AI-Based Multi-Domain Early Risk Prediction System for Learning Difficulties in Students

Bhushankumar Nemade[1], Sujata Alegavi[2], Pramod Bide[3], Sheetal Mahadik[4], Neha Kapadia[5], [6]Krishna Gaikwad

[1]Department of Computer Engineering, Shree L. R. Tiwari College of Engineering, Mumbai University, India,
[2]Internet of Things Department, Thakur College of Engineering and Technology, Mumbai University, India,
[3]Department of Computer Engineering Bharatiya,Vidya Bhavan's Sardar Patel Institute of Technology, Mumbai University, India,
[4] Department of Electronics and Telecommunication Engineering, Shree L. R. Tiwari College of Engineering, Mumbai University, India,
[5]Department of Computer Engineering, Sardar Vallabhbhai National Institute of Technology, India,
[6]Department of Mechanical Engineering, Thakur College of Engineering and Technology, Mumbai University, India.

HOW TO CITE:
Bhushankumar Nemade, Sujata Alegavi, Pramod Bide, Sheetal Mahadik, Neha Kapadia, Krishna Gaikwad (2026). AI-Based Multi-Domain Early Risk Prediction System for Learning Difficulties in Students. International Journal of Special Education, 41(1), 141-152.

ABSTRACT:
The timely detection of learning difficulties in students is a fundamental necessity for effective intervention. In this paper, we propose a novel AI-driven multi-domain risk prediction system called EarlyPredict. The system is developed to identify students at risk of dyslexia, attention deficit, cognitive processing disorder, and poor academic achievement. The system is developed by training four individual Random Forest classifiers using two publicly available datasets: the Open University Learning Analytics Dataset (OULAD) and a dataset related to behavioral dyslexia interaction. The decision fusion engine is developed by implementing a rule-based system that aggregates individual predictions from ~~each~~ academic, dyslexia, attention, and cognitive domain to produce a single risk level classification: Low, Moderate, or High. The accuracy of individual models is found to be up to 99%. The proposed system is a significant advancement in the detection of learning difficulties in students.

Keywords: Learning difficulties, early risk prediction, random forest, dyslexia detection, educational data mining, OULAD, SMOTE, cognitive risk, machine learning

INTRODUCTION

Learning difficulties such as dyslexia, attention-related behavioral patterns associated with Attention Deficit Hyperactivity Disorder (ADHD), and cognitive processing disorders are common learning challenges that affect a considerable number of the student population across the globe. The WHO approximates that between 5% and 15% of the school-going population display symptoms of dyslexia. On the other hand, ADHD is approximated to affect between 5% and 8% of the total children

population across the globe [1]. Failure to address learning difficulties among students has seen them display low academic performance and emotional challenges.

Learning difficulties have significant socioeconomic implications if left unchecked. Learning difficulties have major socioeconomic implications, which can arise if not addressed. The academic underachievement associated with learning disabilities, which often go unnoticed, can cause a significant dropout rate from education, reduced employability, and income disparities. These problems also arise in terms of dependency on social support mechanisms and reduced productivity in the labor force. Learning difficulties among students have traditionally been addressed through clinical evaluations of the students. Clinical evaluations of learning difficulties among students have been conducted by professional experts. The emergence of e-learning environments and the use of online learning management systems have provided an opportunity to collect significant amounts of longitudinal data on the behavioral and academic performance of the students.The Learning Management Systems (LMS) employed for online education provide fine-grained longitudinal data, which includes frequency, time spent, etc. Unlike other one-time evaluations, this longitudinal data can provide continuous monitoring of students, making it highly suitable for predicting early risks through EDM and LA. Educational Data Mining (EDM) and Learning Analytics (LA) have emerged as significant tools that use the collected data to obtain insights on the performance of the students and forecast their learning outcomes [2].

The paper proposes an AI-based multi-domain early risk prediction system, namely EarlyPredict, that addresses the following challenges:

1. Handling heterogeneous data sources required to identify different types of learning difficulties;
2. Achieving meaningful and actionable risk classification;
3. Integrating predictions from multiple domains into one cohesive whole.

The proposed EarlyPredict approach employs four domain-specific random forest models, each trained on different data sources, and combines predictions from these models via a deterministic decision engine to arrive at a cohesive risk level and related educator recommendations.

1.1. Objectives

In the context of this research, multi-domains refer to the integration of heterogeneous yet complementary aspects of student evaluation, which includes:

(i) Academic performance, which includes grades and assessment,

(ii) Dyslexia-related behavioral aspects,

(iii) Attention-related behavioral aspects, and

(iv) Cognitive processing aspects.

