Enhancing the Human Touch: A Data-Driven Analysis of Student Archetypes in an AI-Augmented Classroom

Gomathi Thiyagarajan[1], Avishek Nandi[2], and Avichandra Singh Ningthoujam[3]
[1]Dept. of Computer Application, CMR Institute of Technology, Bengaluru ,
[2]Dept. of Computer Application, Manipal University Jaipur, Jaipur,
[3]Dept. of Computer Application, Manipal University Jaipur, Jaipur
[1]gomathi.t@cmrit.ac.in, [2]avisheknandi10@gmail.com, [3]avichandra0420@gmail.com

***Abstract*—In the contemporary landscape of Indian higher education, the incorporation of Artificial Intelligence (AI) of- fers both unparalleled possibilities and considerable educational challenges. A principal concern is the potential diminution of the essential 'human touch' which is foundational to effective teaching. This study addresses this issue by proposing a data-driven framework to identify distinct student archetypes within an AI-augmented learning environment. We leverage a comprehensive dataset comprising 638 undergraduate engineering students, which includes prior academic records, weekly performance metrics, and system interaction logs, upon which the K-Means clustering algorithm is applied. Our analysis successfully delineates four distinct student archetypes: the 'High-Achieving and Consistent', the 'Diligent but Struggling', the 'Disengaged and At-Risk', and the 'Erratic Performer'. By characterizing these cohorts based on their academic, behavioral, and engagement patterns, we provide educators with actionable insights. These insights empower instructors to transcend monolithic teaching strategies and implement targeted, personalised interventions. Such a strategy ensures that while AI manages scalability, the educator's role is amplified, enabling them to provide a more nuanced, empathetic, and effective human touch where it is most critically needed. This work proposes a symbiotic model wherein AI-driven analytics and human pedagogy converge to foster a more supportive and effective learning ecosystem.**

***Index Terms*—AI in education; educational data mining; K-Means clustering; learning analytics; student archetypes; student-centric learning.**

***ICTIEE Track*—Innovations in Engineering Education for the Future.**

***ICTIEE Sub-Track*— Artificial Intelligence in Education.**

I. Introduction

THE Indian higher education system is navigating a period of profound transformation, shaped by the dual imperatives of expanding access to a growing student population and upholding rigorous standards of instructional quality. In this context, the Incorporation of Artificial Intelligence (AI) into educational practices has been widely regarded as a paradigm-shifting development, offering the potential for personalised learning pathways and efficient feedback mechanisms at a massive scale Wang et al. (2024). AI-driven platforms possess the capability to meticulously track student progress, recommend bespoke learning resources, and automate evaluative tasks, thereby mitigating the substantial administrative workload faced by educators in typically large and heterogeneous classrooms.

This technological advancement, however, is not without its apprehensions. A significant concern articulated by pedagogues and policymakers alike is the potential dilution of the 'human touch'—the empathetic, mentor-driven interaction that constitutes the bedrock of meaningful education Bedenlier et al. (2020). An educational paradigm reliant solely on automation risks insensitivity to the subtle indicators of student disengagement, confusion, or personal challenges that an experienced human instructor can adeptly perceive and address. The quintessential challenge, therefore, lies not in replacing educators with AI, but in augmenting their innate capabilities, empowering them to apply their pedagogical expertise with greater precision and impact.

This paper proposes that a synergy between AI and hu- man instruction can be achieved by employing AI-driven analytics to provide educators with a deeper, more structured understanding of their student body. Rather than perceiving the classroom as a homogeneous collective, it is possible to utilize data analytics to unearth latent patterns of student behavior and academic performance. This process facilitates the identification of distinct 'student archetypes'—recurrent profiles of learners who exhibit analogous academic histories, engagement patterns, and learning trajectories.

The central research question guiding this investigation is formulated as follows: Is it possible to identify statistically significant and pedagogically meaningful student archetypes from multifaceted educational data, and how can such knowledge empower educators to refine and enhance their intervention strategies? To address this question, we present a framework that utilizes established Educational Data Mining (EDM) techniques, specifically K-Means clustering, on a rich, real-world dataset sourced from an undergraduate engineering course. This dataset is notable for its comprehensiveness, integrating not only conventional academic metrics but also granular, temporal data on weekly performance, task-relat

Avichandra Singh Ningthoujam
Manipal University Jaipur, Jaipur Dehmi Kalan, Bagru, Rajasthan, 303007
avichandra0420@gmail.com

time expenditure, and system-level engagement traces.

The principal contributions of this study are enumerated as follows:

1) We demonstrate the efficacy of an unsupervised machine learning model in identifying four distinct and interpretable student archetypes from a complex, multimodal educational dataset.
2) We provide a detailed characterisation of these archetypes, constructing a data-supported narrative for each group that elucidates the interplay between their prior academic standing, in-course behaviour, and eventual academic outcomes.
3) We propose a set of concrete, archetype-specific intervention strategies, thereby illustrating a practical pathway for educators to leverage this data-driven understanding for providing targeted, human-centric support within an AI-augmented pedagogical framework.

