

Multi-label feature aware XGBoost model for student performance assessment using behavior data in online learning environment

Shashirekha Hanumanthappa¹, Chetana Prakash²

¹Department of Computer Science and Engineering, Visvesvaraya Technological University, Mysore, India

²Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere, India

Article Info

Article history:

Received Nov 24, 2023

Revised Mar 11, 2024

Accepted Mar 21, 2024

Keywords:

Ensemble learning

Feature extraction

Feature selection

Machine learning

Multi-label classification

Student performance

ABSTRACT

In light of recent outbreaks like COVID19, the use of online-based learning streams (i.e., e-Learning systems) has increased significantly. Institutional efforts to boost student achievement have made precise predictions of academic success a priority. To analyze student sessions-streams and anticipate academic success, e-learning platforms are starting to combine data mining (DM) with machine-learning (ML) techniques. Recent research highlights the difficulties that ML-based methods have while dealing with unbalanced data. In tackling ensemble-learning, we combine several ML algorithms to select the most appropriate approach for the given data. Current ensemble-based approaches for predicting student achievement, nevertheless, don't do exceptionally well, particularly when it comes to multi-label classification, because they don't factor the relevance of features into their approaches. This study presents multi-label feature aware XGBoost (MLFA-XGB) method that improves upon the previously used ensemble-learning technique. The MLFA-XGB makes use of a robust cross-validation approach for gaining a deeper understanding of feature relationships. The experimental results demonstrate that in comparison with the state-of-the-art ensemble-based student achievement predictive approach, this suggested MLFA-XGB based approach provides much higher accuracy for prediction.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Shashirekha Hanumanthappa

Department of Computer Science and Engineering, Visvesvaraya Technological University

Mysore, India

Email: shashirekha_h2k22@rediffmail.com

1. INTRODUCTION

Educational and industrial training has shifted beyond the traditional offline method towards a digital online-mode, known as an e-learning environment [1], because of the proliferation of internet access alongside the development of technological devices. The importance within an e-learning environment was driven strongly throughout the COVID-19 outbreak when every school was switched to a fully online instructional method. A trustworthy and precise approach for predicting performance among learners [2] is difficult to provide. Academic achievement for learners can be enhanced using individualized curriculum if an efficient evaluation approach is developed by analyzing session recordings from an online learning environment.

The biggest problems with today's e-learning systems [3] originate from the fact that they don't allow for the sharing of material that can be customized to each student's unique preferences and learning approach. In order to better understand each student, educators have placed an emphasis on using adaptive

personalization strategies [4]. Machine-learning (ML) and data-mining (DM) techniques were just recently put to the mission of predicting academic success for individual students. As illustrated in Figure 1, DM has been employed to gain valuable information through the session-streams information of a particular e-learning platform's learners, which in turn has improved decision-making and increased productivity [5]. The applications of DM and ML in a variety of sectors, like organization, information security, and educational opportunities [6] show great promise. Education data-mining (EDM) [7] is a relatively new discipline that aims to improve teaching methods, learner profiles and academic outcomes [8]. Various kinds of data make up the EDM, including records of administrative actions, records of learner session-streams operation, and records of learner academic achievement. EDM datasets were made available in [9], [10], which gathered information from several online resources. They used several ML models along with an ensemble-learning technique to forecast how well students will do throughout the course of study. The results demonstrate that the ensemble approach provides the most accurate predictions. Nevertheless, these approaches failed to create a feature affecting prediction approach, leading to low accuracy in classifying whenever the information considering multi-label classification and imbalanced data problem.

