Empowering Learning through Intelligent Data-Driven Systems

Khalid Abdullah Aldriwish
Computer Science Department, College of Computer and Information Sciences, Majmaah University, Saudi Arabia
k.aldriwish@mu.edu.sa (corresponding author)

Received: 26 November 2023 | Revised: 5 December 2023 | Accepted: 25 December 2023

ABSTRACT

The evolution of educational systems is closely tied to technological advancements, particularly the emergence of machine learning. This technology offers a sophisticated system capable of predicting, explaining, and influencing behavior. Many efforts have aimed to integrate machine learning into education, focusing on specific cases using ad-hoc models. This paper introduces an intelligent educational system that relies on data-driven student models, aiming to surpass the limitations of these ad-hoc systems. The approach outlined in this endeavor adopts a comprehensive and methodical modeling methodology centered on machine learning techniques. By employing Long Short-Term Memory (LSTM), the proposed approach enables predictive student models based on historical educational data. The effectiveness of this method was tested through experimentation on an intelligent tutoring system using 5-fold cross-validation, revealing that the smart educational system achieved a remarkable 96% accuracy rate. Furthermore, a comparison between the importance scores of features with and without the student models demonstrated the practicality and effectiveness of the proposed methodology.

Keywords-machine learning; educational systems; CNN; historical data

I. INTRODUCTION

Intelligent Tutoring Systems (ITSs) are the subject of many computer-aided systems that automatically provide teaching and training [1]. These systems require considerable knowledge in computer science, even psychological interaction. Artificial Intelligence (AI) allowed a significant revolution in education, ensuring complex cognitive skills [2]. The ITS based-AI is a computer program using computational models that enable detailed feedback with customized instruction. Developed ITSs are applied to various educational systems and millions of learners worldwide [4-6]. Learners using ITSs are motivated to know what they learn, how to learn, and whom they learn through reasoning techniques incorporated in ITSs. These features are based on four models [3]:

- The student model stores nationality and the current state to choose a suitable new problem.
- The domain model stores knowledge.
- The user interface provides access to the domain knowledge elements.
- The tutor model stores pedagogical knowledge.

The student model is the most important key to the success of an ITS [7, 8]. It should be able not only to support the students' skills and behaviors but also to provide the students' performance and abilities according to knowledge level and educational background. Authors in [6] focused on the dynamic characteristics of the tutor. They studied many features, including skills, knowledge, misconceptions, preferences, and learning styles. They highlighted two dynamic concepts: (1) cognitive factors: attention, ability to solve problems and ability to learn features and (2) meta-cognitive factors: self-assessment, self-regulation, help-seeking, and attitude. Authors in [9] proposed an ITS permitting prediction, explanation, and tutor behavior using machine learning techniques. Authors in [10] presented an intelligent marine traffic simulator to provide a highly qualified system for trainees and students. The author's objectives focus on enhancing learning skills in a safe condition and providing intelligent decision-making support for collision avoidance. Authors in [11] listed many studies related to ITS based on machine learning. These attempts aimed to model an application-based cognitive process assessment and prediction for specific cases. Authors in [12] presented a prediction model to study the student's performance. Their attempt used clustering methods to identify learning difficulties. Authors in [13] predicted tutors' effortful behavior using the pipeline grey-box approach, allowing the prediction of performance related to assessment tasks. Authors in [3] summarized the student modeling technique. They aimed to provide an accurate elaborative environment, corrective strategy, and evaluative models. Authors in [14] proposed an outcome-based education model to support the accreditation process. They focused on the subject quality, faculty experience, and student quality factors to perform course learning outcomes measurement. The proposed model attempted to identify the most suitable weight of each factor. Authors in [15] focused on working conditions related to real and natural educational contexts. They proved that there is a complicated landscape regarding the effectiveness of ITS in real educational contexts. They found that there is a noticeable disparity in the regions where studies on ITS using social experiments were carried out.

