Educational Performance Prediction with Random Forest and Innovative Optimizers: A Data Mining Approach

Yanli Chen[1], Ke Jin[2]
School of Smart Health College, Chongqing College of Electronic Engineering, Chongqing, 400716, China[1]
College of Fine Arts and Design, Guangzhou University, Guangzhou, Guangdong, 510006, China[2]
College of Fine Arts and Design, Chengdu University, Chengdu, Sichuan, 610106, China[2]

***Abstract*—In the ever-evolving landscape of education, institutions grapple with the intricate task of evaluating individual capabilities and forecasting student performance. Providing timely guidance becomes pivotal, steering students toward specific areas for focused academic enhancement. Within the educational domain, the utilization of data mining emerges as a powerful tool, revealing latent patterns within vast datasets. This study adopts the Random Forest classifier (RFC) for predicting student performance, bolstered by the integration of two innovative optimizers—Victoria Amazonia Optimization (VAO) and Phasor Particle Swarm Optimizer (PPSO). A notable contribution of this research lies in the introduction of these novel optimizers to augment the model's accuracy, elevating the precision of predictions. Robust evaluation metrics, including Accuracy, Precision, Recall, and F1-score, meticulously gauge the model's effectiveness in this context. Remarkably, the results underscore the supremacy of RFC+VAO, showcasing exceptional values for Accuracy (0.934), Precision (0.940), Recall (0.930), and F1-score (0.930). This substantiates the significant contribution of integrating VAO into the Random Forest framework, promising substantial advancements in predictive analytics for educational institutions. The findings not only accentuate the efficacy of the proposed methodology but also herald a new era of precision and reliability in predicting student performance, thereby enriching the landscape of educational data analytics.**

Keywords*—Student performance; Random Forest Classification; victoria amazonia; phasor particle swarm

I. Introduction

Educational institutions, including schools, universities, and training centers, handle vast amounts of data originating from various sources like registration departments, exam centers, and virtual courses, as well as e-learning systems [1], [2]. Within this educational data lie valuable insights that, once uncovered, can significantly improve the effectiveness of the entire educational system [3]. Machine learning (ML) and statistical methods have been increasingly applied to develop intelligent educational systems [4], [5], [6]. These systems aid decision-makers in educational institutions in attaining a thorough grasp of their organization [7]. Forecasting students' performance presents a complex challenge, but doing so can enable lecturers and decision-makers to identify effective strategies for addressing students' underperformance [8].

Additionally, the prediction of students' ultimate examination scores through the consideration of diverse elements like quiz results, homework, and project achievements will offer a holistic evaluation of the student's educational competence [9]. Machine learning methods have proven to be effective when used on problems related to association rules, web mining, classification, clustering, and deep learning in the field of education [10]. Researchers in the education industry continue to be greatly inspired by complicated data, which motivates them to explore techniques such as clustering and classification in order to create very accurate instructional models [11], [12].

Data classification stands out as the most efficient method for conducting data mining research, relying on the classification of data through predictive attribute value [13]. Data quality, which can disrupt algorithms and lead to misclassification, impacting the model's performance, is central to the challenge of classification [14]. By using this predictor, educational institutions can identify underperforming students and offer support to help them attain higher grades, ultimately paving the way for a brighter future [15]. Several established prediction techniques encompass classification, regression, and density estimation [16]. In contemporary data science, aside from enhancing the accuracy of their results, it is now essential to have trust in and a comprehensive understanding of prediction models [17], [18].

It is imperative that strong machine learning technologies be developed so that teachers may make well-informed judgments to reduce the chance of student failure. The objective of this project is to construct a reliable model for forecasting student grades utilizing a dataset associated with student performance. This data can be categorized into personal details (e.g., parent status, family size, and family educational support), educational background (e.g., weekly study time, motivations for pursuing higher education, and extracurricular activities), and general information (e.g., home address and commute time to school).

II. Related Work

In the realm of educational institutions, a multitude of researchers have utilized statistical techniques and machine learning algorithms to predict student performance. In their study, Bharadwaj et al. [19] utilized data from a past student

database, incorporating variables such as student attendance, class participation, seminar involvement, and assignment scores to anticipate semester-end outcomes. Their findings indicated that decision tree analysis yielded the highest accuracy, followed by K-nearest neighbor (KNN) classification [20], whereas Bayesian classification systems displayed the lowest accuracy. Ogunde et al. [21] undertook the development of a system that utilizes the decision tree technique known as Iterative Dichotomiser (ID3) and input data to predict grades. As per the authors, their approach has the potential to be highly efficient in predicting students' ultimate graduation levels. Duzhin and Gustafsson [22] introduced a machine-learning technique to take into account students' prior knowledge. Their method is based on symbolic regression and utilizes historical university scores as non-experimental input data. This classification approach holds promise for assisting the Ministry of Education in improving student performance through early performance predictions. Naïve Bayes [23] exhibits characteristics of conditional independence, making it skilled at determining class conditional probabilities. In their work, Watkins et al. [24] unveiled an approach called SENSE (Student Performance Quantifier using Sentiment analysis) to improve the content of secondary school reports by utilizing natural language processing. Sentiment analysis [25] can have a significant role in influencing student performance.

Table I shows the limitations and proposed solutions of the mentioned literature.

TABLE I. LIMITATIONS AND PROPOSED SOLUTIONS OF MENTIONED WORKS OF LITERATURE

Study	Limitations	Proposed Solutions
Bharadwaj et al. [19]	Limited scope of variables, potential bias in data	Expand variable inclusion, employ diverse data sources
Ogunde et al. [21]	Dependency on the decision tree technique ID3	Explore alternative machine learning algorithms
Duzhin and Gustafsson [22]	Reliance on historical university scores as input data	Incorporate additional non-experimental input data sources
Watkins et al. [24]	Emphasis on secondary school reports, potential bias	Explore the integration of diverse data types and sources

These limitations underscore the need for a more comprehensive and diverse approach to predicting student performance. To overcome these gaps, the current research introduces substantial variations of RFC algorithms. This approach aims to address the limitations identified in prior studies by incorporating a broader set of variables, exploring alternative machine learning algorithms, and diversifying input data sources. By taking these proposed solutions into account, the present study strives to provide a more robust and nuanced prediction model for student performance in the specific context of secondary school education statistics. This acknowledgment and proposed strategy not only build upon the existing body of knowledge but also pave the way for a more comprehensive and effective approach to addressing the limitations identified in previous research.

