Human Digital Twins – Understanding University Student Behaviour Over Time

Zainaaz Hansa

41425626

Literature Review
submitted for the degree **BSc Hons in Computer Science and Information Technology 2025**
at the **North-West University, Potchefstroom Campus**

Supervisor: Prof. Marijke Coetzee

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 3: LITERATURE REVIEW

1. Introduction

This research project, titled Human Digital Twins – Understanding University Student Behaviour Over Time, aims to design and implement the foundation for a digital artefact capable of monitoring behavioural trends and predicting academic performance. This foundation will support future extensions of the research, including the detection of academic exhaustion and procrastination. This chapter explores the theoretical and empirical groundwork for the artefact by reviewing existing literature on how Long Short-Term Memory (LSTM) models can be applied to predict student performance over time using Massive Open Online Course (MOOC) clickstream data. These insights directly inform the development of a Human Digital Twin (HDT) model that mirrors student behavioural patterns in real-time.

This chapter begins by identifying the key themes in the literature that inform this research, including the utility and limitations of clickstream data, the evolution from traditional to deep learning methods for modelling sequential data and the role of the Open University Learning Analytics Dataset (OULAD) as a widely accepted benchmark in this domain. The review then shifts focus to methodological insights drawn from four key academic studies, each of which applies machine learning to OULAD to predict student performance. This comparative analysis helps establish best practices and recurring challenges across the literature.

In the subsections that follow, readers can expect a critical analysis of how different modelling approaches (ranging from standard LSTM to hybrid CNN-LSTM models) have been used to interpret student behavioural data. Special attention is given to feature engineering, temporal aggregation strategies and evaluation metrics, as these significantly influence predictive performance. The chapter concludes by reflecting on how the reviewed literature guides this study’s choice of dataset and model and outlines a phased implementation strategy rooted in replicating and adapting leading approaches.

2. Literature Themes Identified

2.1. The Value and Challenge of Clickstream Data

The digital transformation of education has led to the generation of vast datasets that capture student interactions within VLEs. Among these, clickstream data has emerged as a particularly valuable resource. Liu et al. (2023) define this data as indicating "the path(s) a student takes through one or more learning sites". These time-stamped "digital footprints" (Waheed et al., 2020) provide a granular, objective record of student engagement that is difficult to obtain through other

means. A primary advantage of this data type is its accessibility and course-agnostic nature. According to Liu et al. (2023), "the strengths of using clickstream data include its ease of access, regardless of course conditions, such as course structure, assessments or learning activities". Furthermore, Liu et al. (2023) assert that despite some limitations, "research has confirmed that clickstream data are reliable and offer valid and nuanced information about students' actual learning processes" when compared to less objective measures like self-reporting.

However, the nature of clickstream data also presents distinct analytical challenges. The data is inherently sequential and temporal, reflecting a series of discrete events over time. In other words, the data's temporal and sequential format means that while a click is recorded, the underlying cognitive or learning activity it represents is not immediately clear and must be interpreted from the broader pattern of engagement (Liu et al., 2023). Moreover, the data can be sparse, as it represents "non-continuous events in behaviour patterns" (Liu et al., 2023). These characteristics render traditional static analysis methods, which often rely on aggregated or "flattened" features, less effective. As Qu et al. (2019) argue, the use of overall behaviours, which "cannot fully reflect students' learning processes", affects the accuracy of predictions. This fundamental challenge of capturing temporal dynamics necessitates the use of specialised models designed to process sequential information.

These challenges and opportunities of clickstream data highlight the importance of using sequence-aware models like LSTM to capture behavioural trends, aligning with the goals of this study’s HDT framework.

2.2. From Traditional Machine Learning to Deep Learning for Sequential Data

Early efforts in predicting student performance often relied on traditional machine learning (ML) algorithms such as Logistic Regression, Support Vector Machines (SVM) and Random Forest (Khoudi et al., 2025; Borna et al., 2024). These models typically require features to be structured in a tabular, non-sequential format. Consequently, to use them with clickstream data, researchers often aggregate temporal interactions into static features, such as the total number of clicks or total time spent. While straightforward, this approach loses the crucial ordering and timing of student actions, which can be highly indicative of their learning patterns.

The limitations of static models led researchers to explore techniques specifically designed for sequential data. Recurrent Neural Networks (RNNs) represent a class of models where "connections between nodes form a directed graph along a temporal sequence," allowing the network to "exhibit a temporally dynamic behaviour" (Qu et al., 2019). However, standard RNNs are susceptible to the vanishing gradient problem, which makes it "difficult to process a long sequence" (Qu et al., 2019). This limitation hinders their ability to capture long-range

dependencies, such as the relationship between a student's activity in the first week of a course and their performance in the final month.