By conceptualizing AI as a sophisticated tool for deepening student understanding rather than as a substitute for the instructor, this research charts a course for a more balanced, effective, and humanized future for engineering education.

II. Related Work

The application of data mining and machine learning within the educational domain, a field often referred to as Learning Analytics (LA) and Educational Data Mining (EDM), has experienced substantial growth over the last decade. A significant part of the literature has been dedicated to predicting student performance, frequently employing regression or classification models to identify learners vulnerable to academic failure Pallathadka et al. (2022). For instance, the contribution of Kruger et al. (2022) Kru¨ger et al. (2023) is notable for its use of eXplainable AI (XAI) to not only predict student dropout but also to provide transparent rationales for these predictions, thereby facilitating the design of targeted interventions. Concurrently, researchers have explored a diverse array of features, ranging from demographic and prior academic data to intricate behavioural indicators extracted from Learning Management System (LMS) logs Chen and Liu (2024).

While predictive modelling is undoubtedly valuable, a more nuanced understanding of underlying student behaviours and learning patterns can offer deeper insights for pedagogical enhancement. This realisation has spurred a growing interest in clustering techniques to identify distinct student archetypes or profiles. Clustering, as an unsupervised learning method, groups students based on inherent similarities in their attributes without relying on a predefined outcome variable. The K-Means model remains one of the most prevalent methods in this area due to its computational efficiency and interpretability Zheng et al. (2015). Early work by Speily et al. (2016) Speily et al. (2020) utilised clustering to categorise students based on their interaction patterns in a social learning platform, identifying roles such as ’lurkers’ and ’leaders’. More recent investigations have applied K-Means to analyse engagement trends in Massive Open Online Courses (MOOCs) Edumadze and Govender (2024) and blended learning settings Quinn and Gray (2019), consistently reinforcing the notion that student populations are heterogeneous.

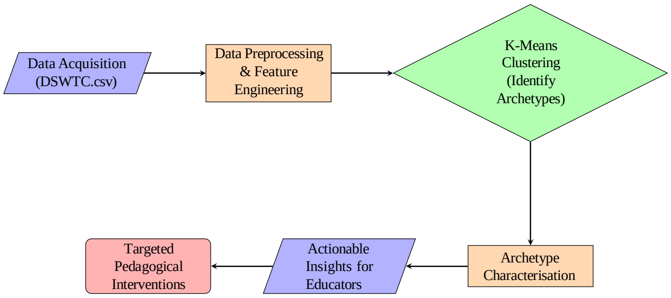
Data Acquisition (DSWTC.csv)
Data Preprocessing & Feature Engineering
K-Means Clustering (Identify Archetypes)
Archetype Characterisation
Actionable Insights for Educators
Targeted Pedagogical Interventions

Fig. 1. System architecture for identifying and utilising student archetypes.

A parallel trend in the field is the move toward integrating multimodal data for a broader understanding of the student. Multimodal Learning Analytics (MMLA) seeks to synthesise data from disparate sources, including traditional log files, video, audio, and physiological sensors Xiao et al. (2025). Although our study does not incorporate physiological data, it adheres to the multimodal philosophy by amalgamating static academic records with dynamic, time-series data on weekly performance and engagement. This approach is congruent with recent scholarly calls for more comprehensive data collection to unravel complex learning processes Cohn et al. (2024).

Crucially, recent scholarship has emphasized the importance of translating these data-driven insights into practical pedagogical actions. It is insufficient to merely identify at- risk students or behavioural clusters; the ultimate objective is to furnish instructors with actionable feedback Susnjak et al. (2022). A systematic review by Seufert et al. (2019) Seufert et al. (2019) highlighted the critical need for LA initiatives to be firmly grounded in pedagogical theory, ensuring that the analytics directly inform and support teaching practices. Our research builds directly upon this principle by defining student archetypes in a manner that naturally suggests clear, targeted intervention strategies for educators, thereby empowering their pedagogical decision-making Alonso-Ferna´ndez et al. (2019).

III. Methodology

The methodology employed in this research was designed to systematically process and analyse the student dataset to uncover meaningful learner archetypes. The overarching architecture of our approach is illustrated in Fig. 1, com- prising phases of data acquisition and preprocessing, feature engineering, application of the K-Means clustering algorithm, and subsequent archetype characterisation and interpretation.

A. Dataset Description

The empirical basis for this study is the DSWTC.csv dataset, which contains anonymised records of 638 undergraduate students. This is a rich, multifaceted dataset that comprises several categories of variables:

- *Demographics and Background* Attributes such as gender, academic scores in 10th and 12th standards.
- *Prior Academic Performance* The Cumulative Point Grade Average (CPGA) serves as an indicator of previous academic standing.

