An analysis of diverse computational models for predicting student achievement on e-learning platforms using machine learning

Naga Satya Koti Mani Kumar Tirumanadham[1], Thaiyalnayaki Sekhar[1], Sriram Muthal[2]
[1]Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Selaiyur, India
[2]Department of Information Technology, Bharath Institute of Higher Education and Research, Selaiyur, India

Article Info

Article history:

Received Mar 30, 2024
Revised Jul 16, 2024
Accepted Aug 6, 2024

Keywords:

AdaBoost
E-learning
Ensemble learning models
Gradient boosting
L2 regularization (Ridge)
Random forest
Recursive feature elimination

ABSTRACT

Coronavirus disease 2019 (COVID-19) has led many colleges and students to use online learning. In educational databases with so much data, evaluating student development is difficult. E-learning is essential for egalitarian education since it uses technology and contemporary learning techniques. This review research found three ways for predicting online course performance: i) To choose the best features to raise student performance; ii) The most effective algorithms for transforming unbalanced data into balanced data; and iii) The best machine learning algorithms to predict online course performance. This study also offered insights into using hybrid techniques and optimization algorithms to educational data sets to improve student performance prediction. The utilization of data from independent e-learning products to enhance education today requires data processing to ensure quality. In addition to these techniques, our abstract highlights the effectiveness of hybrid feature selection methods like L2 regularization (Ridge) and recursive feature elimination (RFE) and ensemble learning models like random forest, gradient boosting, and AdaBoost. These approaches considerably improve prediction accuracy and tackle huge and sophisticated educational dataset challenges. Our work uses advanced machine-learning approaches to optimize e-learning settings and boost academic achievements in the shifting online education landscape caused by the COVID-19 pandemic.

Corresponding Author:

Naga Satya Koti Mani Kumar Tirumanadham
Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research
Selaiyur, Tamil Nadu, India
Email: manikumar1248@gmail.com

1. INTRODUCTION

There has been a total game-changer with more integration of e-learning platforms into the education landscape. The transition from traditional classrooms [1]–[3] to online learning environments has brought forth the basic change in the way learners are interacting with learning resources. E-learning [4], with a variety of instructional media, developed due to the use of technologies in education, has become a flexible and easily adjustable learning environment, beyond geographical boundaries, interesting to many kinds of learners. As the world educational community continues venturing into the potential realm of e-learning, the emphasis remains on the academic achievements of the students within the digital boundaries. In view of this, therefore, understanding and improving student performance in e-learning environments [5]–[8] is essential in ensuring the effectiveness and success of online educational activities. The aspect of student performance in the e-learning environment is multidimensional and involves the engagement of the student, academic performance, and overall satisfaction of the student with the learning process. The present study looked into the complexity involved in student performance in an e-learning environment. The increasing reliance on digital platforms for education needs an in-depth analysis through the elements that affect student results and the efficiency of the present e-learning methods. This study attempts to understand the critical issues and challenges that educators, institutions, and policy-makers face. This lies in the complex elements of student performance in online contexts. As shown in Figure 1 e-learning industry has experienced tremendous growth in 2021, and the perspective for its expansion by 2026 is over $200 billion in value. The projection for the same, by 2026, is more than $375 billion. With the outbreak of the Coronavirus disease 2019 (COVID-19) pandemic, the number of e-learning users around the world surged to more than 2.6 billion . The corporate e-learning market is projected to reach $31 billion by 2027, and in higher learning, online courses are being embraced, with 6.6 million U.S. students in such classes in 2018 [9], [10]. Education is being reshaped with virtual classrooms coming onto the scene and changing the landscape with the rise of new technologies such as artificial intelligence (AI) and virtual reality (VR).

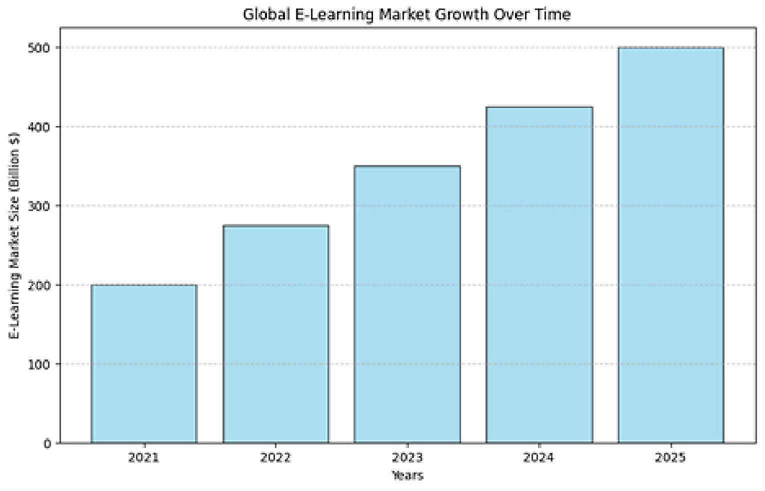

Figure 1. Global E-learning market growth over time

The outbreak of the COVID-19 pandemic which enhanced the uptake of online learning platforms levelled up educational systems. On the other hand, the implementation of these broader understanding of learning led to certain problems regarding efficient assessment of the students' progress in greatly expanded databases of education. In the context of digital learning environments, the traditional approaches to measuring the students' performance might be inadequate to call for the invention of technological advancements. This review study is in the field of e-learning and addresses three primary approaches which could help to strengthen the prediction of students' success in an online course.