Nevertheless, there have been limited attempts to apply classification algorithms within the context of secondary school education statistics. In this research, substantial variations of Random Forest Classification (RFC) classification algorithms have been included to assist educators and parents in predicting the performance of new students and improving next year's outcomes. Additionally, to ensure the utmost reliability in the results, both Victoria Amazonia Optimization (VAO) and Phasor Particle Swarm Optimizer (PPSO) techniques were integrated, leading to the attainment of promising outcomes.

In the subsequent sections, the manuscript navigates through the intricacies of the dataset and methodology, providing a comprehensive understanding. It details the dataset's source, size, and preprocessing steps, highlighting key variables chosen for analysis. The methodology section explains the utilization of the RFC and the integration of VAO and PPSO, introducing a distinctive dual-optimizer approach. The results section presents findings through tables or figures, accompanied by a thorough discussion of evaluation metrics, including Accuracy, Precision, Recall, and F1-score. The analysis extends to comparing different models or variations within the methodology and interpreting results in the context of research questions and existing literature. The conclusion synthesizes key findings, discusses practical implications for educational institutions, acknowledges study limitations, and suggests future research directions. Together, these sections contribute to a coherent narrative, guiding readers through the research process and providing valuable insights into predictive analytics in the context of education.

III. DATASET AND METHODOLOGY

A. Data Gathering

Within this research, a dataset pertaining to education was employed, encompassing 33 distinct attributes thoughtfully selected to provide a precise depiction of students' performance during their academic journey, considering their individual information and circumstances [26]. This dataset compilation was achieved by integrating data obtained from two questionnaire methods and the academic records of the students.

These attributes encompass various aspects related to students, including demographic factors like gender, age, school attended, and type of residence (address). Additionally, they encompass parental characteristics such as parents' cohabitation status ($Pstatus$), educational background, and occupation ($Medu$, $Mjob$, $Fedu$, $Fjob$). The student's guardian, household characteristics such as family size (famsize), the quality of family relationships (famrel), and other characteristics such as the reason for choosing the school (reason), the time it takes to commute to school ($traveltime$), the amount of time spent studying each week ($studytime$), previous academic setbacks (failures), involvement in extracurricular activities (activities), attendance in paid classes ($paidclass$), internet accessibility (internet), attendance in nursery school (school), ambitions for higher education (higher), romantic relationship status (romantic), free time availability after school ($freetime$), socializing preferences (go out), alcohol consumption during working days ($Dalc$) and

weekends (*Walc*), as well as the current health status of the individual (health), the reason for school choice (reason), participation in supplementary educational programs (*schoolsup*), family educational support (famsup). Together with this, 3 other features—Grade 2 (G2), Grade 1 (G1), and Final—display students' grades for each of their three educational assessment periods. The values range from zero, which represents the lowest grade, to twenty, which represents the greatest grade. G3 is the pupils' final grade. As model outputs (dependent variables), these three characteristics were chosen together with the absence number from school. In order to assign grades, the students were split into four groups: 0-12: Subpar; 12-14: Tolerable; 14–16: Good; and 16–20: Outstanding.

In Fig. 1, as anticipated, the cells along the central axis appear in red, indicating a correlation value of 1. The three characteristics, G1, G2, and final, which are all dependent variables and correlate to students' grades, show the highest correlation values among themselves, as seen in the previously mentioned figure.

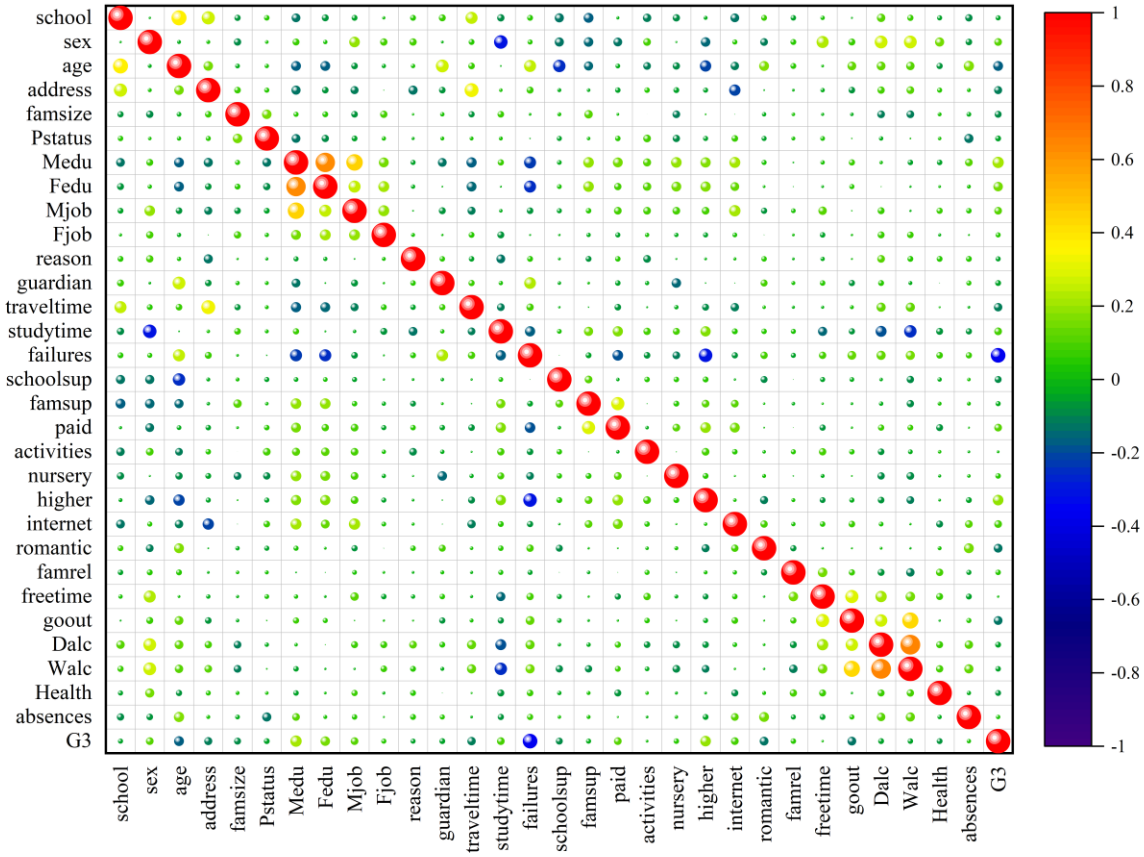

Fig. 1. Correlation matrix for the input and output variables.

