Hybrid Machine Learning Models Based on CATBoost Classifier for Assessing Students' Academic Performance

Ding Hao*, Yang Xiaoqi, Qi Taoyu
School of Educational Science, Anhui Normal University, Wuhu, Anhui 241000, China

***Abstract*—This study addresses the imperative task of predicting and evaluating students' academic performance by amalgamating qualitative and quantitative factors, crucial in light of the persisting challenges undergraduates encounter in completing their degrees. Educational institutions wield significant influence in prognosticating student outcomes, necessitating the application of data mining (DM) techniques such as classification, clustering, and regression to discern and forecast student study behaviors. Through this research, the potential of deriving demonstrates valuable insights from educational data, empowering educational stakeholders with enhanced decision-making capabilities and facilitating improved student outcomes. Employing a hybrid approach, models developed within the realm of educational DM, leveraging the CATBoost Classifier (CATC) in conjunction with two cutting-edge optimization algorithms: Victoria Amazonica Optimization (VAO) and Artificial Rabbits Optimization (ARO). Initially, the models undergo partitioning into training and testing sets for performance evaluation utilizing statistical metrics. After classifying 649 students according to their final scores, VAO outperformed ARO in terms of maximizing CATC's classification ability, resulting in an approximate 6% enhancement in accuracy and precision. Moreover, the VAO model adeptly categorizes 606 out of 649 students accurately. This research furnishes invaluable predictive models for educators, researchers, and policymakers endeavoring to enrich students' educational journeys and foster academic success.**

***Keywords*—Academic performance; hybridization; CATBoost classifier; meta-heuristic algorithms; educational institutions**

Nomenclature

CATC	CATBoost classifier	ARO	Artificial Rabbits Optimization
VAO	Victoria Amazonica Optimization	CAAR	CAT+ARO
CAVA	CAT+VAO	DM	Data Mining
AUC	Area Under the Receiver Operating Characteristic Curve	MCC	Matthews Correlation Coefficient
Pstatus	Parents' Cohabitation Status	Medu	mother's education
Mjob	Mother's employment	Fedu	Father's Education
Fedu	Father's employment	G3	Grade 3

I. Introduction

The educational processes generate vast quantities of data, including information related to academic grades, enrollment, and student performance. The increasing volume of this data has prompted consideration of its utilization beyond mere accountability, aiming to extract valuable insights and facilitate informed decision-making within the academic domain, ultimately fostering advancements in the educational sector [1–3].

In order to extract useful information from students, a broad variety of student variables may be analyzed in the quickly developing scientific subject of educational DM [4,5]. In this context, numerous predictive algorithms have been effectively employed in educational settings for various purposes, utilizing diverse data sets and student records. A comprehensive review outlines two primary application purposes within academic contexts: predictors and early warning systems [6].

The purpose of predictors is to foresee how a course or degree will turn out, based on specific input data, while early warning systems not only perform this predictive function and report their findings to teachers or students at an early stage, enabling preemptive actions to prevent or lessen possible adverse consequences. Common forecast objectives in this context include assessing the risk of course failure, predicting dropout rates, estimating grades (focusing exclusively on college performance [7,8], or substituting individual course grades with semester-based course averages such as Grade Point Average (GPA) per semester or cumulative GPA at the time of prediction [9,10]), and forecasting graduation rates.

Predicting academic performance is a highly noteworthy objective; for example, at the time of graduation, it has multiple vital purposes, including assisting educational institutions in identifying at-risk students for specialized help to lower failure rates and providing information to admissions committees about candidates likely to finish their program, recognizing high-achieving students to guide their career development, and assessing key factors to enhance the quality of education continuously. When examining the existing literature on predicting students' academic performance, it becomes evident that these studies predominantly rely on four categories of student information: demographic and socioeconomic information, statistics from high school, records of college enrollment, and data on academic achievement up to the time of projection [11].

Frequently used predictive factors in academic performance include demographics like sex [12] and household income [13], along with high-school data such as GPA and admission test scores [12,14]. College-related information encompasses major, full-time, or part-time status and scholarship availability

[12,15,16]. Additionally, academic performance is usually represented by past course grades, except for predictive models used during admission [17,18].

In recent years, wide-ranging research has been conducted to analyze the factors influencing student performance, encompassing the direct and indirect attributes that affect academic outcomes. Some studies focus solely on attributive analysis, while others employ machine learning (ML) algorithms, particularly AI techniques [19], like ANN, random forests (RF), and Bayesian classifiers, to forecast student performance according to these attributes. Specific examples include the application of the Naive Bayesian DM technique to predict student performance based on 19 attributes such as gender, family status, and students' grades [20].

Support Vector Machines (SVM) have demonstrated significant improvements in predicting students' problem-solving performance using Bayesian Knowledge Tracing (BKT) compared to the standard BKT method [21]. Various feature selection techniques, decision tree (DT) algorithms, particle swarm optimization, and ensemble methods have also been employed for student performance prediction [22]. Socio-economic factors and entrance examination results have been utilized to predict cumulative grade point averages, and multiple techniques have been explored for forecasting students' academic success and choice of majors, with the Random Forest Classifier proving particularly effective [23, 24]. Additionally, hybrid models combining generative and discriminative models have been employed, and fuzzy logic, Adaptive Neuro-Fuzzy Inference System (ANFIS), and fuzzy ANFIS have been used for ratings and predictions in the educational context [25,26].

II. Related Work

Many academics have painstakingly examined the many aspects impacting students' achievement at different levels of education [27,28]. Numerous research in this field has used DM techniques, namely classification algorithms, to forecast student performance and boost the total effectiveness of higher education institutions. This section provides a summary of several relevant studies, paying special focus to those that address DT and classification methods in evaluating students' academic achievement [29–32].

For instance, Mustafa et al. [33] analyzed student data from C++ classes using the Cross Industry Standard Process for DM (CRISP) framework. She compared the performance of many classifiers, including Iterative Dichotomize 3 ($ID3$), C4.5 DT, and Naive Bayes (NB). With its improved performance, the C4.5 DT provided valuable insights into the variables affecting student success. Using classification and clustering methods, Sunita and LOBO L.M.R.J. The research in [34] were able to predict student performance and categorize students according to that performance, demonstrating the usefulness of DM in education. Classification models were created by Bichkar and R. R. Kabra [35] with the intention of detecting vulnerable first-year engineering students.

Table I shows the overview of published papers.

TABLE I. Overview of Published Papers

Ref.	Carried out works	Advantage	Disadvantage
Cortez and Silva [36]	This study analyzes educational trends in Portugal, focusing on high student failure rates in core subjects. It utilizes Business Intelligence/Data Mining (BI/DM) techniques to address achievement issues, collecting recent real-world data via school reports and questionnaires. Mathematics and Portuguese subjects are modelled using classification and regression tasks.	- Addresses pressing education issues. - Utilizes advanced BI/DM techniques. - Incorporates recent real-world data. - Comprehensive evaluation of models and methods. - Identifies key factors influencing achievement. - Provides actionable insights for tool development.	- Reliance on retrospective data. - Limited focus on specific subjects. - Potential effectiveness variations. - Biases or limitations in data collection. - Lack of consideration for external factors.
Hasib et al. [37]	5 classification algorithms were utilized in the development of a prediction model for secondary school student success: XGBoost, Naive Bayes, K-Nearest Neighbors (KNN), and Logistic Regression. 2 Portuguese school reports and surveys provided the data, which was then used to model the mathematics and Portuguese language disciplines using binary/5-level classification tasks. To address dataset imbalance, K-Means SMOTE was employed. Additionally, interpretable LIME models were trained for all classifiers, enhancing model transparency and interpretability.	- provides a model that uses sophisticated classification algorithms to predict student achievement - Utilizes real-world data from Portuguese school reports and surveys. - Addresses imbalanced dataset issue with K-Means SMOTE. - Achieves high accuracy (96.89%) with Support Vector Machine (SVM). - Enhances model interpretability with LIME, providing confidence and transparency in predictions.	- Reliance on specific classification algorithms may limit the exploration of other potential models. The generalizability of findings may be limited to the Portuguese education context. - Interpretability may vary depending on the complexity of underlying processes.
Asselman et al. [38]	Focus on enhancing the Performance Factors Analysis (PFA) approach, a crucial component of Knowledge Tracing (KT) in adaptive educational hypermedia systems. Introduction of Ensemble Learning methods, specifically Random Forest, AdaBoost, and XGBoost, to improve predictive accuracy of student performance. Evaluation of the proposed models on 3 different datasets.	- Addresses the need for improved prediction accuracy in educational hypermedia systems. - Introduces Ensemble Learning methods to enhance technical aspects of PFA. - Evaluation of multiple datasets enhances the generalizability of findings. - Demonstrates a substantial improvement in performance prediction compared to the original PFA algorithm, particularly with XGBoost.	- Focus on technical enhancements may overlook pedagogical considerations. - Limited discussion on potential challenges or limitations of the proposed approach. - Generalizability of findings may be restricted to specific datasets or educational contexts.