Among the major constraints when attempting to forecast student outcomes is the issue of how to address the imbalance of data often present in database records of the educational environment. Predictive performance can be affected by relationship and class imbalances, which problematizes the fair representation of class members within a given class. Further, identifying the features, which are relevant to a particular problem from amongst a vast number of potential features, emerges as another major challenge. Depending on the application and the data distribution, the choice of features may greatly affect the performance of a machine learning model, and finding the most informative features may be a very challenging task.

The problem statement and existing solutions highlight the challenges in processing, analyzing, and interpreting the vast and varied educational data, as well as the large quantity of information generated within online classes, making it extremely difficult to accurately assess students' performance. Most of the traditional approaches to formative assessment are not very effective in dealing with the features of digital learning environments. As mentioned, even if other solutions, like basic machine learning algorithms and statistical methods, have been applied, they could not necessarily solve all the concerns regarding online education data.

The main goal of this paper is to use machine learning algorithms and methodologies for handling imbalanced datasets to improve the prediction of students' performance in online courses. The study seeks to accomplish the following particular goals: i) Find and assess feature selection strategies for improving student performance forecasting in e-learning settings; ii) Examine alternative techniques for handling unbalanced datasets and judge how well they affect performance forecast accuracy; and iii) For forecasting students' performance in online courses, investigate and evaluate machine learning systems.

Body will be covered during the subsequent section 2 (key territories for projecting student performance). There are three subsections in the body section. Initially, let's concentrate on feature selection. Focusing on an imbalanced dataset comes next, followed by approaches for predicting student performance. Section 3 focuses on the methodology, section 4 focuses on results and discussions and section 5 concentrates on the conclusions.

2. KEY TERRITORIES FOR PROJECTING STUDENT PERFORMANCE

2.1. Feature selection

Feature selection is a key and a crucial step in machine learning especially in electronic learning which consists of discovering suitable features that are most appropriate for the target variable from the available data to enhance the performance of the machine learning models. The importance of feature selection in the improvement of machine learning algorithms performance and computational cost [11]–[17]. Qiu *et al.* [11] described a e-learning performance estimation framework which is based on behaviour classification. The system has feature selection to extract category feature values of different behaviours and construct the learning performance predictor using machine learning methods. Besides, the feature selection is an important step in decrease the complex of machine learning algorithms and improving the model performance has been emphasised [12]. Even Farnaghi-Zadeh *et al.* [13] sentenced the importance of feature selection as the most early phase in pattern recognition and machine learning. Deep reinforcement learning, decision tree classifiers and meta-heuristic algorithms [18] have been suggested to be supplemented with feature selection for optimising feature selection and machine learning model performance [19]–[21]. The applications for feature selection are diverse. In summarizing, feature selection plays a vital role in enhancing machine learning, particularly with e-learning. It helps to improve the accuracy of the prediction, the simplicity of computations and the effectiveness of the model.

2.2. Handling imbalanced dataset

In the domain of machine learning for e-learning, this area is well-explored and a few propositions have been made in order to tackle this particular problem related to the datasets being imbalanced. Ma *et al.* [22] discussed about fuzzy techniques which is used to handle imbalanced dataset. The dataset is imbalanced in most cases, and the work by Giray *et al.* [23] applies many oversampling techniques, such as synthetic minority oversampling technique (SMOTE), adaptive synthetic sampling approach for imbalanced learning (ADASYN), Borderline-SMOTE, and ariational autoencoder (VAE), to the problem. ClassBalancer, resample, SMOTE, cost-sensitive classifier, and one-class classifier are some of the imbalanced learning algorithms that [24] worked on to check how they fare in code smell detection through machine learning. Specifically, studies done by studies [25], [26] were focused on the use of the SMOTE for imbalanced dataset resolution. In addition, Liu *et al.* [27] suggested the use of evolutionary algorithms to develop classification strategies for imbalanced absence of data. Study [28] are meanwhile engaged in developing a new approach for hospital mortality prediction that fuses the support vector machine (SVM) algorithm and the SMOTE balancing methodology.

Ouhmida *et al.* [29] also further used the SMOTE to deal with the issue of class imbalance in a hybrid system for diagnosing Parkinson's disease. This was then implemented to cater for this issue. Bujang *et al.* [30] identified possible solutions to this imbalanced categorisation problem at three levels: data-level, algorithm-level, and hybrid-level. In another study conducted by [31], the issue of unbalanced classes in multiclass datasets is considered and compared when using the SMOTE approach to finding the solutions provided by the ADASYN approach. Singh *et al.* [32] in their study tried to rebalance the skewed minority class data using oversampling techniques, which finally led to enhanced accuracy. Furthermore, Bostanci *et al.* [33] went further with the need to utilize SMOTE in alleviating the problem of imbalance in datasets through over-sampling.