- *Weekly Engagement and Performance* For four consecutive weeks (W1-W4), the dataset captures granular data on the time taken (in hours) and scores achieved for various assessments.
- *System-Level Engagement* The ’TRACEAL:TraceFitness’ variable is a composite metric that quantifies a student’s engagement fitness, derived from system interaction logs.
- *Final Outcome* A binary ’Final Score’ (0 for Fail/Low, 1 for Pass/High) is provided, which we interpret as a categorical indicator of overall course success.

B. Data Preprocessing and Feature Selection

The initial phase of our methodology involved rigorous data preprocessing. To reduce dimensionality and capture overarching behavioural trends, we engineered summary features from the weekly data. Specifically, we calculated the average score and average time taken across all recorded weekly activities for each student. We checked the raw data for errors; we removed any student records that were missing a ’Final Score’ or had incomplete logs (less than 75% attendance) to ensure accuracy. This left us with 638 students. We also used the Interquartile Range (IQR) method to find and fix ’Time taken’ errors, such as cases where the system was left idle for too long. To reduce the number of variables and see general behavior trends. We handle missing data by using the mean for continuous variables and the mode for categorical variables A PyTorch-based implementation was used for the clustering stage to leverage potential GPU acceleration.

For the clustering model, we judiciously selected a set of five key features designed to provide a holistic representation of a student’s profile:

1) *CPGA* Represents a student’s academic history.
2) *Avg_Score* The arithmetic mean of all weekly scores, capturing in-course academic performance.
3) *Avg_Time* The mean of all ’Timetaken’ fields, a proxy for effort.
4) *TRACEAL* A direct measure of student engagement with the learning system.
5) *Score_Zeros:* A count of weekly activities for which a student received a score of zero, a strong indicator of non-submission or disengagement.

Following selection, these features were standardized using Z-score normalization to ensure that every attribute had an equal influence on the distance computations within the clustering algorithm, regardless of its original scale.

C. K-Means Clustering

We employed the K-Means algorithm to group students into distinct cohorts.

K-Means aims to divide n data points into k clusters, allotting each data point to the group whose mean (centroid) is allocating to it.

The algorithm repeatedly allocates data points to the closest centroid and then updates the centroid’s position to the mean of its assigned points, minimising the within-cluster sum of squares (WCSS).

We chose the K-Means algorithm over other methods like DBSCAN or Hierarchical Clustering for two practical reasons

1) we needed to make sure every student was included. Some algorithms, like DBSCAN, treat unique or unusual data points as ”noise” and don’t assign them to a group. In a classroom, those ”outliers” are often the students who are struggling the most or behaving erratically. We couldn’t risk leaving them out of the analysis, so we needed a method that forces every data point into a cluster.
2) K-Means is easier for teachers to understand. It works by finding the ”average” behavior for each group. This creates clear, simple profiles (like ”high effort but low scores”) that an instructor can instantly recognize. In contrast, Hierarchical Clustering produces complex tree diagrams that can be difficult to interpret quickly during a busy semester. K-Means gave us the most straightforward and actionable results.

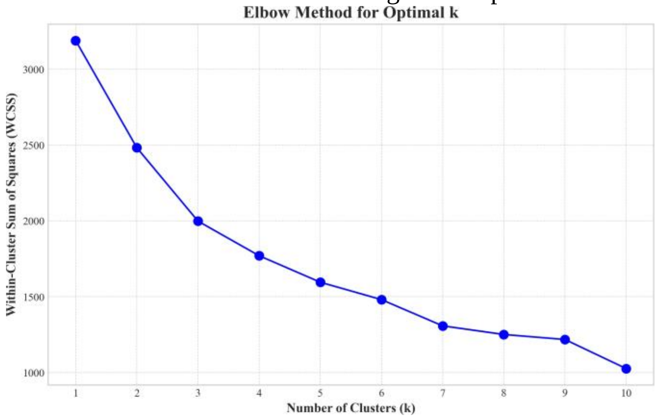

Fig. 2. Elbow Method plot for finding the ideal number of clusters (k). The ”elbow” is visible at k=4.

To ascertain the ideal number of clusters (k), we utilised the elbow method. As shown in Fig. 2, the WCSS was plotted for a range of k values. The ”elbow” point, where the marginal decrease in WCSS begins to diminish, was observed at $k = 4$. This suggests that four clusters offer the best balance between model complexity and the interpretability of the resulting groups. The complete procedure is formalised in Algorithm 1.