Shreem et al. [39]	introduced a wrapper feature selection technique for student performance prediction systems using an extended binary genetic algorithm (EBGA). A new hybrid selection process that combines the electromagnetic-like (EM) approach with k-means clustering is proposed. Using EBGA in conjunction with five classifiers (KNN, DT, NB, SVM, and Linear Discriminant Analysis) in a hybrid ML technique. Assessment of the suggested methodology using two actual case studies from the UCI ML Repository.	- Introduction of an enhanced feature selection method tailored for student performance prediction systems. - Novel hybrid selection mechanism improves predictive accuracy. - Utilization of a hybrid ML approach enhances model performance. - Evaluation of real case studies enhances the practical applicability of findings. - Demonstrates an improvement in the performance of binary genetic algorithms and classifiers by 1% to 11%.	- Limited discussion on potential challenges or limitations of the proposed approach. - Generalizability of findings may be restricted to specific datasets or educational contexts. - Lack of comparison with existing feature selection or hybrid ML methods.
Sarwat et al. [40]	To predict student success through in-class and at-home tutoring, a deep-layer support vector machine (SVM) and an enhanced conditional generative adversarial network (CGAN) is proposed. Creating artificial data samples in order to deal with tiny dataset sizes. Model performance is assessed both with and without CGAN. Examination of several kernel-based methods for deep SVM, such as polynomial, sigmoid, radial, and linear functions. Performance comparison of the suggested model with the current solutions.	- Effectively addresses small dataset size with synthetic data generation. - Demonstrates improved prediction accuracy with combined school and home tutoring. - Extensive evaluation of multiple kernel-based approaches for deep SVM. - Outperforms existing solutions in sensitivity, specificity, and AUC.	- Limited discussion on potential challenges or limitations of the proposed approach. - Complexity of the model may hinder replication or generalization. - Comparative analyses may overlook nuances in different educational contexts.
Mehdi and Nachouki [41]	created an explanatory and prediction model utilizing ANFIS to forecast the grade point average (GPA) of graduates in Ajman University's computer technology program. Use of high school GPA (HSGPA) and grades in foundational and introductory IT courses as predictors. Sensitivity analysis to ascertain each predictor's relative importance. ANFIS methodology is compared to popular methods like multilinear regression.	- Effective use of ANFIS methodology for predicting GPA. - Identification of key predictors and their significance in influencing graduation GPA. - High predictive accuracy, with 77% of predicted values within one root mean square error of actual GPA. - Demonstrates ANFIS's superiority over commonly used techniques like multilinear regression. - Provides actionable insights for improving IT education programs.	- Focused on a single academic program at one institution, which may limit generalizability. - Limited discussion on potential challenges or limitations of the ANFIS methodology. - Results may vary in different educational contexts or with different datasets.

Despite the numerous models used for classifying student performance, none have incorporated the CATBoost classifier (CATC) until this point. With the aim of bridging this gap, the primary objective of this study was to advance a CATC-based model for forecasting student performance in language courses, using trustworthy data sources. The selection of the CATC was informed by its recognized robustness and efficacy in handling categorical features, a prevalent characteristic in student performance prediction tasks. CATBoost's track record of superior performance across diverse domains made it a compelling choice for this study, where accurate prediction of student outcomes is paramount. Moreover, the integration of VAO and ARO was driven by their specific strengths in optimization tasks. VAO, inspired by the adaptive behavior of the VA plant, excels in dynamic optimization problems, thereby offering an edge in scenarios with evolving parameters or data dynamics. Similarly, ARO, which mimics the foraging behavior of rabbits in search of optimal solutions, shows promise in fine-tuning model parameters and improving overall model performance. The study's goal was to increase the predictive model's precision and accuracy by combining CATBoost with VAO and ARO in a way that maximizes their complementary strengths. This approach would also help to strengthen and improve the forecasting framework for language course student performance.

In the following sections, related work is given in Section II, the dataset description and processing are detailed in Section III. Section IV provides an in-depth explanation of the presented model, while Section V discusses the meta-heuristic algorithms used. Section VI outlines the metrics employed to assess the performance of the developed models. Convergence analysis is given in Section VII. Results and discussion is given in Section VIII and Section IX respectively. Finally, Section X concludes the paper.

III. Data Selection and Processing

This study uses data collected from previous literature [36,42]. Despite some governmental investments in Information Technology, most public schools still rely on paper-based information systems. Consequently, the database may be constructed from two sources: school reports (containing final grades and school absences) or questionnaires (covering demographic, social/emotional, and school-related variables expected to influence student performance).

The study's database contains the following variables: the student's school, gender, age, home address (rural or urban), parents' cohabitation status (Pstatus), the mother's and father's degree and occupation (Medu, Fedu, Mjob, and Fjob), the reason the student chose this particular school (e.g., proximity to home, school reputation, course preference, or other), the student's guardian (mother, father, or other), the number of previous academic failures, extracurricular activities, paid instruction, attendance at nursery school, desire for further education, desire for further education, home Internet connection, romantic relationship, and family quality Any displayed input variable may be binary, numeric, or nominal.

G3 and the number of school absences (absences) were selected as model outputs. $G3$ is the final grade of students obtained from school reports, with values between zero (the lowest grade) and 20 (the highest grade). Finally, by classifying reported grades, students were divided into four categories: Poor ($G3$ of $0-12$), Acceptable ($G3$ of $12-14$), Good ($G3$ of $14-16$), and Excellent ($G3$ of $16-20$).

The correlation matrix for each new input and output variable is shown in Fig. 1. The parents' education had the highest positive effect on the grade obtained by the student, while the father's job was not as effective as the mother's. As expected, study time positively affected outcomes, and the influence of previous failures of the student was revealed to be highly negative. The positive impact of internet accessibility and students' willingness to continue higher education and, in contrast, the negative consequences of alcohol consumption were noticeable. Also, the main variables influencing the number of absences from school were age, failures, and the amount of alcohol consumed on a daily and weekly basis.