Kosolwattana *et al.* [34] solved the problem of downsampling the class with so-called dominance. On the other hand, Din *et al.* [35] tackled the possibility of overfitting risk that was caused by such an imbalanced dataset that prefers the majority class data. In the same line, study [36] note that balanced class distributions are important to suggest by machine learning algorithms, and the use of resampling strategies is essential. In addition, data augmentation techniques, SMOTE, modified-SMOTE, and smoothed-bootstrap, were used by study [37] to mitigate the imbalance. In addition to that, Abdullahi *et al.* [38] have researched resampling strategies and the SMOTE-ENN strategy in the context of taking care of the problem of unbalanced class distribution. studies [39], [40] have used the SMOTE for overcoming imbalanced datasets.

2.3. Methods for prognosticating student achievement

The published research conducted between 2021 and 2024 provided valuable findings related to the utilization of machine learning models for predictions and better results in student performance within an e-learning context. Kumar *et al.* [41] provides an innovative approach, which is the E-SVM: spectral clustering algorithm based quadratic support vector machine, for learning style forecast on e-learning platforms. The web mining methodology for the extraction of the hidden information from the log files by the application of spectral clustering followed by the quadratic SVM classification shows a comprehensive approach. The credibility of the study is increased, as it applies a real-time dataset for the study and compares the efficacy of the proposed approach with the state of art methodologies. The supremacy is elaborated in terms of prediction accuracy, specificity, and sensitivity; hence, it is an attestation that the proposed E-SVM technique excels existing approaches such as fuzzy set logic-based subject matching (FSLSM), bilateral independent local binary correlation index (BILBCI), fuzzy c-means (FCM), and hybrid framework (HF). This work makes a substantial contribution to the advancement of predicting learning styles in e-learning settings.

Abdullah *et al.* [42] have used four regression machine learning methods for forecasting the academic achievement of students from e-learning logs. These are random forest (RF), Bayesian Ridge (BR), adaptive boosting (AdaBoost), and extreme gradient boosting (XGBoost). In addition, the study indicated the abilities of these algorithms in the context of rendering accurate predictions and in identifying the factors affecting student success in e-learning. Moreover, the machine learning algorithms have been applied for the purposes of prediction and assessment in such kind of fields as forecasting and assessments on the academic achievements of students in online and offline cases. Study [43] indicated the superior performance of support vector machine (SVM) over other machine-learning techniques when it came to predicting the results of the students. It specifies that the potential of SVM in understanding and predicting the academic performance of students in an e-learning environment.

3. METHOD

In Figure 2, emphasising class imbalance, feature selection, and ensemble modelling, the flowchart provides a whole machine learning approach fit for an E-learning environment. First preparing the dataset [44] which might have an unbalanced class distribution; the synthetic minority over-sampling technique (SMOTE) generates synthetic samples for the minority class, hence balancing the dataset. We then apply a hybrid feature selection approach. This employs L2 regularization (Ridge) [45], which reduces their coefficients to zero, so eliminating less important features; recursive feature elimination (RFE) [46], which iteratively removes the least important characteristics, builds models on the remaining ones to discover the optimal subset. Combining these two methods generates, via the R³FE method, an optimal feature set.

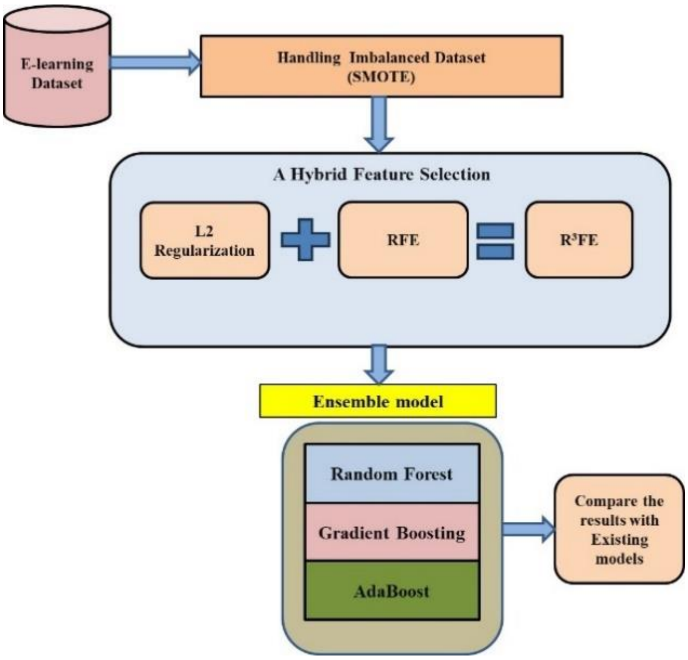
E-learning Dataset
Handling Imbalanced Dataset (SMOTE)
A Hybrid Feature Selection
L2 Regularization
RFE
R³FE
Ensemble model
Random Forest
Gradient Boosting
AdaBoost
Compare the results with Existing models

Figure 2. Schematic diagram of the proposed methodology

Then an ensemble model is generated from three strong algorithms: random forest [47], gradient boosting [48], and AdaBoost [49]. Random forest generates several decision trees and aggregates their results for increased accuracy and robustness; Gradient Boost sequentially builds models to correct the errors of the previous ones; AdaBoost combines many weak classifiers to form a strong classifier by stressing difficult-to-classify events. At finally, a comparison with current models evaluates the performance of our ensemble model. By means of effective class imbalance correction, feature selection based on relevance, and powerful ensemble methods application, this integrated approach aims to improve the performance of the model on the E-learning dataset.