TABLE I
DESCRIPTIVE STATISTICS OF KEY VARIABLES (N=638)

Variable	mean	std	min	max
10th Score	8.76	0.92	5.05	10.00
12th Score	8.44	1.06	5.26	10.04
CPGA	7.57	1.05	2.83	9.56
TRACEAL:TraceFitness	0.50	0.25	0.20	1.00
Final Score	0.43	0.50	0.00	1.00

TABLE II
CORRELATION MATRIX FOR SELECT VARIABLES

	CPGA	Avg Score	TRACEAL	Final Score
CPGA	1.00	0.21	-0.06	-0.03
Avg Score	0.21	1.00	0.09	0.11
TRACEAL	-0.06	0.09	1.00	0.82
Final Score	-0.03	0.11	0.82	1.00
Note: Avg	Score is	the calculated	average of weekly scores.	

```
Algorithm 1 Student Archetype Identification via K-Means
Input: Student dataset $D$, number of clusters $k = 4$, feature set $F = \{$CPGA, Avg–Score, Avg–Time, TRACEAL, Score–Zeros$\}$
Output: Set of $k$ student archetypes $C = \{C_1, ..., C_k\}$
Standardize the feature set $F$ to obtain $F_{std}$.
Randomly initialize $k$ centroids $\mu_1, \mu_2, ..., \mu_k$ from $F_{std}$.
repeat
    // Assignment Step
    for each student data point $x_i \in F_{std}$ do
        Determine the nearest centroid $\mu_j$.
        Assign student $i$ to cluster $C_j$.
    end for
    // Update Step
    for each cluster $C_j$ do
        Update centroid $\mu_j$ as the mean of all points within $C_j$.
    end for
until centroids have converged
return Final cluster assignments $C$.
```

D. Ethical Considerations

Using student data requires us to follow strict ethical rules. All data in this study was anonymized before we used it; we hid personal details like names and IDs to protect student pri- vacy. Also, this system follows a human-in-the-loop approach. These student profiles are tools to help instructors make de- cisions, not labels used for automatic grading. The goal is to support students who are struggling, not to punish them.

IV. RESULTS AND ANALYSIS

The application of our proposed methodology yielded substantial insights into the underlying structure of the student population. This section presents the descriptive statistics of the dataset, a detailed characterization of the identified clusters, and a visual analysis of their distinguishing features.

A. Descriptive Statistics and Correlations

A preliminary examination of the dataset was conducted with N=638 students. Table I furnishes the descriptive statistics for the principal parameters. The mean CPGA of the cohort was 7.57. The mean for the binary ‘Final Score‘ was 0.43, indicating that 43% of the students achieved what was cate- gorised as a successful outcome (Pass/High).

A correlation analysis, presented in Table II, revealed a critical insight. The ‘TRACEAL‘ metric, which quantifies system engagement, demonstrated a very strong positive correlation with the ‘Final Score‘ (r = 0.82), powerfully underscoring the importance of consistent student engagement for academic success. In stark contrast, ‘CPGA‘, representing prior academic performance, showed a negligible correlation with ‘Final Score‘ (r = -0.03). This finding suggests that a student’s historical academic record is a poor predictor of their performance in this specific course context compared to their real-time engagement behaviour.

B. Student Archetype Characterisation

The K-Means algorithm partitioned the student population into four distinct clusters. The standardised centroids of these clusters, detailed in Table III, define the profile of each archetype.

Archetype 1 The High-Achieving and Consistent (26% of students). This group’s defining feature is an exception- ally high TRACEAL score (1.51), indicating outstanding and consistent engagement with the learning system. They also have the lowest number of zero-score submissions. Interestingly, their CPGA and average scores are close to the mean, while their time spent is below average, suggesting they are highly efficient learners who achieve success primarily through consistent engagement rather than innate high aptitude or excessive effort. Fig. 3 shows that this group is composed almost entirely of passing students.

Archetype 2: The Diligent but Struggling (28% of students). This cohort is characterised by the highest average time spent on tasks (0.90) and the highest average weekly scores (0.91). They also enter with a high CPGA (0.54). While the label ”Struggling” may seem counterintuitive given their high scores, it reflects their high-effort learning style; they achieve good results but require significantly more time than other groups. Their TRACEAL score is below average, suggesting their engagement, while time-consuming, might be less effective or focused than that of the first archetype.

Archetype 3: The Disengaged and At-Risk (31% of students). Comprising the largest segment, this group is characterised by low values across the board: low CPGA, low average scores, low time investment, and a low TRACEAL score. They represent a classic profile of disengagement. As Fig. 3 confirms, this group has the highest number of failing students, making them the primary cohort in need of proactive intervention.

Archetype 4: The Erratic Performer (15% of students). This archetype presents the most distinctive profile. Their most prominent feature is an extremely high count of zero-score submissions (1.79), indicating a pattern of missed assignments

TABLE III
CLUSTER CENTROIDS AND ARCHETYPE DEFINITIONS (STANDARDISED VALUES)

Archetype	CPGA	Avg_Score	Avg_Time	TRACEAL	Score_ Zeros	Students (%)
1: High-Achieving & Consistent	-0.12	-0.04	-0.40	1.51	-0.48	26%
2: Diligent but Struggling	0.54	0.91	0.90	-0.42	-0.35	28%
3: Disengaged & At-Risk	-0.39	-0.48	-0.60	-0.58	-0.14	31%
4: Erratic Performer	-0.01	-0.66	0.21	-0.56	1.79	15%

