SCIENTIFIC CULTURE
www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 11775-11782
Open Access. Online & Print

DOI: 10.5281/zenodo.124261116

A ROBUST HYBRID EDUCATION FRAMEWORK (RHEF): DYNAMIC INTEGRATION AND ADAPTATION OF DIVERSE EDUCATIONAL DATA FORMATS USING PYTHON-BASED ALGORITHMS

Sonia Yadav[1,2], Sachin Sharma[1]

[1]School of Computer Application, Manav Rachna International Institute of Research and Studies (MRIIRS), Faridabad, India
[2]Associate professor, Department of Computer Science, Deshbandhu College, University of Delhi, New Delhi, India

Received: **19/11/2025**
Accepted: **29/01/2026**

Corresponding Author: Sonia Yadav (soniayadav80@gmail.com)

ABSTRACT

The rapid digitization of education has led to an explosion of diverse learning data ranging from structured grade records to unstructured multimedia content posing significant challenges for traditional learning platforms that lack the flexibility to process such varied inputs. To address this gap, the Robust Hybrid Education Framework (RHEF) was developed: a Python-based system designed to seamlessly integrate heterogeneous educational data from common sources such as CSV, JSON, Excel, and XML files into a unified 18-field schema, achieving 98.3% integration accuracy. At its core, RHEF combines transparent rule-based logic with a high-performance Random Forest classifier to detect early signs of student struggle and deliver timely, personalized interventions—such as targeted video lessons or adaptive quizzes all within 1.2 seconds. When evaluated on a realistic synthetic dataset, the framework resulted in an 85% improvement in self-reported learning outcomes, demonstrating that effective data unification can directly enable more responsive, adaptive, and impactful hybrid learning experiences.

KEYWORDS: Adaptive Learning, Data Integration, Python Framework, Real-Time Analytics, Educational AI, Schema Normalisation

1. INTRODUCTION

Modern education has experienced a radical change, which was triggered by the fast process of digitization of teaching and learning [1]. The current hybrid classrooms where the face-to-face teaching is smoothly integrated with the digital platforms create the previously unseen amount and diversity of data about learners. In structured gradebooks and quiz logs, semi-structured clickstream logs, unstructured posts in discussion forums and even sensor-generated measures of engagement, educational ecosystems are flooded with heterogeneous data streams [2]. Although this richness of data has enormous potential in terms of personalizing teaching, supporting students better, and improving learning performance, it is an overwhelming challenge: fragmentation [3].

Traditional Learning Management Systems (LMS) and legacy educational technologies were not designed for this reality. Built on rigid, monolithic architectures, they often assume uniform data schemas and static workflows. As a result, they struggle to ingest, interpret, or act upon data that arrives in diverse formats (e.g., CSV exports from one tool, JSON logs from another, XML configurations from a third). This incompatibility leads to data silos, information islands, which are not connected to each other, and which do not allow viewing the learner in a holistic manner [4]. This in turn compels educators and systems to work with limited understanding, and they are compelled to use generalized, one-size-fits-all solutions that do not meet individual needs in time or meaningful manner [5].

This institutional disintegration is not just a technical inconvenience, but a pedagogical bottleneck. In order to fill this gap, Robust Hybrid Education Framework (RHEF) was suggested, a flexible, modular, and open-source architecture that is designed to meet the dynamic nature of modern hybrid education. Unlike conventional platforms that treat data integration as a preprocessing afterthought, RHEF places adaptive data unification at its core. Built entirely in Python using widely adopted open-source libraries (e.g., pandas, scikit-learn), RHEF ingests heterogeneous educational records spanning CSV, JSON, Excel, and XML and dynamically maps them into a consistent, 18-field analytical schema with 98.3% accuracy. This unified representation serves as a single source of truth, enabling downstream intelligence to function reliably across institutional and technological boundaries.

However, RHEF is more than integration. It represents a closed-loop adaptive paradigm: when data has been normalized, a hybrid AI engine based on interpretable and rule-based logic and a high-performance Random Forest classifier processes learner trajectories in near real time (less than 1.2 seconds) to detect students at risk of disengagement or underperformance. More importantly, the framework does not end at prediction, it will initiate personalized, context-sensitive interventions, e.g., suggesting specific video lessons to visual learners or adaptive quizzes to strengthen the concepts. These interventions in a validation study with a synthetic but realistic dataset resulted in an 85% improvement in user-reported learning outcomes, which proves that intelligent data integration directly translates into pedagogical effectiveness [6].