B. Random Forest Classifier (RFC)

Breiman's suggested random forest model [27] is composed of a collection of tree predictors. Each tree is constructed following the procedure below:

1) In the bootstrap phase, a local training set is created by randomly selecting a subset from the training dataset [28]. The remaining samples in the training dataset are designated as the out-of-bag (OOB) set, and they serve the purpose of evaluating the goodness-of-fit of the random forest model.

2) In the expansion phase, the tree's growth involves partitioning the local training set at each node based on a single variable's value. This variable is selected from a randomly sampled subset of variables, and the division, known as the optimal split, is determined using the Classification and Regression Tree (CART) method.

3) Every tree is allowed to grow to its maximum extent, with no pruning being employed.

The bootstrap and growth phases make use of random variables [29]. It is assumed that these variables are independent across different trees and follow an identical distribution. Consequently, each tree can be considered an independent sample drawn from the entire ensemble of tree predictors for a specific training dataset. During the prediction phase, an instance is processed through each tree within the forest until it reaches a terminal node that assigns it a class. The predictions from the trees are then subjected to a voting procedure, where the forest selects the class that receives the highest number of votes. In cases of ties, the final decision is made through a random selection. To introduce the feature contribution method in the upcoming section, a probabilistic interpretation of the prediction process in the forest needs to be

established. The collection of classes is denoted as C = $\{C_1, C_2, ..., C_K\}$, and the set Δ_k is used to represent.

$$\Delta_k = \{(P_1, \dots, P_K): \sum_{K=1}^{K} P_k = 1 \; and \; P_k \geq 0\} \quad (1)$$

An element in the set Δ_k can be seen as a probability distribution that covers the classes in C. Consider, for example, ek, an element in Δ_k, where its value is 1 at position k, indicating that it is a probability distribution focused on class Ck. When a tree labeled as t predicts that an instance i pertains to class Ck, express this as $\widehat{Y_{\iota,t}} = e_k$. This forges a link between the tree's predictions and the set Δ_k, which denotes probability distributions across C.

$$\widehat{Y_k} = \frac{1}{T}\sum_{t=1}^{T} \widehat{Y_{\iota,t}} \quad (2)$$

Within this context, with T denoting the overall number of trees in the forest, the predicted value $\widehat{Y_k}$ falls within the set Δ_k. The random forest's prediction, for instance, i aligns with class Ck when the $k-th$ coordinate of $\widehat{Y_\iota}$ is the most substantial.

C. Victoria Amazonica Optimization (VAO)

The VAO approach is primarily preoccupied with the dispersal of the initial populace, comprising both Leaves and Flowers and their respective potential to propagate or expand across the external façade [30]. The algorithm being examined is mainly characterized as a metaheuristic algorithm based on swarm local search. However, its sole drawback is its susceptibility to getting stuck in local optima. Moreover, it demonstrates exceptional speed and robustness, rendering it extremely well-suited for a wide spectrum of optimization challenges. The present study utilizes the scientific nomenclature, ξ, to depict the circular expansion of the entity's diameter as it grows circularly. The augmentation, as mentioned earlier, is succeeded by the quantum of the geographical area that they could potentially acquire through the exertion of physical force on fellow entities, driven by their augmenting potency and thorny projections. The aforementioned competition is commonly known by its designations of intra-competition or Γ for the formulation.

Furthermore, there exist three commonly encountered obstacles that impede the growth of vegetation. The mortality of beetles within the floral structure, inadequate or absent pollination by beetle species, and a reduction in ambient temperature are factors that contribute to suboptimal reproductive success in plants. All of the constituents mentioned above can exert negative effects on the given procedure, and collectively, they are denoted as φ herein. A higher value of the parameter ω corresponds to a plant with less vigor. Pests, such as water lily Aphids, have the potential to inflict damage upon the plant by feeding on its leaves and resulting in the formation of perforations. The symbol denoted by Θ is deemed representative of the hazard quotient in the present exposition. The conditions for plant growth and expansion become increasingly favorable as the value of Θ decreases.

Subsequently, the occurrence of mutation arises as a result of cross-pollination between the beetles within the pond and a distinct variety of water lilies. The present phenomenon is denoted as Hybrid Mutation and is symbolized by the η. As posited in [30], this alteration has the potential to manifest in either a positive or negative trajectory, with an incidence of 0.2% for each succession of offspring. The optimal leaf specimen can be delineated by its superior size and robust physical attributes, designated as α. Moreover, the VAO algorithm is delineated below in the pseudo-code form.

$$VOA = \sum_{i=1}^{n}\sum_{j=1}^{n}(xij[\xi ij, \Gamma ij] + \Theta + \varphi) \times (\eta) \quad (3)$$

Algorithm 1 pseudo code of VOA
Start
Developing population of plants xi $(i = 1,2, \dots, n)$
Determine Expansion ξi in xi
Determine Intra Competition Γi in xi
Determine the Drawback coefficient of φ in xi (random range in [0.1 to 0.3])
Determine the Drawback coefficient of Θ in xi (random range in [0.1 to 0.3])
Determine Hybrid Mutation Rate of $\eta = 0.2$
While Max iterations are not satisfied
For $i = 1$ to n plants
For $j = 1$ to n plants
If $\xi i > \xi j$ or Γi > Γj for xi $(i = 1,2, \dots, n)$
Plant i goes planet j
End if
Apply hybrid mutation η
Apply Drawback coefficient φ and Θ
Evaluate new solutions by cost function and update expansion
End
End
Sort and rank plants and find the current global best
Developing new generation
End of while
End

D. Phasor Particle Swarm Optimization (PPSO)

1) The parameter's setting: In consideration of the enhanced PSO algorithms utilized in prior research, the regulation and guidance of a system or process can be achieved through the implementation of appropriate control methods. A range of strategies must be included in the PSO parameters in order to properly optimize a specific issue. The objective of this work is to improve the efficiency of optimization in order to increase the convergence capabilities [31]. The PPSO generates PSO control parameters by using suitable and efficient phasor angle functions to achieve the aforementioned goals. To effectively implement a range of strategies in PPSO, an individual scalar phasor angle is assigned to each particle. These phasor angles are used to describe the PSO control parameters using mathematical functions that include both cos and sin. $\overrightarrow{X_\iota}\angle\theta_i$, where θ_i is the phasor angle and $(\overrightarrow{X_\iota})$ is the magnitude vector used to represent the ith particle as an example.

