IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Digital Behavior Profiling of Students to Measure Academic Performance Using Machine Learning

Yashvant Vilas Sathe
Sem – IV | Masters Degree Programme (Data Science)
yashsathe007@gmail.com

Abstract: *The primary objective of this research project is to develop and evaluate a Machine Learning-based Digital Behaviour Profiling System using the Open University Learning Analytics Dataset (OULAD) — a publicly available, real-world dataset containing anonymised demographic data, VLE (Virtual Learning Environment) interaction logs, and academic performance records of 32,593 students across 22 course presentations at the Open University, UK. The system is designed to analyse students' digital engagement patterns, construct interpretable behaviour profiles through unsupervised clustering, and accurately predict final academic outcomes through supervised classification modelling.*

Keywords: Digital Behavior Profiling of Students

I. OBJECTIVE

1.1 Overall Objective

The primary objective of this research project is to develop and evaluate a Machine Learning-based Digital Behaviour Profiling System using the Open University Learning Analytics Dataset (OULAD) — a publicly available, real-world dataset containing anonymised demographic data, VLE (Virtual Learning Environment) interaction logs, and academic performance records of 32,593 students across 22 course presentations at the Open University, UK. The system is designed to analyse students' digital engagement patterns, construct interpretable behaviour profiles through unsupervised clustering, and accurately predict final academic outcomes through supervised classification modelling.

1.2 Specific Objectives

1. To extract, clean, and engineer composite digital behavioural features from the OULAD VLE clickstream logs, assessment records, and student registration data.
2. To apply unsupervised clustering algorithms to segment OULAD students into interpretable behaviour profile categories based on their digital engagement patterns.
3. To train and comparatively evaluate multiple supervised ML classifiers — including Random Forest, XGBoost, SVM, ANN, and Logistic Regression — for predicting the four OULAD outcome classes: Distinction, Pass, Fail, and Withdrawn.
4. To identify the most predictive behavioural features using SHAP (SHapley Additive exPlanations) value analysis.
5. To design and validate an Early Warning System that flags at-risk (Fail/Withdrawn) students as early as 30–50% into the course timeline.
6. To evaluate model performance using accuracy, precision, recall, F1-score, and ROC-AUC, benchmarked against published OULAD studies.

II. INTRODUCTION / BACKGROUND

The proliferation of digital learning environments has fundamentally transformed how academic institutions collect, store, and potentially leverage student behavioural data. The Open University Learning Analytics Dataset (OULAD), published in 2017 by Kuzilek, Hlosta, and Zdrahal in Nature Scientific Data, represents one of the most comprehensive

DOI: 10.48175/IJARSCT-36886

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

and widely cited real-world educational datasets available for public research. The dataset contains anonymised records from 32,593 students enrolled in 22 presentations of 7 modules at the Open University (OU), UK — one of the world's largest distance-learning institutions that, as of 2022/23, was actively teaching 150,619 students.

OULAD uniquely combines three categories of data: demographic and registration information (including gender, region, disability status, prior education level, and IMD band); assessment records (TMA and CMA scores across each course presentation); and, most critically, Virtual Learning Environment (VLE) interaction logs — comprising 10,655,280 entries recording daily summaries of student clicks on course materials, forums, quizzes, resources, and learning activities. This rich combination of demographic, behavioural, and outcome data makes OULAD ideally suited for the development and validation of a digital behaviour profiling system for academic performance prediction.

Under India's National Education Policy (NEP 2020) and the broader global push for data-driven personalised education, the ability to identify academically at-risk students early in their course journey — and to understand the behavioural patterns that differentiate high-performing from struggling learners — is of critical institutional and societal importance. OULAD provides the empirical foundation required to develop such a system with scientific rigour and reproducibility, grounded in real student interaction data rather than simulated proxies.

2.1 OULAD Dataset Overview

The OULAD is structured across seven interrelated CSV files: studentInfo.csv (student demographics and final results), studentRegistration.csv (module enrolment dates), studentAssessment.csv (assessment submission scores), assessments.csv (assessment metadata), courses.csv (course metadata), vle.csv (VLE material descriptions), and studentVle.csv (daily clickstream interaction logs). The target variable — final_result — takes four values: Distinction (highest performance), Pass, Fail, and Withdrawn (student disengaged before course completion). The class distribution across the full dataset is approximately: Withdrawn 31.5%, Pass 34.4%, Fail 16.1%, Distinction 18.0%.

2.2 Motivation and Problem Statement

Despite the availability of rich VLE interaction data, a substantial proportion of distance-learning students at the Open University withdraw from or fail their courses each academic year. Studies indicate MOOC dropout rates can range from 80% to 95%. The OULAD data enables the design of early warning systems that can identify students at risk of withdrawal or failure weeks before the end of the course, allowing institutions to deploy timely, targeted support interventions. The present study addresses this challenge by building a comprehensive end-to-end ML pipeline that integrates behaviour profiling, predictive modelling, and early warning alert generation using real OULAD data.

III. RELATED WORKS / LITERATURE SURVEY

3.1 Foundational OULAD Studies

Kuzilek, Hlosta, and Zdrahal (2017) introduced the OULAD in Nature Scientific Data, describing the dataset's structure, collection methodology, and intended research applications. This foundational paper established the dataset as the primary benchmark for learning analytics research in distance education, enabling reproducible, comparable studies across research groups worldwide.

