy Logic and Recurrent Neural Network for Dynamic Student Achievement Prediction," *Informatica*, vol. 49, no. 26, 2025.

[49] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive models for educational purposes: A systematic review," *Big Data Cogn. Comput.*, vol. 8, no. 12, p. 187, 2024.

[50] W. J. Van Der Linden and R. K. Hambleton, "Item Response Theory: Brief History, Common Models, and Extensions," in *Handbook of Modern Item Response Theory*, W. J. Van Der Linden and R. K. Hambleton, Eds., New York, NY: Springer New York, 1997, pp. 1–28. doi: 10.1007/978-1-4757-2691-6_1.

[51] D. J. Hand, "Handbook of modern item response theory," *Biometrics*, vol. 54, no. 4, p. 1680, 1998.

[52] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[53] T. Chen, "XGBoost: A Scalable Tree Boosting System," *Cornell Univ.*, 2016.

[54] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[55] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big Data*, vol. 1, no. 1, pp. 51–59, Mar. 2013, doi: 10.1089/big.2013.1508.

[56] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950, doi: 10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3.

[57] R. Pelánek, "Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques," *User Model. User-Adapt. Interact.*, vol. 27, no. 3, pp. 313–350, 2017.

[58] X. Guo, R. Wang, X. Sun, and Q. Zhang, "Research on Student Performance Prediction Based on SVM Optimized by Hybrid SSA.," *IAENG Int. J. Comput. Sci.*, vol. 52, no. 7, 2025.

[59] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*, vol. 1857, in Lecture Notes in Computer Science, vol. 1857. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. doi: 10.1007/3-540-45014-9_1.

[60] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2025.

[61] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1, pp. 1–39, Feb. 2010, doi: 10.1007/s10462-009-9124-7.

[62] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

[63] L. Breiman, "Stacked regressions," *Mach. Learn.*, vol. 24, no. 1, pp. 49–64, Jul. 1996, doi: 10.1007/BF00117832.