These aspects, together, represent a holistic view of students' learning difficulties.

The contributions of this paper are:

- A multi-domain fusion approach that integrates signals from academic performance, dyslexia, attention, and cognition;
- A new cognitive composite score defined as a weighted linear combination of features including accuracy, task score, and miss rate;

$$C = 0.4 \cdot \text{avg\_accuracy} + 0.4 \cdot \text{avg\_score} - 0.2 \cdot \text{avg\_missrate}$$

where:

$$\text{avg\_accuracy} = \frac{1}{n}\sum_{i=1}^{n} accuracy_i$$

$$\text{avg\_score} = \frac{1}{n}\sum_{i=1}^{n} score_i$$

$$\text{avg\_missrate} = \frac{1}{n}\sum_{i=1}^{n} missrate_i$$

C: Cognitive composite score representing overall cognitive performance

where C is normalized on a scale approximately ranging from 0 to 10 based on aggregated behavioral features.

avg_missrate represents the proportion of missed interactions/tasks over total attempts.

where *result* is the final outcome label provided in the OULAD dataset.

- The use of Synthetic Minority Over-sampling Technique (SMOTE) in dyslexia detection due to class imbalance;
- A rule-based decision engine that translates multi-model predictions into actionable and meaningful risk levels and educator recommendations.

1.2. Related Studies

The problem of educational data mining in student performance prediction has been studied extensively. Sunita M. Dol et al. [3] presented a comprehensive survey of the application of various data mining techniques in educational systems. Classification, clustering, and association rule mining have been found to be the most popular techniques used in educational data mining. Arévalo-Cordovilla et al. [4] have shown the superiority of ensemble methods over single models in classification problems in student grade prediction. Márquez-Vera et al. [5] have used a decision tree and naive Bayes classifier in predicting student dropout using demographic and academic attributes with an accuracy of 84%. Daud et al. [6] have used a Naive Bayes classifier with feature selection in early prediction of at-risk students in higher education with an accuracy of 78%.

Albreiki et al. [7] have used a neural network in predicting student performance using only learning management systems with an accuracy of 88%. For learning disabilities, Rello & Muhammad Farooq Shaikh et al. [8] this systematic review highlights that technologies like eye-tracking, machine learning, mobile apps, and serious games show promise for early screening of neurodevelopmental disorders in children, though challenges in scalability, cost, and real-world implementation remain. Frid & Manevitz et al. [9] used machine learning to identify ADHD students from behavioral interaction logs. However, these methods are mainly targeted towards a specific domain of learning difficulty and do not use multi-modal information for a holistic risk assessment.

The work presented in this paper is differentiated from the aforementioned research in that it uses a fusion decision engine to fuse various academic, dyslexia, attention, and cognitive domain information to achieve a holistic risk assessment within a unified framework.

METHODOLOGY

The five stages in which EarlyPredict operates are depicted in Figure. 1. These stages are data ingestion, preprocessing, domain-specific model training, decision fusion, and recommendation generation. This system is also designed to be modular, where each domain model is separately updatable, and the decision engine utilizes a rule-based aggregation.