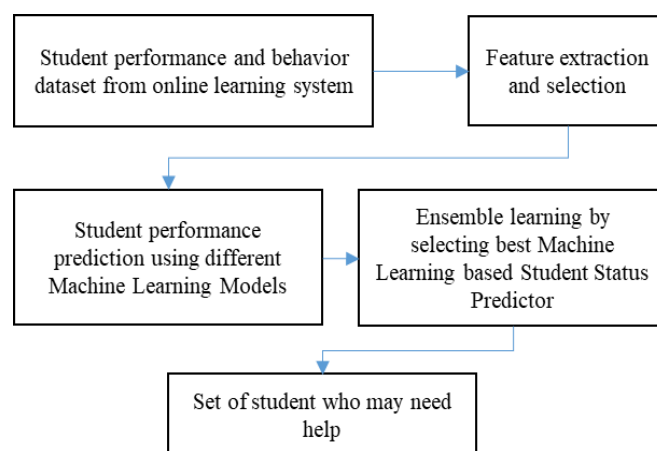


Figure 1. Architecture of proposed

The primary aim of this study is to develop a robust learner prediction approach which predicts accurately student grades throughout a given course by developing multi-label feature aware (MLFA)-XGB to effectively analyze learner session-streams e-learning information. The model is very effective in classifying the performance of students considering three different classes such as weak, average, and good. The proposed model achieves much higher accuracy than current ML and ensemble-based approaches.

Manuscript organization. In section 2 the detailed survey of various existing methodologies and their limitations have been highlighted. The proposed methodology is discussed in section 3. The experiment study using student session stream data is given in section 4. The last section significance of work is given and future research direction for enhancing student performance prediction outcomes.

2. LITERATURE SURVEY

In the literature survey section, several studies have been discussed, each focusing on predicting student performance using ML algorithms in various educational settings. These studies leverage different datasets, feature selection techniques, and ML algorithms to forecast academic achievement and enhance the quality of education. In [11], [12], the primary objective was to predict student performance at different stages of course delivery using ML. Two separate datasets representing course delivery at 20% and 50% completion stages were meticulously analyzed. The study begins with a feature analysis to gain insights into the dataset's nature, which informs the selection of ML algorithms and their parameters. A systematic approach based on the Gini index and p-value is proposed to choose a suitable ensemble learner from six potential ML algorithms. The experimental results indicate that the proposed ensemble models achieve high accuracy and low false positive rates for both datasets at all stages.

In [13]–[15], this research focuses on undergraduate datasets from two distinct universities and aims to predict student achievement at two points during course delivery. It follows a similar approach to [16],

with the selection of ML algorithms and parameter optimization. A multi-split methodology based on the Gini index and p-value is employed to optimize a bagging ensemble learner from six foundation ML algorithms. The experimental findings show that the suggested bagging ensemble models deliver good accuracy for the target group in both datasets. Figure 1 shows the proposed architecture.

The central focus of Shahzad *et al.* [17] was on predicting student performance during online interactive sessions using a dataset collected from digital electronics education and design suites. The dataset captures student interactions during online lab work, including text editing, keystrokes, time spent in activities, and exam scores per session. The research introduces a prediction model consisting of 86 statistical features, categorized into three broad groups: activity type, timing statistics, and peripheral activity count. Feature selection is used to retain influential features, and five popular classifiers, including random forest (RF) and support vector machine (SVM), are employed. The model aims to predict whether a student's performance will be low or high. Three different scenarios for model evaluation are considered, and the results demonstrate exceptional classification accuracy, with RF achieving the best performance at 97.4%. In [18]–[20], this study focuses on predicting final exam grades of undergraduate students using their midterm exam grades as source data. It employs various ML algorithms, including K-nearest neighbors (KNN), RF, SVM, naïve Bayes (NB), and logistic regression (LR), to make predictions. The dataset comprises academic achievement grades of 1854 students in a Turkish Language-I course. The proposed model, based on only three parameters (midterm exam grades, department data, and faculty data), achieved a classification accuracy of 70-75%. This study is essential for establishing a learning analysis framework in higher education and aiding in decision-making processes, particularly for identifying students at high risk of failure.