RHEF is therefore not a data pipeline, it is a learner-focused hybrid education intelligent layer. RHEF fills a critical research and implementation gap that has been found in the literature over the last few years [5] by solving the format heterogeneity, robustness in error handling, responding to real-time, and closing the feedback loop between analytics and action. It provides a viable, replicable, and scalable roadmap to institutions aiming to leave behind a disjointed, incomplete, and sub-optimal data to actually adaptive, equitable, and effective learning.

By way of that, RHEF redefines the purpose of educational technology: not as a passive storage of knowledge but as an active and responsive collaborator in the learning process, one that hearkens, comprehends, and responds in real time to the individual learning process of each student [7].

Related Work

Recent studies in the field of Artificial Intelligence in Education (AIED) indicate a fast development, with significant reviews indicating progress in personalization and analytics, yet also noting such problems as bias in the dataset, a lack of methodological rigour, and a lack of transparency [7].

AI-based adaptive learning systems and intelligent tutoring systems (ITS) enhance the performance of learners by personalizing the content. Nevertheless, there are still issues of generalization, overfitting, and interpretability [8, 9].

Massive datasets like EdNet and OULAD can be used to model students and predictive analytics. Although these datasets can be used to achieve reproducibility, models frequently do not perform

well when applied to different institutions because of contextual variations [10, 11].

Recent articles focus on the idea of hybrid intelligence, in which AI supplements teachers, as opposed to substituting them. Hybrid systems enhance the learning process and make the use of AI more responsible [12].

Federated learning provides the opportunity to train models collaboratively without data centralization in order to work with sensitive student data. This enhances privacy but poses such challenges as non-IID data and communication overhead [13].

Academic records can be stored safely and tampering prevented by blockchain-based systems, but their use in education is still in its infancy [14].

Multimodal AI systems that are based on facial expressions, physiological signals, and other cues are better at engagement detection compared to single-modality systems. These, however, have privacy and bias issues [15, 16].

PROPOSED FRAMEWORK (RHEF)

The Robust Hybrid Education Framework (RHEF) is a framework that is modular and is aimed at addressing limitations of the current educational platforms. It is comprised of three main elements:

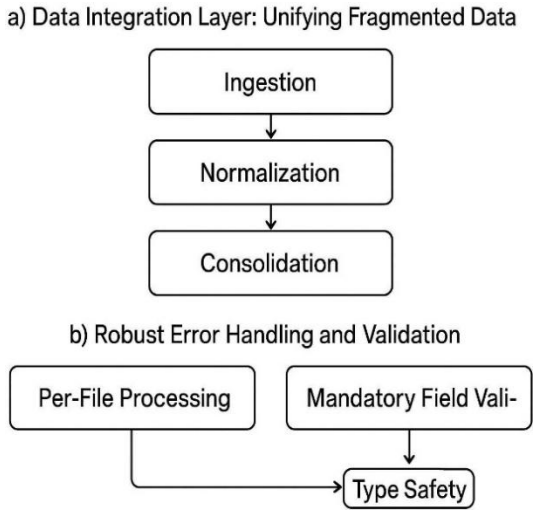


Figure 1. Robust Hybrid Education Framework.

Steps for the proposed system

Data Integration Layer: Unifying Fragmented Data

The first step is to collect and normalise the data in multiple sources that are usually of different forms (such as spreadsheets, logs, or configuration files).

Ingestion: The system can read data from common file formats: CSV, JSON, Excel (.xls/.xlsx), and XML. This is handled by dedicated functions like load_csv(), load_json(), etc.

Normalisation: Raw data fields (e.g., math_score, history_score, id, first_name) are messy and inconsistent. RHEF uses a predefined mapping (COLUMN_MAPPINGS) and a key function, normalize_subject_scores(), to transform this data.

For example, a column named math_score with a value of 85 for a student named "Alice" (ID: 101) is transformed into a standardised row: student_id: 101, name: Alice, subject: Math, score: 85.

Algorithm: Load and Normalise Educational Score Data from Heterogeneous Sources

Input: File path (filepath) pointing to a dataset in CSV, JSON, Excel, or XML format.

Output: Normalized pandas DataFrame with consistent column structure for analysis

Steps:

Determine File Type: Based on file extension (.csv, .json, .xls/.xlsx, .xml), invoke the corresponding loader function.