The PSO-TVAC in [32] and a contemporary PSO-TVAC [33] are similar in that their inertia weight values are zero. Below is an outline of the suggested particle movement model

for PPSO. Still, this technique may be improved by combining ideas from other enhanced PSO techniques.

$$V_i^{it} = p\left(\theta_i^{it}\right) \times \left(pbest_i^{it} - x_i^{it}\right) + g(\theta_i^{it}) \times (Gbest_i^{it} - x_i^{it}) \quad (4)$$

After examining several $g\left(\theta_i^{it}\right)$ and $p\left(\theta_i^{it}\right)$ functions, the PPSO algorithm selected the following functions.

$$p\left(\theta_i^{it}\right) = \left|cos\theta_i^{it}\right|^{2\times sin\theta_i^{it}} \quad (5)$$

$$g\left(\theta_i^{it}\right) = \left|sin\theta_i^{it}\right|^{2\times cos\theta_i^{it}} \quad (6)$$

The proposed functions, which depend solely on the phasor angles of the particles, can enable behaviors such as reversal of values, simultaneous increase or decrease of values, reaching of large values, and attainment of identical values. The aforementioned behaviors give rise to adaptive search traits, promoting a balance between local and global searches. Consequently, PPSO is an adaptive and non-parametric algorithm that excels at evading local optima and circumventing premature convergence, a shortcoming often associated with the PSO.

2) Formulation of PSO: The velocity of individual particles is computed in every iteration of the algorithm utilizing the subsequent formula.

$$V_i^{it} = \left|cos\theta_i^{it}\right|^{2\times sin\theta_i^{it}} \times \left(pbest_i^{it} - x_i^{it}\right) + \left|sin\theta_i^{it}\right|^{2\times cos\theta_i^{it}} \times (Gbest_i^{it} - x_i^{it}) \quad (7)$$

Then, the following equation is used to update the particle's position:

$$\vec{x}_i^{it+1} = \vec{x}_i^{it} + \vec{V}_i^{it} \quad (8)$$

Afterward, in a manner similar to the traditional PSO method, the locations of the Global Best (Gbest) and Personal Best (Pbest) are determined.

Subsequently, an update will be made to the particles' maximum velocities and phasor angles as follows:

$$\theta_i^{it+1} = \theta_i^{it} + T(\theta) \times (2\pi) = \theta_i^{it} + \left|\cos(\theta_i^{it}) + \sin(\theta_i^{it})\right| \times (2\pi) \quad (9)$$

$$V_{i,max}^{it+1} = W(\theta) \times (X_{max} - X_{min}) = \left|cos\theta_i^{it}\right|^2 \times (X_{max} - X_{min}) \quad (10)$$

It should be noted that the empirical formulae used in Eq. (4) to Eq. (7) and Eq. (8) to Eq. (10) were selected after a wide range of functions were tested. It would be impossible to list every function that was evaluated for this reason because there were so many of them.

E. Performance Criteria

When evaluating classifier performance, there exists a variety of evaluation criteria. Accuracy, a widely used measure, evaluates classifier effectiveness by determining the percentage of correctly predicted samples. In addition to Accuracy, Precision and Recall are commonly used metrics. Recall calculates the ratio of correctly predicted positive instances to the total actual positive instances, while precision assesses the probability of positive predictions being correct. Combining Precision and Recall results in a composite metric called the F1-score.

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

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$Recall = TPR = \frac{TP}{P} = \frac{TP}{TP+FN} \quad (13)$$

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

In these formulas, TP represents a positive prediction that matches the actual positive outcome. FP signifies a positive prediction when the actual outcome is negative. TN denotes a negative prediction that aligns with the actual negative outcome. FN stands for a negative prediction when the actual outcome is positive.

IV. Result and Discussion

A. Convergence

The suggested models' convergence curve is shown in Fig. 2, which provides a visual depiction of the algorithm's development in the direction of its goal. This curve delineates the accuracy performance metric against the number of iterations, unveiling crucial insights into the optimization process. The curve's shape and behavior become instrumental in gauging convergence efficiency; a steep descent signifies rapid convergence, while plateaus or erratic fluctuations may indicate challenges in reaching the optimal solution.

Convergence curves serve as pivotal tools in evaluating algorithm performance, refining parameters and comprehending the trade-offs between speed and accuracy in diverse computational tasks. Within this context, Fig. 2 specifically examines and illustrates the convergence curves of RFC+VAO and RFC+PPS. Notably, the accuracy curve of RFC+VAO commences from a more advantageous point compared to RFC+PPS and achieves its optimal result more swiftly. This observation implies that RFC+VAO outperforms RFC+PPS as iterations progress, suggesting its superior convergence efficiency in this computational task.

B. Comparison of Developed Models

The results in Table II reveal the performance metrics of the presented models, including RFC+VAO, RFC+PPS, and RFC, based on various index values: Accuracy, Precision, Recall, and F1-Score. These metrics are crucial for assessing the models' effectiveness in predicting student performance. RFC+VAO achieves an impressive accuracy of 0.934, indicating its correct predictions of student performance in the majority of cases. With a precision score of 0.940, it demonstrates a high level of precision, suggesting accurate predictions when it anticipates student success. The recall value of 0.930 shows that the model effectively identifies a substantial portion of students who will perform well. The F1-Score of 0.930 underscores its effectiveness in achieving a balance between precision and recall.

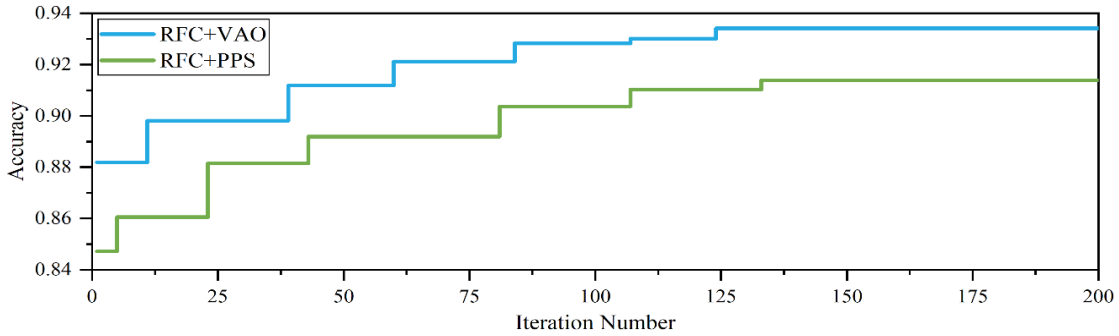

Fig. 2. Convergence curve of hybrid models.