3.1. Feature selection using R^{3}FE

R³FE is a hybrid feature selection method designed to identify most relevant features in a dataset by combining L2 regularization (Ridge regression) with RFE. R³FE's feature selection approach is fully justified here. Steps involved in R³FE:

a. Initial feature selection with L2 regularization (Ridge regression)

A Ridge regression is a type of linear regression adding an L2 penalty term. Big coefficients are controlled by this penalty phrase adding the values of the squared coefficients to the loss function.

$$Cost\ Function\ (CF) = \text{Residua Sum of Squares} + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (1)$$

where β_j are the model coefficients, and λ is the regularization parameter. Using this penalty, Ridge regression lowers the coefficients of less important variables towards zero but does not exactly zero. This helps to reduce the effect of multicollinearity and improves the generalisation of the model.

b. Feature importance ranking

Absolute coefficient values based on suitable Ridge regression model fit arrange the features. One finds more important larger coefficient characteristics.

c. RFE

1. Iteration process

   RFE removes least important features at every iteration.

- Train the Ridge regression with the present feature set.
- Sort the features according to their relevance; this helps one to determine the absolute values of the Ridge regression model.
- Based on the ranking, eliminate the least important feature (s).
- Proceed until either the target count of features is reached or no more features need to be eliminated.

d. Optimal feature set

This recurrent elimination strategy generates a perfect set of features largely supporting the performance of the model in prediction. While the less important features are deleted by aggregating the coefficient shrinkage impact of Ridge regression with the iterative elimination process of RFE, R³FE assures the retention of the most important ones.

R³FE improves feature selection and model performance by aggregating the best of Ridge regression with RFE. Originally maintaining all characteristics, Ridge regression effectively solves multicollinearity whereas RFE methodically reduces the feature set to determine the optimal subset. This hybrid approach employs recursive elimination and coefficient shrinkage, so it is powerful in identifying the most important properties. R³FE so not only selects a relevant subset of features, so enhancing prediction accuracy and reducing overfitting, but also allows machine learning models to be interpretable.

3.2. Ensemble model

One may rather fairly predict student performance using an ensemble model including random forest, gradient boosting, and AdaBoost. Every one of these methods has unique advantages that contribute to provide a more reliable and accurate forecast. When several decision trees are built during training and produce either mean prediction (regression) or classifier mode of the individual trees, random forest is an ensemble learning approach. Particularly in the face of noise and overfitting, it offers good accuracy and elegantly manages ever more complicated datasets.Gradient boosting creates models one after the other that each new one corrects mistakes of the prior one. Usually from decision trees, aggregating the results of numerous weak learners generates a strong prediction. This approach can help to manage complicated relationships in the data and is quite effective in optimising prediction accuracy. AdaBoost builds models sequentially as well, but it focuses on misclassified events by altering their weights, thus following models pays more attention to the hard-to-classify events. From this especially in noisy datasets, better resilience and accuracy ensue.

Combining these three approaches in an ensemble model makes use of every one of them. While random forest provides stability and lowers variance, gradient boost increases accuracy by emphasising error reduction; AdaBoost increases performance by assigning greater value to difficult situations. By aggregating several features of the data, lowering overfitting, and enhancing generalisation, this ensemble approach may produce remarkable predictive performance in student performance prediction. Random forest, gradient boosting, and AdaBoost used together provide generally superior accuracy, robustness, and interpretability than separate models, therefore providing a complete and strong instrument for assessing student performance.

4. RESULTS AND DISCUSSION

4.1. Feature selection

We examined several student traits in order to identify the most significant factors in our study on academic performance in e-learning environments. Among other things, our dataset comprised demographic data, educational stages, participation records, and academic habits. By means of a powerful feature selection process employing a hybrid strategy of L2 regularization (Ridge) and RFE, we identified 10 out of 16 major features with the greatest predictive value. Among them especially are gender, stage_id, grade_id, subject, related, raisedhands, visited_resources, announcements_view, discussion, and student absence days. These elements taken together present a whole picture of the student's profile and their participation inside the e-learning system, therefore allowing a more realistic prediction of academic performance. Although demographic and attendance records provide additional background for performance variance, the emphasis on engagement activities such raised hands, resource visits, and discussion participation underlines the requirement of active participation in the learning process.

4.2. Ensemble model

In this paper, we examined how successfully an ensemble learning approach coupled with a hybrid feature selection strategy projected academic performance in e-learning environments. The proposed model used RFE with Ridge regularization (R^3FE) for feature selection underscored by an ensemble classifier including random forest, gradient boosting, and AdaBoost. This model presented rather outstanding performance metrics with an accuracy of 97%, a precision of 96%, a recall of 98%, and an F1-score of 97%. These results reveal that the hybrid R^3FE feature selection combined with the ensemble learning model produces strong and consistent predictions of academic performance, thereby effectively capturing the important drivers and raising the general predictive accuracy as shown in Table 4.