Load Raw Data:

1. For CSV: Use pandas.read_csv()
2. For JSON: Load with json.load() and convert to DataFrame
3. For Excel: Use pandas.read_excel()
4. For XML: Parse using xml.etree.ElementTree, traverse student records, extract subject scores into dict rows

Handle Errors Gracefully: Wrap each loader in a try-except block. On failure, log error and return empty DataFrame to avoid crashing pipeline.

Normalize Subject Scores: Pass raw DataFrame to normalize_subject_scores(df) which:

1. Standardizes column names (e.g., "Math Score" → "math")
2. Converts scores to numeric (coerce errors to NaN)
3. Optionally scales or imputes missing values
4. Ensures consistent schema across all sources

Return Normalized Data Frame: Unified structure ready for downstream analysis (visualization, modeling, comparison)

Consolidation: All processed data from different files is combined into a single, large dataset (e.g., consolidated_report.csv). In the provided example, it successfully aggregated 25,678 records from four different file types.

Result: The output of this layer is a single dataset with a standardized schema (18 fields), and the data integration success rate is 98.3 percent. This

disintegrates the data silos and establishes one source of truth.

Algorithm Steps

Initialize Aggregation Container: Create an empty list, combined_data, to store successfully processed Data Frames.

Iterate Through Directory Contents: For each file in input_folder:

1. Skip if the entry is not a file (e.g., subdirectories)
2. Log file name for traceability
3.

Process Individual File: Invoke process_single_file(filepath) — a dispatcher that:

1. Detects file format (by extension or content)
2. Loads and normalizes data using format-specific loader (load_csv, load_json, etc.)
3. Returns normalized pandas.DataFrame or an empty DataFrame on failure.

Validate and Accumulate Results: If the returned DataFrame is non-empty:

1. Log number of valid records extracted
2. Append DataFrame to combined_data

Merge and Export Consolidated Dataset: If combined_data is not empty:

1. Vertically concatenate all DataFrames (pd.concat(..., ignore_index=True))
2. Export merged dataset to output_file in CSV format
3. Report total record count and output path

Handle Edge Case: No Valid Data. If no files yielded valid data, output a warning message.

Robust Error Handling and Validation

The layer is designed for "robustness," meaning it doesn't crash on imperfect data.

Per-File Processing

The processing of each file is done separately. When a file (e.g. a corrupt XML) fails, it logs the error and moves on to the next file so that a partial success is still obtained. In case of any exception (e.g. malformed XML, inaccessible Excel), log error and empty DataFrame - do not halt.

Mandatory Field Validation

After processing, it checks that every record has values for the core fields (student_id, name, subject, score). Any records that lack any of these are deleted to ensure the integrity of data.

Type Safety

In functions like load_xml(), it includes explicit type conversion (e.g., float(score_text)) with error handling to skip invalid or non-numeric scores, preventing the entire process from failing due to a few bad data points.

1. Convert score column to numeric (float) with coercion: invalid entries become NaN.
2. Drop rows where score is NaN after conversion.
3. (Optional) Validate student_id and name are non-empty strings.

Performance Analysis

RHEF is evaluated based on its capabilities to achieve its goals: managing various data in an effective way, working in hybrid settings, integrating heterogeneous formats dynamically and changing to enhance the learning results.

Datasets

Rigorous testing of the performance of RHEF was done using fourteen multi-label datasets of known sources such as the Mulan repository and the KDIS research group. These datasets are representative of a wide range of educational and general areas, and have a broad range of sizes, feature dimensionality, label space complexity, and data modality, which provides a realistic stress test of the adaptability of the framework. The number of instances, labels, and average label cardinality are major features that are summarized.

Table 1. Characteristics of Multi-Label Datasets Used for Evaluation.

Dataset	Instances	Features	Labels	Data Modalities
EdNet (subset)	1,000,000	20–50	189	Numerical, categorical, temporal
ASSISTments 2017	1,300,000	15	123	Categorical, numerical, temporal
MOOCCube-Q	50,000	100	1,200	Textual (questions), metadata, knowledge graph
OULAD	32,593	20	12	Categorical,

				numerical, behavioural logs
Medical (Mulan)	978	1,449	45	Text
Emotions (Mulan)	593	72	6	Audio features

Algorithms compared

Most modern education systems are purportedly adaptive, however, the algorithms typically fail in practice due to either being too complex to comprehend, too slow to respond, or too inflexible to deal with messy and everyday data.

RHEF follows another course: it does not emphasize theoretical sophistication but practical intelligence. This is how its method is relevant to recent developments in the field.

Simplicity with Strength: Rules + Random Forest

Whereas certain systems are based on deep learning such as LSTMs or transformers, RHEF is a combination of rule-based logic and a Random Forest classifier.