In comparison, RFC+PPSO exhibits a respectable accuracy of 0.914, implying its effective performance in predicting student success. It achieves a precision score of 0.910, indicating a solid ability to make accurate predictions. With a recall value of 0.910, RFC+PPS effectively identifies a substantial portion of students who will perform well, although slightly lower than RFC+VAO. The F1-Score of 0.910 showcases RFC+PPS's ability to maintain a good balance between precision and recall. As for RFC, without the additional optimizers, it still demonstrates a reasonable accuracy of 0.889. With a precision value of 0.890, RFC maintains a good level of precision. The recall value of 0.890 indicates its effectiveness in identifying students with good performance, although slightly lower than RFC+VAO. The F1 Score of 0.890 underscores RFC's balanced performance between precision and recall.

In summary, the results in Table II highlight the positive impact of incorporating optimization techniques, such as VAO and PPS, into the Random Forest Classifier (RFC). RFC+VAO outperforms RFC+PPS and RFC in all metrics, showcasing its effectiveness in predicting student performance. The high precision and recall values for RFC+VAO and RFC+PPS indicate their potential for early identification of students who may excel, which is crucial for educational institutions aiming to provide timely guidance and support to improve overall academic performance.

TABLE II. RESULT OF PRESENTED MODELS

Model	Index values			
	Accuracy	*Precision*	*Recall*	*F1 _core*
RFC+VAO	0.934	0.940	0.930	0.930
RFC+PPS	0.914	0.910	0.910	0.910
RFC	0.889	0.890	0.890	0.890

Table III presents a thorough evaluation of the developed models' performance based on various grade categories, namely Excellent, Good, Acceptable, and Poor. The models, including RFC+VAO, RFC+PPSO, and RFC, are assessed in terms of Precision, Recall, and F1-score for each category. For RFC+VAO, the "Excellent" category reveals a precision of 0.97, while the recall is 0.82, resulting in an F1-score of 0.89. In the "Good" category, the model shows a precision of 0.87 and a recall of 0.90, leading to an F1-score of 0.89, indicating a well-balanced prediction. The "Acceptable" category exhibits a precision of 0.83 and a recall of 0.89, resulting in an F1-score of 0.86. In the "Poor" category, the model performs exceptionally well with a precision and recall of 0.97, yielding an F1-score of 0.97, highlighting its high accuracy.

Turning to RFC+PPS, the "Excellent" category displays a precision of 0.91 and a recall of 0.80, resulting in an F1 score of 0.85. In the "Good" category, it achieves a precision of 0.81 and a recall of 0.87, leading to an F1-score of 0.84. For the "Acceptable" category, the model has a precision of 0.82 and a recall of 0.81, resulting in an F1-score of 0.81. Similar to RFC+VAO, in the "Poor" category, it attains a precision and recall of 0.97, resulting in an F1-score of 0.97. As for RFC, it exhibits a precision of 0.82 and a recall of 0.78 in the "Excellent" category, resulting in an F1 score of 0.79. In the "Good" category, it has a precision of 0.80 and a recall of 0.80, leading to an F1-score of 0.80. For the "Acceptable" category, it showcases a precision of 0.73 and a recall of 0.84, resulting in an F1-score of 0.78. In the "Poor" category, it attains a precision of 0.97 and a recall of 0.94, resulting in an F1-score of 0.96. These results offer a detailed breakdown of the performance of each model across different grade categories. RFC+VAO and RFC+PPS consistently outperform RFC, particularly in the "Excellent" and "Good" categories, where they exhibit higher precision and recall values, signifying the positive impact of optimization techniques on accurate grade-level predictions.

To comprehensively evaluate the model's proficiency in predicting student performance and facilitate meaningful comparisons, Fig. 3 presents a column chart representing the four grades under consideration. This visual representation provides a clear indication of which model closely aligns with the measured values for each grade, thereby highlighting superior performance.

Upon examination of accuracy, both hybrid models, RFC+VAO and RFC+PPS, stand out by correctly predicting 227 out of 233 instances, while RFC closely follows by predicting 220 instances. In the categories of "Acceptable" and "Good" grades, the models demonstrate comparable performance, with RFC+VAO showing a slight edge in both instances. However, in predicting "Excellent" grades, all models perform closely, with RFC+VAO exhibiting a slightly superior performance.

This visual assessment not only aids in discerning the models' accuracy across different grade categories but also emphasizes the nuanced distinctions in performance, particularly highlighting the marginal superiority of RFC+VAO in certain instances.