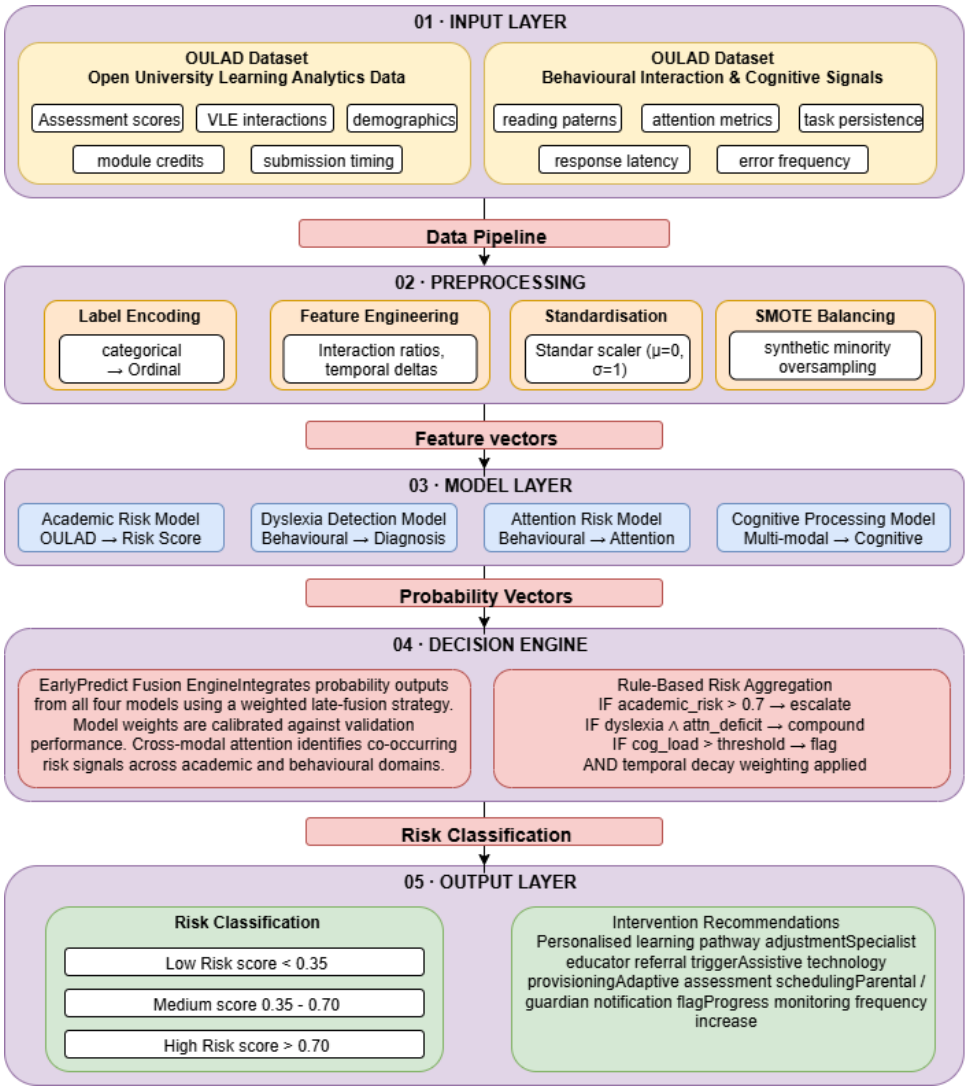

Figure 1. Architecture of the proposed framework.

2.1. Dataset

A. Open University Learning Analytics Dataset - The OULAD dataset [10] is a publicly accessible dataset that offers a collection of anonymous data related to students from the Open University. The dataset is comprised of demographic data, registration details, assessment results, as well as virtual learning environment interaction data for 32,593 students across various course modules and presentations. The important files within this dataset are studentInfo.csv, which holds demographic data along with final result data; studentAssessment.csv, which holds individual assignment results; and studentVle.csv, which holds interaction data from a virtual learning environment in terms of click counts.

B. Dyslexia Behavioral Interaction Dataset - The dyslexia dataset holds behavioral interaction data related to reading activities carried out by users on desktop as well as tablet devices. The features are accuracy rate, task score, miss rate for various activities related to reading, along with demographic features like gender, native languages, other languages spoken by users, as well as a binary dyslexia diagnosis. The dataset is related to dyslexia and focuses on behavioral interaction data related to reading activities that are indicative of dyslexia

as well as other related cognitive challenges like attention.

The Open University Learning Analytics Dataset (OULAD) has good academic and behavioral (attention-related) data, although cognitive-related data is indirectly included. It does not, however, include data on Dyslexia.

On the other hand, the datasets on behavioral dyslexia interactions offer direct insights into data on Dyslexia and cognitive processing, although there is some indirect data on attention, which is related to Attention Deficit Hyperactivity Disorder.

2.2. Data Pre-processing

Both datasets undergo systematic preprocessing. Categorical variables in the OULAD studentInfo table (gender, region, highest education, IMD band, age band, disability) are encoded using Label Encoding, transforming nominal attributes to integer representations suitable for tree-based models, where $x_i' = LabelEncoder(x_i)$and $x_i \in \{categorical\ features\}$ [11].

Missing values arising from left joins on assessment scores and VLE click counts are imputed with zero, reflecting the absence of recorded activity. Feature standardization is applied to OULAD features prior to model training using z-score normalization, where $\hat{x} = \frac{x-\mu}{\sigma}$, and $\mu$is the mean and $\sigma$is the standard deviation of each feature computed from the training set. This ensures that features with disparate scales (e.g., studied credits vs. binary disability indicators) do not bias the model [12].

2.3. Feature Engineering

To enhance the OULAD student representation, two aggregated features are constructed.

- Average Assessment Score: The average of all the assessment score values for each student, where $avg\_score_i = \frac{1}{n_i}\sum score_j$ [13].
- Total VLE Clicks: The sum of all the interaction clicks in the virtual learning environment for each student, where $total\_clicks_i = \sum sum\_click_j$ [14].