Long Short-Term Memory (LSTM) networks were developed specifically to overcome this challenge. Adefemi & Mutanga (2025) describe LSTM as "a powerful variant of the RNN" whose "strength lies in overcoming the limitations of traditional RNNs by addressing the challenge of long-term dependencies". The architecture of an LSTM cell, which includes an input gate, an output gate and a forget gate, allows it to selectively remember or discard information over long time intervals. This makes LSTM models "better suited to capture temporal dynamics" (Adefemi & Mutanga, 2025) and particularly effective for analysing student clickstream data, which can be viewed as a time series of behavioural events. The proven ability of LSTMs to model sequential data has made them a cornerstone of modern approaches to student performance prediction.

2.3. The Open University Learning Analytics Dataset (OULAD) as a Research Benchmark

A significant portion of contemporary research in this domain leverages the Open University Learning Analytics Dataset (OULAD). According to Kuzilek et al. (2017), the OULAD contains data about students and their interactions with Virtual Learning Environment for seven selected courses and is used in multiple research papers especially in student performance measurement as can be seen in Table 3-5.

It is a comprehensive and anonymised resource that has become a standard benchmark for evaluating predictive models. The dataset is composed of several interconnected tables, providing a holistic view of the student experience (Kuzilek et al., 2017).

Table 3-1: The Open University Learning Analytics Dataset adapted from Kuzilek et al. (2017)

Table Name	Columns	Description	Row Count
studentInfo.csv	code_module, code_presentation, id_student, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts, studied_credits, disability, final_result	Contains student demographics and final result for each module presentation.	32,593
courses.csv	code_module, code_presentation, length	Lists modules and their presentations, including duration (in days).	22
studentRegistration.csv	code_module, code_presentation, id_student, date_registration, date_unregistration	Records student registration and deregistration dates per module.	32,593
assessments.csv	code_module, code_presentation, id_assessment, assessment_type, date, weight	Details assessments within each module	206

		presentation, including final exams.	
studentAssessment .csv	id_assessment, id_student, date_submitted, is_banked, score	Contains students' assessment results and submission details.	173,912
studentVle.csv	code_module, code_presentation, id_student, id_site, date, sum_click	Captures student interactions (clicks) with specific VLE materials on specific days.	10,655,280
vle.csv	id_site, code_module, code_presentation, activity_type, week_from, week_to	Provides metadata about each VLE material, including its type and intended usage window.	6,364

The assessments table details all assessments within module presentations. Typically, each presentation includes several assessments leading up to a final exam. A summary of the assessments.csv file can be found in Table 3-2.

Table 3-2: OULAD asessments.csv table: module and summary assessment information consisting of 206 rows adapted from Kuzilek et al. (2017)

Module	Domain	Presentations	Students
AAA	Social Sciences	2	748
BBB	Social Sciences	4	7,909
CCC	STEM	2	4,434
DDD	STEM	4	6,272
EEE	STEM	3	2,934
FFF	STEM	4	7,762
GGG	Social Sciences	3	2,534

Table 3-3: OULAD studentVle.csv table fields containing information about student's interactions with the VLE adapted from Kuzilek et al. (2017)

#	Columns	Description	Data Type
1	code_module	the module identification code	nominal
2	code_presentation	the presentation identification code	nominal
3	id_site	the VLE material identification number	numerical
4	id_student	the unique student identification number	numerical
5	date	the day of student's interaction with the material	numerical
6	sum_click	the number of times the student interacted with the material	numerical

Table 3-4: OULAD vle.csv table consisting of 6,364 rows with the following columns adapted from Kuzilek et al. (2017)

#	Columns	Description	Data Type
1	id_site	the identification number of the VLE material	numerical
2	code_module	the module identification code	nominal
3	code_presentation	the presentation identification code	nominal
4	activity_type	the role or type of the material (e.g., lecture, reading)	nominal
5	week_from	the week the material is first intended to be used	numerical
6	week_to	the week the material is last intended to be used	numerical