Pongpaichet *et al.* [21] introduces a ML approach to predict student performance in an online learning environment via the Maharat platform at Taif University, following online learning training standards in Saudi Arabia. Feature extraction is performed using hybrid optimization, and the SVM technique is applied for predictions. The primary objective is to forecast academic achievement and assess the quality assurance of online training programs. Descriptive-analytical methods are used to analyze sample opinions about quality assurance. This study bridges the gap between online learning standards and student performance prediction, contributing to enhancing the quality of online education. Several researchers [22], [23] propose a multi-output hybrid ensemble model that utilizes data from the superstar learning communication platform (SLCP) to predict grades. It uses the XGB model to predict mid-term and final grades, achieving an accuracy of 78.37%, surpassing comparison models. Additionally, the gradient-boosting model is employed to predict homework and experiment grades, outperforming comparison models in mean squared error. This multi-output hybrid ensemble model provides insights into how grade predictions can improve both student learning quality and teacher teaching effectiveness [24].

In summary, these studies collectively employ ML algorithms to predict student performance and enhance the quality of education. They leverage various datasets, feature selection techniques, and ensemble learning methods to achieve high accuracy in predicting academic achievements, ultimately contributing to the improvement of educational processes and outcomes. Each study offers unique insights and methodologies, catering to different educational settings and objectives. However, considering multi-label classification the current method exhibits poor accuracies. The proposed work is aimed at designing an effective method for improving accuracies in performing multi-label classification.

3. MULTI-LABEL FEATURE AWARE XGBOOST MODEL FOR STUDENT PERFORMANCE ASSESSMENT

Here, we introduce an enhanced ML approach called MLFA-XGB. Which has been developed specifically for the purpose of EDM in the context of learner session-streams as described in Figure 2. The MLFA-XGB algorithm represents an advancement over the conventional MLFA-XGB approach by incorporating a more efficient selecting features process.

The XGB approach represents an enhanced iteration of the previous gradient-boosting approach [25]. It involves the aggregation of less effective classifiers to form a robust classifier, resulting in improved classification results. Let us consider a dataset denoted as E , that represents an ongoing stream of learning session information. This dataset consists of o examples, where each sample is represented by a pair (y_j, z_j) . Here, y_j represents a vector of n features, and z_j represents a label associated with the example. The variable \hat{z}_j is utilized to denote the expected result generated by the approach in the following manner.

$$\hat{z}_j = \sum_{l=1}^L g_l(y_j), g_l \in G \quad (1)$$

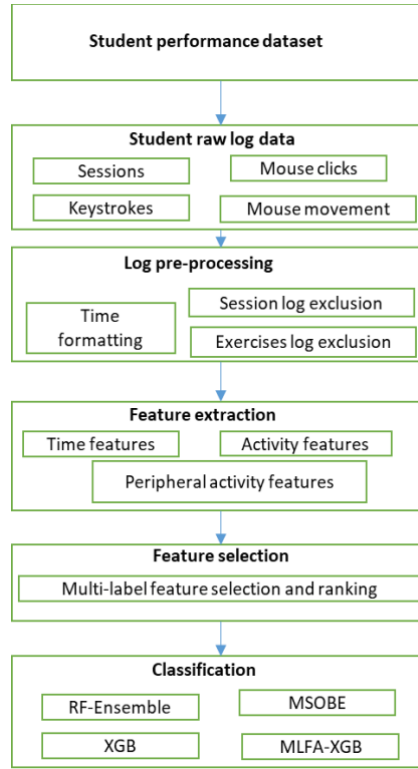


Figure 2. Architecture of proposed

The term g_l refers to an independent regression-tree, while $g_l(y_j)$ denotes the corresponding prediction results generated by the l^{th} tree for the j^{th} sample as shown in (2). The construction of any ensemble-tree is achieved by means of a summation method. The anticipated results for the j^{th} sample during the u^{th} iteration, denoted as $\hat{z}_j^{(u)}$, necessitates the inclusion of g_u in order to minimize the specified function. The evaluation of β is given as given in (4).