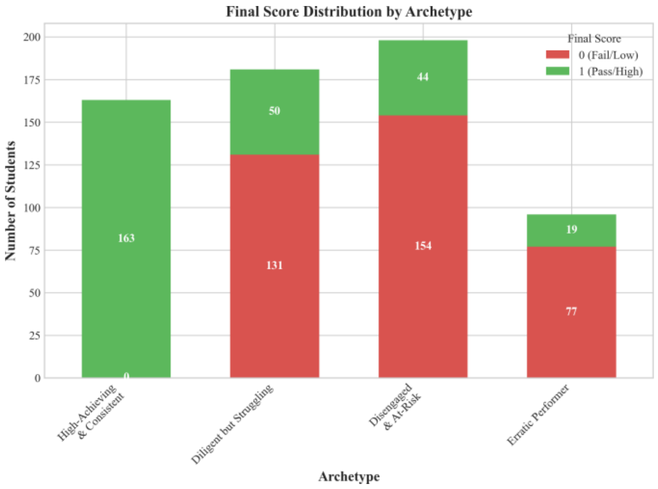

Fig. 3. Final Score Distribution by Archetype

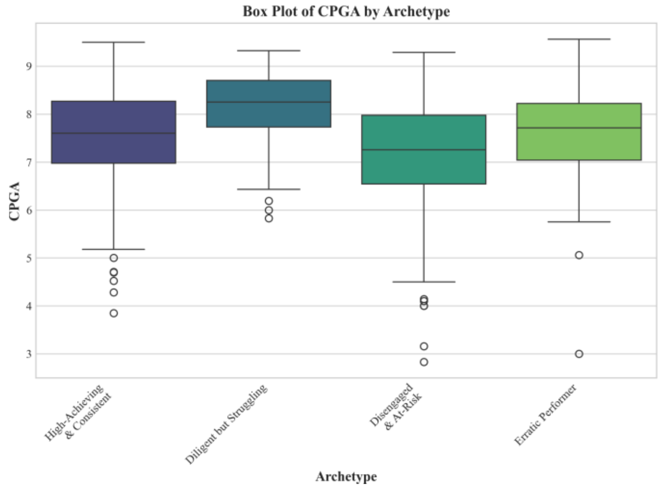

Fig. 5. Box Plot of CPGA by Archetype

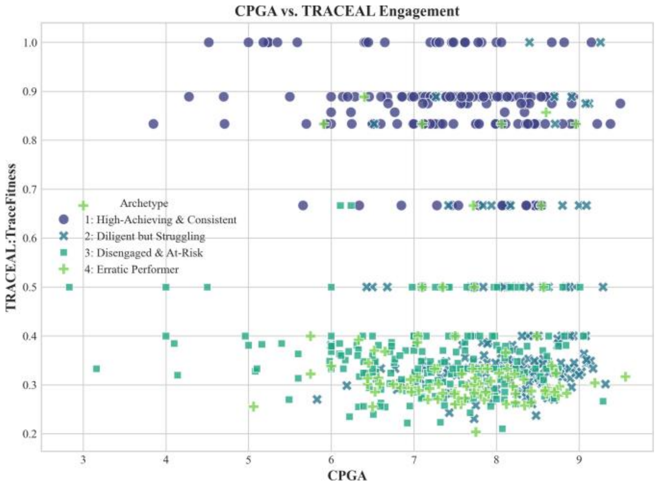

Fig. 4. CPGA vs. TRACEAL Engagement

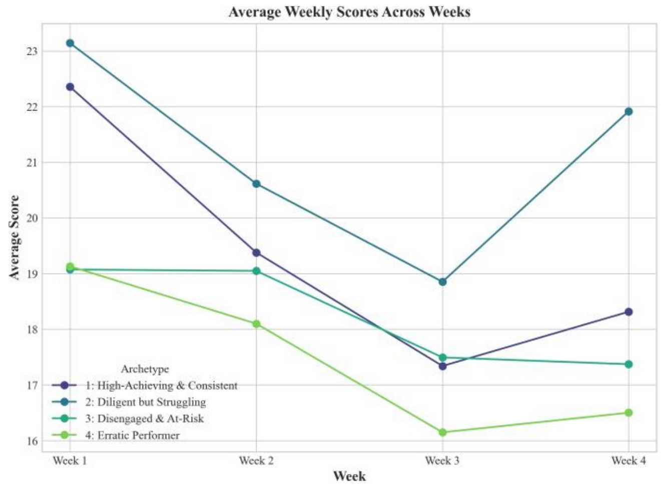

Fig. 6. Average Weekly Scores Across Weeks

or non-participation. Despite this, their CPGA and average time spent are near the mean, and their average scores are only moderately low. This suggests a capable but inconsistent student who engages sporadically. Their outcomes are mixed, as seen in Fig. 3, highlighting their unpredictable nature.