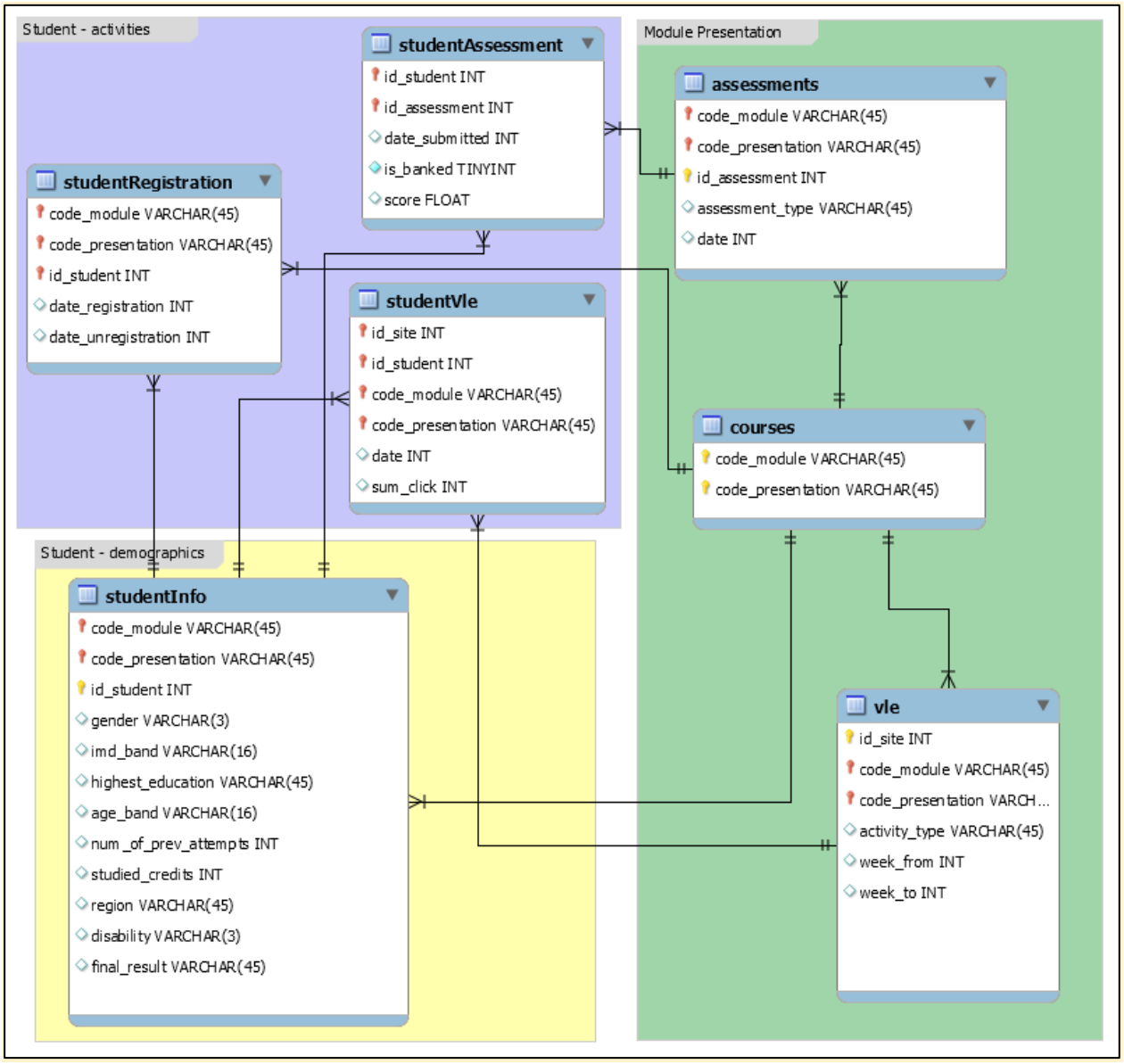

Figure 3-1: Detailed OULAD relationships adapted from Kuzilek et al. (2017)

The widespread use of OULAD across numerous studies, including all four key papers selected for this review, underscores its importance. Its public availability and rich, multi-faceted data allow researchers to develop, test and compare new models against state-of-the-art results on a common ground, fostering transparency and cumulative progress in the field.

3. Literature Methodological Approaches using OULAD

The following subsections provide a detailed analysis of four key studies that have used the OULAD dataset to predict student performance. These papers were selected for their methodological relevance, their use of machine learning and their direct applicability to the goals of this research

3.1. Bidirectional LSTM with Enriched Features (Souai et al., 2022)

The study by Souai et al. (2022) focuses on applying deep learning to "model the learning behaviours of students in a Virtual Learning Environment, predict their performance and prevent students at-risk from failure". A key innovation in their work is the use of a Bidirectional Long-Short Term Memory (BLSTM) model. Souai et al. (2022) defines a BLSTM as "a sequence-processing model made up of two LSTMs, one of which takes the input in one way and the other in the opposite direction," which successfully improves the "quantity of data available to the network, providing the algorithm with more context".