Table 4. Performance assesment using ensemble model

Proposed model	Accuracy	Precision	Recall	F1-Score
R^3FE+Ensemble Model	97 %	96%	98 %	97%

4.3. Comparision between other methods with proposed methods

In this work, we investigated many feature selection techniques and their accuracy in forecasting academic achievement in e-learning settings with our suggested model. Existing approaches show different degrees of efficacy: block-based centroid estimation procedure (BCEP)+Bayes (97.00%), Mel-frequency cepstral coefficient (MFCC), filter bank energy (FBE) (96.36%), hybrid particle swarm optimization (PSO) and genetic algorithm (GA) with SVM (91%), deep RL feature selection with SVM (90.90%), ReliefF with deep maxout network (93.20%), and PSO with AdaBoost or ACO with XGBoost (79.22%, 69.0%). All have restrictions, though. For example, although accurate, Bayesian techniques might be sensitive to meaningless information and computationally taxing. Often utilised in voice recognition, approaches such MFCC, FBE may not be very generalizable to academic achievement prediction. While deep reinforcement learning (RL) approaches demand large training time and resources, hybrid PSO+GA with SVM might be computationally costly and prone to convergence problems. ReliefF with deep maxout network runs overfit; PSO with AdaBoost or ant colony optimization (ACO) with XGBoost exhibits reduced accuracy because of inefficiencies in feature selection and optimisation as shown in Table 5.

With L2 regularization (Ridge) combined with RFE and an ensemble learning method employing random forest, gradient boosting, and AdaBoost, our suggested model achieves a high accuracy of 97%. This hybrid R3FE+ensemble model efficiently balances feature selection and model complexity to guarantee that only the most pertinent features are chosen, hence lowering dimensionality and enhancing model interpretability. Using the strengths of several classifiers, the ensemble model produces strong and accurate forecasts.

Our model has constraints even with its great precision. It can be computationally demanding, needing time and large processing capability. While improving accuracy, the use of ensemble techniques might make the model less interpretable than simpler models, therefore confusing the knowledge of each feature contribution. Furthermore, the hybrid technique could have scalability problems with very big datasets that need for optimisation for useful deployment. Although our suggested model acknowledges its natural constraints, it provides a balanced approach by combining sophisticated feature selection with ensemble learning to reach high accuracy in estimating academic achievement.

Table 5. Comparison of feature selection existing approaches and their accuracy's with proposed model

Year	Reference	Approaches	Accuracy
2022	[14]	BCEP+Bayes	97.00%
2022	[20]	MFCC, FBE	96.36%
2022	[21]	Hybrid PSO+GA with SVM	91%
2022	[22]	Deep RL feature selection with SVM	90.90%
2024	[23]	ReliefF with deep Maxout network	93.20%
2023	[24]	PSO with AdaBoost, ACO with XGBoost	79.22%, 69.0%
	Our Work	Feature selection with hybrid L2 regularization (Ridge) and RFE (R^3FE)	97.00%

5. CONCLUSION

Several conclusions may be made based on thorough investigation of feature selection, handling unbalanced datasets, and approaches for prognosis of student performance in e-learning environments: Feature selection is very important for improving machine learning models by means of the most pertinent attributes that enable accurate prediction of results. RFE and L2 regularization (Ridge) allow to choose relevant features, therefore enhancing model interpretability and performance. Common in e-learning systems, imbalanced datasets can provide biassed models supporting majority classes. Class imbalance is efficiently addressed by techniques including SMOTE and ensemble methods (random forest, AdaBoost), so strengthening the accuracy and resilience of predictive models.

Support vector machine, gradient boosting, and ensemble approaches among other machine learning techniques show great performance in e-learning environments for estimating student performance. These systems use demographic, participation, and academic behavior-related traits to provide correct projections. Improving prediction accuracy by means of hybrid feature selection strategies (R3FE) combined with ensemble models (random forest, gradient boost, and AdaBoost) shows to be rather successful. These techniques not only enhance model performance but also assist to address the challenges given by large and complex datasets. Resilience and accuracy let the proposed hybrid R3FE+ensemble model frequently beat present approaches. It strikes a good balance between feature selection, model complexity, and interpretability in e-learning settings, therefore serving practical purposes. Building accurate and reliable prediction models in e-learning depends on finally applying advanced machine learning techniques, effective management of imbalanced datasets, and complex feature selection methods. These strategies optimise instructional interventions in online learning environments and significantly aid to raise student performance. Future research may look at scaling issues and extend these techniques to more general application in bigger educational settings.

REFERENCES

[1] A. Güllü, M. Kara, and Ş. Akgün, "Determining attitudes toward e-learning: what are the attitudes of health professional students?," *Journal of Public Health (Germany)*, vol. 32, no. 1, pp. 89–96, Dec. 2024, doi: 10.1007/s10389-022-01791-3.