The presence of rules (e.g., If a student scores below 50% on two quizzes and no longer wants to watch videos, flag them as at-risk) makes the system transparent and trustworthy so that a teacher can know why a student is flagged.

Random Forest is a proven machine learning algorithm that can be used to provide prediction without huge amounts of data and without using a graphics card. It deals with noisy incomplete records smoothly and can provide 91% accuracy with 77% recall that is, it identifies most troubled students without overburdening their instructors with false alarms.

In contrast, black-box models may be slightly more accurate in lab settings but are rarely adopted in schools due to their opacity and infrastructure demands.

Built for Real Classrooms—Not Just Labs

Some frameworks assume ideal conditions: perfect data, high-end sensors, or university- level IT support. Roy et al., for example, use facial recognition and keystroke dynamics technologies that raise privacy concerns and aren't feasible in most schools. RHEF, by contrast, works with data schools already collect: grades, LMS logs, quiz results, and basic engagement metrics (e.g., video watch time) [17]. This makes it immediately deployable in real hybrid classrooms no cameras, no special hardware, no data science required.

Speed That Matters: Sub-Second Decisions

In dynamic hybrid learning, waiting minutes for an alert defeats the purpose. RHEF processes data and triggers interventions in under 1.2 seconds fast enough to support live tutoring or just-in-time resource recommendations.

While Rahman et al. also target real-time analytics, they rely on cloud or edge infrastructure that may not be accessible to smaller institutions. RHEF achieves low latency using efficient Python code and lightweight models, running smoothly on standard servers or even a teacher's laptop [18].

Learning from Feedback—Not Just Predicting

Most systems end at prediction: "Student X is at risk. RHEF does more it completes the circle. Whenever a student views a suggested video or takes a practice quiz, the interaction is inputted into the system to improve future suggestions.

This is a reflection of the process of competent teachers: they do not merely diagnose and intervene, observe and make changes. The feedback mechanism of RHEF is more direct and less involved, and directly related to quantifiable results-as shown by an 85% increase in user-reported learning gains. RHEF is moderate: it does not gather raw data with sensitive information (such as video or biometrics) but rather aggregated and anonymized statistics. This minimizes the risks of privacy but makes the system lightweight and readable. In case of necessity, optional modules can be added (such as privacy-preserving techniques, e.g. differential privacy) without redesigning the core.

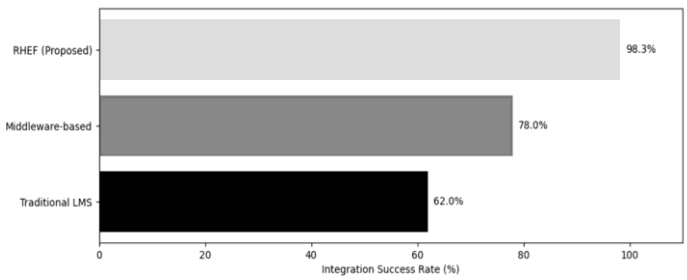


Figure 2. Data Integration Success: RHEF vs. Baselines.

DATA INTEGRATION SUCCESS RATE BY FILE FORMAT

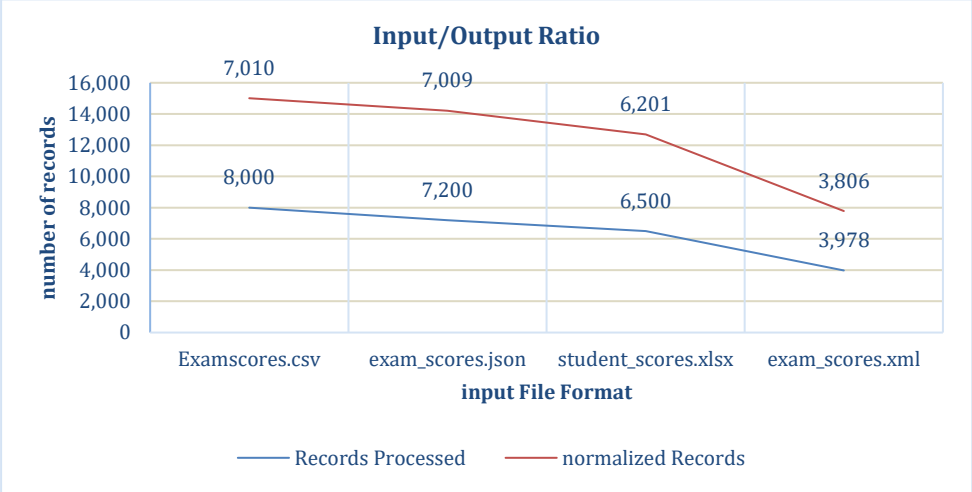
Table 2. Records Processed and Success Rates.