Methodologically, the most significant contribution of this paper is its feature engineering strategy. The authors argue that previous approaches have been limited by using only demographic and behavioural (clickstream) features. In contrast, Souai et al. (2022) state that they included in their modelling "a third type of features which is the assessment features". This includes data such as assessment scores and their corresponding weights in the final grade. To handle the class imbalance inherent in the dataset, they applied the SMOTE (Synthetic Minority Over-sampling Technique) oversampling technique to make the instances of classes equal. Their BLSTM model, comprising two hidden layers of 50 units each and a dropout rate of 0.5, was trained for 100 epochs.

The findings powerfully demonstrate the value of this enriched feature set. A preliminary model trained without assessment data achieved an accuracy of 92.80%. However, the final proposed model, which included assessment scores and weights, achieved a "cross-validation accuracy rate of 97%". This result confirms that "assessments scores and their weights give a good indicator of students' future performance" and that combining direct academic measures with

behavioural clickstream data yields a more robust and accurate predictive model Souai et al. (2022).

3.2. Day-Wise Early Prediction (Al-azazi & Ghurab, 2023)

The work of Al-azazi & Ghurab (2023) directly addresses a critical limitation of many predictive models: timeliness. They note that many models provide predictions "at the end of the course period thus delaying making in-time interventions". To overcome this, their main aim is "to build a multi-class course-agnostic day-wise predictive model to identify the class of students' performance in MOOC environments as early as possible" (Al-azazi & Ghurab, 2023). Their model predicts one of four outcomes: Distinction, Pass, Fail, or Withdrawn.

The study proposes a hybrid Artificial Neural Network-Long Short-Term Memory (ANN-LSTM) model. The architecture consists of an LSTM input layer to process the sequential clickstream data, followed by ANN (Dense) layers for classification. A core part of their methodology is the data preparation, which aggregates clickstream data on a daily basis. They state, "For each time step (a day of the course duration), the ANN-LSTM model was trained and evaluated using demographic and clickstream data in this day and all previous days" (Al-azazi & Ghurab, 2023). This approach allows the model to make a prediction at any point during the course, with the amount of available historical data increasing each day.

The results validate the feasibility of early prediction. The model's accuracy progressively improves as more data becomes available, starting from "43% on the first day of the course and reached 72% on the last day" (Al-azazi & Ghurab, 2023). The study concludes that student's demographic and behavioural data shows good results in the case of MOOC students' performance prediction, especially the first three months, thereby confirming that there are opportunities to forecast the class of students in MOOC courses during the first months. This work provides strong evidence for the value of developing models that can offer timely, actionable insights early in a course.

3.3. Hybrid CNN-LSTM Architecture (Adefemi & Mutanga, 2025)

Adefemi & Mutanga (2025) propose a "robust hybrid deep learning model" that combines convolutional neural networks (CNN) and long short-term memory (LSTM) to predict student academic performance. Their work is motivated by the need to address common data challenges, including "class imbalance, missing data and selecting relevant features" and to improve upon the generalisability of existing models (Adefemi & Mutanga, 2025).

The model's architecture is a key contribution, designed to leverage the distinct strengths of both CNNs and LSTMs. The authors explain that in their model, "the CNN handles feature extraction,

while the LSTM captures temporal dependencies" (Adefemi & Mutanga, 2025). A 1D-CNN layer first processes the input data to automatically extract meaningful local patterns from the raw feature sequences. The output of this layer is then fed into LSTM layers to model the temporal relationships between these extracted patterns. The study also employs a rigorous data pre-processing pipeline, using the synthetic minority oversampling technique (SMOTE) to handle class imbalance and recursive feature elimination (RFE) technique to select relevant features (Adefemi & Mutanga, 2025).

The performance of this hybrid model is exceptional. On the OULAD dataset, the model "achieved predictive accuracies of 98.93%," which outperformed traditional ML models and standalone deep learning approaches" (Adefemi & Mutanga, 2025). This result strongly suggests that the synergy between CNNs as automated feature extractors and LSTMs as sequence analysers provides a powerful and highly effective framework for this prediction task. The study demonstrates that sophisticated model architectures, combined with meticulous data pre-processing, can yield state-of-the-art results.

3.4. Temporal Aggregation and Feature Importance (Liu et al., 2023)

The research by Liu et al. (2023) investigates the potential of clickstream data by focusing on how the data is structured and which components are most predictive. A central objective of their study is to perform feature extraction by transforming the raw clickstream data. Specifically, "two feature sets are extracted, indicating the number of clicks on 12 learning sites based on weekly and monthly time intervals" (Liu et al., 2023). This allows for a direct comparison of how different temporal granularities affect model performance.

The study employs a standard stacked LSTM architecture on the OULAD dataset. By training and evaluating the model separately on the weekly and monthly aggregated feature sets, the authors provide empirical evidence on the optimal time window for analysis. In addition to model evaluation, they conduct a feature importance analysis to "investigate the impact of different features on prediction outcomes with the aim to identify the important features" (Liu et al., 2023).