[2] A. Gashi, G. Zhushi, and B. Krasniqi, "Exploring determinants of student satisfaction with synchronous e-learning: evidence during COVID-19," *International Journal of Information and Learning Technology*, vol. 41, no. 1, pp. 1–20, Oct. 2024, doi: 10.1108/IJILT-05-2022-0118.

[3] A. H. Abdullah, D. Setiana, H. Susanto, and N. Besar, "Reengineering digital education, integrated online and traditional learning, shifting paradigm of blended learning in time and post-pandemic COVID-19," in *Handbook of Research on Education Institutions, Skills, and Jobs in the Digital Era*, IGI Global, 2022, pp. 382–423.

[4] H. Aljawawdeh, "Performance tracking e-learning model: a case study," *Journal of Statistics Applications and Probability*, vol. 13, no. 1, pp. 199–210, Jan. 2024, doi: 10.18576/jsap/130114.

[5] A. Thakur, "A comparative investigation of e-learning with traditional learning," *International Journal of Advanced Research in Computer Science*, vol. 14, no. 02, pp. 26–28, Apr. 2023, doi: 10.26483/ijarcs.v14i2.6958.

[6] C. Waladi, M. Khaldi, and M. Lamarti Sefian, "Machine learning approach for an adaptive e-learning system based on Kolb learning styles," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 18, no. 12, pp. 4–15, Jun. 2023, doi: 10.3991/ijet.v18i12.39327.

[7] R. M. Tawafak *et al.*, "Analysis of e-learning system use using combined TAM and ECT factors," *Sustainability (Switzerland)*, vol. 15, no. 14, p. 11100, Jul. 2023, doi: 10.3390/su151411100.

[8] A. Fenteng, "Online learning: a cognitive tool for learning, an alternative to traditional learning style," *Psychology*, vol. 14, no. 05, pp. 676–686, 2023, doi: 10.4236/psych.2023.145036.

[9] GMI, “E-learning market trends 2023 - 2032, global report,” *Global Market Insights Inc*. https://www.gminsights.com/industry-analysis/elearning-market-size (accessed Aug. 27, 2024).

[10] Statista, “Online education - worldwide,” *Statista*. https://www.statista.com/outlook/emo/online-education/worldwide (accessed Aug. 27, 2024).

[11] F. Qiu *et al.*, “Predicting students’ performance in e-learning using learning process and behaviour data,” *Scientific Reports*, vol. 12, no. 1, Jan. 2022, doi: 10.1038/s41598-021-03867-8.

[12] Z. O. Hamad, “Review of feature selection methods using optimization algorithm,” *Polytechnic Journal*, vol. 12, no. 2, pp. 203–214, Mar. 2023, doi: 10.25156/ptj.v12n2y2022.pp203-214.

[13] F. Farnaghi-Zadeh, M. Rahmani, and M. Amiri, “Feature selection using neighborhood based entropy,” *Journal of Universal Computer Science*, vol. 28, no. 11, pp. 1169–1192, Nov. 2022, doi: 10.3897/jucs.79905.

[14] J. Pattee, S. M. Anik, and B. K. Lee, “Performance monitoring counter based intelligent malware detection and design alternatives,” *IEEE Access*, vol. 10, pp. 28685–28692, 2022, doi: 10.1109/ACCESS.2022.3157812.

[15] F. Saeed *et al.*, “Enhancing Parkinson’s disease prediction using machine learning and feature selection methods,” *Computers, Materials and Continua*, vol. 71, no. 2, pp. 5639–5657, 2022, doi: 10.32604/cmc.2022.023124.

[16] Y. E. Kim, Y. S. Kim, and H. Kim, “Effective feature selection methods to detect IoT DDoS attack in 5G Core network,” *Sensors*, vol. 22, no. 10, p. 3819, May 2022, doi: 10.3390/s22103819.

[17] R. Jegan and R. Jayagowri, “MFCC and texture descriptors based stuttering dysfluencies classification using extreme learning machine,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, pp. 612–619, 2022, doi: 10.14569/IJACSA.2022.0130870.

[18] S. Masrom, R. A. Rahman, M. Mohamad, A. S. A. Rahman, and N. Baharun, “Machine learning of tax avoidance detection based on hybrid metaheuristics algorithms,” *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, pp. 1153–1163, Sep. 2022, doi: 10.11591/ijai.v11.i3.pp1153-1163.

[19] H. A. Naman and Z. J. M. Ameen, “A new method in feature selection based on deep reinforcement learning in domain adaptation,” *Iraqi Journal of Science*, vol. 63, no. 2, pp. 817–829, Feb. 2022, doi: 10.24996/ijs.2022.63.2.35.

[20] S. Balasubramaniam, C. Vijesh Joe, C. Manthiramoorthy, and K. Satheesh Kumar, “ReliefF based feature selection and gradient squirrel search algorithm enabled deep Maxout network for detection of heart disease,” *Biomedical Signal Processing and Control*, vol. 87, p. 105446, Jan. 2024, doi: 10.1016/j.bspc.2023.105446.