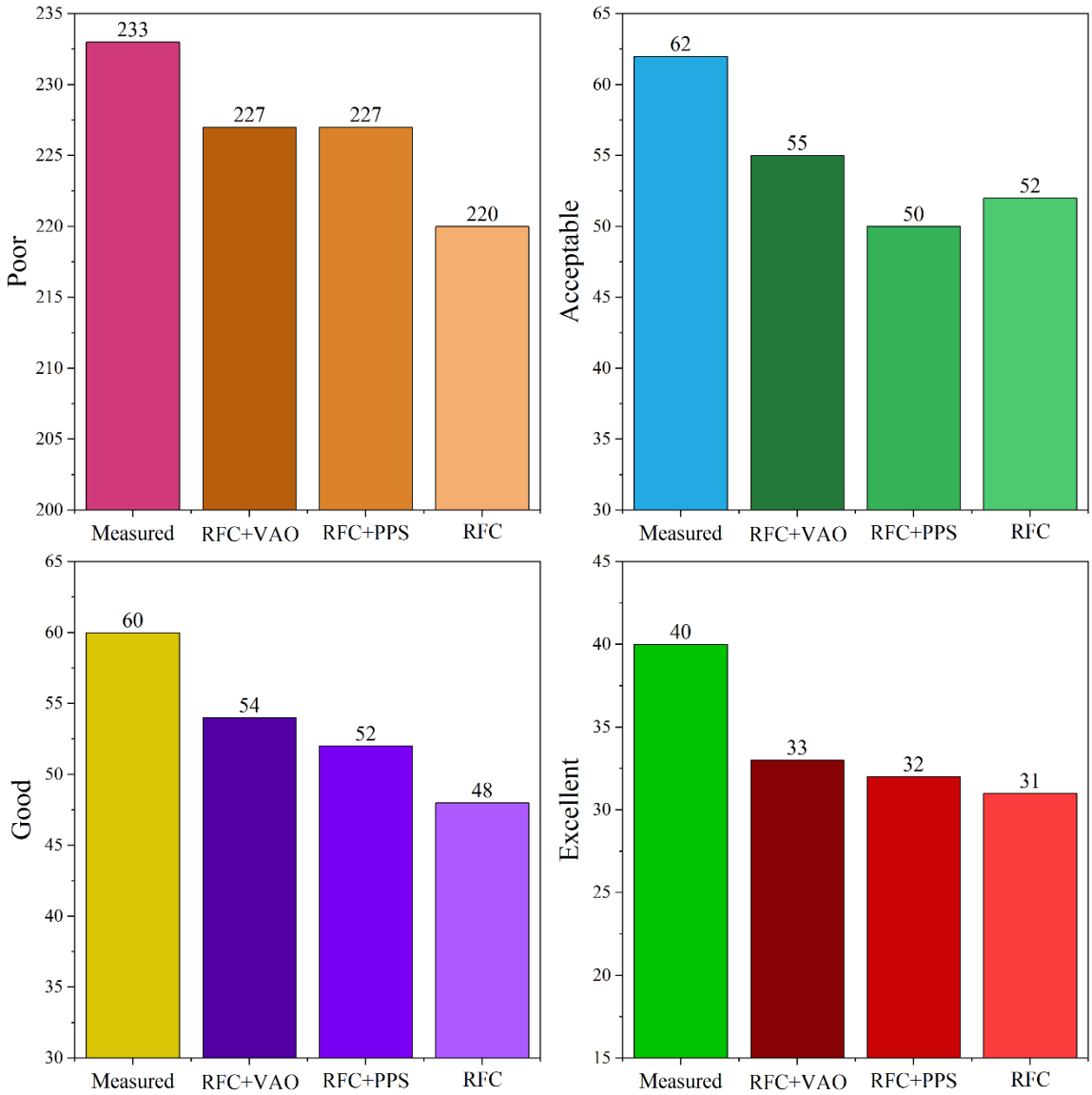

Fig. 3. A column chart displaying the association between the observed and anticipated values.

TABLE III. PERFORMANCE EVALUATION INDICES FOR THE DEVELOPED MODELS BASED ON GRADES

Model	Grade	Index values		
		Precision	***Recall***	***F1-score***
RFC+VAO	Excellent	0.97	0.82	0.89
	Good	0.87	0.9	0.89
	Acceptable	0.83	0.89	0.86
	Poor	0.97	0.97	0.97
RFC+PPS	Excellent	0.91	0.8	0.85
	Good	0.81	0.87	0.84
	Acceptable	0.82	0.81	0.81
	Poor	0.97	0.97	0.97
RFC	Excellent	0.82	0.78	0.79
	Good	0.8	0.8	0.8
	Acceptable	0.73	0.84	0.78
	Poor	0.97	0.94	0.96

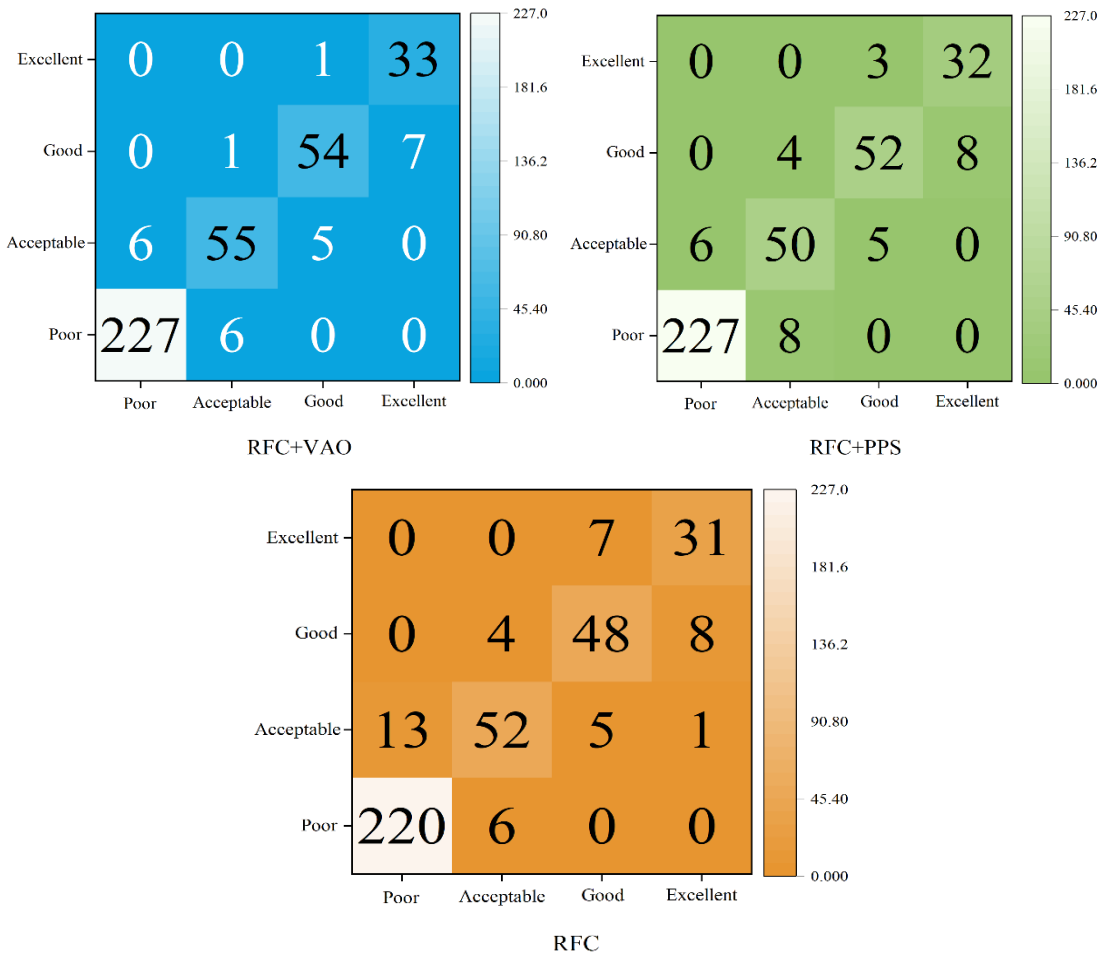

Fig. 4. Confusion matrix for each model's accuracy.