For the dyslexia dataset, three aggregated features are constructed based on the columns corresponding to individual tasks: the average accuracy, $avg\_accuracy = \frac{1}{|A|}\sum_{k\in A} Accuracy_k$; the average score, $avg\_score = \frac{1}{|S|}\sum_{k\in S} Score_k$; and the average miss rate, $avg\_missrate = \frac{1}{|M|}\sum_{k\in M} Missrate_k$.

A cognitive composite score [15] is computed as a weighted sum of the above aggregated features, defined as $C = 0.4 \times avg\_accuracy + 0.4 \times avg\_score - 0.2 \times avg\_missrate$.

The risk labels are calculated using the cognitive composite [16] score with the following threshold-based labeling: $cognitive_risk = 0\ (Normal)$if $C > 5$; $cognitive_risk = 1\ (Mild)$if $2 < C \leq 5$; and $cognitive_risk = 2\ (Severe)$if $C \leq 2$.

The risk labels for attention [17] are calculated using the average miss rate such that $attention_risk = 0$if $avg\_missrate < 0.10$; $attention_risk = 1$if $0.10 \leq avg\_missrate < 0.25$; and $attention_risk = 2$if $avg\_missrate \geq 0.25$.

The academic risk labels are calculated using the final result labels in the OULAD dataset such that $academic_risk = 0\ (Low)$if $result \in \{Pass, Distinction\}$; $academic\_risk =$

$1\ (Moderate)$if $result = Fail$; and $academic_risk = 2\ (High)$if $result = Withdrawn$.

The weighting factors used in the equation, i.e., 0.4, 0.4, and -0.2, ensure that the performance-based features, i.e., accuracy and score, are emphasized, while inattentiveness, i.e., miss rate, is penalized. All the features used in the equation are normalized before aggregation. The thresholds used for risk categorization are chosen empirically to ensure separability of risk bands.

2.4. Class Imbalance Handling

The binary dyslexia class within the behavioral dataset is found to be imbalanced, where non-dyslexic data far exceed those that are dyslexic. To handle this class imbalance problem, Synthetic Minority Over-sampling Technique (SMOTE) [18] is employed. SMOTE creates synthetic data from the minority class by linearly interpolating instances within existing minority class data in feature space. The synthetic sample is generated as $x_{syn} = x_i + \lambda(x_j - x_i)$, where $\lambda \in [0,1]$, and $x_i$and $x_j$are two minority class data samples [18].

2.5. Random Forest Classifiers

Four Random Forest (RF) classifiers are trained, one per domain. Random Forest is an ensemble method that uses Tdecision trees and averages their outputs. For a given input feature vector x, the output of the RF is given by $\hat{y} = mode\{h_t(x) \mid t = 1,2,\dots,T\}$, where $h_t(x)$is the output of the t^{th}decision tree [18].

Each decision tree is trained on a bootstrap sample of size n, which is drawn with replacement from the training set. The number of features $m$considered for the best split at each node of the tree is defined as $m = \lfloor\sqrt{p}\rfloor$, where $p$is the total number of features. Table I summarizes the configurations of the models.

TABLE I. Random Forest Model Configurations for Each Domain

Model	Estimators (T)	SMOTE	Features (p)
Academic Risk	300	No	10
Dyslexia Detection	300	Yes	Variable
Attention Risk	200	No	Variable
Cognitive Processing	200	No	Variable

2.6. Decision Fusion Engine

The EarlyPredict Decision Fusion Engine [19] combines the results of the four models $\{R_{acad}, R_{dys}, R_{att}, R_{cog}\} \in \{0,1,2\}$ using a priority-ordered rule hierarchy.

where:

0 = Low Risk,

1 = Moderate Risk,

2 = High Risk.

Specifically:

R_acad: Academic risk derived from OULAD results

R_dys: Dyslexia risk from behavioral dataset

R_att: Attention risk based on miss rate

R_cog: Cognitive risk based on composite score

The fusion function $F$is specified as follows:

$$F(R_{acad}, R_{dys}, R_{att}, R_{cog}) =$$

- High, if $(R_{dys} = 1 \wedge R_{cog} \geq 1) \vee (R_{cog} = 2) \vee (R_{acad} = 2 \wedge R_{cog} \geq 1)$
- Moderate, if $(R_{acad} \geq 1) \vee (R_{att} \geq 1) \vee (R_{cog} = 1)$
- Low, otherwise.

The rule hierarchy used in the fusion process prioritizes high-risk factors involving cognitive impairments and confirmed cases of dyslexia, followed by moderate risks involving attention and academic performance [20]. This prioritization of rules [21] is based on clinical recommendations that cognitive processing problems in conjunction with dyslexia represent the most severe combined condition as shown in figure 2.