$$G = \{g(y) = x_{t(y)}\} \quad (2)$$

$$O^{(u)} = \sum_{j=1}^o m \left(z_j, \hat{z}_j^{(u-1)} + g_u(y_j) \right) + \beta(g_l) \quad (3)$$

$$\beta(g_l) = \delta U + \frac{1}{2} \mu \|x\|^2 \quad (4)$$

The regularization-variable is denoted by δ and μ , while the leaf's-size is represented by U . Additionally, the ranking for various leaves is denoted by x . The (3) can be reduced by employing the technique of removing the stable variable using the second-order Taylor's expanding, which can be expressed in the following manner.

$$O^{(u)} = \sum_{j=1}^o \left[h_j g_j(y_j) + \frac{1}{2} i_j g_u(y_j)^2 \right] + \beta(g_l) \quad (5)$$

The variable h_j is used to denote the initial order-gradient with regard to m , and the variable i_j is used to denote the next order-gradient with regard to m . In attaining more optimal performance with less fluctuation the work introduces a cross-entropy loss function aware gradient boosting tree as defined in (6). The parameter \hat{z}_j is calculated in (7), and for activation sigmoid operation is computed as given in (8). Then, the work introduces a K-fold cross validation for selecting and ranking feature with less training error using (9).

$$M = - \sum_{j=1}^o \left[z_j \log(\hat{z}_j) + (1 - z_j) \log(1 - \hat{z}_j) \right] \quad (6)$$

$$\hat{z}_j = \frac{1}{1 + \exp(-a_j)} \quad (7)$$

$$\frac{\partial \hat{z}_j}{\partial a_j} = \hat{z}_j (1 - \hat{z}_j) \quad (8)$$

$$CV(\sigma) = \frac{1}{SM} \sum_{s=1}^S \sum_{k=1}^K \sum_{j \in G_{-k}} P(b_j, \hat{g}_{\sigma}^{-k(j)}(y_j, \sigma)) \quad (9)$$

In (9), to select ideal $\hat{\sigma}$ for optimizing the student prediction model is attained as follows. Where M defines size of training dataset considered, $P(\cdot)$ defines loss function, and $\hat{g}_{\sigma}^{-k(j)}(\cdot)$ defines a function to compute coefficients. The proposed MLFA-XGB based student performance prediction model achieves better classification accuracy in comparison with existing ensemble-based classifier as shown in result section.

$$\hat{\sigma} = CV_s(\sigma) \quad (10)$$

4. RESULT AND ANALYSIS

This section delves into the examination of student performance prediction by employing the presented MLFA-XGB approach alongside additional established ML-based approaches for learner prediction [10], [15]. The evaluation of performance in this study utilizes the e-learning dataset obtained using [10]. The dataset preference is predicated upon the findings presented in a comparative study [9], [10], [15]. The ML approach utilized in this study for the purpose of predicting student performance has been developed with the Python 3 framework. The accuracy, specificity, sensitivity, and F1-score are metrics used for validating models. The proposed work namely MLFA-XGB is compared with existing methodologies namely multi-split optimization bagging ensemble (MSOBE) [10], RF-ensemble [15], and XGB [9], [10].

The specificity performance is given in Figure 3. The results show that the MSOBE achieves much less specificity, the XGB model achieves better performance than RF-ensemble and MSOBE. On the other side, the proposed MLFA-XGB achieves much better specificity performance than other existing student performance classification methods. The sensitivity performance is given in Figure 4. The results show that the MSOBE achieves much less sensitivity, the XGB model achieves better performance than RF-ensemble and MSOBE. On the other side, the proposed MLFA-XGB achieves much better sensitivity performance than other existing student performance classification methods.