In Fig. 4, the confusion matrix visually depicts the relationship between observed and predicted classes, where the horizontal axis represents observed classes, and the vertical axis corresponds to predicted classes. Notably, the main diagonal cells in the matrix stand out with higher values, indicating successful predictions by the models.

Taking RFC+VAO as an example, it showcases a robust ability to predict the majority of observation classes accurately. For instance, in a scenario where 233 students were in the "Poor" class, RFC+VAO demonstrated a remarkable accuracy of 97.40%, accurately predicting 227 students. Merely six students were misclassified into the "Poor" category, underscoring the precision of the model.

This high precision extends to other classes as well, with accuracies of 88.70%, 90%, and 82.50% for the "Acceptable," "Good," and "Excellent" classes, respectively. It is worth noting that these figures, while slightly lower, distinguish RFC+VAO's performance from other model configurations.

Comparatively, RFC+PPS achieves accuracies of 97.40%, 80.64%, 86.66%, and 80% for the "Poor," "Acceptable," "Good," and "Excellent" classes, respectively. Meanwhile, RFC delivers accuracies of 94.42%, 83.87%, 80%, and 77.5% for the same classes. This comprehensive breakdown offers a nuanced understanding of the models' predictive performance across various classes, emphasizing RFC+VAO's notable precision and distinctions from other model configurations.

V. Discussion

A. Future Study

In future research, there are several key directions for enhancing predictive modeling in academic settings. The refinement of optimizers, specifically the VAO and PPSO techniques, should involve further fine-tuning to optimize their predictive performance. Additionally, the integration of additional data sources, such as socio-economic factors, health records, or extracurricular activities, is recommended to enrich the model and improve predictive accuracy.

A crucial aspect is the suggestion to conduct a longitudinal analysis, tracking academic trajectories over multiple semesters or years. This would provide insights into the model's stability and its ability to adapt to changes in student performance patterns over time. Lastly, a comparative analysis with other optimization algorithms would contribute valuable insights into the relative efficiency and effectiveness of the proposed VAO and PPSO optimizers within the educational data analytics context.

B. Limitations

The study acknowledges its focus on secondary school education statistics, cautioning against the direct generalization of findings to other educational systems due to potential variations in structures and demographics. The dependence on dataset availability and quality is recognized, emphasizing the need to address biases in data collection for robust outcomes. The study also acknowledges the sensitivity of machine learning algorithms to parameter changes and advocates for sensitivity analyses to assess the model's robustness. Ethical considerations, including transparency, fairness, and accountability, are highlighted to ensure the responsible and ethical deployment of predictive analytics in education. Overall, these considerations contribute to a nuanced understanding of the study's limitations and underscore the importance of ethical and context-aware applications of predictive models in diverse educational contexts.

C. Comparison with Papers

Table IV compares the present research paper with previously published studies, focusing on the predictive models and their respective accuracy levels. The present paper employs a RFC with VAO, achieving a notable accuracy of 93.4%. In contrast, previous studies predominantly used DTC or NBC and reported lower accuracy levels ranging from 69.94% to 82%. The methodological advancement in the present paper, incorporating VAO, suggests a promising improvement in predictive accuracy, with potential implications for more precise student performance predictions in educational settings.

TABLE IV. COMPARISON BETWEEN THE PRESENTED AND PUBLISHED PAPERS

Article	Model	Index values
		Accuracy
Edin Osmanbegovic et al. [34]	NBC	76.65%
Kabakchieva [35]	DTC	72.74%
Nguyen and Peter [36]	DTC	82%
Bichkar and R. R. Kabra [37]	DTC	69.94%
Present paper	RFC+VAO	93.4%

VI. CONCLUSION

In this extensive study, the focus was on predictive modeling for student performance using a dataset derived from the educational landscape. The goal was to enhance the predictive accuracy of the Random Forest Classifier (RFC) by integrating innovative optimization techniques, namely, Victoria Amazonia Optimization (VAO) and Phasor Particle Swarm Optimizer (PPS). The results shed light on the effectiveness of these models in predicting student performance across various grade categories. The analysis revealed that both RFC+VAO and RFC+PPS models consistently outperformed the standard RFC. This superiority was evident not only in predicting student grades but also in distinguishing between different academic performance levels. RFC+VAO and RFC+PPS consistently exhibited higher precision, recall, and F1 scores, particularly in the "Excellent" and "Good" grade categories. This underscores the impact of optimization techniques in improving model accuracy and their potential to enhance student support systems. The models excelled in identifying students falling within the "Excellent" and "Good" grade categories, which is vital for educational institutions aiming to provide timely guidance and support for academic excellence. RFC+VAO, in particular, demonstrated a slight advantage in predicting "Excellent" grades, indicating the potential of the Victoria Amazonia Optimization technique in fine-tuning model performance. Furthermore, the confusion matrix in this analysis highlighted the models' proficiency in classifying observations, with the main diagonal consistently containing higher values, confirming the model's precision in predicting various class categories. In summary, this research underscores the promising potential of machine learning models, especially when combined with optimization techniques, in educational data analysis. It provides a foundation for institutions to utilize these models as valuable tools in student performance prediction and support systems. The accurate prediction of a student's academic trajectory benefits not only the students themselves but also empowers educational institutions to implement tailored strategies and interventions. As the educational landscape evolves, the integration of machine learning and optimization techniques will play a pivotal role in ensuring academic success for students, ultimately shaping a brighter future for the education sector. The findings presented in this article encourage further exploration and real-world testing to refine and optimize these models for effective utilization in educational institutions.

FUNDING

2021-2022 Key Project of Higher Education Research of Chongqing Higher Education Society. Number (CQGJ21A004).

REFERENCES

[1] S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," in IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2020, p. 32019.

[2] R. Alamri and B. Alharbi, "Explainable student performance prediction models: a systematic review," IEEE Access, vol. 9, pp. 33132–33143, 2021.

[3] P. M. Arsad and N. Buniyamin, "A neural network students' performance prediction model (NNSPPM)," in 2013 IEEE International Conference on Smart Instrumentation, Measurement, and Applications (ICSIMA), IEEE, 2013, pp. 1–5.

[4] D. Kabakchieva, "Student performance prediction by using data mining classification algorithms," International Journal of Computer Science and Management Research, vol. 1, no. 4, pp. 686–690, 2012.