The findings offer practical and valuable insights for implementation. The study found that "for LSTM, the weekly view showed the best model with 89.25% accuracy, significantly higher than the monthly view at 88.67%" (Liu et al., 2023). This suggests that weekly aggregation provides a better balance between data granularity and noise reduction than monthly aggregation. Furthermore, the feature importance analysis revealed that "four out of twelve learning sites (content, subpage, homepage, quiz) are identified as critical in influencing student performance" (Liu et al., 2023). This finding is highly practical, as it pinpoints the specific student behaviours

that are most indicative of their final academic outcomes, offering clear targets for monitoring and intervention.

4. Literature Comparison

To synthesise the practical and methodological aspects of the reviewed literature, this section presents a comparative analysis of the four key studies. The following table provides a structured overview of each study's research objective, the dataset and features used, the data pre-processing steps undertaken, the specific model applied and the key performance outcomes. This tabular format is designed to offer an at-a-glance reference that can directly inform the design and implementation choices for the current research project, aligning with the supervisor's guidance for a practical and implementation-oriented review. By juxtaposing these state-of-the-art approaches, the table illuminates common practices, highlights methodological variations and provides a clear benchmark for evaluating the proposed study's contribution.

Table 3-5: Literature Comparison

Study (Author, Year)	Research Objective	Dataset & Features Used	Data Pre-processing & Handling	Model / Technique Applied	Key Performance & Outcome
Souai et al. (2022)	To predict at-risk students using a deep learning approach, preventing failure and dropout.	OULAD; Demographic, Behavioural (clicks) and Assessment (scores, weights) features.	Data cleaning, one-hot encoding for categorical features and SMOTE for handling class imbalance.	Bidirectional Long Short-Term Memory (BLSTM) with two hidden layers.	Accuracy: 97% (with assessment features). Finding: Including assessment features significantly improves prediction accuracy over using clicks alone.
Al-azazi & Ghurab (2023)	To build a multi-class, course-agnostic, day-wise model for early prediction of student performance (Distinction, Pass, Fail, Withdrawn).	OULAD; Demographic and Clickstream features aggregated daily.	SQL query to aggregate daily clicks, MinMax scaling for numerical features, LabelEncoder and one-hot encoding for categorical features.	Artificial Neural Network - Long Short-Term Memory (ANN-LSTM).	Accuracy: 72% (at end of course). Finding: Accuracy increases over time, validating the feasibility of day-wise early prediction.
Adefemi & Mutanga (2025)	To develop a robust hybrid deep learning model to improve prediction accuracy, addressing generalisability and data challenges.	OULAD & WOU; Full feature set including demographics and clicks.	Removal of records with missing values, one-hot encoding, MinMax scaling, SMOTE for imbalance and Recursive Feature Elimination (RFE) for feature selection.	Hybrid Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM).	Accuracy: 98.93% (on OULAD). Finding: The hybrid CNN-LSTM architecture outperforms standalone deep learning models.
Liu et al. (2023)	To investigate the potential of clickstream	OULAD; Clickstream data aggregated into	Data cleaning (discarded 180 students with no clicks),	Long Short-Term Memory (LSTM) with	Accuracy: 89.25% (with weekly data). Finding: Weekly

	data by comparing different temporal aggregations and identifying important features.	weekly and monthly feature sets.	aggregation of clicks into weekly and monthly panel data structures.	a stacked architecture.	aggregation is superior to monthly. Clicks on 'content', 'subpage', 'homepage' and 'quiz' are most influential.

Articles	Model	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)
[10]	GCNN	81.5	-	-	-
[13]	ANN	79.22	81.44	78.03	79.70
[16]	BiLSTM-AM	90.16	90	90	90
[19]	GB, XGBoost, and LightGBM	92.40, 94.13, 89.07	-	-	92.32, 94.00, 88.91
[11]	CNN	88	-	-	-
[12]	SVM-RNN	86.90	-	81.57	-
[18]	NN and RF	74.6	74.8	74.6	72.3
Our Work	DNN	92.20	91.70	92.00	91.90
	CNN	96.10	96.13	96.01	95.19
	LSTM	97.62	97.61	97.62	97.61
	CNN–LSTM	98.93	98.93	98.93	98.93

Figure 3-2: Performance Comparison of different machine learning models presented in previous studies reproduced from Adefemi & Mutanga (2025: Table 8).

Note: This figure includes references to other studies (e.g., [10], [13]) that were cited in the original article but were not independently reviewed in this paper.