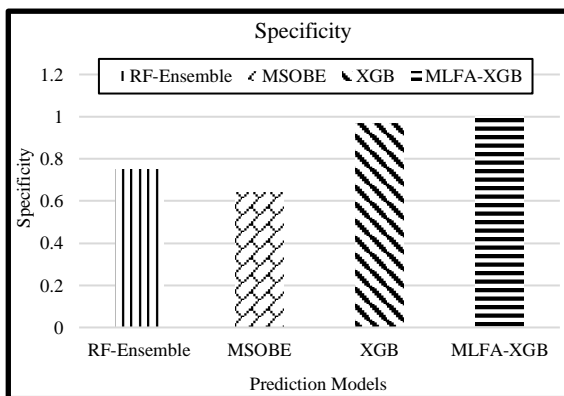


Figure 3. Specificity performance

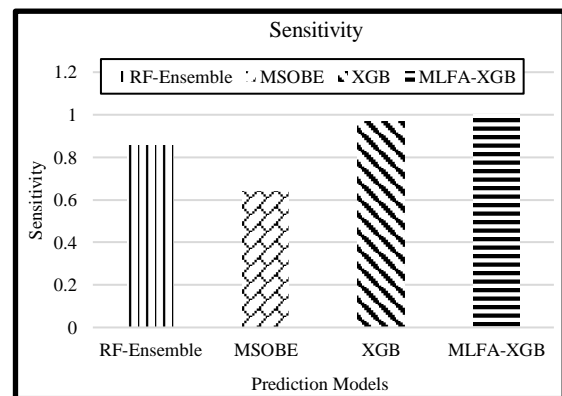


Figure 4. Sensitivity performance

The accuracy performance is given in Figure 5. The results show that the MSOBE achieves much less accuracy, the XGB model achieves better performance than RF-ensemble and MSOBE. On the other side, the proposed MLFA-XGB achieves much better accuracy performance than other existing student performance classification methods.

The F1-score performance is given in Figure 6. The results show that the MSOBE achieves much less F1-score, the XGB model achieves better performance than RF-ensemble and MSOBE. On the other side, the proposed MLFA-XGB achieves much better accuracy performance than other existing student performance classification methods.