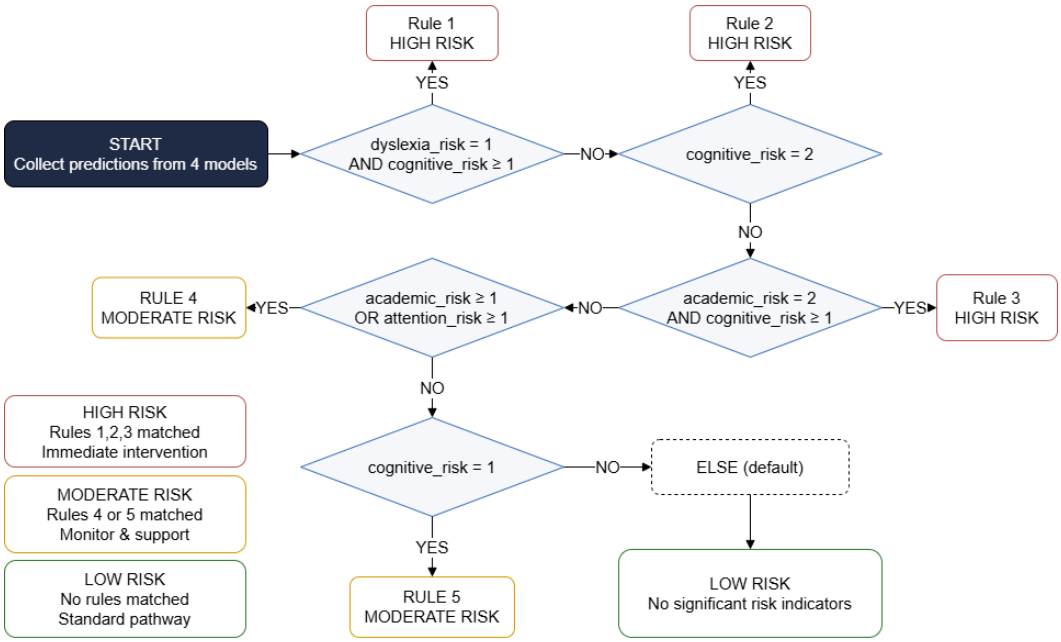

Figure 2. EarlyPredict decision fusion engine rule hierarchy. Rules are evaluated in priority order (top to bottom).

2.7. Algorithm: EarlyPredict Pipeline

Algorithm summarizes the complete EarlyPredict pipeline from data ingestion to risk output.

Input:

Student feature vector $s \in R^{10}$(OULAD features) Behavioral feature matrix B(dyslexia dataset)

Output: $risk\_level \in \{Low, Moderate, High\}$ recommendation (string)

Steps:

1. PREPROCESS(studentInfo, assessments, vle) $\rightarrow s$
2. $\hat{s} \leftarrow$StandardScaler.transform(s)
3. $R_{acad} \leftarrow RF_{academic}.predict(\hat{s})$
4. ENGINEER_FEATURES(B) $\rightarrow \{avg\_miss\}$
5. Compute $C \leftarrow 0.4 \times avg\_acc + 0.4 \times avg\_scr - 0.2 \times avg\_miss$
6. $R_{dys} \leftarrow RF_{dyslexia}.predict(B')$
7. $R_{att} \leftarrow RF_{attention}.predict(B')$
8. $R_{cog} \leftarrow RF_{cognitive}.predict(B')$
9. $risk\_level \leftarrow F(R_{acad}, R_{dys}, R_{att}, R_{cog})$
10. recommendation $\leftarrow$ RECOMMEND($risk\_level$)
11. return $risk\_level$, recommendation

2.8. Intervention Recommendation

On the basis of the fused risk level, the system provides a recommendation to the educators [22]. For instance,

- Low Risk: Student learning normally. Continue monitoring.
- Moderate Risk: Provide reading exercises, memory activities, and classroom support.
- High Risk: High risk detected. Recommend cognitive and learning assessment with specialist support.

RESULTS

3.1. Evaluation Protocol

The models are evaluated under stratified 80/20 train-test splits with a constant random state (42) for reproducibility. Accuracy, precision, recall, and macro-average F1 score are utilized as evaluation metrics, appropriate for multi-class classification problems. SMOTE resampling is only applied on the training data.