C. Visual Analysis of Archetypes

To further elucidate these profiles, a series of visualisations were generated, as presented in Fig. 3. Fig. 3 clearly shows the pass/fail distribution, confirming that the ’High-Achieving’ group almost universally succeeds, while the ’Disengaged’ group overwhelmingly fails. The other two groups show mixed results. The scatter plot in Fig. 4 visually separates the ’High-Achieving’ group with its high TRACEAL scores, while the other three archetypes cluster at lower engagement levels. The CPGA shows less clear separation, reinforcing the correlation analysis. The box plot in Fig. 5 illustrates the distribution of prior academic performance (CPGA), showing significant overlap between groups, with the ’Diligent’ group having a slightly higher median. The time-series plot of weekly scores (Fig. 6) reveals different trajectories: ’High-Achieving’ and ’Diligent’ students show a performance dip in Week 3

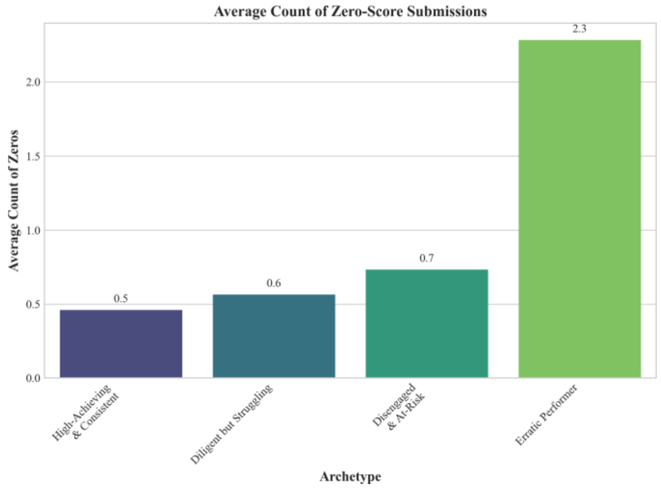

Fig. 7. Average Count of Zero-Score Submissions

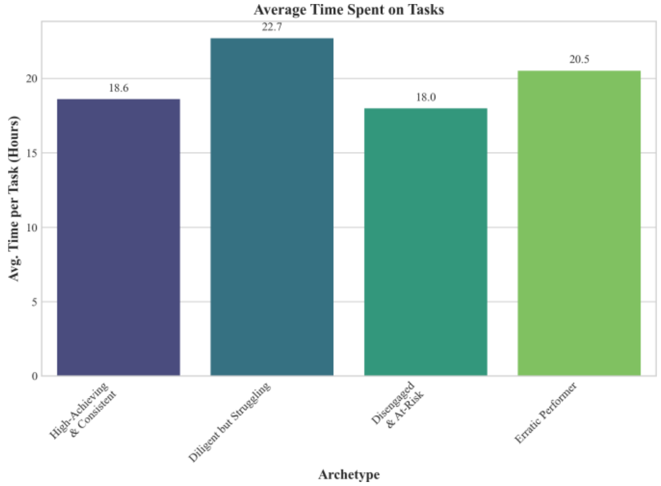

Fig. 8. Average Time Spent on Tasks

before recovering, while ’Erratic Performers’ show a dramatic drop. Finally, the bar charts starkly highlight the behavioural differences. Fig. 7 shows that the ’Erratic Performer’ archetype is defined by a high number of zero-score submissions, while Fig. 8 confirms the ’Diligent but Struggling’ group spends the most time on tasks.

D. Student Progress Parameters

We measured student progress by analyzing changes in their weekly grades, the amount of time they invested in tasks, and how consistently they submitted their assignments.

1) Tracking Progress Over Time: We didn’t just look at a snapshot; we watched how scores changed week by week in 3. This highlighted a key difference in resilience: the ’Diligent’ group managed to recover after a bad week, whereas the ’Erratic’ group’s performance remained low and did not improve.
2) Time vs. Results We dug into the ’Time Spent’ data to see if effort matched the output. We found that putting in more hours didn’t always lead to higher grades, which is shown in Fig. 4. This was crucial for distinguishing students who are trying hard but getting stuck from those who simply aren’t engaging.
3) Reliability Check We looked at how often students received a zero on a submission. This clarified that for the ’Erratic’ group, the main issue wasn’t a lack of skill, but a lack of consistency they were simply missing too many assignments to succeed.

V. Discussion and Implications

The delineation of these four student archetypes offers a powerful analytical lens through which educators can bet- ter comprehend the complex dynamics of their classrooms and, consequently, tailor their pedagogical support. This data-informed approach facilitates a strategic shift from reactive to proactive teaching methodologies, thereby enhancing the educator’s ’human touch’ by directing it towards areas where it can yield the most significant impact.