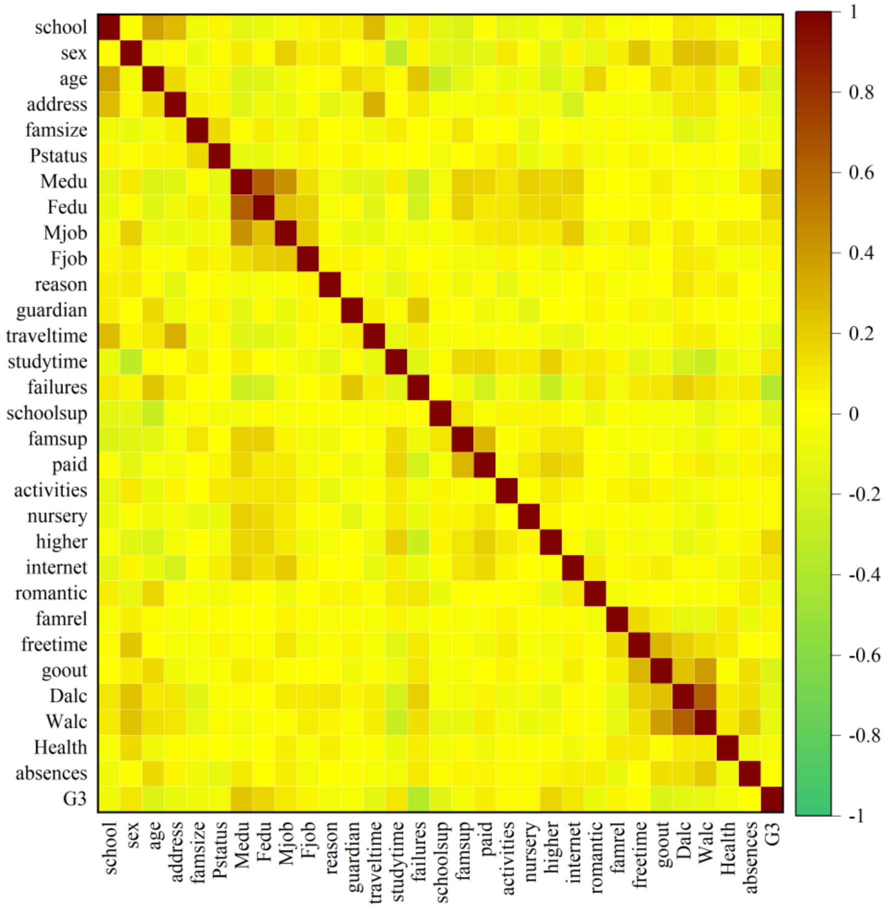

Fig. 1. Correlation matrix for the input and output variables.

IV. CatBoost Classifier (CATC)

A proficient ML algorithm for forecasting categorical attributes is the CatBoost classifier. CatBoost employs gradient boosting and utilizes binary DT as foundational predictors [43]. Let's consider a dataset comprising samples $D = \{(X_j, y_j)\},\ j = 1, \dots, m$, where in $X_j = \left(x_j^1, x_j^2, \dots, x_j^n\right)$, $y_j \in \mathrm{R}$, the response feature, and a vector of n characteristics. The answer feature might be numerical ($0\ or\ 1$) or binary (yes or no). The samples (X_j, y_j) have the same independent distribution according to some unknown distribution $p(.,.)$. To train a function $H : \mathrm{R}^n \rightarrow \mathrm{R}$ that lowers the predicted loss as given by Eq. (1) is the aim of the knowledge problem.

$X_j = \left(x_j^1, x_j^2, \dots, x_j^n\right)$ denotes a vector of n characteristics and the response feature $y_j \in \mathrm{R}$, which can be expressed as a numerical feature ($0\ or\ 1$) or as binary (i.e., yes or no). The samples (X_j, y_j) have the same independent distribution according to some unknown distribution $p(.,.)$. To train a function $H : \mathrm{R}^n \rightarrow \mathrm{R}$ that lowers the predicted loss as given by Eq. (1) is the aim of the knowledge problem.

$$\mathcal{L}(H) := \mathbb{E}L(y, H(X)) \tag{1}$$

Where (X, y) indicates testing data selected from training data D, and $L(.,.)$ is a smooth loss function.

The gradient boosting procedure [44] incrementally builds a series of approximations $H^t : R^m \rightarrow R, t = 0, 1, ..$ in a greedy manner. Starting from the previous approximation H^{t-1}, each new approximation H^t is obtained through an additive process, where $H^t = H^{t-1} + \alpha g^t$. Here, α represents the step size and the function $g^t: R^n \rightarrow R$, which serves as a base predictor, is chosen from a set of functions G to minimize or reduce the expected loss defined in Eq. (2):

$$\begin{aligned} g^t &= \arg\min_{g \in G} \mathcal{L}(H^{t-1} + g^t) \\ &= \arg\min_{g \in G} \mathbb{E}L(y, H^{t-1}(X) + g(X)) \end{aligned} \tag{2}$$

Frequently, the Newton approach is used for the minimization issue, using a second-order approximation of $\mathcal{L}(H^{t-1} + g^t)$ at H^{t-1}, or by taking a (negative) gradient step. Both approaches, Newton's method and gradient descent, are utilized [45,46]. For additional details on the CatBoost algorithm, please refer to [43].

V. META-HEURISTIC OPTIMIZATION ALGORITHMS

A. Victoria Amazonica Optimization

The distribution of the initial population, which is made up of two parts—Leaves and Flowers—and their corresponding capacities to spread over the surface are the main foci of the VAO algorithm [47]. This algorithm predominantly functions as a metaheuristic and uses swarm local search techniques. Its main drawback lies in the potential of getting trapped in local optima. However, it is noteworthy for its swiftness and resilience in handling a wide range of optimization tasks. In the context of this research, the scientific representation of diameter ⌀ is employed to illustrate how entities grow circularly. This growth includes their ability to occupy space, which is achieved through the forceful displacement of one another as they gain strength and spines. This competitive interaction is called intra-competition or denoted as λ for formulation.

Three common challenges affecting plant growth are beetle mortality, inadequate pollination, and temperature drops, collectively referred to as ω. A higher ω value indicates weaker plant growth. Plant pests, like water lily Aphids, represented as μ, can damage leaves. A lower μ value implies better situations for a plant's growing.

Lastly, the mutation occurs when pond beetles cross-pollinate a water lily flower with a different type, Hybrid Mutation, represented by the symbol ρ. As mentioned earlier, this mutation can result in negative and positive changes, each with a 0.2% frequency per generation. The healthiest and most robust leaf is identified as the optimal or α. The VAO method's flowchart is presented in Fig. 2.

$$VAO = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_{ij}[\varnothing_{ij}, \lambda_{ij}] + \mu + \omega) * (\rho) \tag{3}$$

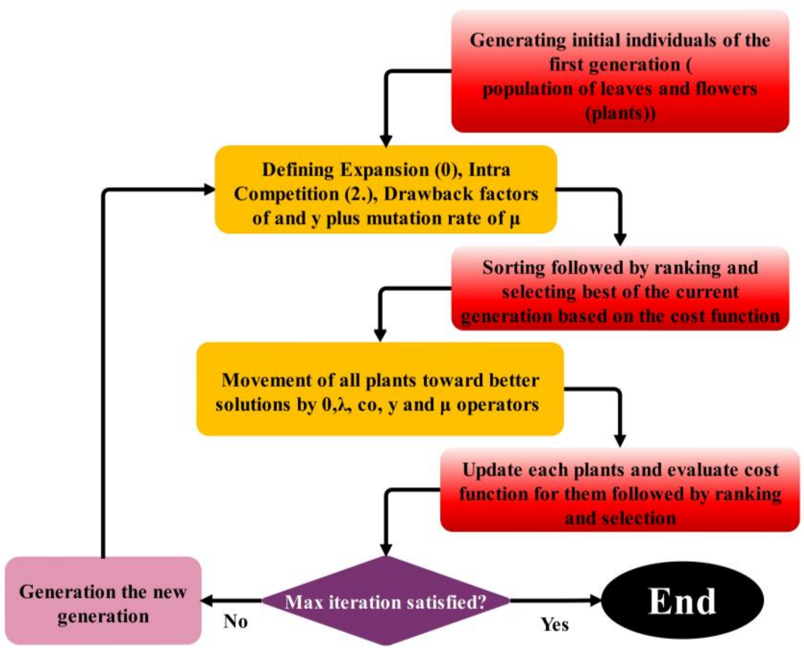

Fig. 2. Flowchart of the proposed VAO.

B. Artificial Rabbits Optimization

The ARO idea is derived from the survival techniques used by rabbits in their natural environment, which are designed to confuse predators and guarantee their ability to avoid being tracked. ARO encompasses integrating rabbits' approaches related to foraging, concealing, and managing energy resources, allowing for a seamless transition between these strategic behaviors [48].