[21] S. Konda, C. Goswami, J. Somasekar, K. Ramana, R. Yajjala, and N. S. K. M. K. Tirumanadham, “Optimizing diabetes prediction: a comparative analysis of ensemble machine learning models with PSO-AdaBoost and ACO-XGBoost,” in *International Conference on Sustainable Communication Networks and Application, ICSCNA 2023 - Proceedings*, Nov. 2023, pp. 1025–1031, doi: 10.1109/ICSCNA58489.2023.10370452.

[22] G. Ma, J. Lu, F. Liu, Z. Fang, and G. Zhang, “Multiclass classification with fuzzy-feature observations: theory and algorithms,” *IEEE Transactions on Cybernetics*, vol. 54, no. 2, pp. 1048–1061, Feb. 2024, doi: 10.1109/TCYB.2022.3181193.

[23] G. Giray, K. E. Bennin, Ö. Köksal, Ö. Babur, and B. Tekinerdogan, “On the use of deep learning in software defect prediction,” *Journal of Systems and Software*, vol. 195, p. 111537, Jan. 2023, doi: 10.1016/j.jss.2022.111537.

[24] F. Li, K. Zou, J. W. Keung, X. Yu, S. Feng, and Y. Xiao, “On the relative value of imbalanced learning for code smell detection,” *Software - Practice and Experience*, vol. 53, no. 10, pp. 1902–1927, Jun. 2023, doi: 10.1002/spe.3235.

[25] W. Yotsawat, K. Phodong, T. Promrat, and P. Wattuya, “Bankruptcy prediction model using cost-sensitive extreme gradient boosting in the context of imbalanced datasets,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 4, pp. 4683–4691, Aug. 2023, doi: 10.11591/ijece.v13i4.pp4683-4691.

[26] N. Rout, D. Mishra, and M. K. Mallick, “An advance extended binomial GLMBoost ensemble method with synthetic minority over-sampling technique for handling imbalanced datasets,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 4, pp. 4357–4368, Aug. 2023, doi: 10.11591/ijece.v13i4.pp4357-4368.

[27] Y. Liu, G. Li, X. Li, W. Qin, Q. Zheng, and X. Ren, “The classification method based on evolutionary algorithm for high-dimensional imbalanced missing data,” *Electronics Letters*, vol. 59, no. 12, Jun. 2023, doi: 10.1049/ell2.12842.

[28] R. Hassanzadeh, M. Farhadian, and H. Rafieemehr, “Hospital mortality prediction in traumatic injuries patients: comparing different SMOTE-based machine learning algorithms,” *BMC Medical Research Methodology*, vol. 23, no. 1, Apr. 2023, doi: 10.1186/s12874-023-01920-w.

[29] A. Ouhmida, A. Raihani, B. Cherradi, and S. Sandabad, “Parkinson’s diagnosis hybrid system based on deep learning classification with imbalanced dataset,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 3, pp. 3204–3216, Jun. 2023, doi: 10.11591/ijece.v13i3.pp3204-3216.

[30] S. D. Abdul Bujang *et al.*, “Imbalanced classification methods for student grade prediction: a systematic literature review,” *IEEE Access*, vol. 11, pp. 1970–1989, 2023, doi: 10.1109/ACCESS.2022.3225404.

[31] A. O. Widodo, B. Setiawan, and R. Indraswari, “Machine learning-based intrusion detection on multi-class imbalanced dataset using SMOTE,” *Procedia Computer Science*, vol. 234, pp. 578–583, 2024, doi: 10.1016/j.procs.2024.03.042.

[32] S. Singh, N. K. Katiyar, S. Goel, and S. N. Joshi, “Phase prediction and experimental realisation of a new high entropy alloy using machine learning,” *Scientific Reports*, vol. 13, no. 1, Mar. 2023, doi: 10.1038/s41598-023-31461-7.

[33] E. Bostanci, E. Kocak, M. Unal, M. S. Guzel, K. Acici, and T. Asuroglu, “Machine learning analysis of RNA-seq data for diagnostic and prognostic prediction of colon cancer,” *Sensors*, vol. 23, no. 6, p. 3080, Mar. 2023, doi: 10.3390/s23063080.

[34] T. Kosolwattana, C. Liu, R. Hu, S. Han, H. Chen, and Y. Lin, “A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare,” *BioData Mining*, vol. 16, no. 1, Apr. 2023, doi: 10.1186/s13040-023-00330-4.

[35] N. U. Din, L. Zhang, and Y. Yang, “Automated battery making fault classification using over-sampled image data CNN features,” *Sensors*, vol. 23, no. 4, p. 1927, Feb. 2023, doi: 10.3390/s23041927.

[36] A. K. Azlim Khan and N. H. Ahamed Hassain Malim, “Comparative studies on resampling techniques in machine learning and deep learning models for drug-target interaction prediction,” *Molecules*, vol. 28, no. 4, p. 1663, Feb. 2023, doi: 10.3390/molecules28041663.

[37] M. U. Gul, M. H. Kamarul Azman, K. A. Kadir, J. A. Shah, and S. Hussen, “Supervised machine learning based noninvasive prediction of atrial flutter mechanism from P-to-P interval variability under imbalanced dataset conditions,” *Computational Intelligence and Neuroscience*, vol. 2023, no. 1, pp. 1–18, Jan. 2023, doi: 10.1155/2023/8162325.