Wasif et al. (2019) conducted one of the early comprehensive ML evaluations on OULAD, comparing Gaussian Naive Bayes, Random Forest, Logistic Regression, and Multilayer Perceptron models for binary pass/fail classification. Their Random Forest model achieved the highest performance with 89% accuracy, 89% precision, 88% recall, and 88% F1-score, establishing Random Forest as a strong baseline for OULAD prediction tasks.

3.2 Machine Learning Approaches on OULAD

Torkhani and Rezgui (2023) conducted a comprehensive comparative evaluation of ML and deep learning techniques on OULAD for multi-class student performance prediction. Their study implemented Logistic Regression, Decision

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Trees, Random Forests, SVMs, CNNs, RNNs, and LSTM networks. Among deep learning models, LSTM achieved the highest accuracy of 83.41% and precision of 82.20%. Random Forest and optimised Decision Tree models provided the strongest balance between accuracy and recall among traditional ML approaches.

Jha, Ghergulescu, and Moldovan (2019) demonstrated that models based on students' VLE interaction data achieved the best performance for both dropout and result prediction on OULAD, with AUC scores ranging from 0.91 to 0.93 using Gradient Boosting Machine. Their study confirmed that clickstream behavioural features are substantially more predictive than demographic features alone.

Adnan et al. (2021) developed a prediction framework using Deep Feed Forward Neural Networks to predict the four-class OULAD outcome (Distinction, Pass, Fail, Withdrawn) at five different time points during the course, demonstrating that early prediction (at 30–50% course completion) achieves accuracy comparable to end-of-course predictions when VLE interaction features are incorporated.

3.3 Clustering and Behaviour Profiling on OULAD

Research exploring the OULAD dataset for student engagement and self-regulated learning (SRL) analysis has employed K-Means, Expectation-Maximisation (EM), and Agglomerative Clustering algorithms to identify student behaviour profiles. Results consistently demonstrate a positive correlation between student engagement intensity with VLE resources and academic performance, confirming that greater interaction with learning resources corresponds to better academic outcomes and is indicative of self-regulated learning behaviour.

Kovanović et al. (2015) applied clustering to VLE time-on-task data and identified distinctive engagement archetypes. Their study introduced the concept of 'intensive learners' and 'sporadic learners' — clusters that map closely to the Proactive Learner and Last-Minute Learner profiles developed in the present study. The Silhouette Scores reported for these cluster solutions ranged from 0.62 to 0.74.

3.4 Feature Engineering from VLE Data

Liu et al. (2022) demonstrated that predictive accuracy on OULAD is substantially enhanced when raw clickstream counts are transformed into engineered temporal features capturing study consistency, deadline proximity behaviour, and resource diversity. Their study achieved accuracy improvements of 6–12 percentage points over raw feature baselines using engineered composite metrics, directly supporting the feature engineering approach adopted in the present study.

Rizvi et al. (2019) established that time-on-task estimated from VLE daily click summaries is among the most consistently predictive behavioural variables across OULAD courses, reinforcing the inclusion of session duration proxies in the feature engineering pipeline.

3.5 Early Warning Systems Using OULAD

Adnan et al. (2021) demonstrated that at-risk student prediction at 30% course completion using OULAD VLE data achieves AUC scores of 0.80–0.85, rising to 0.88–0.92 by 50% completion. This establishes a validated benchmark for early warning system performance that the present study targets to match or exceed.

Waheed et al. (2020) developed a neural network-based early prediction system for self-paced OULAD courses, achieving an AUC of 0.91 for at-risk identification using only the first 30% of course clickstream data. Their study confirmed that temporal engagement consistency is the most critical early predictor of eventual course withdrawal.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

3.6 Gradient Boosting and Ensemble Methods

Niyogisubizo et al. applied ensemble methods combining XGBoost and Random Forest for OULAD prediction, achieving accuracy improvements of 3–5% over single-model baselines. The TRV-SVM model by Kumar et al. (2021) achieved 93.8% accuracy for at-risk prediction and 93.5% for marginal student identification on OULAD, while reducing training time by 59% — establishing a high-performance benchmark for SVM-based approaches. Alsariera et al. (2022) confirmed Random Forest as consistently achieving the highest accuracy among classical classifiers for OULAD pass/fail prediction.

3.7 Demographic vs. Behavioural Feature Comparison

Multiple OULAD studies have demonstrated that including VLE behavioural data substantially outperforms demographic-only models. IEEE publication by researchers in 2022 showed that Decision Tree, Random Forest, Gradient Boosting, and KNN classifiers each showed greater accuracy when VLE interaction data was included alongside demographic features, confirming that digital behaviour is the primary signal for academic performance prediction rather than student background characteristics.

3.8 Research Gaps Addressed by This Study

The review of existing OULAD literature reveals five key gaps addressed by the present study: (1) most studies perform binary classification (pass vs. withdraw) rather than full four-class prediction; (2) no existing OULAD study integrates both unsupervised behaviour clustering and supervised performance prediction within a single unified pipeline; (3) few studies provide a fully reproducible end-to-end system architecture; (4) SHAP-based interpretability analysis of OULAD behavioural features is underrepresented in the literature; and (5) most published results use single-course OULAD subsets rather than multi-course aggregated datasets.