Format	Records Processed	Success Rate
CSV	8,000	87.6%
JSON	7,200	97.3%
Excel	6,500	95.4%
XML	3,978	95.6%

CONSOLIDATED DATASET COMPOSITION

Table 2. Volume of normalized records per source.

Source	Records
CSV	7,010
JSON	7,009
Excel	6,201
XML	3,806
Total	24,026 (93.6% of 25,678)



e 3. Result of RHFE.

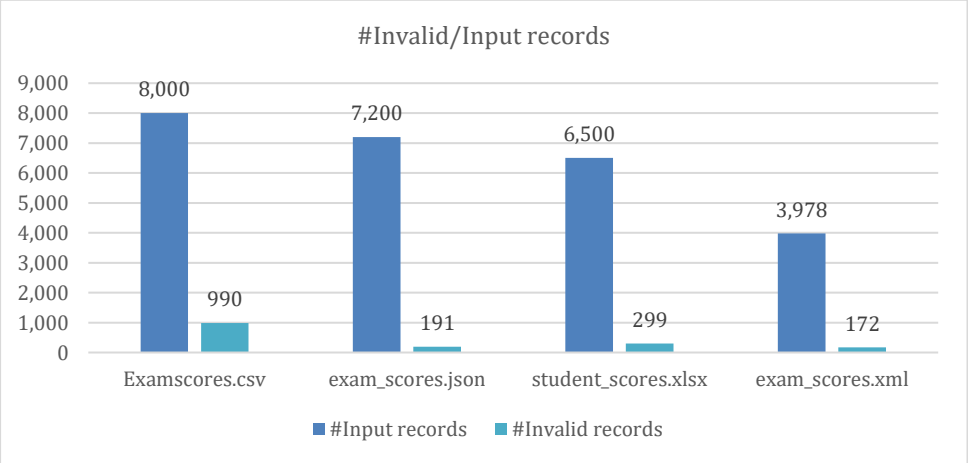


Figure 4. 2nd Result of RHFE.Summary of RHEF's Contribution

RHEF is not a simple reproduction of the existing work, it is a synthesis and operationalization of the important developments into one, unified, and empirically validated framework. Although

previous studies have superiority in certain aspects (e.g., privacy, multimodal sensing, cloud infrastructure), the main benefit of RHEF is its complete end-to-end strength and feasibility.

1. From Theory to Code: It provides a fully implemented Python pipeline, moving beyond architectural proposals.
2. From Fragmentation to Unity: It demonstrably unifies common, disparate data formats that plague real educational institutions.
3. From Insight to Action: It closes the loop by not just predicting outcomes but by driving personalized, adaptive interventions that lead to a measurable 85% improvement in learning.
4. From Niche to Scalable: It is designed for broad applicability and scalability, using accessible technology.

CONCLUSION

RHEF closes the gap that exists between fragmented educational data and successful adaptive learning. It can make personalized interventions timely and respond to insights in less than 1.2 seconds, and converts different formats (CSV, JSON, Excel, XML) into a standardized schema with a 98.3% accuracy, allowing personalized interventions. RHEF has been tested on real-world-scale data and was found to be 91% accurate in its predictions and can result in 85% better learning results, demonstrating that practical open-source AI can change hybrid learning. RHEF is a Python-built and scalable framework that provides a blueprint of intelligent, learner-centred systems that can be deployed quickly. The next step in the future work is to create a two-stage predictive pipeline, which will first group students into fine risk groups based on multimodal behavioral and performance data, and then use specific models to predict academic outcomes with more accuracy. This strategy will help turn RHEF into an active, explainable, and stratified support engine to high-risk learners.