Fig. 4 visualizes the relationship between students’ overall grades (CPGA) and their engagement levels (TraceFitness). There is a clear engagement gap. The High-Achieving and Diligent students cluster tightly at the top, showing that consistent interaction with the learning platform correlates with higher grades. Conversely, the Disengaged and At-Risk students are scattered across the bottom, indicating that low engagement is a strong predictor of lower academic outcomes. Fig. 5 box plot reveals the spread of final grades for each group. Interestingly, the Diligent but Struggling group per- formed impressively well, with a median score rivaling the High-Achievers, proving that their persistence paid off. The Erratic Performers showed the most volatility their box is tall, meaning their grades swung wildly from high to low. As ex- pected, the Disengaged group consistently sat at the bottom of the grade distribution. Fig. 6 shows the performance trends over time. This line chart tracks average scores week by week. Everyone started strong in Week 1, but the groups diverged as the difficulty likely increased. The Diligent students (blue line) showed resilience, bouncing back significantly in Week 4 after a dip. In contrast, the Disengaged and Erratic groups (green lines) flatlined; once their performance dropped in Week 2, they failed to recover, highlighting a lack of resilience in their study habits. Fig. 7 shows the chart measures the average hours spent on tasks, serving as a proxy for effort. It highlights why the Diligent group succeeds: they work the hardest, spending an average of 22.7 hours on tasks to overcome their struggles. Notably, the Erratic performers also spent a lot of time (20.5 hours), but their lower grades suggest this time wasn’t spent efficiently. Fig. 8 shows the consistency and missed work; this bar chart pinpoints exactly where the Erratic Performers fail: reliability. While most groups rarely missed an assignment (averaging below 0.7 zero scores), the Erratic group averaged 2.3 zero scores per student. This indicates that their primary challenge isn’t necessarily a lack of skill, but rather a habit of completely skipping assignments.

A. Pedagogical Interventions

The distinct profiles of the identified archetypes suggest the need for differentiated intervention strategies. We propose the

TABLE IV
COMPARISON WITH STATE-OF-THE-ART (SOTA) WORKS IN STUDENT ANALYSIS

Study	Methodology	Dataset Focus	Key Contribution	Our Approach Alignment
Kruger et al. (2023) Kru¨ger et al. (2023)	Gradient Boosting, SHAP (XAI)	E-learning platform data (clicks, quizzes)	Provides explainable predictions for student dropout.	Aligned in using behavioural data but we focus on clustering for profiles.
Edumadze et al. (2024) Edumadze and Govender (2024)	K-Means, Sequential Pattern Mining	MOOC data (video interactions, forum posts)	Identified engagement patterns in MOOCs (e.g., ’auditing’).	Similar methodology, but our dataset includes formal grades and is not MOOC.
Quinn (2019) Quinn and Gray (2019)	SVM, Decision Tree, Naive Bayes	Moodle quiz logs, demographics	Performance prediction in a blended learning environment.	Complements classification by providing unsupervised profiles.
This Study	K-Means Clustering	Prior academics, weekly scores/time, system trace	Identifies four interpretable archetypes for targeted intervention.	Integrates prediction-relevant features into a profiling framework.

following targeted approaches:

- *For the High-Achieving & Consistent:* This cohort is highly engaged and efficient. Intervention should fo- cus on enrichment and challenge. Educators can pro- vide advanced material, research opportunities, or peer-mentoring roles to foster leadership and deeper learning, acknowledging their exemplary engagement.
- *For the Diligent but Struggling:* This group achieves good results but invests significant time. They could benefit from guidance on study efficiency and time management. One-on-one consultations could help identify conceptual bottlenecks that consume excessive time. Acknowledging their hard work while offering strategies to work smarter, not just harder, is key.
- *For the Disengaged & At-Risk:* This group requires immediate and proactive human intervention. Automated alerts are insufficient. A personal outreach from the instructor is crucial to understanding the root causes of disengagement (which could be academic, personal, or motivational) and to building a supportive connection.
- *For the Erratic Performer:* The high number of non-submissions is the critical red flag. Interventions should focus on consistency, accountability, and time management. Breaking down large assignments and setting smaller, regular deadlines could help. The goal is to guide them towards sustained effort rather than sporadic bursts of activity.

B. Comparison with State-of-the-Art

Our research advances the existing body of literature on student profiling. Table IV provides a comparative analysis of our approach against other contemporary studies. A key differentiator of our work is the use of a multifaceted dataset that reveals the paramount importance of a dynamic engagement metric (TRACEAL) over static academic history (CPGA) in a formal undergraduate engineering course. We frame our findings explicitly serving to enhance, not substitute for, the educator’s role. *Limitations and Future Work*

It is pertinent to recognise the constraints of this study. The dataset was sourced from a single course within one institution, which may circumscribe the generalizability of the specific archetypes identified. Using this in other schools faces practical challenges. To work, schools need a modern Learning Management System (LMS) that can track detailed logs, which older systems might not do. There is also a data literacy’ gap; for these tools to work, schools must train faculty so they understand the data and don’t rely too much on what the computer says. While our data is specific to one field, the student habits we analyzed often called ’digital body language’ are universal. Every student leaves a digital footprint through their login frequency, time management, and adherence to deadlines. Because nearly all modern university courses rely on digital platforms to track progress, this framework can be easily adapted by educators in any discipline to better understand and support their own students. Future research should apply this framework across diverse courses and insti- tutional contexts. Additionally, the ’TRACEAL’ feature, while powerful, was treated as an atomic input; deconstructing its constituent components could yield further insights.