IV. METHODOLOGY

4.1 Research Design

This study employs a quantitative, experimental research design grounded in the OULAD real-world dataset. The research follows a structured ML pipeline comprising data ingestion, exploratory data analysis, feature engineering, unsupervised behaviour clustering, supervised multi-class classification modelling, model evaluation, and early warning system validation. A mixed-methods validation approach is adopted, combining quantitative performance metrics with interpretive SHAP feature analysis.

4.2 Dataset Description — OULAD

Objective

The Open University Learning Analytics Dataset (OULAD) was downloaded from the official repository at https://analyse.kmi.open.ac.uk/open_dataset under a CC-BY 4.0 open licence. The dataset contains:

- 32,593 unique students across 22 course presentations of 7 modules (AAA, BBB, CCC, DDD, EEE, FFF, GGG).
- 10,655,280 VLE interaction log entries (studentVle.csv) representing daily click summaries per student per VLE material.
- 173 unique types of VLE learning materials/activities including course resources, forums, quizzes, wikis, HTML pages, and subpages.
- 4 final outcome classes: Distinction (18.0%), Pass (34.4%), Fail (16.1%), Withdrawn (31.5%).
- Demographic features: gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, disability, studied_credits.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

For this study, the DDD course module (two presentations: 2013J and 2014J) was selected as the primary analysis corpus, consistent with multiple benchmark studies in the OULAD literature, yielding a working dataset of 6,272 student-course records after preprocessing and quality filtering.

4.3 Data Preprocessing

Raw OULAD data underwent a six-stage preprocessing pipeline: (1) merging of studentInfo, studentVle, studentAssessment, and assessments tables via student_id and code_module/code_presentation keys; (2) removal of records with missing final_result values; (3) aggregation of daily VLE click logs into per-student per-material-type totals and temporal distribution statistics; (4) imputation of missing assessment scores using the course median; (5) application of SMOTE (Synthetic Minority Over-Sampling Technique) to address class imbalance, particularly for the Fail class (16.1%); and (6) MinMaxScaler normalisation of all numerical features to the [0,1] range.

4.4 Feature Engineering

Twenty-eight composite behavioural features were engineered from the raw OULAD tables, grouped into five categories:

Feature Category	Engineered Features	Source Table
Engagement Volume	total_clicks, resource_clicks, forum_clicks, quiz_clicks, video_clicks	studentVle
Temporal Consistency	active_days_ratio, weekly_std_dev, pre_deadline_spike_ratio	studentVle
Assessment Behaviour	avg_score, submission_delay_days, attempts_per_TMA, score_trend	studentAssessment
Learning Depth	material_diversity_index, repeat_visit_ratio, content_progression_rate	studentVle + vle
Demographic Context	imd_band_encoded, prior_attempts, studied_credits, disability_flag	studentInfo

Table 1: Engineered Feature Categories from OULAD Tables

4.5 Unsupervised Behaviour Clustering

K-Means clustering with k=5 was applied to the 28 engineered features (demographic context features excluded for pure behaviour profiling). The optimal k was determined using the Elbow Method (inertia curve) and confirmed by Silhouette Score analysis. MiniBatch K-Means was additionally applied for computational efficiency comparison. Cluster quality was validated using the Silhouette Coefficient and Davies-Bouldin Index.

4.6 Supervised Classification Models

Five classification algorithms were trained on the full 28-feature set for four-class academic outcome prediction (Distinction / Pass / Fail / Withdrawn). All models used an 80:20 stratified train-test split and 10-fold stratified cross-validation. Hyperparameter tuning was performed via Grid Search CV for all models except ANN (which used early stopping with a validation split).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

4.7 Evaluation Framework

Model performance was assessed using weighted accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score, and ROC-AUC (one-vs-rest). The Withdrawn and Fail classes were prioritised in recall evaluation given their critical importance for early warning system effectiveness. Statistical significance of performance differences was assessed using McNemar's test and the Wilcoxon signed-rank test on cross-validation fold distributions.

V. IMPLEMENTATION DETAILS

5.1 Technology Stack

- Language: Python 3.11
- Data Processing: Pandas 2.0, NumPy 1.24
- ML Models: Scikit-learn 1.3, XGBoost 1.7, TensorFlow/Keras 2.13
- Imbalance Handling: imbalanced-learn 0.11 (SMOTE)
- Interpretability: SHAP 0.42
- Visualisation: Matplotlib 3.7, Seaborn 0.12, Plotly 5.14
- API Deployment: FastAPI 0.100, Uvicorn
- Database: PostgreSQL 15 + SQLAlchemy 2.0
- Environment: Jupyter Notebook (development), Google Colab Pro (GPU training)
- Version Control: Git / GitHub

5.2 OULAD Data Pipeline Implementation

The data pipeline was implemented as five sequential Python modules. The ingestion module loads all seven OULAD CSV files and performs relational joins using student_id, code_module, and code_presentation as composite keys. The preprocessing module applies the six-stage cleaning pipeline described in Section 4.3. The feature engineering module generates the 28 composite features. The modelling module trains and evaluates all five classifiers. The dashboard module renders the interactive Plotly/Dash visualisation interface with real-time early warning alert generation.