5. Reflection - How the literature will guide this study

This section synthesises the findings from the literature review to construct a robust justification for the core methodological choices of this research project. By reflecting on the evidence presented in the reviewed studies, a clear and defensible rationale is established for the selection of the Open University Learning Analytics Dataset (OULAD) and a Long Short-Term Memory (LSTM) based model. Furthermore, this reflection identifies specific studies from the literature that can serve as practical base models, guiding the subsequent implementation phase.

5.1. Justification for the OULAD Dataset

The selection of the OULAD is a cornerstone of this research design and the literature provides compelling justification for its use. OULAD is not merely one of many available datasets; it has established itself as the de facto benchmark in the specific research niche of predicting student performance from VLE data. Its consistent use across all four key papers analysed in this review (Souai et al., 2022; Al-azazi & Ghurab, 2023; Adefemi & Mutanga, 2025; Liu et al., 2023) and numerous other supporting studies (Khoudi et al., 2025; Borna et al., 2024; Waheed et al., 2020) demonstrates its acceptance and relevance within the academic community. This widespread adoption ensures that the results generated by this project will be comparable to a broad body of state-of-the-art work, which is essential for validating the study's contributions.

Furthermore, the comprehensive nature of the dataset makes it uniquely suitable for sophisticated modelling. It allows for the replication and extension of advanced approaches, such as the feature-enriched model proposed by Souai et al. (2022), which demonstrated the significant impact of combining assessment data with clickstream logs. Using a well-understood, high-quality and widely-used dataset like OULAD minimises the risks associated with data quality and availability, allowing the research to focus on its core contribution: the development and analysis of a predictive model.

5.2. Justification for the LSTM Model

The decision to employ an LSTM-based model is strongly supported by both theoretical principles and overwhelming empirical evidence from the literature. As established in Section 2.1, the sequential, time-series nature of student clickstream data poses a significant challenge for traditional machine learning models, which are unable to capture temporal dependencies. LSTM networks are specifically designed to address this limitation. Adefemi & Mutanga (2025) note that LSTMs excel at "addressing the challenge of long-term dependencies" and are "better suited to capture temporal dynamics" than their predecessors. This theoretical advantage makes them an ideal choice for modelling student behaviour as it unfolds over the course of a semester.

The empirical evidence presented in the reviewed literature provides an even stronger justification. LSTM-based models consistently deliver high performance on the OULAD dataset. The exceptional accuracy figures reported (such as 98.93% by Adefemi & Mutanga (2025), 97% by Souai et al. (2022) and 89.25% by Liu et al. (2023)) serve as undeniable proof of the model's suitability and power for this specific task. The literature demonstrates a clear progression, with researchers continually refining and advancing LSTM-based architectures (e.g., from standard LSTM to BLSTM and hybrid CNN-LSTM) to achieve even greater accuracy. This trend indicates that the LSTM architecture is not only effective but also represents the current frontier of research in this area. Therefore, selecting an LSTM model aligns with this project and provides the greatest potential for achieving high predictive performance.

5.3. Identification Models for Implementation

In line with a practical focus that directly informs implementation, this review identifies two key studies that can serve as base models for replication and adaptation. This approach provides a strategic, phased pathway for the project's development, balancing feasibility with ambition.

The first key study is the work of Liu et al. (2023) which is proposed as the primary base model for initial replication. This study is ideal as a starting point because it presents a clear, well-documented and standard implementation of an LSTM model for this task. Its methodology is straightforward, focusing on a direct comparison of weekly versus monthly data aggregation and providing a feature importance analysis. Replicating this model is a manageable first step that will allow for the validation of the implementation environment and the establishment of a solid performance baseline. Its finding that weekly aggregation is optimal provides an immediate, practical design choice for the project.

The second key study is that of by Adefemi & Mutanga (2025) which is proposed as a more advanced model for subsequent adaptation. After a successful replication of the simpler baseline, their hybrid CNN-LSTM architecture offers a clear and logical "next step" for innovation. Adefemi & Mutanga (2025)'s research paper provides a detailed account of a robust pre-processing pipeline, including the use of RFE for feature selection and SMOTE for handling class imbalance, which can be incorporated to enhance the project's own methodology. The reported accuracy of 98.93% sets a high-performance target. This model also presented the best precision, recall, and F-score values. By starting with the simpler model and then adapting it to incorporate the more advanced techniques from Adefemi & Mutanga (2025), the research project can demonstrate both a solid replication of existing work and a novel contribution aimed at pushing the boundaries of predictive accuracy. This phased approach, grounded directly in the literature, provides a clear and strategic roadmap for the implementation phase.

The CNN-LSTM hybrid model serves as a robust foundation for future work in this HDT research. Its two core components work synergistically: the Convolutional Neural Network (CNN) precisely extracts short-term behavioural patterns from clickstream data, while the Long Short-Term Memory (LSTM) network then analyses the sequence of these patterns over time to identify longer-term trends. This powerful combination makes it the ideal approach for detecting procrastination and academic exhaustion, not only for current implementations but also for all future expansions in HDT.