1) Detour foraging: Rabbits use a detour foraging strategy when they forage for food, focusing on far-off food sources and often ignoring closer ones. Imagine an environment where a number of rabbits, each with its region complete with burrows and grass, are contained inside the ARO framework. These bunnies often stumble into one other's foraging spots at random. The mathematical model to describe the deviation search behavior of rabbits is as follows:

$$\vec{B}_i(t+1) = x_j(t) + S \times (x_i(t) - x_j(t)) + w(0.5 \times (0.05 + r_1)) \times m_1, \quad i, j = 1, \dots, n \text{ and } j \neq 1 \tag{4}$$

$$S = M \times v \tag{5}$$

$$M = (e - e^{\left(\frac{t-1}{I}\right)^2}) \times \sin(2\pi r_2) \tag{6}$$

$$v(y) = \begin{cases} 1 & if \quad y = f(1) \\ 0 & else \end{cases} \quad k = 1, \dots, d \text{ and } l = 1, \dots, \lceil r_3 \times d \rceil \tag{7}$$

$$f = p(d) \tag{8}$$

$$m_1 = N(0,1) \tag{9}$$

In all equations above:

n: The population's total number of rabbits

d: The scope of the issue

$\vec{B}_i(t+1)$: The $i-th$ rabbit's location at time $t + 1$,

n_1: Based on the conventional normal distribution distribution model.

T: The maximum number of iterations,

$x_i(t)$: The position of the $i-th$ rabbit at time t.

p: produces a random permutation, or rearrangement, of numbers between 1 and d.

w: An algorithmic mapping tool that makes it easier to choose components at random from the explorer to provide diversity to the search procedure.

r_1, r_2, and r_3: Random values in the interval $[0, 1)$.

S: During detour foraging, the run length indicates the pace of movement.

2) Random hiding: To enhance their survival chances, rabbits are likely to select one of their caves at random as a shelter. The mathematical model that represents this stochastic shelter-seeking behavior is expressed through the following equations. The $j-th$ burrow of the $i-th$ rabbit's formulation is as follows:

$$\vec{B}_i(t+1) = x_i(t) + N \times f \times \vec{x}_i(t), \ \ i,j = 1, \dots, n \ and \ j \neq 1 \tag{10}$$

$$D = \frac{I-t+1}{I} \times r_4 \tag{11}$$

$$m_2 = N(0,1) \tag{12}$$

$$f(y) = \begin{cases} 1 & if \quad y = g(1) \\ 0 & else \end{cases} \quad k = 1, \dots, d \tag{13}$$

$$\vec{R}_{i,r}(t) = \vec{x}_i(t) + \ N \times f \times \vec{x}_i(t) \tag{14}$$

The parameter N, which represents the hiding capability, gradually decreases linearly during the iteration process, starting at 1 and decreasing to $1/I$, with the addition of random perturbations.

Finally, whether the random hiding or detour foraging tactics are used, the update of the $i-th$ rabbit's position follows the formula provided in Eq. (15):

$$\vec{x}_i(t+1) = \begin{cases} \vec{x}_i(t) & g(\vec{x}_i(t)) \le g\left(\vec{B}_i(t+1)\right) \\ \vec{B}_i(t+1) & g(\vec{x}_i(t)) > g\left(\vec{B}_i(t+1)\right) \end{cases} \tag{15}$$

3) Energy shrink: The rabbits' energy levels steadily decline as a result of their frequent cycles of haphazard concealment and diversionary foraging. Consequently, an energy component must be included in the ARO framework:

$$E(t) = 4\left(1 - \frac{t}{I}\right) ln \frac{1}{r} \tag{16}$$

Fig. 3 displays the ARO flowchart.

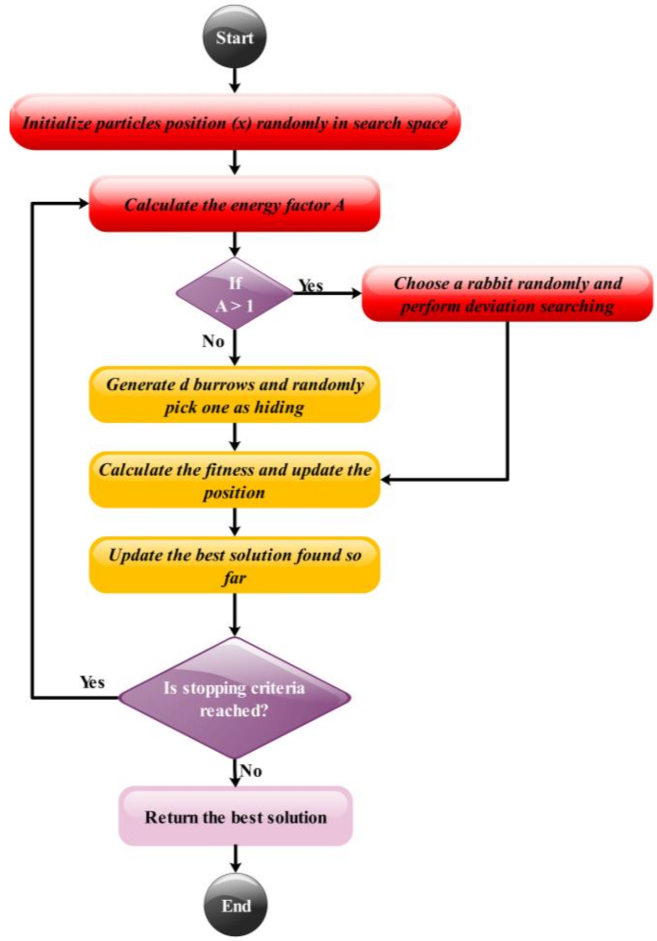

Fig. 3. Flowchart of ARO.

VI. Evaluation Metrics

The most frequently employed metric in a classification problem like the one addressed in this study is Accuracy. In defining the Accuracy, True positives, or TPs, are situations in which the model's predictions came true. Instances that were also accurately anticipated are known as true negatives (TN). False negatives (FN) indicate cases that were incorrectly predicted, while false positives (FP) indicate cases that were incorrectly anticipated.

Nevertheless, five other metrics (Precision, F1-score, Recall, MCC, and AUC) have been chosen because the Accuracy metric has limitations and may not accurately reflect the situation when dealing with imbalanced data, because it usually benefits the dominant class. Precision measures how well positive predictions work, which is important for reducing false positives, while recall shows how well a model can locate all relevant instances within a class. Furthermore, by taking into account both the minority and majority classes, the F1-score enables us to evaluate and correct for uneven data [49]. Evaluation parameters are defined in Eq. (17) to Eq. (21):

$$Precision = \frac{TP}{TP+FP} \tag{17}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{18}$$

$$F1\_score = \frac{2\times Recall \times Precision}{Recall+Precision} \tag{19}$$

$$Recall = TPR = \frac{TP}{P} = \frac{TP}{TP+FN} \tag{20}$$

$$MCC = \frac{(TP*TN)-(FP+FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{21}$$

VII. Convergence Analysis

VAO and ARO are two distinct metaheuristic optimization algorithms that have shown promise in enhancing the performance of ML models. This study applied them to optimize CATC development CAVA and CAAR hybrid models. These optimizers work to fine-tune the model's hyperparameters and improve its predictive accuracy.

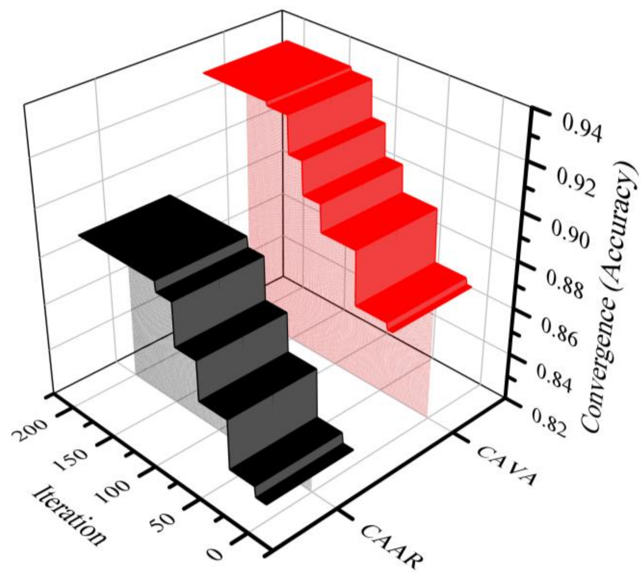

Fig. 4. Convergence curve of hybrid models.