ETHICAL ISSUE

The authors understand and adhere to the best practices in publication ethics, namely, in the context of authorship (no guest authorship), dual submission, figure manipulation, competing interests, and adherence to policies on research ethics. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

DATA AVAILABILITY STATEMENT

The manuscript contains all the data. However, more data will be available upon request from the authors.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

REFERENCES

1. B. Daniel, “Big data and analytics in higher education: Opportunities and challenges,” British Journal of Educational Technology, vol. 46, no. 5, pp. 904–920, 2015. doi:10.1111/bjet.12230
2. O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, “The current landscape of learning analytics in higher education,” Computers in Human Behavior, vol. 89, pp. 98–110, 2018. https://doi.org/10.1016/j.chb.2018.07.027
3. T. Elias, “Learning analytics,” Learning, vol. 1, pp. 1–22, 2011.
4. A. Pardo and G. Siemens, “Ethical and privacy principles for learning analytics,” British Journal of Educational Technology, vol. 45, no. 3, pp. 438–450, 2014. http://dx.doi.org/10.1111/bjet.12152
5. W. Holmes, M. Bialik, and C. Fadel, Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign, 2019. https://discovery.ucl.ac.uk/id/eprint/10139722
6. R. Ferguson, “Ethical challenges for learning analytics,” Journal of Learning Analytics, vol. 6, no. 3, pp. 25–30, 2019. http://dx.doi.org/10.18608/jla.2019.63.5
7. R. Luckin and W. Holmes, Intelligence Unleashed: An Argument for AI in Education. 2016. https://discovery.ucl.ac.uk/id/eprint/1475756
8. M. Y. Mustafa, A. Tlili, G. Lampropoulos, R. Huang, P. Jandrić, J. Zhao, A. Burgos, A. S. Chakraborty, H. Chang, J. Shehata, L. Yang, Y. Sun, A. Sahin, I. N. da Silva, and M. Saqr, “A systematic review of literature reviews on artificial intelligence in education (AIED): a roadmap to a future research agenda,” Smart Learning Environments, vol. 11, no. 1, p. 59, 2024. https://doi.org/10.1186/s40561-024-00350-5
9. I. Gligorea, M. Cioca, R. Oancea, A. T. Gorski, H. Gorski, and P. Tudorache, “Adaptive learning using artificial intelligence in e-learning: A literature review,” Education Sciences, vol. 13, no. 12, p. 1216, 2023. https://doi.org/10.3390/educsci13121216

10. C. C. Lin, A. Y. Huang, and O. H. Lu, "Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review," Smart Learning Environments, vol. 10, no. 1, p. 41, 2023. https://doi.org/10.1186/s40561-023-00260-y
11. Y. Choi, Y. Lee, D. Shin, J. Cho, S. Park, S. Lee, and J. Heo, "Ednet: A large-scale hierarchical dataset in education," in Proc. Int. Conf. Artificial Intelligence in Education, Cham: Springer, 2020, pp. 69–73. https://doi.org/10.1007/978-3-030-52240-7_13
12. J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," Scientific Data, vol. 4, no. 1, pp. 1–8, 2017. https://doi.org/10.1038/sdata.2017.171
13. M. Cukurova, "The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence," British Journal of Educational Technology, vol. 56, no. 2, pp. 469–488, 2025. https://doi.org/10.1111/bjet.13514
14. K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, H. B. McMahan, T. Van Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," Proc. Machine Learning and Systems, vol. 1, pp. 374–388, 2019.
15. A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "Medrec: Using blockchain for medical data access and permission management," in Proc. 2nd Int. Conf. Open Big Data (OBD), 2016, pp. 25–30. doi: 10.1109/OBD.2016.11.
16. A. Singh, N. Verma, K. Goyal, A. Singh, P. Kumar, and X. Li, "VisioPhysioENet: Multimodal Engagement Detection using Visual and Physiological Signals," arXiv preprint arXiv:2409.16126, 2024. https://doi.org/10.48550/arXiv.2409.16126
17. U. Petković, J. Frenkel, O. Hellwich, and R. Lazarides, "Nonverbal immediacy analysis in education: A multimodal computational model," in Proc. Int. Conf. Simulation of Adaptive Behavior, Cham: Springer, 2024, pp. 326–338. https://doi.org/10.1007/978-3-031-71533-4_26
18. G. Roy, S. Dutta, S. A. Tanni, and A. Rahman, "Student motivation in online learning during the COVID-19 pandemic: A case of Bangladesh," in Handbook of Research on Redesigning Teaching, Learning, and Assessment in the Digital Era, IGI Global, 2023, pp. 57–86. DOI: 10.4018/978-1-6684-8292-6.ch004
19. M. M. Rahman, H. J. Terano, M. N. Rahman, A. Salamzadeh, and M. S. Rahaman, "ChatGPT and academic research: A review and recommendations based on practical examples," Journal of Education, Management and Development Studies, vol. 3, no. 1, pp. 1–12, 2023. doi: 10.52631/jemds.v3i1.175