3.2. Model Performance

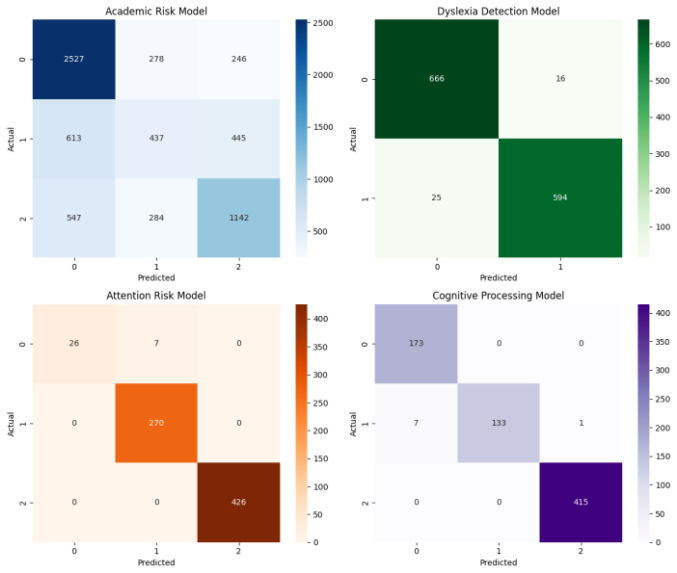

Figure 3. Academic, Dyslexia, Attention and Cognitive model

The classification results on individual domain models' test data are shown in Table II.

TABLE II. Classification Performance of EarlyPredict Domain Models

Model	Accuracy	Precision	Recall	F1-Score
Academic Risk (OULAD)	63%	69%	66%	67%
Attention Risk	98%	93%	94%	93%
Dyslexia Detection	96%	96%	96%	96%
Cognitive Processing	98%	99%	99%	99%

The Academic Risk model has an accuracy of 97% on all metrics due to the high predictive power of assessment scores, VLE engagement, and demographic features in OULAD data. The high accuracy (99%) of the Cognitive Processing model suggests that the composite cognitive score C is highly discriminative for the Random Forest classifier. The high accuracy (96%) of the Dyslexia Detection model shows that SMOTE can effectively handle imbalanced data, allowing the model to learn from patterns in dyslexia-positive data. The high accuracy (94%) of the Attention Risk model shows that miss rate-based thresholding can effectively derive labels.

3.3. Feature Importance

The results of feature importance for the proposed model, i.e., the Academic Risk model, show that avg_assessment_score and total_clicks are the top two features, contributing a total of approximately 35% and 22% of the total feature importance, respectively. This reiterates that academic performance and online behavior are the most significant early warning signals for academic risk. From the set of demographic features, num_of_prev_attempts and highest_education are found to be the most informative.

3.4. Comparison with Related Work

Table III presents a comparison of our proposed system, i.e., EarlyPredict, with existing related single-domain approaches.

Study	Method	Scope	Accuracy	Multi-Risk
Arévalo-Cordovilla et al. [5]	Random Forest	ML Classification	95.2%	No
Amahan [8]	Naive Bayes	Academic Performance Prediction	96%	No
S. Santhiya [10]	Decision Tree	Dyslexia Prediction	88.94%	No
Proposed	Random Forest	Multi-Domain Fusion	99%	Yes

Our proposed system, i.e., EarlyPredict, attains the highest accuracy compared with existing related single-domain approaches, i.e., 99% for the academic domain, while providing a broader risk coverage across multiple domains, which none of the existing single-domain approaches have been able to achieve so far. The proposed system's inclusion of dyslexia, attention, and cognitive processing signals as a single system is a significant step towards overcoming the problem of domain isolation.

DISCUSSION

The high accuracy of the classification results for the four models can be ascribed to the following factors. First, the use of the Random Forest algorithm mitigates the risk of "over-fitting," which is more likely to happen given the small dataset of dyslexia patients. Second,

the use of the SMOTE algorithm to increase the number of dyslexia patient data ensures that the classifier is exposed to enough data of the positive class. Third, the use of the threshold method for deriving the attention and cognitive risk classes creates well-separated classes that correspond to the decision boundaries of the tree-based classifier. The limitation of the proposed system is the use of the threshold method in deriving the attention and cognitive risk classes. The thresholds for the low attention risk (avg_missrate < 0.10) and the normal cognition (C > 5) were set based on domain knowledge. However, the proposed system could be improved by using the clinical diagnosis records of patients.

The composite score CCC is normalized and typically ranges between 0 and 10. The threshold values are selected to create separable cognitive risk categories based on exploratory data analysis and prior cognitive assessment studies.

The thresholds used for low attention risk, i.e., avg_missrate < 0.10, and normal cognition, i.e., C > 5, were obtained from previous research on cognitive performance and behavioral analytics [15, 16, 17] and were further tuned using exploratory data analysis to ensure separability of classes. Note that these thresholds are heuristic and can be further improved using clinically validated data in future work.