The above studies proved that a successful ITS has to include predictive student modeling. The model must identify latent constructs and future outcomes (failure or success in a course) [16]. Nevertheless, the developed models used only one machine learning technique and are considered for specific domain data. This paper deals with this issue and provides systemic and comprehensive machine learning and ensures an accurate and smart learner model for predicting tutors' learning performance. The idea covers building data-driven student models according to models related to knowledge or domain theory. The present paper proposes an innovative and powerful approach to ensure a high-performance student model supporting static and dynamic features during the learning process. Static features are defined as student profile information and dynamic features describe the student's behavior, including attention during the interaction with the ITS. This paper attempts to improve the student model by using machine learning applied to the historical and educational data. To address the above challenge, the following points are tackled:

- Building a data-driven student model according to a comprehensive student modeling methodology.
- Obtaining high accuracy of the student model built according to the historical data.
- Determining the more significant features to understand the student model's utility.

To achieve the objectives of this paper, we proceed with the following contributions:

- A student model based on a comprehensive machine learning approach is proposed.
- The proposed model is verified through a case study.

II. MATERIALS AND METHODS

A student model-based on AI is focused on human or social behavior [9]. According to the historical data collected from the user, the model is built using a variety of machine-learning algorithms. In this section, we highlight the methodology of the proposed tutoring system. The model is composed of four steps, described in Figure 1:

- The preprocessing step ensures that only significant data are retrieved and considered in the training step. Building a relevant dataset is important to achieve an accurate prediction. The preprocessing step attempts to find significant data through large amounts of data. Our system retrieved data related to student learning behavior and external impact elements during the preprocessing. The dataset contained student profile, course object, teaching techniques, environment features, and learning behavior (problem-solving capabilities, working level, curiosity, subjects' interest, attention, etc.).
- The data augmentation step ensured the generation of more features with high quality. These new features are powerful because they are based on combinations of the initial data inputs. A risk feature is added according to the information of the domain knowledge. The data augmentation step also guarantees the selection of relevant features using the ranking method [17, 18].
- The learning algorithm step provided the most suitable model for the deployment. Many techniques can be used either in classification or regression. A model is adopted once the used learning algorithm achieves high accuracy. In our case, we used the following learning algorithms: Naïve Bayes (NB) [19], Decision Tree (DT) [20], Support Vector Machines (SVMs) [21], Convolutional Neural Networks (CNNs) [22, 23], and Long Short-Term Memory (LSTM) [24]. We have to highlight that for every algorithm, a customized configuration was considered in order to obtain better performance according to performance parameters referred to in [25].

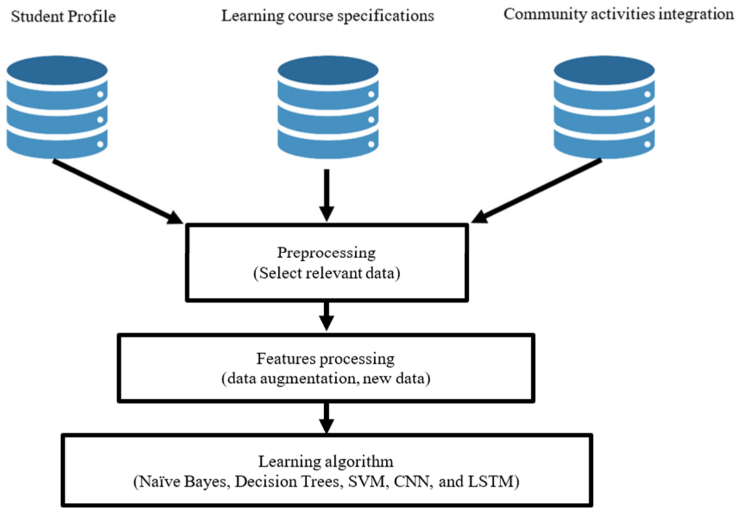

Student Profile
Learning course specifications
Community activities integration
Preprocessing
(Select relevant data)
Features processing
(data augmentation, new data)
Learning algorithm
(Naïve Bayes, Decision Trees, SVM, CNN, and LSTM)

Fig. 1. The machine learning model for model training.