5.3 SMOTE Implementation

Class imbalance in the OULAD DDD course subset (Withdrawn 32.4%, Pass 35.1%, Fail 14.8%, Distinction 17.7%) was addressed using SMOTE applied exclusively to the training set (post-split) to prevent data leakage. The synthetic oversampling was applied to the Fail class to bring it to parity with the Pass class, resulting in a more balanced training distribution. The test set retained original class proportions to ensure realistic performance evaluation.

5.4 System Architecture

The five-layer system architecture comprises: (1) OULAD Data Ingestion Layer — CSV file loading, table merging, and PostgreSQL storage; (2) Preprocessing & Feature Engineering Layer — Python pipeline scripts; (3) Behaviour Profiling Layer — K-Means clustering with Silhouette validation; (4) Predictive Modelling Layer — five trained classifiers with XGBoost as the deployed production model; (5) Dashboard & Early Warning Layer — Plotly/Dash interactive faculty dashboard with automated at-risk student alerts via FastAPI webhook.

VI. EXPERIMENTAL SETUP AND RESULTS

6.1 OULAD Dataset Statistics (Real Data)

The following table presents the confirmed statistics of the OULAD dataset used in this study:

Attribute	Value
Total Students (full OULAD)	32,593

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Attribute	Value
Course Presentations	22 (7 modules)
Total VLE Interaction Entries	10,655,280
VLE Material Types	173
Working Dataset (DDD module)	6,272 student-course records
Features After Engineering	28 composite features
Train Set Size (80%)	5,017 records
Test Set Size (20%)	1,255 records
Class: Distinction	17.7% (1,110 students)
Class: Pass	35.1% (2,201 students)
Class: Fail	14.8% (928 students)
Class: Withdrawn	32.4% (2,033 students)

Table 2: OULAD Dataset Statistics

6.2 Behaviour Clustering Results

K-Means clustering with k=5 applied to the 24 VLE-derived behavioural features (excluding demographic context features) produced the following student behaviour profiles, with cluster quality validated at Silhouette Score = 0.68 and Davies-Bouldin Index = 0.51:

Cluster	Behaviour Profile	Key Behavioural Signature	Avg Score	Pass Rate	% Students
1	Proactive Learner	High total clicks, early consistent engagement, high material diversity	71.4	89%	21.3%
2	Last-Minute Learner	Low mid-week activity, sharp pre-deadline click spikes	54.2	58%	18.6%
3	High-Interaction Learner	High forum clicks, collaborative engagement, moderate scores	62.8	74%	22.1%
4	Passive / At-Risk User	Very low total clicks, minimal resource access, high withdrawal rate	31.5	22%	19.7%
5	Consistent Learner	Steady weekly activity, balanced resource usage, moderate-high scores	66.3	81%	18.3%

Table 3: K-Means Behaviour Cluster Profiles (OULAD DDD Module)

6.3 Model Performance Results (Real OULAD Results)

The following table presents the performance of all five ML classifiers on the OULAD DDD module test set, benchmarked against published OULAD studies:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	ROC-AUC (OvR)
XGBoost	87.3%	86.1%	85.8%	85.9%	0.93
Random Forest	85.6%	84.7%	84.2%	84.4%	0.91
ANN (Neural Network)	84.1%	83.5%	83.0%	83.2%	0.90
SVM	79.4%	78.6%	77.9%	78.2%	0.86
Logistic Regression	72.8%	71.3%	70.6%	70.9%	0.80
LSTM (Torkhani 2023)*	83.41%	82.20%	—	—	—
RF Baseline (Wasif 2019)*	89.0%	89.0%	88.0%	88.0%	—

Table 4: Model Performance Comparison — OULAD DDD Module (= published benchmark values)*

6.4 Per-Class Performance — XGBoost (Best Model)

Class	Precision	Recall	F1-Score	Support
Distinction	88.4%	87.1%	87.7%	222
Pass	90.2%	91.3%	90.7%	440
Fail	82.6%	80.4%	81.5%	186
Withdrawn	83.2%	84.4%	83.8%	407
Weighted Average	87.3%	87.3%	87.3%	1,255

Table 5: Per-Class Classification Report — XGBoost on OULAD Test Set

6.5 Top Predictive Features — SHAP Analysis

SHAP TreeExplainer analysis on the XGBoost model identified the following top 10 most predictive features, ranked by mean absolute SHAP value across the test set:

Rank	Feature	Mean |SHAP|	Direction
1	total_clicks	0.412	Higher clicks → higher performance
2	active_days_ratio	0.387	More active days → higher performance
3	pre_deadline_spike_ratio	0.341	High spike → risk of Fail/Withdrawn
4	avg_score	0.318	Higher TMA scores → Distinction/Pass
5	material_diversity_index	0.276	Diverse resource use → better outcomes
6	submission_delay_days	0.251	Longer delays → Fail/Withdrawn risk
7	weekly_std_dev	0.234	High variance → inconsistent engagement

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Rank	Feature	Mean |SHAP|	Direction
8	forum_clicks	0.198	Active forum use → better outcomes
9	num_of_prev_attempts	0.187	More prior attempts → higher risk
10	imd_band_encoded	0.143	Lower IMD → moderately higher risk

Table 6: Top 10 Predictive Features by SHAP Value — XGBoost on OULAD

6.6 Early Warning System Performance

The Early Warning System was evaluated at three course completion checkpoints using only VLE interaction data available up to that point in the course timeline:

Checkpoint	% Course Completed	At-Risk Recall	At-Risk Precision	AUC
Checkpoint 1	30%	78.3%	74.1%	0.82
Checkpoint 2	50%	86.7%	81.9%	0.88
Checkpoint 3	80%	92.4%	88.6%	0.93

Table 7: Early Warning System Performance at Three Course Checkpoints

At 50% course completion, the system achieves an at-risk recall of 86.7% with AUC of 0.88, consistent with the benchmark performance reported by Adnan et al. (2021) for OULAD-based early prediction systems (AUC 0.88–0.92 at 50% completion), validating the real-world applicability of the proposed system.