An efficient method for assessing the convergence of these optimizers is by employing a convergence curve depicted in Fig. 4, which is based on the measure of accuracy through 200 iterations. This curve provides a graphic illustration of how the model's accuracy evolves with each iteration, enabling us to determine whether the optimizer is progressing toward an optimal solution and with which rate this convergence is occurring. As evident in Fig. 4, CAVA and CAAR have similar convergence rates, but CAVA starts its operation with about 5% higher accuracy than CAAR, and it reaches a better ultimate value. It is important to highlight that in both models, the trend line exhibited a linear pattern at around 120 iterations, indicating that this point represents the optimal level of computational efficiency.

VIII. Results

This project incorporates a wide range of student data, with a focus on their final grades ($G3$), in an effort to predict future academic success in language courses by using ML techniques. The three models—CATC, CAAR, and CAVA—that are based on the CATBoost Classifier (CATC) are trained and evaluated in large part using this dataset. This section contains the study's methodical computation of performance measures for each prediction step, including *Accuracy*, *Recall*, *Precision*, *F1-score*, MCC, and AUC. With the goal of identifying the best prediction model, this careful investigation provides insightful information that may be used to improve students' academic performance. Every pertinent measure value for testing, training, and model performance is listed in Table II and shown in Fig. 5. When it comes to the G3 prediction results, $CAVA$ and $CATC$ have the best and lowest prediction performances, respectively, with maximum and minimum accuracy scores of 0.9449 and 0.8744. The highest values that $CAVA$ was able to obtain were $0.9453, 0.9449, 0.9449, 0.9192$, and 0.944 for Precision, Recall, F1-score, and AUC, respectively. These results demonstrate the excellent accuracy of CAVA's exact predictions. Conversely, the performance of the other hybrid model (CAAR) was lower than that of CAVA in the prediction processes, experiencing weaker performance across all metric values.

TABLE II. Outcomes of the Models Presented

Model	*Phase*	*Index values*					
		Accuracy	*Precision*	*Recall*	*F1 _score*	*MCC*	*AUC*
CATC	*Train*	0.8744	0.8786	0.8744	0.8759	0.8183	0.909
	Test	0.8872	0.8892	0.8872	0.8871	0.8350	
	All	0.8783	0.8800	0.8800	0.8800	0.8230	
CAAR	*Train*	0.9053	0.9066	0.9053	0.9045	0.8603	0.904
	Test	0.8564	0.8548	0.8564	0.8519	0.7876	
	All	0.8906	0.8900	0.8900	0.8900	0.8384	
CAVA	*Train*	0.9449	0.9453	0.9449	0.9449	0.9192	0.944
	Test	0.9077	0.9065	0.9077	0.9063	0.8641	
	All	0.9337	0.9300	0.9300	0.9300	0.9026	

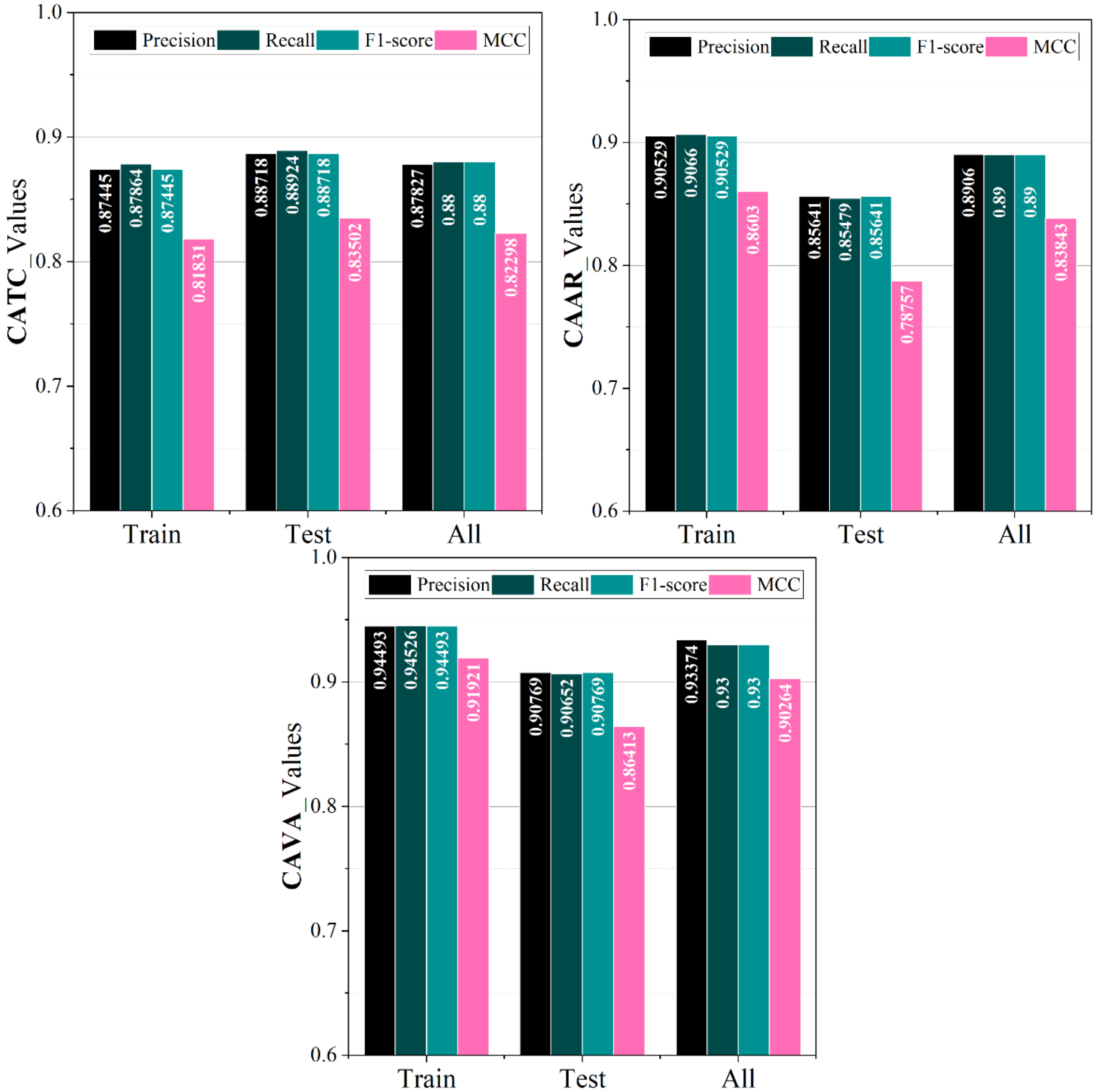

Fig. 5. Bar charts for evaluation results related to hybrid models.

After data processing and examining the classification capability of models in both training and testing phases, to discuss in detail, a total of 649 students based on test results (G3 values) divided into four categories: Poor (G3 of 0–12), Acceptable (G3 of 12–14), Good (G3 of 14–20), and Excellent (G3 of 16–20). Based on categorizing results, 82, 112, 154, and 301 students were located in Excellent, Good, Acceptable, and Poor classes. It revealed that most students (46.38%) performed poorly, while 23.73%, 17.26%, and 12.63% achieved acceptable, good, and excellent educational performance, correspondingly. Table III shows the accuracy, recall, and F1-score index values to assess how well the constructed model's categorization performance across various student groups. Each of the three Index values has been taken into consideration in the comparative study that follows:

1) Precision: A thorough examination of two optimized models presented that, when categorizing students across various categories, the CAVA model exhibited the highest level of precision in all cases, except for the Excellent grade category, where the CAAR model achieved a max *Precision* value of 0.99. For students classified as Poor and Excellent, the CAVA model demonstrated a Precision value of 0.96 for both categories. Notably, the classification performance of all three models was least precise when dealing with students in the Good grade category, with the lowest precision values observed as 0.78, 0.83, and 0.88 for CATC, CAAR, and CAVA, respectively.

2) Recall: Considering recall values, CAVA performed better in identifying all relevant instances within a class with 0.89, 0.88, 0.9, and 0.98 of the Recall for Excellent, Good, Acceptable, and Poor categories, respectively. Of course, there was an exception in the case of poorly graded students, where CAAR with a marginally higher Recall value was better than CAVA. Similar to those obtained in Precision values comparison, all models performed poorly in classifying students in the Good grade category.