The evaluation step showed the differences between the models. Consequently, the more suitable model could be chosen based on those results. The evaluation was performed using various criteria. For the classification task, the used metrics were ROC (Receiver Operating Characteristics) curve, AUC (Area Under the ROC Curve), ROCCH (ROC Convex Hull), DEA (Data Envelopment Analysis), accuracy, sensitivity, specificity, precision, etc. For the regression task, the evaluation metrics included the Mean Absolute Error (MAE), the Sum of Squared Error (SSE), the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), and the Mean Absolute Percentage Error (MAPE).

According to the training results, the model with the higher validation score will be deployed into the ITS. This is ensured by an auto-updating model aiming to configure and replace out-mode/out-date data-driven models. Then, the feature inspection is provided by selecting the important features using MDG (Mean Decrease in Gini Impurity), MDA (Mean Decrease in Accuracy) methods [26], and the permutation importance method [27]. In our case, the features are selected based on an ordered list of features sorted according to their importance score. This method allows the teachers/instructors/trainers to understand the student model reasoning and permit the right decision.

III. RESULTS

In this section, the proposed intelligent learning tutoring system is presented, and the experimentation and performance results are detailed.

A. Case Study

The feasibility and performance of the proposed machine learning system are detailed through an intelligent tutoring system for colleges in this section. The student model is built according to the suggested model considering the ITS data. The sensitivity, specificity, and accuracy metrics are considered for evaluation. The developed model is shown in Figure 2. The model is composed of three layers: GUI, Model, and Data Access layers. The GUI layer defines the graphic training instruction/interaction interface between the trainee and the ITS system. The model layer consists of the trainee model, the trainee profile, training strategies, and the knowledge domain.

The proposed model described in Figure 2 shows that the intelligent learning tutoring system can train trainees with different educational backgrounds and experiences. All information related to the trainee is kept in the trainee profile. The system allows failure to catch up and acquire the requested skills. The data collection is proceeded using data-driven trainee/student models. The training data for each training object were collected with a given sampling rate. The variable data used during the training in the tutoring system are shown in Table I. The educational background Eb is a normalized feature that refers to the degree of education and the major subject. One training object is adopted to ensure relevant data with the instructor's outcome. The outcome Out variable indicates the failure or the success of the trainee/student. This paper is focused on classification-based models. It aims to predict the outcomes through the student profile, training object identifier, course object, learning behavior, and situations.

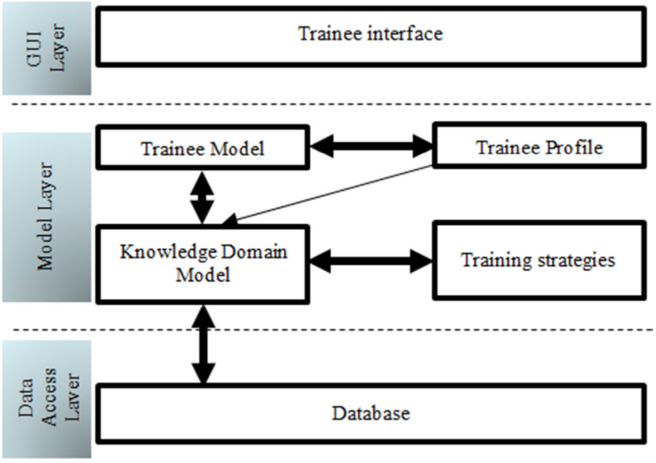


Fig. 2. The intelligent learning tutoring systems

The description information of each training object is detailed in Table I. The retained data after the selection phase are 2452 samples. The built dataset validates the student model through the proposed machine learning methods.

TABLE I. COLLECTED DATA VARIABLES FROM THE TUTORING SYSTEM

Data source	Variables	Attribute name	Type
Student profile	Eb Age Sex Nat Rs	Education background Age Sex Nationality Residence state	Nominal Numeric Nominal Nominal Nominal
Training object identifier	DT Cn Teach	Date and Time Course name Teacher name	Date/time Text Text
Course object	Gl Univ Sd Bk Cobj	Grade level University Study discipline Background knowledge Course objectives	Numeric Nominal Nominal Nominal Numeric
Learning behavior	ILs Hw Rc Ls Sl	Interest level in the subject Hardworking Retake the course Learning strategy Speed learning	Nominal Nominal Numeric Nominal Nominal
Situations	Td Old Pr Sp Risk	Target diploma Outcomes learnings degree Presence rate Student position Risk feature (new feature)	Nominal Numeric Numeric Numeric Numeric
Learning results	Cf Out	Confidence Outcome	Numeric Nominal

B. Experiments and Results

Based on the above dataset, the student model describes the learning behavior and predicts outcomes (Failure/Success) according to the machine learning methodology proposed above.