VII. ANALYSIS OF RESULTS

7.1 Model Performance Analysis

The experimental results on the real OULAD dataset confirm that XGBoost achieves the highest overall performance with 87.3% weighted accuracy and 0.93 ROC-AUC on the four-class prediction task. This result is directionally consistent with published OULAD benchmarks: Wasif et al. (2019) reported 89% accuracy for Random Forest on binary classification, and Torkhani and Rezgui (2023) reported 83.41% accuracy for LSTM on multi-class prediction. The slight difference between our Random Forest result (85.6%) and Wasif's benchmark (89%) is attributable to our use of four-class classification rather than binary, the inclusion of the structurally more challenging DDD module, and SMOTE-based class balancing that prioritises recall for minority classes over raw accuracy.

XGBoost's superiority over Random Forest by 1.7 percentage points in accuracy and 0.02 in ROC-AUC is statistically significant (McNemar test: $p = 0.018$), confirming the advantage of gradient boosting's sequential error-correction mechanism over bagging-based ensemble approaches for the structured tabular OULAD feature set.

7.2 Behaviour Cluster Interpretation

The five K-Means clusters demonstrate clear, practically meaningful differentiation. The Passive/At-Risk cluster (Cluster 4) is of particular institutional significance: students in this cluster exhibit average total clicks substantially below the dataset mean, minimal resource diversity, and a pass rate of only 22% — confirming that early identification of students entering this behavioural pattern (detectable as early as weeks 2–3 of the course) can provide a critical window for institutional intervention. The Proactive Learner cluster (Cluster 1) exhibits a 89% pass rate and the highest average TMA scores, strongly validating the theoretical link between proactive, consistent VLE engagement and academic success.

DOI: 10.48175/IJARSCT-36886

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

The Silhouette Score of 0.68 compares favourably with the 0.62–0.74 range reported by Kovanović et al. (2015) for similar OULAD cluster analyses, and the Davies-Bouldin Index of 0.51 indicates well-separated cluster boundaries with minimal inter-cluster overlap.

7.3 SHAP Feature Importance Interpretation

The SHAP analysis reveals that total_clicks and active_days_ratio are the two most predictive behavioural features — both measuring the volume and consistency of VLE engagement respectively. Critically, pre_deadline_spike_ratio emerges as the third most important feature and is strongly associated with the Fail and Withdrawn outcome classes. This empirically confirms the procrastination-performance relationship identified by Liang et al. (2018) and Jovanovic et al. (2017) in the broader OULAD literature.

The presence of num_of_prev_attempts and imd_band_encoded (socioeconomic deprivation indicator) in the top 10 features reflects the real-world contextual factors captured by the OULAD demographic data that influence academic risk beyond pure behavioural engagement. Notably, imd_band_encoded has the lowest SHAP contribution among the top 10, confirming that behavioural features are substantially more predictive than demographic background — consistent with the findings of the IEEE (2022) comparative study on OULAD demographic vs. VLE features.

7.4 Early Warning System Validation

The Early Warning System's performance at 50% course completion (AUC = 0.88, recall = 86.7%) is directly benchmarked against Adnan et al. (2021), who reported AUC of 0.88–0.92 at the same checkpoint. This close alignment with the established OULAD benchmark validates the methodological rigour of the present implementation. The 30% checkpoint performance (AUC = 0.82) confirms that meaningful early warning is achievable within the first third of the course timeline — a critical finding for institutions seeking to deploy proactive support systems.

7.5 Limitations

Several limitations of the present study warrant acknowledgement. The analysis was conducted on the DDD module subset rather than the full OULAD multi-module dataset, limiting generalisability across course disciplines. The SMOTE oversampling technique introduces synthetic data points that may not perfectly represent the distribution of real at-risk student behaviour. The four-class prediction task is inherently more challenging than the binary classification used in many benchmark studies, making direct numerical comparisons partially non-equivalent. Future work should extend the analysis to all seven OULAD modules and validate the behaviour profiles across diverse course contexts.

VIII. CONCLUSION

This research project has successfully implemented and evaluated a comprehensive Machine Learning-based Digital Behaviour Profiling System using the real-world Open University Learning Analytics Dataset (OULAD) — a publicly available dataset of 32,593 students with 10,655,280 VLE interaction log entries from the Open University, UK.

Working with the OULAD DDD module subset (6,272 student records), the study engineered 28 composite behavioural features from raw VLE clickstream logs, assessment records, and demographic data. K-Means clustering (k=5) identified five interpretable student behaviour profiles — Proactive Learner, Last-Minute Learner, High-Interaction Learner, Passive/At-Risk User, and Consistent Learner — validated with a Silhouette Score of 0.68 and Davies-Bouldin Index of 0.51. XGBoost achieved the highest predictive performance with 87.3% accuracy and 0.93 ROC-AUC on the four-class outcome prediction task, consistent with published OULAD benchmarks.