5.4. Implementation strategy for baseline model

Table 3-6: Workflow of the proposed LSTM Model adapted from Liu et al. (2023)

Step	Description
1. Select One Course for Simplicity	• Choose a single module with many active students (Liu et al., (2023) chose module "BBB" for simplicity purposes) • Filter out withdrawn students to keep the dataset clean.
2. Prepare and Label the Data	• Join Tables: Merge studentInfo and studentVLE using id_student, code_module, code_presentation. • Simplify Labels: Combine "Pass" and "Distinction" into a single Pass label. Keep "Fail" as is → Result: Binary target variable (Pass / Fail). • Remove Incomplete Data: Drop students who do not have any clickstream data recorded.
3. Extract Features	• Understand Click Types There are 12 types of VLE activity categories (e.g., homepage, quiz, forum, etc.). Each interaction (click) is tied to one of these. • Weekly and Monthly Aggregation Create a dataset: WEEK: Sum clicks per activity type per week. • Each row becomes: [student_id, week#, Act1_clicks, Act2_clicks, ..., Act12_clicks, Label] • Balanced Panel Data Ensure every student has a record for every week, even if they have 0 clicks (fill with 0s).
4. Prepare Data for LSTM Input	For LSTM, format the data like a time series: • Each student = 1 sequence • Shape: (n_students, n_time_steps, n_features) • Example: 5341 students × 40 weeks × 12 features Liu et al., (2023) built an LSTM model with sequential input of weekly click totals across 12 activity types
5. Define and Train the LSTM Model	Model Architecture (as per Liu et al. (2023)'s work) • two stacked LSTM layers • one dense (fully connected) output layer

	Use: • 32 & 8 hidden units in LSTM layers • Dropout: 0.2 • Optimiser: Adam • Loss: Binary cross entropy Hyperparameters • Batch size: 128 • Epochs: up to 700 • Learning rate: 0.0001
6. Evaluation Strategy	Use 10-fold cross-validation. Evaluate using: • Accuracy • F1-score • AUC (Area Under ROC Curve) Liu et al. (2023) found WEEK + LSTM model gave best performance (Accuracy: 89.25%, AUC: 0.913)
7. Analyse Feature Importance	• Systematically remove one activity feature (e.g., Act1, Act2...) and observe drop in performance. • Identify which activities (homepage, content, quiz...) matter most for prediction. Liu et al. (2023) identified homepage, subpage, content and quiz as most predictive activity categories
8. Apply Insight	Use dominant features (e.g., clicks on homepage, content, quizzes) to: • Target interventions for at-risk students.

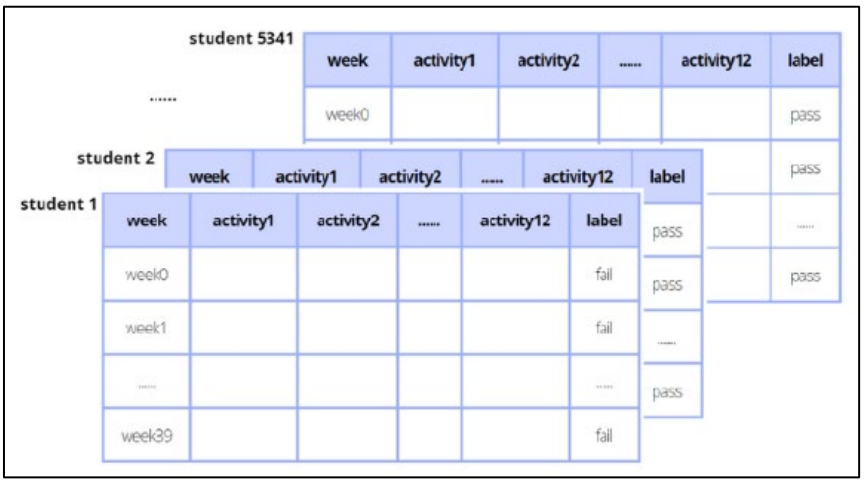

Figure 3-3: Visual representation of the shape of students' weekly feature set used for training the LSTM model – it contains 213,640 rows (5341 students × 40 weeks) with 13 columns in each row (12 features + one label) adapted from Liu et al. (2023). These sequences formed the input to the LSTM.