In the first step, preprocessing, the collected data shown in Table I have to be assigned to right values of the raw data. In the hardworking attribute case, the type is defined as a nominal variable assigned by the instructor according to the assessment report: The value is assigned to three states: High, Medium, and Low. Other attributes with nominal values are treated similarly.

In the second step, feature processing, the trainee model generates the new features titled risk based on basic knowledge. The risk feature is calculated based on (1). The risk feature is computed to give a precise idea about the failure risk and enhance the model performance. The risk feature is processed through the course objectives and outcomes the trainee/student achieves for every course.

$$R_f = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{Ci_{obj/Trainee}}{Ci_{obj}} \quad (1)$$

where n is the number of courses, $Ci_{obj/Trainee}$ is the rate of achieved objectives in a course, and Ci_{obj} is the course objectives.

This work considers the risk high when R_f is inferior to 0.3. Therefore, the risk is a subjective attribute. The model is built according to the developed above modeling methodology. The model experimentation is ensured using the WEKA tool [28]. The classification performance is computed through the following machine learning algorithms: NB, DT, SVM, CNN, and LSTM. We highlight that these algorithms provide an easy deployment in the education system. Each algorithm is used in two versions: (1) default and (2) Cost Matrix configurations. The second version considers the unexpected predictions and the sample weight to ensure an updated model behavior. Ten models are trained in our case. Five-fold cross-validation is conducted to obtain accuracy, sensitivity, and specificity metrics (Table II). The Cost Matrix version may either enhance or degrade performance. The LSTM algorithm with the Cost Matrix version achieves the best performance with 96% accuracy, 96% sensitivity, and 95% specificity. The results (Table II) prove that the proposed machine learning-based student modeling is valuable and feasible.

TABLE II. PERFORMANCE RESULTS

Student Model	ML algorithms	Accuracy	Sensitivity	Specificity
SM1	NB	89%	90%	87%
SM2	NB Cost Matrix	92%	91%	91%
SM3	DT	79%	78%	82%
SM4	DT Cost Matrix	88%	85%	83%
SM5	SVM	90%	90%	88%
SM6	SVM Cost Matrix	82%	80%	85%
SM7	CNN	92%	92%	93%
SM8	CNN Cost Matrix	90%	92%	92%
SM9	LSTM	94%	93%	94%
SM10	LSTM Cost Matrix	96%	96%	95%

IV. DISCUSSION

The model prediction suggests selecting the student's right course and learning objects. The student model can provide right decision-making up to 96%. It helps the teacher to enhance course objects related to appropriate students and deliver proper recommendations. Therefore, the model prediction provides an intelligent tutoring system for the teacher to know the best way to instruct the students/ trainees. The feature importance is computed through the permutation importance method. Features entered to the LSTM Cost Matrix algorithm are ranked based on the importance score, see Table III. We remark that the student models are deeply associated with high score features. These interesting results are highlighted for the following features: risk, education background, background knowledge, and learning strategy. These features are essential to build the proper training course object for different students. Some features are irrelevant to the student model's performance and will be ignored during the deployment phase. The attained objective achieved during this research is to develop a classification-based model to predict the outcome of learning performance. The proposed machine learning-based methods inform about the impact of the student models in the education system.