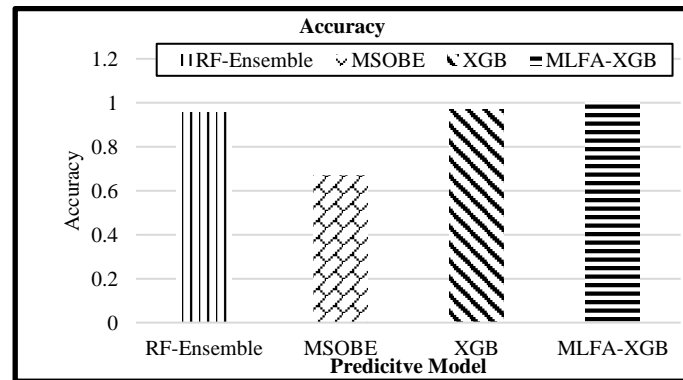


Figure 5. Accuracy performance

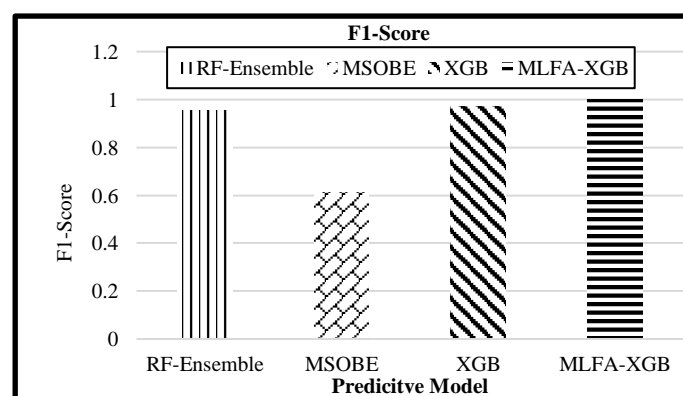


Figure 6. F-measure performance

5. CONCLUSION

The present study introduces a novel ensemble ML approach which demonstrates enhanced efficiency through the modification of the XGBoost algorithm. Notably, this method exhibits robust performance even in scenarios where the training information suffers from imbalanced class distribution. In this study, we offer a novel and efficient cross-validation method that aims to determine the specific features that have significant effects on the correctness of a prediction approach. The utilization of the CV method involves the implementation of a proficient feature ranking method, which aims to enhance the accuracy of predictions by minimizing the prediction-error. The research study was carried out utilizing a dataset consisting of conventional student-session streaming information. The MLFA-XGB method exhibits notable enhancements in terms of precision, accuracy, specificity, sensitivity, and F-measure performance when compared to existing student performance predictive approaches using RF-ensemble, MSOBE, and XGB-based approaches. The future work would be focused in enhancing the model further and also further validate the model under more diverse dataset.




REFERENCES

- [1] A. E. Tatar and D. Düşteğör, "Prediction of academic performance at undergraduate graduation: Course grades or grade point average?," *Applied Sciences*, vol. 10, no. 14, 2020, doi: 10.3390/app10144967.
- [2] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student engagement level in an e-learning environment: clustering using k-means," *American Journal of Distance Education*, vol. 34, no. 2, pp. 137–156, 2020, doi: 10.1080/08923647.2020.1696140.
- [3] S. A. Priyambada, T. Usagawa, and M. ER, "Two-layer ensemble prediction of students' performance using learning behavior and domain knowledge," *Computers and Education: Artificial Intelligence*, vol. 5, 2023, doi: 10.1016/j.caeai.2023.100149.
- [4] A. Kumar, A. Kaur, P. Singh, M. Driss, and W. Boulila, "Efficient multiclass classification using feature selection in high-dimensional datasets," *Electronics*, vol. 12, no. 10, 2023, doi: 10.3390/electronics12102290.
- [5] A. Al-Zawqari, D. Peumans, and G. Vandersteen, "A flexible feature selection approach for predicting students' academic performance in online courses," *Computers and Education: Artificial Intelligence*, vol. 3, 2022, doi: 10.1016/j.caeai.2022.100103.
- [6] K. Jawad, M. A. Shah, and M. Tahir, "Students' academic performance and engagement prediction in a virtual learning environment using random forest with data balancing," *Sustainability*, vol. 14, no. 22, 2022, doi: 10.3390/su142214795.