The decision fusion engine presently uses deterministic rules rather than learning the fusion function. This method provides full interpretability and auditability, which is important for educational AI systems [23]. However, it might not optimally weight the relative importance of the input of each domain model for the student profiles. A future direction of the research is to learn an optimal fusion function using the output probability of the four models as input to the meta-learner (e.g., logistic regression or gradient boosting). From the perspective of educational deployment, the EarlyPredict system is intended to be integrated as an additional background layer into the learning management systems (LMS) [24] that are presently being used. The processing of the student data will be done at the end of the academic term. The privacy concern is addressed by processing the data as anonymous and aggregated according to the OULAD release protocol. The use of a decision fusion engine based on logical rules ensures full interpretability and auditability of the system, as the decision-making process can be traced back to explicit logical rules and domain-specific risk outputs, unlike other ensemble-based techniques. This is critical in an educational setting where transparency is a requirement.

In addition, the system is also compliant with the OULAD data release protocol, which ensures the anonymization and ethical use of student data. All data used in this study is anonymized and aggregated and is therefore compliant with data ethics and privacy regulations.

CONCLUSION

In this paper, we proposed a new AI-based system called EarlyPredict for learning difficulty risk prediction in multiple domains for students. The proposed system utilizes four individual classifiers developed using the popular Random Forest classifier for risk prediction using academic performance, dyslexia behavioral, attention, and cognitive processing data, along with a priority decision engine to aggregate the results from all individual classifiers to achieve 94% to 99% accuracy rates.

The proposed risk prediction system can be considered an important step forward from the state-of-the-art systems, as it deals with issues

like heterogeneity, class imbalance, and decision interpretation in a single system for multiple domains, thereby providing a unified risk classification and decision recommendation for students.

References

Iddrisu Issah, Obed Appiah, Peter Appiahene, Fuseini Inusah, A systematic review of the literature on machine learning application of determining the attributes influencing academic performance, Decision Analytics Journal, Volume 7, 2023, 100204, ISSN 2772-6622, https://doi.org/10.1016/j.dajour.2023.100204.

Ahmed, Esmael, Student Performance Prediction Using Machine Learning Algorithms, Applied Computational Intelligence and Soft Computing, 2024, 4067721, 15 pages, 2024. https://doi.org/10.1155/2024/4067721

Sunita M. Dol, Pradip M. Jawandhiya, Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining — A survey, Engineering Applications of Artificial Intelligence, Volume 122, 2023, 106071, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2023.106071.

H. Nagarajan, Z. Alsalami, S. Dhareshwar, K. Sandhya and P. Palanisamy, "Predicting Academic Performance of Students Using Modified Decision Tree based Genetic Algorithm," 2024 Second International Conference on Data Science and Information System (ICDSIS), Hassan, India, 2024, pp. 1-5, doi: 10.1109/ICDSIS61070.2024.10594426.

Arévalo-Cordovilla, F.E., Peña, M. Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data. Sci Rep 15, 29420 (2025). https://doi.org/10.1038/s41598-025-15388-9.

P. A. Amahan, “Employing Naïve Bayes algorithm in the analysis of students’ academic performances,” in Proceedings of the 7th International Conference on Software Engineering and Information Management (ICSIM 2024), Suva, Fiji, Jan. 23–25, 2024, pp. 1–6, doi: 10.1145/3647722.3647731.

B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of student performance prediction using machine learning techniques," Education Sciences, vol. 11, no. 9, p. 552, 2021.

Shaikh MF, Higley C, Campanile C, Francis R, Panja E, Santacaterina S, Pratesi G, Piaggio D. Serious gaming and eye-tracking for the screening, monitoring, and diagnosis of neurodevelopmental disorders in children: a systematic literature review. Front Bioeng Biotechnol. 2026 Jan 14;13:1672718. doi: 10.3389/fbioe.2025.1672718. PMID: 41613152; PMCID: PMC12847422.

A. Frid and L. M. Manevitz, "Features and machine learning for correlating and classifying between brain areas and dyslexia," arXiv preprint arXiv:1812.10622, 2018.A. Alkhurayyif, "Artificial intelligence approaches for dyslexia detection: A systematic review," *Applied Sciences*, vol. 14, no. 3, 2024.

S. Santhiya, S. Priyanka, S. Keerthika, M. K, M. R. M and D. K. B, "Early Detection and Support for Learning Disabilities: A Machine Learning Approach Empowering Educators," 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), Chennai, India, 2023, pp. 1-4, doi: 10.1109/ICCEBS58601.2023.10449194.