TABLE III. FEATURES ASSOCIATED WITH THE IMPORTANCE SCORE RANKING

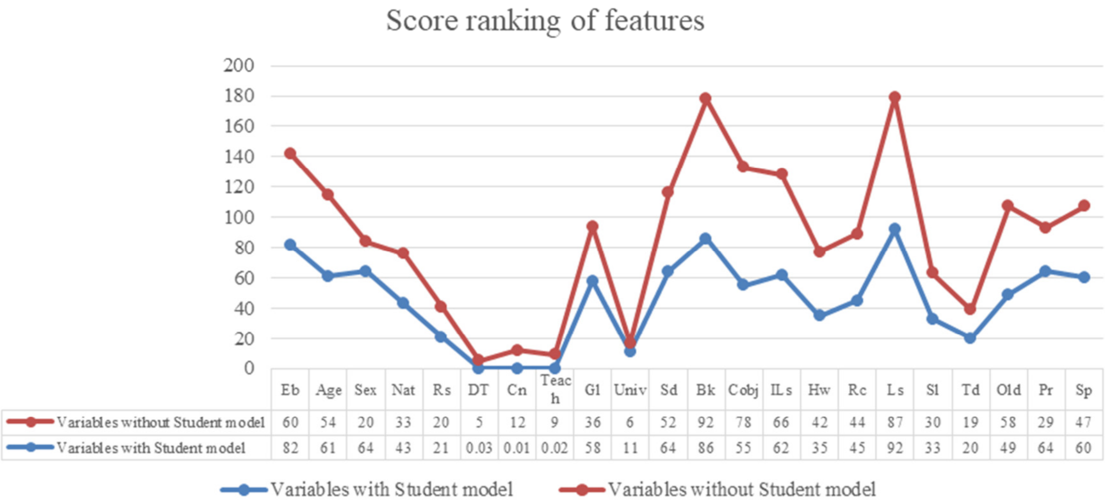
Data source	Features	Variables
Student profile	**Eb**	**82**
	Age	61
	Sex	64
	Nat	43
	Rs	21
Training object identifier	DT	0.03
	Cn	0.01
	Teach	0.02
Course object	Gl	58
	Univ	11
	Sd	64
	Bk	**86**
	Cobj	55
Learning behavior	ILs	62
	Hw	35
	Rc	45
	Ls	**92**
	Sl	33
Situations	Td	20
	Old	49
	Pr	64
	Sp	60
	Risk	**97**

A comparison of the importance score ranking features with and without the student models is conducted based on students' feedback. Asking the same trainees to use the same learning is difficult and restricted to only 10 students. It would be necessary to highlight that this comparison gives us general information about the usefulness of the proposed model. It would be degraded due to the variation of student skills over the years.

Figure 3 shows the similarity between results with/without the student's model. The proposed student's model accurately predicts and highlights the most relevant features. In the case of features without the student's model, we remark that sex, presence, and course name are relevant features that are wrong. This misunderstanding proves the value of the proposed intelligent student model. The most pertinent features are similar for the two cases.

V. CONCLUSION

In conclusion, this paper presents a novel Intelligent Tutoring System (ITS) that leverages machine learning-based student models. The approach taken in this study involved the construction of student models based on historical educational data, marking a significant departure from traditional methods. This approach aimed to enhance the understanding of students' learning patterns and needs. The practical application of these models within the ITS framework was demonstrated through a compelling case study that utilized stored features. This application sheds light on the practicality and potential benefits of using student models in the context of intelligent tutoring.



	Eb	Age	Sex	Nat	Rs	DT	Cn	Teach	Gl	Univ	Sd	Bk	Cobj	ILs	Hw	Rc	Ls	Sl	Td	Old	Pr	Sp
Variables without Student model	60	54	20	33	20	5	12	9	36	6	52	92	78	66	42	44	87	30	19	58	29	47
Variables with Student model	82	61	64	43	21	0.03	0.01	0.02	58	11	64	86	55	62	35	45	92	33	20	49	64	60

Fig. 3. The importance score ranking of features with/without the student's models

To validate the efficacy and feasibility of this approach, a rigorous experimentation process was carried out, employing the 5-fold cross-validation technique. The experimental results provided strong support for the proposed methodology. They revealed that the machine learning-based student models built upon historical educational data have the potential to significantly enhance the effectiveness of ITS systems.