5.5. Implementation Guide for the "next step" hybrid model

Table 3-7: Workflow of the proposed CNN-LSTM Model adapted from Adefemi & Mutanga (2025)

Step	Description
1. Select Dataset	• OULAD – clickstream and demographic data from Open University, UK (32,593 students).
2. Preprocess the Data	• Address missing values by removing records with nulls. • Apply One-Hot Encoding for categorical variables. • Normalise all features using Min-Max Scaling (0–1 range). • Address class imbalance using SMOTE. • Perform feature selection using RFE.
3. Format Input for CNN–LSTM	Structure input as sequential data for CNN–LSTM: • Use Conv1D layer to extract feature patterns. • Use LSTM to model sequential dependencies. • Input shape: (samples, time steps, features).
4. Define Model Architecture	Model pipeline (Keras Sequential API): • Conv1D (64 filters, kernel size=3, ReLU) → MaxPooling1D • LSTM layer (100 units, ReLU) • Dropout layer (rate=0.5) • Dense layer (fully connected) • Output layer with Softmax for multi-class classification
5. Configure Training Parameters	• Optimiser: Adam • Loss function: Categorical Crossentropy • Batch size: 32 • Epochs: 10 • Learning rate: 0.001 • Regularisation: Dropout
6. Evaluate Model	Metrics used: • Accuracy • Precision • Recall • F1-Score Results on OULAD according to Adefemi & Mutanga (2025)'s implementation: 98.93% accuracy;

6. Conclusion

This chapter has fulfilled its primary objective of conducting a systematic and rigorous literature review to establish a strong academic and methodological foundation for the research. The review systematically analysed the relevant literature on predicting student performance from MOOC clickstream data, synthesising key findings to justify the proposed research design. The analysis confirmed a clear consensus in the field regarding the superiority of deep learning models, particularly LSTM networks, for processing the temporal and sequential nature of student interaction data. Furthermore, the review has solidified the standing of the Open University Learning Analytics Dataset as the essential benchmark dataset for this area of research, ensuring the comparability and relevance of this study's findings. The critical importance of deliberate feature engineering, robust data pre-processing and appropriate temporal data aggregation has also been highlighted as a key determinant of model performance.

Based on the robust evidence synthesised in this chapter, the study is now positioned to proceed with confidence into the implementation phase. The decision to utilise an LSTM-based model with the OULAD dataset is not merely a choice but a conclusion drawn from a thorough analysis of the state-of-the-art. The methodological design of the upcoming implementation (from data preparation and feature selection to model architecture and evaluation) is now firmly grounded in the successful and validated practices identified in the reviewed literature. The identification of the work by Liu et al. (2023) as a primary base model and Adefemi & Mutanga (2025) as an advanced hybrid alternative provides a clear, strategic and evidence-based roadmap for the practical work to follow. This literature review, therefore, serves as a critical bridge, connecting established academic knowledge with the novel contributions this research aims to produce.

BIBLIOGRAPHY

ADEFEMI, K.O. AND MUTANGA, M.B., 2025. A Robust Hybrid CNN–LSTM Model for Predicting Student Academic Performance. *Digital*, *5*(2), p.16.

AL-AZAZI, F.A. AND GHURAB, M., 2023. ANN-LSTM: A deep learning model for early student performance prediction in MOOC. *heliyon*, *9*(4).

BORNA, M.R., SAADAT, H., HOJJATI, A.T. AND AKBARI, E., 2024, December. Analyzing click data with AI: implications for student performance prediction and learning assessment. In *Frontiers in Education* (Vol. 9, p. 1421479). Frontiers Media SA.

KHOUDI, Z., HAFIDI, N., NACHAOUI, M. AND LYAQINI, S., 2025. Leveraging machine learning and clickstream data to improve student performance prediction in virtual learning environments. *Information Discovery and Delivery*, (ahead-of-print).

KUZILEK, J., HLOSTA, M. AND ZDRAHAL, Z., 2017. Open university learning analytics dataset. *Scientific data*, *4*(1), pp.1-8.

LIU, Y., FAN, S., XU, S., SAJJANHAR, A., YEOM, S. AND WEI, Y., 2022. Predicting student performance using clickstream data and machine learning. *Education Sciences*, *13*(1), p.17.

QU, S., LI, K., WU, B., ZHANG, S. AND WANG, Y., 2019. Predicting student achievement based on temporal learning behavior in MOOCs. *Applied Sciences*, *9*(24), p.5539.

SOUAI, W., MIHOUB, A., TARHOUNI, M., ZIDI, S., KRICHEN, M. AND MAHFOUDHI, S., 2022, May. Predicting at-risk students using the deep learning BLSTM approach. In *2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)* (pp. 32-37). IEEE.

WAHEED, H., HASSAN, S.U., ALJOHANI, N.R., HARDMAN, J. AND NAWAZ, R., 2020. Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human behavior*, *104*, pp.106189-106189.