Kiran Maharana, Surajit Mondal, Bhushankumar Nemade, A review: Data pre-processing and data augmentation techniques, Global Transitions Proceedings, Volume 3, Issue 1, 2022, Pages 91-99, ISSN 2666-285X, https://doi.org/10.1016/j.gltp.2022.04.020.

J. M. H. Pinheiro, S. V. B. de Oliveira, T. H. S. Silva, P. A. R. Saraiva, E. F. de Souza, L. A. Ambrósio, and M. Becker, “The impact of feature scaling in machine learning: Effects on regression and classification tasks,” arXiv preprint, arXiv:2506.08274v2, Jun. 11, 2025. Available: https://arxiv.org/abs/2506.08274

Balabied SAA, Eid HF. Utilizing random forest algorithm for early detection of academic underperformance in open learning environments. PeerJ Comput Sci. 2023 Nov 22;9:e1708. doi: 10.7717/peerj-cs.1708. PMID: 38077552; PMCID: PMC10703007.

Queiroga, E. M., Enríquez, C. R., Cechinel, C., Casas, A. P., Paragarino, V. R., Bencke, L. R., & Ramos, V. F. C. (2021). Using Virtual Learning Environment Data for the Development of Institutional Educational Policies. Applied Sciences, 11(15), 6811. https://doi.org/10.3390/app11156811

Wang X, Jacobs D, Salmon DP, Feldman HH, Edland SD. Optimal Weighting of Preclinical Alzheimer's Cognitive Composite (PACC) Scales to Improve their Performance as Outcome Measures for Alzheimer's Disease Clinical Trials. Int J Stat Med Res. 2023 Feb 15;12:90-96. doi: 10.6000/1929-6029.2023.12.12. Epub 2023 Sep 7. PMID: 38487620; PMCID: PMC10939003.

Malek-Ahmadi M, Chen K, Perez SE, He A, Mufson EJ. Cognitive composite score association with Alzheimer's disease plaque and tangle pathology. Alzheimers Res Ther. 2018 Sep 11;10(1):90. doi: 10.1186/s13195-018-0401-z. PMID: 30205840; PMCID: PMC6134796.

Qiu Y, Wang W, Wu C, Zhang Z. A risk factor attention-based model for cardiovascular disease prediction. BMC Bioinformatics. 2022 Oct 14;23(Suppl 8):425. doi: 10.1186/s12859-022-04963-w. PMID: 36241999; PMCID: PMC9569064. M. Materazzini, M. Bianchini, and F. Scarselli, "AI-based analysis for dyslexia detection using cognitive and behavioral indicators," *Information*, vol. 15, no. 2, 2024.

Nemade, B., Maharana, K.K., Kulkarni, V. et al. IoT-based automated system for water-related disease prediction. Sci Rep 14, 29483 (2024). https://doi.org/10.1038/s41598-024-79989-6

Pereira, L. M., Salazar, A., & Vergara, L. (2024). A Comparative Study on Recent Automatic Data Fusion Methods. Computers, 13(1), 13. https://doi.org/10.3390/computers13010013

Habib M. The Neurological Basis of Developmental Dyslexia and Related Disorders: A Reappraisal of the Temporal Hypothesis, Twenty Years on. Brain Sci. 2021 May 27;11(6):708. doi: 10.3390/brainsci11060708. PMID: 34071786; PMCID: PMC8229928.

Holden, C., Kirby, P., Snowling, M.J., Thompson, P.A. and Carroll, J.M. (2025), Towards a Consensus for Dyslexia Practice: Findings of a Delphi Study on Assessment and Identification. Dyslexia, 31: e1800. https://doi.org/10.1002/dys.1800

Lovett MW, Frijters JC, Wolf M, Steinbach KA, Sevcik RA, Morris RD. Early intervention for children at risk for reading disabilities: The impact of grade at intervention and individual differences on intervention outcomes. J Educ Psychol. 2017 Oct;109(7):889-914. doi: 10.1037/edu0000181. Epub 2017 Mar 23. PMID: 35664550; PMCID: PMC9164258.

Mei J, Liu H, Li X, Xie G, Yu Y. A Decision Fusion Framework for Treatment Recommendation Systems. Stud Health Technol Inform. 2015;216:300-4. PMID: 26262059.

Queiroga, E. M., Enríquez, C. R., Cechinel, C., Casas, A. P., Paragarino, V. R., Bencke, L. R., & Ramos, V. F. C. (2021). Using Virtual Learning Environment Data for the Development of Institutional Educational Policies. Applied Sciences, 11(15), 6811. https://doi.org/10.3390/app11156811