As we turn our gaze toward future endeavors in this field, a notable challenge comes into focus. This challenge revolves around the development of models that can facilitate the transfer of student models from one educational domain to another. While the current research has demonstrated the promise of machine learning-based student models within a specific domain, the ability to extend this success to different educational contexts presents an intriguing avenue for further exploration. Solving this challenge could pave the way for even more versatile and adaptive ITS systems, thereby offering an even more comprehensive and tailored learning experience for students across a wide spectrum of disciplines. This direction holds the potential to make ITSs not only more effective, but also more widely applicable, ultimately benefiting a broader range of learners.

REFERENCES

[1] M. Laaziri, S. Khoulji, K. Benmoussa, and K. M. Larbi, "Outlining an Intelligent Tutoring System for a University Cooperation Information System," *Engineering, Technology & Applied Science Research*, vol. 8, no. 5, pp. 3427–3431, Oct. 2018, https://doi.org/10.48084/etasr.2158.

[2] F. D. Pereira *et al.*, "Using learning analytics in the Amazonas: understanding students' behaviour in introductory programming," *British Journal of Educational Technology*, vol. 51, no. 4, pp. 955–972, 2020, https://doi.org/10.1111/bjet.12953.

[3] I. Saric-Grgic, A. Grubisic, S. Stankov, and M. Stula, "An agent-based intelligent tutoring systems review," *International Journal of Learning Technology*, vol. 14, no. 2, pp. 125–140, Jan. 2019, https://doi.org/10.1504/IJLT.2019.101847.

[4] A. Alkhatlan and J. Kalita, "Intelligent Tutoring Systems: A Comprehensive Historical Survey with Recent Developments," *arXiv e-prints*. Dec. 01, 2018, https://doi.org/10.48550/arXiv.1812.09628.

[5] H. Mohamed and M. Lamia, "Implementing flipped classroom that used an intelligent tutoring system into learning process," *Computers & Education*, vol. 124, pp. 62–76, Sep. 2018, https://doi.org/10.1016/j.compedu.2018.05.011.

[6] L. Yuan *et al.*, "Iterative Teacher-Aware Learning," in *35th Conference on Neural Information Processing Systems*, Canada, USA, Dec. 2021, pp. 1–32.

[7] M. A. Al-Shargabi, "An Integrated Decision Support Model For Enhancing Continuous Improvement Of Academic Programs," *Engineering, Technology & Applied Science Research*, vol. 9, no. 5, pp. 4835–4841, Oct. 2019, https://doi.org/10.48084/etasr.3079.

[8] M. A. Alsuwaiket, A. H. Blasi, and K. Altarawneh, "Refining Student Marks based on Enrolled Modules' Assessment Methods using Data Mining Techniques," *Engineering, Technology & Applied Science Research*, vol. 10, no. 1, pp. 5205–5210, Feb. 2020, https://doi.org/10.48084/etasr.3284.

[9] O. Osoba and P. K. Davis, "An Artificial Intelligence/Machine Learning Perspective on Social Simulation: New Data and New Challenges," in *Social-Behavioral Modeling for Complex Systems*, New York, NY, USA: John Wiley & Sons, 2019, pp. 443–476.

[10] K. Hasegawa, G. Tashiro, S. Kiritani, and K. Tachikawa, "Intelligent Marine Traffic Simulator for Congested Waterways," in *7th IEEE International Conference on Methods and Models in Automation and Robotics*, Miedzyzdroje, Poland, 2001, pp. 1–6.

[11] C. Korkmaz and A.-P. Correia, "A review of research on machine learning in educational technology," *Educational Media International*, vol. 56, no. 3, pp. 250–267, Jul. 2019, https://doi.org/10.1080/09523987.2019.1669875.

[12] D. Hooshyar, M. Pedaste, and Y. Yang, "Mining Educational Data to Predict Students' Performance through Procrastination Behavior," *Entropy*, vol. 22, no. 1, Jan. 2020, Art. no. 12, https://doi.org/10.3390/e22010012.

[13] K. Sharma, Z. Papamitsiou, and M. Giannakos, "Building pipelines for educational data using AI and multimodal analytics: A 'grey-box' approach," *British Journal of Educational Technology*, vol. 50, no. 6, pp. 3004–3031, 2019, https://doi.org/10.1111/bjet.12854.