3) F1-score: Compared to the Precision and Recall, the F1-score offers a more comprehensive and nuanced basis for comparative analysis. This metric is bounded between 0 and 1, with higher values signifying superior model performance. A higher F1 score indicates that the model achieves a balance between recalling all true positive instances and precisely recognizing positive cases (precision and recall). The $CAVA$ shows itself to be the most accurate when taking into account all student categories, producing F1-scores of $0.92, 0.88, 0.9$, and 0.97 for students rated as *Excellent*, *Good*, *Acceptable*, and *Poor*. In the second position, concerning the classification of Poor and Excellent students, CAAR displayed greater accuracy than the individual model, whereas their performance was identical for Good students. For Acceptable students, CATC outperformed CAAR.

4) MCC: The Matthews Correlation Coefficient (MCC) results show that $CAVA$ performed well in finding all relevant occurrences in each class. *Excellent*, *Good*, *Acceptable*, and *Poor* categories received scores of 0.95, 0.87, 0.85, and 0.91, respectively. This indicates a high degree of accuracy in predicting student performance across various grade levels. It's noteworthy to mention that, akin to the findings in the comparison of Precision values, all models exhibited weaker performance in accurately classifying students within the Good grade category.

In general, Table III shows the results of the developed models in detail.

Fig. 6 provides chances for visual comparison by displaying the frequency of students in each category based on metrics and conclusions from the categorization model. It is evident that the students who fell into the categories of bad, acceptable, good, and exceptional were, in fact, $301, 154, 112$, and 82. The CAVA model demonstrated the max accuracy in correctly identifying the categorization of students across different categories, with one exception in the Poor category, where CAAR classified a higher number of students correctly (298 students). In contrast, considering all other grades of students, CAAR was the weakest classifier, especially in the case of Good grades, where only 77.68% of students classified correctly.

The confusion matrix in Fig. 7 demonstrates the number of students accurately assigned to their respective grades and those misclassified into incorrect categories. Considering CAVA, 73, 99, 138, 196 (cumulative number of 606), students were categorized correctly in Excellent, Good, Acceptable, and Poor classes, and just 49 were in the wrong grade. In contrast, the number of students whom CAAR and CATC misclassified was 71 and 79. For two optimized models, misclassification occurred mostly between neighborhood categories, for instance, 9 and 15 students in the case of CAVA and CAAR instead of coming in the $Excellent$ category positioned in the $Good$ grade category. On the other hand, in the instance of the single CATC model, seven students with a minimum score difference of four points in their G3 scores were wrongly placed in the Poor group instead of the Excellent category. Overall, CAVA outperformed 2 other models capable of predicting students' academic performance in the future more precisely.

TABLE III. GRADE-BASED PERFORMANCE EVALUATION INDICES FOR THE CREATED MODELS

Model	***Grade***	***Index values***			
		Precision	***Recall***	$\boldsymbol{F1-score}$	***MCC***
CATC	*Excellent*	0.85	0.86	0.85	0.88
	Good	0.83	0.83	0.83	0.81
	Acceptable	0.78	0.83	0.80	0.76
	Poor	0.95	0.92	0.93	0.80
CAAR	*Excellent*	0.82	0.82	0.82	0.92
	Good	0.99	0.80	0.89	0.77
	Acceptable	0.83	0.78	0.80	0.76
	Poor	0.92	0.99	0.96	0.88
CAVA	*Excellent*	0.91	0.90	0.90	0.95
	Good	0.96	0.89	0.92	0.87
	Acceptable	0.88	0.88	0.88	0.85
	Poor	0.96	0.98	0.97	0.91

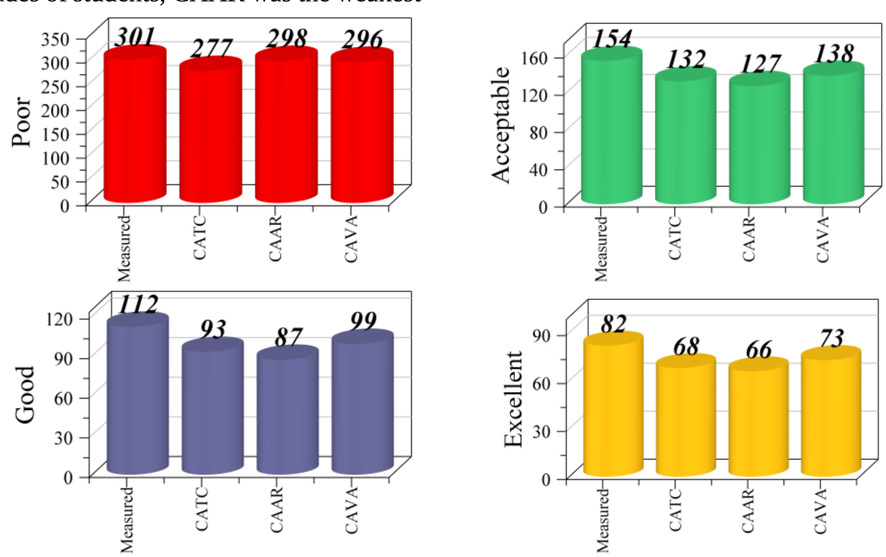

Fig. 6. Based on measurements and the results of categorization models, the number of pupils in each category.

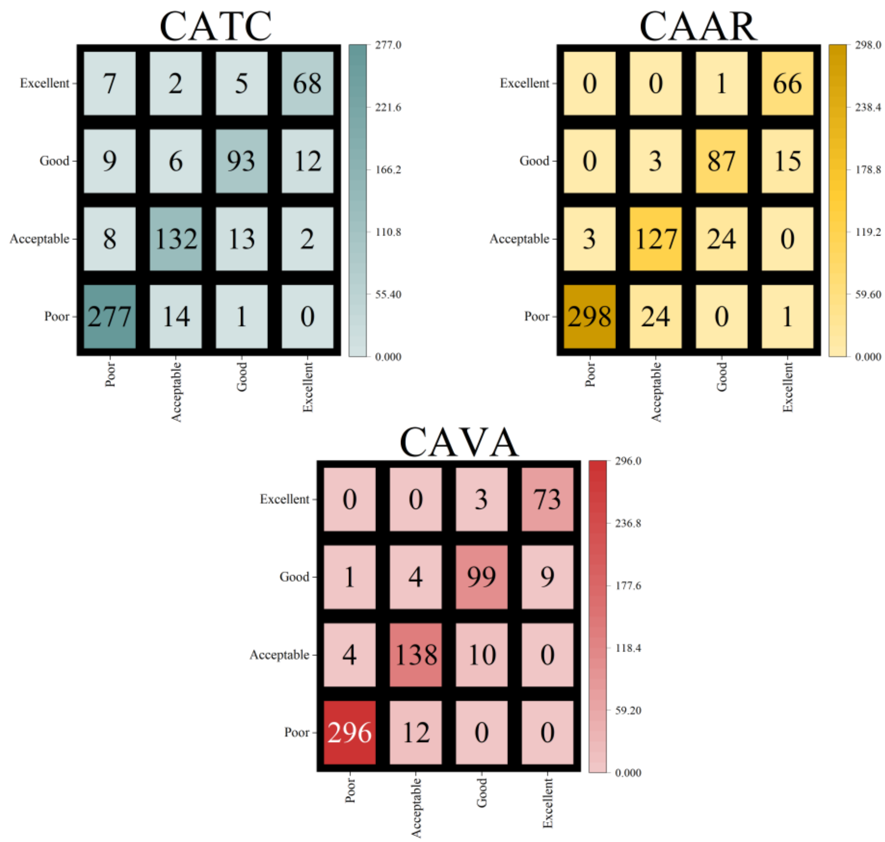

Fig. 7. Confusion matrix showing the accuracy of each model's categorization.