The logical progression of this work involves implementing this framework within a live classroom environment. The development of an instructor-facing dashboard that provides real-time updates on student archetype classifications would be a significant step, enabling timely and data-informed interventions. Subsequently, a controlled study could be designed to quantitatively assess the outcomes of these archetype-driven interventions on learner engagement and academic achievements.

CONCLUSION

In an educational era increasingly defined by technological integration, the role of the human educator is not being diminished but rather fundamentally transformed. This study has presented a practical, data-driven framework designed to enhance and strategically focus the educator’s ”human touch”. Through the application of K-Means clustering on a multimodal dataset of 638 students, we have identified four distinct and pedagogically relevant student archetypes: the ‘High-Achieving & Consistent’, the ‘Diligent but Struggling’, the ‘Disengaged & At-Risk’, and the ‘Erratic Performer’.

A key finding of this work is that real-time student engagement is a far more potent indicator of success than historical academic performance. This underscores the value of observing and responding to current student behaviour. The data-derived profiles provide educators with a structured, nuanced understanding of their students, allowing for the deployment of their pedagogical expertise where it is most needed. We advocate for a symbiotic model of AI-educator collaboration, where AI performs complex analytics to unearth deep insights, and educators leverage these insights to cultivate a more empathetic, supportive, and ultimately more human learning experience. As the Indian higher education sector continues its digital transformation, such human-centric applications of AI will be indispensable in ensuring that technology serves to enrich, rather than supplant, the core mission of education.

REFERENCES

Cristina Alonso-Ferna´ndez, Ana Rus Cano, Antonio Calvo-Morata, Manuel Freire, Iva´n Mart´ınez-Ortiz, and Baltasar Ferna´ndez-Manjo´n. Lessons learned applying learning an- alytics to assess serious games. Computers in Human Behavior, 96:65–74, 2019.

Svenja Bedenlier, Melissa Bond, Katja Buntins, Olaf Zawacki-Richter, and Michael Kerres. Learning by Doing? Reflec- tions on Conducting a Systematic Review in the Field of Educational Technology, pages 111–127. Springer Fachme- dien Wiesbaden, Wiesbaden, 2020.

Ming Chen and Zhi Liu. Predicting performance of students by optimizing tree components of random forest using genetic algorithm. Heliyon, 10(12):e32570, 2024.

Clayton Cohn, Eduardo Davalos, Caleb Vatral, Joyce Horn Fonteles, Hanchen David Wang, Meiyi Ma, and Gautam Biswas. Multimodal methods for analyzing learning and training environments: A systematic literature review. CoRR, abs/2408.14491, 2024.

Joseph K. E. Edumadze and Desmond W. Govender. The community of inquiry as a tool for measuring student en- gagement in blended massive open online courses (moocs): a case study of university students in a developing country. Smart Learning Environments, 11(1):19, 2024.

Joa˜o Gabriel Corre^a Kru¨ger, Alceu de Souza Britto, and Jean Paul Barddal. An explainable machine learning approach for student dropout prediction. Expert Systems with Applications, 233:120933, 2023.

Harikumar Pallathadka, Shikha Jain, Suraj Kamble, and Korakod Tongkachok. Educational data mining: A comprehen- sive review and future challenges. ECS Transactions, 107: 16129–16136, April 2022.

Rory Quinn and Geraldine Gray. Prediction of student academic performance using moodle data from a further education setting. Irish Journal of Technology Enhanced Learning, 5, October 2019.

Sabine Seufert, Christoph Meier, Matthias Soellner, and Ro- man Rietsche. A pedagogical perspective on big data and learning analytics: A conceptual model for digital learning support. Technology, Knowledge and Learning, 24:599–619, 2019.

Omid Speily, Alireza Rezvanian, Ardalan Ghasemzadeh, Ali M. Saghiri, and S. Mehdi Vahidipour. Lurkers Versus Posters: Investigation of the Participation Behaviors in Online Learning Communities, pages 269–298. January 2020.

Teo Susnjak, G. S. Ramaswami, and Anuradha Mathrani. Learning analytics dashboard: a tool for providing action- able insights to learners. International Journal of Educa- tional Technology in Higher Education, 19(1):12, 2022.

Shan Wang, Fang Wang, Zhen Zhu, Jingxuan Wang, Tam Tran, and Zhao Du. Artificial intelligence in education: A systematic literature review. Expert Systems with Applications, 252:124167, 2024.

Jun Xiao, Ming Chen, Yue Yang, et al. An exploratory multimodal study of the roles of teacher-student interaction and emotion in academic performance in online classrooms. Education and Information Technologies, 30:15507–15527, 2025.

Saijing Zheng, Mary Beth Rosson, Patrick C. Shih, and John M. Carroll. Understanding student motivation, behav- iors and perceptions in moocs. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work Social Computing (CSCW ’15), pages 1882–1895. ACM, 2015.