- [7] P. Pujar, A. Kumar, and V. Kumar, "Plant leaf detection through machine learning based image classification approach," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 1, pp. 1139–1148, 2024, doi: 10.11591/ijai.v13.i1.pp1139-1148.
- [8] S. H. Sreedhara, V. Kumar, and S. Salma, "Efficient big data clustering using adhoc fuzzy c means and auto-encoder CNN," *Lecture Notes in Networks and Systems*, vol. 563, pp. 353–368, 2023, doi: 10.1007/978-981-19-7402-1_25.
- [9] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Systems*, vol. 200, 2020, doi: 10.1016/j.knosys.2020.105992.
- [10] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Multi-split optimized bagging ensemble model selection for multi-class educational data mining," *Applied Intelligence*, vol. 50, no. 12, pp. 4506–4528, 2020, doi: 10.1007/s10489-020-01776-3.
- [11] M. L. Nistal, "An experience of continuous assessment in telecommunication technologies engineering: New costs for the teacher," *Revista Iberoamericana de Tecnologías del Aprendizaje*, vol. 8, no. 2, pp. 90–95, 2013, doi: 10.1109/RITA.2013.2258225.
- [12] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, 2022, doi: 10.1186/s40561-022-00192-z.
- [13] M. N. Alsubaie, "Predicting student performance using machine learning to enhance the quality assurance of online training via Maharat platform," *Alexandria Engineering Journal*, vol. 69, pp. 323–339, 2023, doi: 10.1016/j.aej.2023.02.004.
- [14] H. Xue and Y. Niu, "Multi-output-based hybrid integrated models for student performance prediction," *Applied Sciences*, vol. 13, no. 9, 2023, doi: 10.3390/app13095384.
- [15] G. B. Brahim, "Predicting student performance from online engagement activities using novel statistical features," *Arabian Journal for Science and Engineering*, vol. 47, no. 8, pp. 10225–10243, 2022, doi: 10.1007/s13369-021-06548-w.
- [16] E. Alhazmi and A. Sheneamer, "Early predicting of student's performance in higher education," *IEEE Access*, vol. 11, pp. 27579–27589, 2023, doi: 10.1109/ACCESS.2023.3250702.
- [17] R. Shahzad *et al.*, "Multi-agent system for student's cognitive assessment in e-learning environment," *IEEE Access*, vol. 12, pp. 15458–15467, 2024, doi: 10.1109/ACCESS.2024.3356613.
- [18] 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, 2021, doi: 10.1109/TE.2020.3008751.
- [19] P. Jiang and X. Wang, "Preference cognitive diagnosis for student performance prediction," *IEEE Access*, vol. 8, pp. 219775–219787, 2020, doi: 10.1109/ACCESS.2020.3042775.
- [20] N. A. Butt, Z. Mahmood, K. Shakeel, S. Alfarhood, M. Safran, and I. Ashraf, "Performance prediction of students in higher education using multi-model ensemble approach," *IEEE Access*, vol. 11, pp. 136091–136108, 2023, doi: 10.1109/ACCESS.2023.3336987.
- [21] S. Pongpaichet, K. Nirunwiroj, and S. Tuarob, "Automatic assessment and identification of leadership in college students," *IEEE Access*, vol. 10, pp. 79041–79060, 2022, doi: 10.1109/ACCESS.2022.3193935.
- [22] J. Figueroa-Canas and T. Sancho-Vinuesa, "Early prediction of dropout and final exam performance in an online statistics course," *Revista Iberoamericana de Tecnologías del Aprendizaje*, vol. 15, no. 2, pp. 86–94, 2020, doi: 10.1109/RITA.2020.2987727.
- [23] J. L. P.-Lujan, C. T. Calafate, J. L. P.-Yague, and J. C. Cano, "Assessing the impact of continuous evaluation strategies: tradeoff between student performance and instructor effort," *IEEE Transactions on Education*, vol. 59, no. 1, pp. 17–23, 2016, doi: 10.1109/TE.2015.2418740.
- [24] A. Smith, S. L.-Munk, A. Shelton, B. Mott, E. Wiebe, and J. Lester, "A multimodal assessment framework for integrating student writing and drawing in elementary science learning," *IEEE Transactions on Learning Technologies*, vol. 12, no. 1, pp. 3–15, 2019, doi: 10.1109/TLT.2018.2799871.
- [25] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-Aug, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

BIOGRAPHIES OF AUTHORS



Mrs. Shashirekha Hanumanthappa    currently working as Asst. Professor in the Department of Computer Science and Engineering, Visvesvaraya Technological University Centre for Post Graduation Studies, Mysuru. She has completed M.Tech. in Computer Science and Engineering from UBDT College of Engineering (Kuvempu University), Davanagere, Karnataka, India in the year 2008. Her field of interest is big data, artificial intelligence, and machine learning. She can be contacted at email: shashivtu@gmail.com.



Dr. Chetana Prakash    holds Doctor of Philosophy (Ph.D.) in Computer Science and Engineering and she is currently working as Professor in the Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere. She has teaching experience of more than 30 Years. Her field of interest is speech signal processing, data mining, image processing, fuzzy techniques, IoT, and data analytics. She can be contacted at email: chetana.p.m@gmail.com.