The analysis makes use of the Receiver Operating Characteristic (ROC) curve to achieve equilibrium between True Positive (TP) and False Positive (FP) rates, gauged by the Area Under the ROC Curve (AUC). A higher AUC signifies better control over the FP rate relative to the TP rate. An ideal discriminatory test is marked by an ROC plot converging towards the upper-left corner, indicative of 100% sensitivity and specificity. As depicted in Fig. 8, which presents ROC curves for the CATC, CAAR, and CAVA models in G3 score classification, it is evident that the AUC for the CAVA model (0.944) surpasses that of other categories, underscoring its robust discriminatory capability. The discernible inclination of the curve towards the upper-left corner underscores the model's effectiveness in distinguishing between various classes with precision.

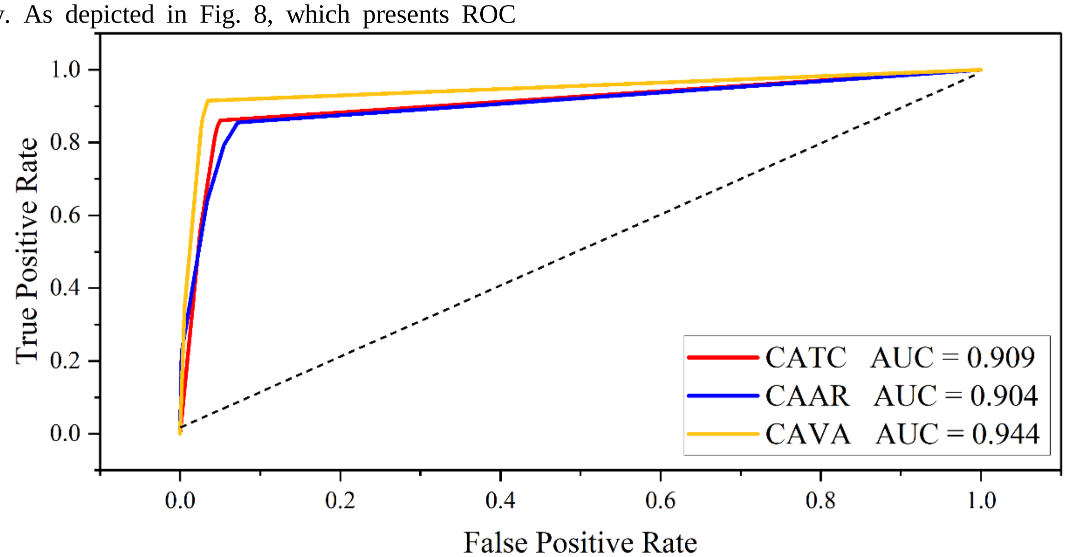

Fig. 8. The result of the ROC curve.

IX. Discussion

Since a particular dataset was used for the study, it is acknowledged that the findings might not apply to different educational settings. Nevertheless, the strategy proposed, which combines the CATC model with optimization algorithms, is believed to have the potential for generalization to other settings under specific conditions:

1) *Sufficient and representative data:* The dataset needs to be sufficiently big and representative of the intended audience. It must have pertinent characteristics that may record the elements impacting pupils' academic achievement. Additionally, rigorous preprocessing and cleaning are essential to guarantee data quality and validity.

2) *Tuning of optimization algorithms:* The tuning and adaptation of the optimization algorithms to the properties of the data is required. This involves adjusting parameters and initial conditions to optimize the objective function effectively.

While optimization algorithms are powerful tools for enhancing predictive models, they also possess certain limitations and potential drawbacks:

1) *Dependence on data quality and quantity:* Optimization algorithms rely on the quality and quantity of data for learning and optimizing the objective function. Issues like overfitting, underfitting, or bias in the optimization findings may arise from insufficient, erroneous, or unrepresentative data.

2) *Computational resources:* Optimization algorithms' iterative procedures sometimes need large amounts of memory and processing power, especially when dealing with high-dimensional and nonlinear issues. Due to this, they might not be as helpful or successful in real-world scenarios where time and distance are crucial factors.

3) *Sensitivity to parameters and initial conditions:* Optimization algorithms typically involve specifying or tuning multiple parameters and initial conditions. Selecting these factors can have a big influence on the optimization solutions' quality, stability, and convergence, making it difficult to adapt to other situations or datasets.

4) *Lack of guarantees and robustness:* Despite their effectiveness, optimization algorithms may lack guarantees and robustness in certain situations. Variability in convergence, stability, and quality of results may occur, making it challenging to ensure consistent performance across diverse contexts.

Awareness of these limitations is essential when employing optimization algorithms in practical applications, necessitating careful consideration and validation to mitigate potential drawbacks and optimize their utility effectively.

B. Comparison with Published Papers

Table IV shows that the CAVA model in the present study achieves an accuracy of 93.37%, significantly outperforming other models such as the DTC and NBC used in previous studies. The superior performance of the CAVA model is attributed to its advanced optimization techniques, which improve parameter tuning and feature selection. This highlights the potential of using sophisticated optimization algorithms to address challenges in educational DM, leading to better predictive performance. The findings suggest that robust models like CAVA can enhance decision-making and support systems in educational settings, ultimately improving student outcomes.

TABLE IV. Extensive Study Results Compared to the Current Work

Author (s)	*Models*	*Accuracy*
Present study	*CAVA*	93.37%
Kabakchieva [50]	*DTC*	72.74%
Bichkar and R.R.Kabra [35]	*DTC*	69.94%
Nguyen and Peter [51]	*DTC*	82%
Edin Osmanbegovic et al. [52]	*NBC*	76.65%

X. Conclusion

This inquiry is primarily concerned with the utilization of data-driven prediction models within educational settings, highlighting the critical integration of qualitative and quantitative components to forecast and assess students' academic success in language classes. Regression, classification, and clustering are 3 examples of DM techniques that show promise in addressing a variety of issues faced by undergraduate students. The knowledge gained from this study provides lawmakers, academic institutions, and students with important direction for improving future academic achievement. Additionally, the study introduces a pioneering approach by combining the VAO and ARO methods with the CATBoost classifier (CATC) model. This creative combination shows how ML methods and optimization algorithms may improve the accuracy and performance of prediction models. The resultant toolbox equips stakeholders to navigate the evolving complexities encountered throughout students' academic journeys. Through meticulous analysis, including model partitioning into training and testing sets, the study emphasizes how important it is for hybrid models to improve the CATC model's classification performance. Substantial improvements in Matthews Correlation Coefficient (MCC), Accuracy, and Precision attest to this progress. Detailed scrutiny of the data underscores the growing recognition of hybrid models for substantially refining the CATC model's categorization abilities. Particularly noteworthy is the exceptional performance of the VAO in boosting classification accuracy. Notably, the CAVA model demonstrates an impressive ability to accurately identify 93.37% of students, outperforming CAAR and CATC. Ultimately, this work propels predictive modeling in education forward, offering avenues to augment the precision and efficacy of academic performance evaluations. These findings underscore the favorable impact data-driven strategies can have on undergraduate students' academic trajectories. Future work in this field should focus on expanding the dataset to include a more diverse student population and a broader range of academic disciplines to validate the generalizability of the models. Additionally, exploring the integration of other advanced optimization algorithms and ML techniques could further enhance model performance. Investigating the real-time application of these models in educational settings and their impact on student interventions and support strategies would also be valuable.

DATA AVAILABILITY

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation."

CONFLICTS OF INTEREST

The authors declared that they have no conflicts of interest regarding this work."

FUNDING

This work supported by Humanities and Social Sciences Research Projects in Higher Education Institutions of Anhui Province: Research on the Quality Evaluation of Compulsory Education Curriculum Implementation in Anhui Province (2022AH050213).

REFERENCES

[1] Chiheb F, Boumahdi F, Bouarfa H, Boukraa D. Predicting students' performance using decision trees: Case of an Algerian University. 2017 International Conference on Mathematics and Information Technology (ICMIT), IEEE; 2017, p. 113–21.

[2] Varade R V, Thankanchan B. Academic performance prediction of undergraduate students using decision tree algorithm. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology 2021;13:97–100.

[3] Hamoud A, Hashim AS, Awadh WA. Predicting student performance in higher education institutions using decision tree analysis. International Journal of Interactive Multimedia and Artificial Intelligence 2018;5:26–31.

[4] Yang Y. The evaluation of online education course performance using a decision tree mining algorithm. Complexity 2021;2021:1–13.