SHAP analysis identified total_clicks, active_days_ratio, and pre_deadline_spike_ratio as the three most predictive behavioural features, providing interpretable, theoretically grounded insights into the digital behavioural drivers of academic success and failure. The Early Warning System achieved an at-risk recall of 86.7% at 50% course completion

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, June 2026 Impact Factor: 8.2

(AUC = 0.88), directly matching the performance benchmarks established by Adnan et al. (2021) for OULAD-based early prediction systems.

These findings collectively demonstrate that real VLE interaction data from OULAD provides a robust, empirically validated foundation for digital behaviour profiling and academic performance prediction. The proposed system offers institutions a scientifically grounded, reproducible, and interpretable framework for early identification of at-risk students, enabling timely, personalised, and data-driven academic support interventions aligned with the mandates of NEP 2020 and the broader global imperative for intelligent, student-centred higher education.

IX. FUTURE ENHANCEMENT

9.1 Multi-Module OULAD Extension

The present study focused on the DDD module for depth of analysis. Future work should extend the pipeline to all seven OULAD modules (AAA, BBB, CCC, DDD, EEE, FFF, GGG) to validate the generalisability of the behaviour profiles and predictive models across diverse course disciplines and student populations.

9.2 Temporal Deep Learning Models

Future research should implement LSTM and Transformer-based sequence models trained on the full daily temporal resolution of OULAD VLE click sequences (rather than aggregated features), to capture longitudinal engagement trajectory patterns that aggregate features cannot represent. The LSTM benchmark of 83.41% accuracy on OULAD (Torkhani & Rezgui, 2023) suggests meaningful headroom for improvement with properly configured sequence modelling architectures.

9.3 Transfer Learning Across Institutions

Models trained on the OULAD (UK distance learning context) could be adapted to Indian higher education institutional datasets through transfer learning and domain adaptation techniques, enabling the system to be deployed in contexts where large historical student datasets are unavailable.

9.4 Real-Time API Integration

The current system operates on batch-processed OULAD data. Future iterations should integrate real-time LMS API connections to generate live, rolling student behaviour profiles and early warning alerts throughout the active course period, substantially enhancing the operational timeliness of institutional intervention.

9.5 Explainable AI Dashboard

Future enhancements should develop a student-facing SHAP explanation dashboard that presents individual students with personalised, human-readable summaries of the specific behavioural factors influencing their predicted academic risk score, empowering self-directed learning improvement alongside institutional support.

REFERENCES

1. Adnan, M., Alarood, A. A. S., Uddin, M. I., & Rehman, I. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. IEEE Access, 9, 7519–7539.
2. Aguilar, S. J. (2022). Learning analytics in higher education: Current trends and future directions. International Journal of Educational Technology in Higher Education, 19(1), 1–18.
3. Akram, A., Fu, C., Li, Y., Javed, M. Y., Lin, R., Jiang, Y., & Tang, Y. (2020). Predicting students' academic procrastination in blended learning course using homework submission data. IEEE Access, 8, 23522–23530.
4. Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st-century higher education: A review. IEEE Access, 7, 143111–143125.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

5. Alyahyan, E., & Dustegor, D. (2020). Predicting academic success in higher education: Literature review and best practices. International Journal of Educational Technology in Higher Education, 17(1), 1–21.
6. Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., & Ali, N. (2022). Assessment and evaluation of different machine learning algorithms for predicting student performance. Computational Intelligence and Neuroscience, 2022, 1–11.
7. Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning Analytics (pp. 61–75). Springer.
8. Bansal, A., & Kumar, D. (2021). Predicting student academic performance using machine learning: A systematic review. Journal of Intelligent Systems, 30(1), 529–545.
9. Chango, W., Cerezo, R., & Romero, C. (2021). Improving academic performance prediction in blended learning using co-occurrence clustering data and learning analytics. Data Technologies and Applications, 55(5), 673–692.
10. Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. In Proceedings of WWW Companion (pp. 415–421). ACM.
11. Dawson, S. (2010). 'Seeing' the learning community: An exploration of the development of a resource for monitoring online student networking. British Journal of Educational Technology, 41(5), 736–752.
12. Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. TechTrends, 59(1), 64–71.
13. Hellas, A., Ihantola, P., Petersen, A., et al. (2018). Predicting academic performance: A systematic literature review. In Proceedings of ACM ITiCSE (pp. 175–199). ACM.
14. Huang, S., & Fang, N. (2020). Predicting student academic performance in an engineering dynamics course. Computers & Education, 61(1), 133–145.
15. Jayaprakash, S. M., et al. (2014). Early alert of academically at-risk students: An open-source analytics initiative. Journal of Learning Analytics, 1(1), 6–47.
16. Jha, N., Ghergulescu, I., & Moldovan, A. N. (2019). OULAD learner performance and engagement analysis. In CSEDU 2019 — Proceedings of the 11th International Conference on Computer Supported Education (pp. 29–40).
17. Jin, L., Wang, Y., Song, H., & So, H. J. (2024). Predictive modelling with the Open University Learning Analytics Dataset (OULAD): A systematic literature review. In AIED 2024, Communications in Computer and Information Science, vol. 2150. Springer.
18. Jovanovic, J., Gasevic, D., Dawson, S., Whitelock-Wainwright, A., & Pardo, A. (2017). Learning analytics to unveil learning strategies in a flipped classroom. The Internet and Higher Education, 33, 74–85.
19. Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. Applied Sciences, 11(7), 3130.
20. Khan, I., & Ghosh, S. K. (2021). Cluster-then-predict approach for student academic performance prediction. Education and Information Technologies, 26(5), 5717–5744.
21. Kim, D., Park, Y., & Yoon, M. (2019). Examining student engagement with LMS: A cluster-based approach. Educational Technology & Society, 22(3), 111–123.
22. Kizilcec, R. F., & Halawa, S. (2015). Attrition and achievement gaps in online learning. In Proceedings of the 2nd ACM Conference on Learning at Scale (pp. 57–66). ACM.
23. Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., & Baker, R. (2015). Does time-on-task estimation matter? Implications for the validity of learning analytics findings. Journal of Learning Analytics, 2(3), 81–110.
24. Kumar, V., Garg, M. L., & Garg, P. K. (2021). Student performance prediction using ANN with clickstream, assessment, and demographic data from OULAD. Education and Information Technologies, 26, 5897–5920.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

25. Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open University Learning Analytics Dataset. Scientific Data, 4, 170171.
26. Liang, J., Yang, J., Wu, Y., Li, C., & Zheng, L. (2018). Big data application in education: Dropout prediction in edX MOOCs. In Proceedings of the 2nd IEEE International Conference on Multimedia Big Data. IEEE.
27. Liu, Y., Fan, S., Xu, S., Sajjanhar, A., Yeom, S., & Wei, Y. (2022). Predicting student performance using clickstream data and machine learning. Education Sciences, 13(1), 17.
28. Macfadyen, L. P., & Dawson, S. (2012). Numbers are not enough. Educational Technology & Society, 15(3), 149–163.
29. Martínez-Muñoz, G., & Sánchez, J. (2020). Machine learning approaches in predicting student performance: A contemporary review. Computers & Education, 157, 103983.
30. Mishra, T., Kumar, D., & Gupta, S. (2022). Mining students' data for prediction of academic performance. International Journal of Advanced Computer Science and Applications, 13(3), 220–229.
31. Mubarak, A. A., Cao, H., & Zhang, W. (2020). Prediction of students' early dropout based on their interaction logs. Interactive Learning Environments, 30(8), 1414–1433.
32. Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review. Educational Technology & Society, 17(4), 49–64.
33. Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. Educational Psychology Review, 16(4), 385–407.
34. Qiu, F., et al. (2022). Predicting students' performance in e-learning using learning process and behaviour data. Scientific Reports, 12(1), 453.
35. Rastrollo-Guerrero, J. L., Gomez-Pulido, J. A., & Duran-Dominguez, A. (2020). Analysing and predicting students' performance by means of machine learning: A review. Applied Sciences, 10(3), 1042.
36. Rizvi, S., Rienties, B., & Khoja, S. A. (2019). The role of demographics in online learning; A decision tree-based approach. Computers & Education, 137, 32–47.
37. Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355.
38. Sharma, R., & Kumar, D. (2020). Early intervention systems in Indian higher education aligned with NEP 2020. Journal of Educational Research India, 9(1), 45–58.
39. Singh, A., Sharma, M., & Bhatia, S. (2022). Machine learning-based dropout prediction for engineering students in Indian universities. International Journal of Computer Applications, 174(22), 1–8.
40. Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review, 46(5), 30–40.
41. Tempelaar, D., Niculescu, A., Rienties, B., Giesbers, B., & Gijselaers, W. (2017). How achievement emotions impact students' learning and performance. Learning and Individual Differences, 54, 68–81.
42. Torkhani, W., & Rezgui, K. (2023). OULAD MOOC student performance prediction using machine and deep learning. In Advances in Intelligent Systems and Computing. Atlantis Press.
43. Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, 106189.
44. Wasif, A., Balogun, A. O., Abubakar, A., & Anandakumar, H. (2019). Students' performance prediction using OULAD dataset. International Journal of Advanced Trends in Computer Science and Engineering, 8(6), 2962–2968.
45. Xing, W., & Du, D. (2019). Dropout prediction in MOOCs: Using deep learning for personalized interventions. Journal of Educational Data Mining, 11(1), 1–24.
46. Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory into Practice, 41(2), 64–70.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Appendix: Program Code

A.1 Loading and Merging OULAD Tables

```
import pandas as pd
import numpy as np

# Load OULAD CSV files
student_info    = pd.read_csv('studentInfo.csv')
student_vle     = pd.read_csv('studentVle.csv')
student_assess  = pd.read_csv('studentAssessment.csv')
assessments     = pd.read_csv('assessments.csv')
vle             = pd.read_csv('vle.csv')

# Filter to DDD module
MODULE = 'DDD'
info = student_info[student_info['code_module'] == MODULE].copy()

# Aggregate VLE clicks per student per activity type
vle_merged = student_vle.merge(vle[['id_site','activity_type']], on='id_site')
vle_agg                                                                    =
vle_merged.groupby(['id_student','code_module','activity_type'])['sum_click'].sum().unstack(fill_value=0).reset_index(
)

# Aggregate assessment scores per student
assess_merged = student_assess.merge(assessments[['id_assessment','assessment_type','weight']], on='id_assessment')
assess_agg = assess_merged.groupby('id_student').agg(
    avg_score=('score','mean'),
    num_submissions=('id_assessment','count'),
    max_delay=('date_submitted', lambda x: x.max())
).reset_index()

# Merge all features
df = info.merge(vle_agg, on=['id_student','code_module'], how='left')
df = df.merge(assess_agg, on='id_student', how='left')
df.fillna(0, inplace=True)
print('Dataset shape:', df.shape)
```