[5] F. Masoumi, S. Najjar-Ghabel, A. Safarzadeh, and B. Sadaghat, "Automatic calibration of the groundwater simulation model with high parameter dimensionality using sequential uncertainty fitting approach," Water Supply, vol. 20, no. 8, pp. 3487–3501, Dec. 2020, doi: 10.2166/ws.2020.241.

[6] Behnam Sedaghat, G. G. Tejani, and S. Kumar, "Predict the Maximum Dry Density of Soil based on Individual and Hybrid Methods of Machine Learning," Advances in Engineering and Intelligence Systems, vol. 002, no. 03, 2023, doi: 10.22034/aeis.2023.414188.1129.

[7] M. Chitti, P. Chitti, and M. Jayabalan, "Need for interpretable student performance prediction," in 2020 13th International Conference on Developments in eSystems Engineering (DeSE), IEEE, 2020, pp. 269–272.

[8] B.-H. Kim, E. Vizitei, and V. Ganapathi, "GritNet: Student performance prediction with deep learning," arXiv preprint arXiv:1804.07405, 2018.

[9] I. Khan, A. R. Ahmad, N. Jabeur, and M. N. Mahdi, "A Conceptual Framework to Aid Attribute Selection in Machine Learning Student Performance Prediction Models.," International Journal of Interactive Mobile Technologies, vol. 15, no. 15, 2021.

[10] H. Al-Shehri et al., "Student performance prediction using support vector machine and k-nearest neighbor," in 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 2017, pp. 1–4.

[11] Z. Xu, H. Yuan, and Q. Liu, "Student performance prediction based on blended learning," IEEE Transactions on Education, vol. 64, no. 1, pp. 66–73, 2020.

[12] F. Ünal, "Data mining for student performance prediction in education," Data Mining-Methods, Applications and Systems, vol. 28, pp. 423–432, 2020.

[13] Y. Su et al., "Exercise-enhanced sequential modeling for student performance prediction," in Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[14] H. Hassan, S. Anuar, and N. B. Ahmad, "Students' performance prediction model using meta-classifier approach," in Engineering Applications of Neural Networks: 20th International Conference, EANN 2019, Xersonisos, Crete, Greece, May 24-26, 2019, Proceedings 20, Springer, 2019, pp. 221–231.

[15] P. Shruthi and B. P. Chaitra, "Student performance prediction in the education sector using data mining," 2016.

[16] P. Chaudhury and H. K. Tripathy, "An empirical study on attribute selection of student performance prediction model," International Journal of Learning Technology, vol. 12, no. 3, pp. 241–252, 2017.

[17] H. Chanlekha and J. Niramitranon, "Student performance prediction model for early identification of at-risk students in traditional classroom settings," in Proceedings of the 10th International Conference on Management of Digital EcoSystems, 2018, pp. 239–245.

[18] H. Lu and J. Yuan, "Student performance prediction model based on discriminative feature selection," International Journal of Emerging Technologies in Learning (Online), vol. 13, no. 10, p. 55, 2018.

[19] B. K. Bhardwaj and S. Pal, "Data Mining: A prediction for performance improvement using classification," arXiv preprint arXiv:1201.3418, 2012.

[20] M. M. R. Khan, M. A. B. Siddique, and S. Sakib, "Non-intrusive electrical appliances monitoring and classification using K-nearest neighbors," in 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET), IEEE, 2019, pp. 1–5.

[21] A. O. Ogunde and D. A. Ajibade, "A data mining system for predicting university students' graduation grades using ID3 decision tree algorithm," Journal of Computer Science and Information Technology, vol. 2, no. 1, pp. 21–46, 2014.

[22] F. Duzhin and A. Gustafsson, "Machine learning-based app for self-evaluation of teacher-specific instructional style and tools," Educ Sci (Basel), vol. 8, no. 1, p. 7, 2018.

[23] K. M. Hasib et al., "A survey of methods for managing the classification and solution of data imbalance problem," arXiv preprint arXiv:2012.11870, 2020.

[24] J. Watkins, M. Fabielli, and M. Mahmud, "Sense: a student performance quantifier using sentiment analysis," in 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–6.

[25] K. M. Hasib, N. A. Towhid, and M. G. R. Alam, "Online review based sentiment classification on Bangladesh airline service using supervised learning," in 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), IEEE, 2021, pp. 1–6.

[26] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[27] F. Livingston, "Implementation of Breiman's random forest machine learning algorithm," ECE591Q Machine Learning Journal Paper, pp. 1–13, 2005.

[28] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption," Energy Build, vol. 147, pp. 77–89, 2017.

[29] B. T. Pham et al., "A novel hybrid soft computing model using random forest and particle swarm optimization for estimation of undrained shear strength of soil," Sustainability, vol. 12, no. 6, p. 2218, 2020.

[30] S. M. H. Mousavi, "Victoria Amazonica Optimization (VAO): An Algorithm Inspired by the Giant Water Lily Plant," arXiv preprint arXiv:2303.08070, 2023.

[31] S. S. Gilan, H. B. Jovein, and A. A. Ramezanianpour, "Hybrid support vector regression–Particle swarm optimization for prediction of compressive strength and RCPT of concretes containing metakaolin," Constr Build Mater, vol. 34, pp. 321–329, 2012.

[32] A. Ratnaweera, S. K. Halgamuge, and H. C. Watson, "Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients," IEEE Transactions on Evolutionary Computation, vol. 8, no. 3, pp. 240–255, 2004.

[33] M. Ghasemi, J. Aghaei, and M. Hadipour, "New self-organizing hierarchical PSO with jumping time‐varying acceleration coefficients," Electron Lett, vol. 53, no. 20, pp. 1360‐1362, 2017.

[34] E. Osmanbegovic and M. Suljic, "Data mining approach for predicting student performance," Economic Review: Journal of Economics and Business, vol. 10, no. 1, pp. 3–12, 2012.

[35] D. Kabakchieva, "Student performance prediction by using data mining classification algorithms," International Journal of Computer Science and Management Research, vol. 1, no. 4, pp. 686–690, 2012.

[36] N. T. Nghe, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," in 2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports, IEEE, 2007, pp. T2G-7.

[37] R. R. Kabra and R. S. Bichkar, "Performance prediction of engineering students using decision trees," Int J Comput Appl, vol. 36, no. 11, pp. 8–12, 2011.