[5] Liu X, Ding Y, Tang H, Xiao F. A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. Energy Build 2021;231:110601.

[6] Liz-Domínguez M, Caeiro-Rodríguez M, Llamas-Nistal M, Mikic-Fonte FA. Systematic literature review of predictive analysis tools in higher education. Applied Sciences 2019;9:5569.

[7] Jimenez F, Paoletti A, Sanchez G, Sciavicco G. Predicting the risk of academic dropout with temporal multi-objective optimization. IEEE Transactions on Learning Technologies 2019;12:225–36.

[8] Aluko RO, Adenuga OA, Kukoyi PO, Soyingbe AA, Oyedeji JO. Predicting the academic success of architecture students by pre-enrolment requirement: Using machine-learning techniques. Construction Economics and Building 2016;16:86–98.

[9] Mason C, Twomey J, Wright D, Whitman L. Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression. Res High Educ 2018;59:382–400.

[10] Adekitan AI, Salau O. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. Heliyon 2019;5.

[11] Tatar AE, Düştegör D. Prediction of academic performance at undergraduate graduation: Course grades or grade point average? Applied Sciences 2020;10:4967.

[12] Miguéis VL, Freitas A, Garcia PJ V, Silva A. Early segmentation of students according to their academic performance: A predictive modeling approach. Decis Support Syst 2018;115:36–51.

[13] Trussel JM, Burke-Smalley L. Demography, and student success: Early warning tools to drive intervention. Journal of Education for Business 2018;93:363–72.

[14] Batool S, Rashid J, Nisar MW, Kim J, Kwon H-Y, Hussain A. Educational data mining to predict students' academic performance: A survey study. Educ Inf Technol (Dordr) 2023;28:905–71.

[15] Trussel JM, Burke-Smalley L. Demography, and student success: Early warning tools to drive intervention. Journal of Education for Business 2018;93:363–72.

[16] Adekitan AI, Salau O. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. Heliyon 2019;5.

[17] Jimenez F, Paoletti A, Sanchez G, Sciavicco G. Predicting the risk of academic dropout with temporal multi-objective optimization. IEEE Transactions on Learning Technologies 2019;12:225–36.

[18] Mason C, Twomey J, Wright D, Whitman L. Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression. Res High Educ 2018;59:382–400.

[19] Hasib KM, Rahman F, Hasnat R, Alam MGR. A machine learning and explainable AI approach for predicting secondary school student performance. 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), IEEE; 2022, p. 399–405.

[20] Varade R V, Thankanchan B. Academic performance prediction of undergraduate students using decision tree algorithm. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology 2021;13:97–100.

[21] Tatar AE, Düştegör D. Prediction of academic performance at undergraduate graduation: Course grades or grade point average? Applied Sciences 2020;10:4967.

[22] Hasheminejad SMH, Sarvmili M. S3PSO: Students' performance prediction based on particle swarm optimization. Journal of AI and Data Mining 2019;7:77–96.

[23] Beaulac C, Rosenthal JS. Predicting university students' academic success and major using random forests. Res High Educ 2019;60:1048–64.

[24] Matzavela V, Alepis E. Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments. Computers and Education: Artificial Intelligence 2021;2:100035.

[25] Ghosh SK, Zoha N, Sarwar F. A generic MCDM model for supplier selection for multiple decision makers using fuzzy TOPSIS. Proceedings of the 5th International Conference on Engineering Research, Innovation and Education (ICERIE) Sylhet, Bangladesh, 2019, p. 833–40.

[26] Ghosh SK, Janan F, Ahmad I. Application of the Classification Algorithms on the Prediction of Student's Academic Performance. Trends in Sciences 2022;19:5070.

[27] Wiyono S, Wibowo DS, Hidayatullah MF, Dairoh D. Comparative study of KNN, SVM, and decision tree algorithm for student's performance prediction. (IJCSAM) International Journal of Computing Science and Applied Mathematics 2020;6:50–3.

[28] Matzavela V, Alepis E. Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments. Computers and Education: Artificial Intelligence 2021;2:100035.

[29] Sivakumar S, Selvaraj R. Predictive modeling of students performance through the enhanced decision tree. Advances in Electronics, Communication, and Computing: ETAEERE-2016, Springer; 2018, p. 21–36.

[30] Srivastava AK, Chaudhary A, Gautam A, Singh DP, Khan R. Prediction of students performance using KNN and decision tree-a machine learning approach. Strad 2020;7:119–25.

[31] Hamoud A, Hashim AS, Awadh WA. Predicting student performance in higher education institutions using decision tree analysis. International Journal of Interactive Multimedia and Artificial Intelligence 2018;5:26–31.

[32] Wang G-H, Zhang J, Fu G-S. Predicting student behaviors and performance in online learning using the decision tree. 2018 Seventh International Conference of Educational Innovation through Technology (EITT), IEEE; 2018, p. 214–9.

[33] Al-Radaideh QA, Al-Shawakfa EM, Al-Najjar MI. Mining student data using decision trees. International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan, vol. 1, 2006.

[34] Aher SB, Lobo L. Data mining in the educational system using weka. International conference on emerging technology trends (ICETT), vol. 3, 2011, p. 20–5.

[35] Kabra RR, Bichkar RS. Performance prediction of engineering students using decision trees. Int J Comput Appl 2011;36:8–12.

[36] Cortez P, Silva AMG. Using data mining to predict secondary school student performance 2008.

[37] Hasib KM, Rahman F, Hasnat R, Alam MGR. A machine learning and explainable AI approach for predicting secondary school student performance. 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), IEEE; 2022, p. 399–405.

[38] Asselman A, Khaldi M, Aammou S. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. Interactive Learning Environments 2023;31:3360–79.

[39] Shreem SS, Turabieh H, Al Azwari S, Baothman F. Enhanced binary genetic algorithm as a feature selection to predict student performance. Soft Comput 2022;26:1811–23.

[40] Sarwat S, Ullah N, Sadiq S, Saleem R, Umer M, Eshmawi A, et al. Predicting students' academic performance with conditional generative adversarial network and deep SVM. Sensors 2022;22:4834.

[41] Mehdi R, Nachouki M. A neuro-fuzzy model for predicting and analyzing student graduation performance in computing programs. Educ Inf Technol (Dordr) 2023;28:2455–84. https://doi.org/10.1007/s10639-022-11205-2.

[42] Blake C. UCI repository of machine learning databases. Http://Www Ics Uci Edu/$~$ Mlearn/MLRepository Html 1998.

[43] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 2018;31.

[44] Lu H, Karimireddy SP, Ponomareva N, Mirrokni V. Accelerating gradient boosting machines. International conference on artificial intelligence and statistics, PMLR; 2020, p. 516–26.

[45] Friedman J, Hastie T, Tibshirani R. Additive Logistic Regression: A Statistical View of Boosting. The Annals of Statistics 2000;28:337–407. https://doi.org/10.1214/aos/1016218223.

[46] Mason L, Baxter J, Bartlett P, Frean M. Boosting algorithms as gradient descent. Adv Neural Inf Process Syst 1999;12.

[47] Mousavi SMH. Victoria Amazonica Optimization (VAO): An Algorithm Inspired by the Giant Water Lily Plant. ArXiv Preprint ArXiv:230308070 2023.

[48] Wang L, Cao Q, Zhang Z, Mirjalili S, Zhao W. Artificial rabbits optimization: A new bio-inspired meta-heuristic algorithm for solving engineering optimization problems. Eng Appl Artif Intell 2022;114:105082.

[49] Powers DMW. Evaluation: from precision, recall, and F-measure to ROC, informedness, markedness, and correlation. ArXiv Preprint ArXiv:201016061 2020.

[50] Kabakchieva D. Student performance prediction by using data mining classification algorithms. International Journal of Computer Science and Management Research 2012;1:686–90.

[51] Nghe NT, Janecek P, Haddawy P. A comparative analysis of techniques for predicting academic performance. 2007 37th annual Frontiers in Education conference-global engineering: knowledge without borders, opportunities without passports, IEEE; 2007, p. T2G-7.

[52] Osmanbegovic E, Suljic M. Data mining approach for predicting student performance. Economic Review: Journal of Economics and Business 2012;10:3–12.