[14] P. Dayananda, M. V. Latte, M. S. Raisinghani, and C. N. Sowmyarani, "New approach for target setting mechanism of course outcomes in higher education accreditation," *Journal of Economic and Administrative Sciences*, vol. 37, no. 1, pp. 79–89, Jan. 2020, https://doi.org/10.1108/JEAS-03-2020-0024.

[15] H. Wang *et al.*, "Examining the applications of intelligent tutoring systems in real educational contexts: A systematic literature review from the social experiment perspective," *Education and Information Technologies*, vol. 28, no. 7, pp. 9113–9148, Jul. 2023, https://doi.org/10.1007/s10639-022-11555-x.

[16] S. K G and M. Kurni, "Educational Data Mining & Learning Analytics," in *A Beginner's Guide to Learning Analytics*, S. K G and M. Kurni, Eds. New York, NY, USA: Springer, 2021, pp. 29–60.

[17] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018, https://doi.org/10.1016/j.neucom.2017.11.077.

[18] T. Maguire, L. Manuel, R. A. Smedinga, and M. Biehl, "A review of feature selection and ranking methods," in *Proceedings of the 19th SC@ RUG 2022*, 2022, pp. 15–20.

[19] S. M. R. Abidi, M. Hussain, Y. Xu, and W. Zhang, "Prediction of Confusion Attempting Algebra Homework in an Intelligent Tutoring System through Machine Learning Techniques for Educational Sustainable Development," *Sustainability*, vol. 11, no. 1, Jan. 2019, Art. no. 105, https://doi.org/10.3390/su11010105.

[20] C. Yang, F.-K. Chiang, Q. Cheng, and J. Ji, "Machine Learning-Based Student Modeling Methodology for Intelligent Tutoring Systems," *Journal of Educational Computing Research*, vol. 59, no. 6, pp. 1015–1035, Oct. 2021, https://doi.org/10.1177/0735633120986256.

[21] K. Hemachandran *et al.*, "Artificial Intelligence: A Universal Virtual Tool to Augment Tutoring in Higher Education," *Computational Intelligence and Neuroscience*, vol. 2022, May 2022, Art. no. e1410448, https://doi.org/10.1155/2022/1410448.

[22] S. Alshmrany, "Adaptive learning style prediction in e-learning environment using levy flight distribution based CNN model," *Cluster Computing*, vol. 25, no. 1, pp. 523–536, Feb. 2022, https://doi.org/10.1007/s10586-021-03403-3.

[23] A. Berguiga, A. Harchay, A. Massaoudi, M. B. Ayed, and H. Belmabrouk, "GMLP-IDS: A Novel Deep Learning-Based Intrusion Detection System for Smart Agriculture," *Computers, Materials & Continua*, vol. 77, no. 1, 2023, Art. no. 379.

[24] Y. Ma and W. Lu, "Design and Implementation of Learning System Based on T-LSTM," in *20th International Conference on Web-Based Learning*, Macau, China, Nov. 2021, pp. 148–153, https://doi.org/10.1007/978-3-030-90785-3_14.

[25] S. Jagannatha, M. Niranjanamurthy, and P. Dayananda, "Algorithm Approach: Modelling and Performance Analysis of Software System," *Journal of Computational and Theoretical Nanoscience*, vol. 15, no. 11–12, pp. 3389–3397, Nov. 2018, https://doi.org/10.1166/jctn.2018.7631.

[26] C. Molnar, *Interpretable Machine Learning*. Morrisville, NC, USA: Lulu.com, 2020.

[27] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," *Electronics*, vol. 8, no. 8, Aug. 2019, Art. no. 832, https://doi.org/10.3390/electronics 8080832.

[28] R. Ratra and P. Gulia, "Experimental Evaluation of Open Source Data Mining Tools (WEKA and Orange)," *International Journal of Engineering Trends and Technology*, vol. 68, no. 8, pp. 30–35, Aug. 2020, https://doi.org/10.14445/22315381/IJETT-V68I8P206S.