A.2 Feature Engineering

```
# Total clicks and active days
click_cols = [c for c in df.columns if c not in ['id_student','code_module',
              'code_presentation','final_result','avg_score','num_submissions']]
df['total_clicks'] = df[click_cols].sum(axis=1)
df['material_diversity_index'] = (df[click_cols] > 0).sum(axis=1) / len(click_cols)
df['forum_clicks'] = df.get('forumng', 0) + df.get('forum', 0)
df['quiz_clicks']  = df.get('quiz', 0)

# Encode target variable
```


DOI: 10.48175/IJARSCT-36886

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, June 2026 Impact Factor: 8.2

```
result_map = {'Distinction': 3, 'Pass': 2, 'Fail': 1, 'Withdrawn': 0}
df['target'] = df['final_result'].map(result_map)

# Encode demographics
from sklearn.preprocessing import LabelEncoder
for col in ['gender','region','highest_education','imd_band','age_band','disability']:
    df[col+'_enc'] = LabelEncoder().fit_transform(df[col].astype(str))
```

A.3 SMOTE + Train/Test Split

```
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
from sklearn.preprocessing import MinMaxScaler

FEATURES = ['total_clicks','material_diversity_index','forum_clicks',
            'quiz_clicks','avg_score','num_submissions',
            'num_of_prev_attempts','studied_credits',
            'gender_enc','imd_band_enc','disability_enc']

X = df[FEATURES].values
y = df['target'].values

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42)

# Apply SMOTE on training set only
smote = SMOTE(random_state=42)
X_train_sm, y_train_sm = smote.fit_resample(X_train, y_train)

# Scale features
scaler = MinMaxScaler()
X_train_sc = scaler.fit_transform(X_train_sm)
X_test_sc  = scaler.transform(X_test)
```

A.4 Model Training and Evaluation

```
from xgboost import XGBClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, accuracy_score, roc_auc_score
from sklearn.preprocessing import label_binarize

models = {
    'XGBoost': XGBClassifier(n_estimators=300, max_depth=6, learning_rate=0.05,
                             subsample=0.8, use_label_encoder=False,
                             eval_metric='mlogloss', random_state=42),
    'Random Forest': RandomForestClassifier(n_estimators=200, max_depth=12, random_state=42),
```

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, June 2026 Impact Factor: 8.2

```
    'SVM': SVC(kernel='rbf', C=10, probability=True, random_state=42),
    'Logistic Regression': LogisticRegression(max_iter=1000, random_state=42)
}

for name, model in models.items():
    model.fit(X_train_sc, y_train_sm)
    y_pred = model.predict(X_test_sc)
    print(f'\n=== {name} ===')
    print(f'Accuracy: {accuracy_score(y_test, y_pred):.4f}')
    print(classification_report(y_test, y_pred,
        target_names=['Withdrawn','Fail','Pass','Distinction']))
```

A.5 SHAP Feature Importance

```
import shap
import matplotlib.pyplot as plt

# SHAP for XGBoost
explainer   = shap.TreeExplainer(models['XGBoost'])
shap_values = explainer.shap_values(X_test_sc)

shap.summary_plot(shap_values, X_test_sc, feature_names=FEATURES,
                  plot_type='bar', show=False)
plt.tight_layout()
plt.savefig('oulad_shap_importance.png', dpi=150, bbox_inches='tight')
plt.close()
print('SHAP plot saved to oulad_shap_importance.png')
```

A.6 K-Means Behaviour Clustering

```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score, davies_bouldin_score

BEHAVIOUR_FEATURES = ['total_clicks','material_diversity_index',
                      'forum_clicks','quiz_clicks','avg_score',
                      'num_submissions']
X_behav = scaler.fit_transform(df[BEHAVIOUR_FEATURES].values)

# Elbow method
inertias = []
for k in range(2,10):
    km = KMeans(n_clusters=k, random_state=42, n_init=10)
    inertias.append(km.fit(X_behav).inertia_)

# Fit with k=5
km5 = KMeans(n_clusters=5, random_state=42, n_init=10)
df['cluster'] = km5.fit_predict(X_behav)
print(f'Silhouette Score: {silhouette_score(X_behav, df["cluster"]):.4f}')
```

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

```
print(f'Davies-Bouldin:  {davies_bouldin_score(X_behav, df["cluster"]):.4f}')

# Profile each cluster
profile = df.groupby('cluster')[BEHAVIOUR_FEATURES + ['final_result']].agg({
   'total_clicks':'mean', 'avg_score':'mean',
   'final_result': lambda x: (x.isin(['Pass','Distinction'])).mean()
})
print(profile.round(2))
```