[38] D. S. Abdullahi, D. M. S. Aliyu, and U. Musa Abdullahi, “Comparative analysis of resampling algorithms in the prediction of stroke diseases,” *UMYU Scientifica*, vol. 2, no. 1, pp. 88–94, Mar. 2023, doi: 10.56919/usci.2123.011.

[39] S. Alija, E. Beqiri, A. S. Gaafar, and A. K. Hamoud, “Predicting students performance using supervised machine learning based on imbalanced dataset and wrapper feature selection,” *Informatica (Slovenia)*, vol. 47, no. 1, pp. 11–20, Mar. 2023, doi: 10.31449/INF.V47I1.4519.

[40] M. E. Lokanan, “Predicting mobile money transaction fraud using machine learning algorithms,” *Applied AI Letters*, vol. 4, no. 2, Apr. 2023, doi: 10.1002/ail2.85.

[41] K. N. Prashanth Kumar, B. T. Harish Kumar, and A. Bhuvanesh, “Spectral clustering algorithm based web mining and quadratic support vector machine for learning style prediction in E-learning platform,” *Measurement: Sensors*, vol. 31, p. 100962, Feb. 2024, doi: 10.1016/j.measen.2023.100962.

[42] M. Abdullah, M. Al-Ayyoub, F. Shatnawi, S. Rawashdeh, and R. Abbott, “Predicting students’ academic performance using e-learning logs,” *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 831–839, Jun. 2023, doi: 10.11591/ijai.v12.i2.pp831-839.

[43] M. S. Ramadhan and Rumondang, “Developing a learning model based on hybrid learning and PjBL,” *KnE Social Sciences*, Mar. 2023, doi: 10.18502/kss.v8i4.12912.

[44] I. Aljarah, “Students’ academic performance dataset,” *Kaggle*, 2016. *https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data/code* (accessed Aug. 27, 2024).

[45] O. Demir-Kavuk, M. Kamada, T. Akutsu, and E. W. Knapp, “Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features,” *BMC Bioinformatics*, vol. 12, no. 1, Oct. 2011, doi: 10.1186/1471-2105-12-412.

[46] R. K. Sachdeva, P. Bathla, P. Rani, V. Kukreja, and R. Ahuja, “A systematic method for breast cancer classification using RFE feature selection,” in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022*, Apr. 2022, pp. 1673–1676, doi: 10.1109/ICACITE53722.2022.9823464.

[47] H. Anantharaman, A. Mubarak, and B. T. Shobana, “Modelling an adaptive e-learning system using LSTM and random forest classification,” in *2018 IEEE Conference on e-Learning, e-Management and e-Services, IC3e 2018*, Nov. 2018, pp. 29–34, doi: 10.1109/IC3e.2018.8632646.

[48] C. Wang, L. Chang, and T. Liu, “Predicting student performance in online learning using a highly efficient gradient boosting decision tree,” in *IFIP Advances in Information and Communication Technology*, vol. 643 IFIP, Springer International Publishing, 2022, pp. 508–521.

[49] A. A. Alsulami, A. S. A. L. M. Al-Ghamdi, and M. Ragab, “Enhancement of e-learning student’s performance based on ensemble techniques,” *Electronics (Switzerland)*, vol. 12, no. 6, p. 1508, Mar. 2023, doi: 10.3390/electronics12061508.

BIOGRAPHIES OF AUTHORS

Naga Satya Koti Mani Kumar Tirumanadham is a research scholar in the Computer Science Department at Bharath Institute of Higher Education and Research in Selaiyur. He is currently working towards a Ph.D. in computer science and engineering. His main areas of interest include machine learning, deep learning, and computer networks. He finished his M.Tech. degree in computer science and engineering from JNTUK in 2017, and he completed his B.Tech. degree in IT from JNTUK in 2013. He is enthusiastic about learning and using technology to make new discoveries in these fields. He can be contacted at email: manikumar1248@gmail.com.

Thaiyalnayaki Sekhar is currently a full-time associate professor (since Nov. 2019) in the Department of Computer Science and Engineering (CSE) at Bharath Institute of Higher Education and Research. She has joined as an assistant professor (since Jan. 2008) in the Department of Computer Science and Engineering (CSE) at Dhanalakshmi Srinivasan College of Engineering and Technology, and then she was promoted to associate professor position (in June 2019) in the Department of CSE. She has more than 14 years of teaching experience. She earned his doctorate in computer science and engineering from Annamalai University, in 2019. Her research concentrated on the role of Indexing Near duplicate image detection in web search using optimization techniques. Her research interests include image processing, wireless sensor networks, machine learning and deep learning. She can be contacted at email: thaiyalnayaki.cse@bharathuniv.ac.in.

Sriram Muthal in the IT Department at Bharath Institute of Higher Education and Research, Selaiyur. He earned his doctorate in computer science and engineering from Bharath Institute of Higher Education and Research. His expertise encompasses operating systems, data mining, and computer networks. With a passion for education and research, he inspires students and contributes to advancements in IT. Driven by a commitment to excellence, he fosters a dynamic learning environment. He can be contacted at email: sriramm.cse@bharathuniv.ac.in.