Designing and evaluating a big data analytics approach for predicting students' success factors

Journal of Big Data

METHODOLOGY

Open Access

Designing and evaluating a big data analytics approach for predicting students' success factors

Kiran Fahd[1] and Shah J. Miah[1*]

*Correspondence:
shah.miah@newcastle.edu.au

[1] Business Analytics, Newcastle Business School, University of Newcastle, Hunter Street, Newcastle, NSW 2300, Australia

Abstract

Reducing student attrition in tertiary education plays a significant role in the core mission and financial well-being of an educational institution. The availability of big data source from the Learning Management System (LMS) can be analysed to help with the attrition issues. This study aims to use an integrated Design Science Research (DSR) methodology to develop and evaluate a novel Big Data Analytical Solution (BDAS) as an educational decision support artefact. The BDAS as DSR artefact utilises Artificial Intelligence (AI) approaches to predict potential students at risk. Identifying students at risk helps to take timely intervention in the learning process to improve student academic progress for increasing their retention rate. To evaluate the performance of the predictive model, we compare the accuracy of the collection of representational AI algorithms in the literature. The study utilized an integrated DSR methodology founded on the similarities of DSR and design based research (DBR) to design and develop the proposed BDAS employing an specific evaluation framework that works on real data scenarios. The BDAS does not only aimto replace any existing practice but also support educators to implement a variety of pedagogical practices for improving students' academic performance.

Keywords: Design Science Research (DSR), Big Data, Big Data Analytical Solution (BDAS), Machine Learning (ML), Deep Learning (DL), DSR evaluation, Artificial Intelligence (AI)

Introduction

Despite the increasing demand for higher qualifications in the industry, a greater number of students discontinue their studies without completing a degree in comparison to the past, according to the statistical analysis of HE degrees completion in Australia [1]. On average 23% of the enrolled students in the tertiary sector left without completing the course [1, 2]. Student attrition is a challenging issue of higher education (HE) providers. The HE providers compete to acquire the students and find strategies to retain them. The tertiary institutions have been attentive towards the student numbers revolving around the declined enrolment, increased competition, retention rate, or attrition rate. Attrition is a natural part of higher education that can be defined by the number of

non-completing students who leave their degree programs before finishing according to expected pre-schedule [3]. Several studies claimed that the attrition trend is significantly increased in Australia.

The incremental change in the attrition rate, as shown in Fig. 1, has multiple consequences ranging from social, and economic [4]. Student attrition not only negatively impacts the social interaction of individuals, but also results in negative financial consequences for students, institutions, and the economy. Students, not completing their education degree, fails to find better or appropriate career opportunities. HE providers lose revenue and reputation if students leave before finishing their education. Student attrition not only costs HE providers but the government as well. Non-completing students are unable to peruse progressive careers to earn a well-paid income. Consequently, this may bring students into a situation of not being able to pay back their study loans [5]. According to the Parliament of Australia [6], the total amount of outstanding study loans of approximately 3 million Australians was $68.7 billion in 2020 and approximately 16% of which is not expected to be repaid. Existing studies have been introduced the area of curriculum design [7, 8] and student performance improvement (given in the next section), but student attrition has not been given much attention. Considering these factors, the Department of Education, Science, and Training (DEST) has emphasised student attrition recently as one of the indicating factors to improve the performance of HE providers [9, 10]. This has opened a persistent opportunity for the researchers to study HE student attrition and measure different factors and strategies [11] to reduce student attrition.

In the relevant literature [12], student academic progress is considered one of the key determinants of student attrition. The providers can extend academic support to students through quality learning and teaching to enhance their academic performance. Early and timely identification of students at risk by using any Information System (IS) can support the HE providers to take appropriate measures effectively to enhance student academic progress [13–15]. For example, an Educational Decision Support System (DSS) can be considered a paramount IS to support the appropriate relevant decision

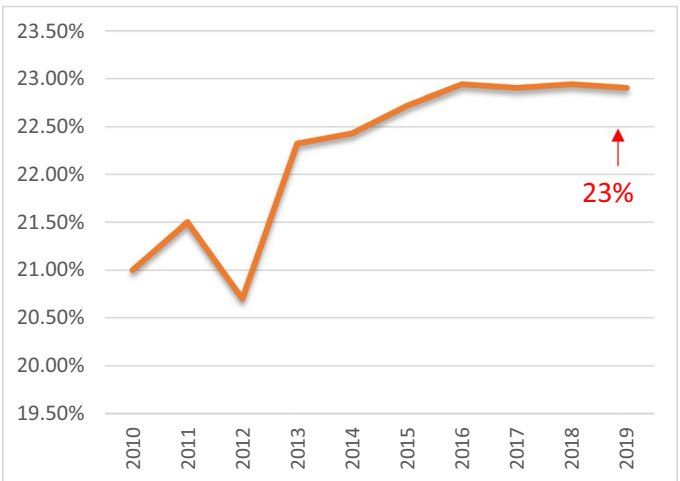

Fig. 1 Statistical analysis of TEQSA data for student attrition trend (adopted from [1])

[16]. The management can arrange useful early interventions that can help students to cope well in their academics and improve their academic progress. This can increase the probability of not going into the path of leaving studies leading to a low attrition rate.

Big data is defined by multiple Vs (e.g. volume, variety, velocity etc.) characteristics [17]. However, three innate characteristics of big data are Velocity defines the rate at which data is generated, Volume defines the vast scale of the data and Variety defines various sources and different formats of the data [18]. In HE, educational big data is gathered from different educational management activities, academic or non-academic activities of student. Voluminous and different set of student data is generated from educational information systems like Student management information systems, LMS or administrative management system such as demographic and socio-economic data, personal, social, enrolment data, academic attributes-based data, and LMS log data [19]. Big data analytics processes large heterogeneous datasets and supports data visualization, adaptive learning, and feedback systems to provide valuable insight for educators [20–22] and widely adopted in educational sector. Big data analytics can be classified as descriptive analytics, diagnostics analytics, decisive analytics, prescriptive analytics, and predictive analytics [23]. Machine Learning (ML), Cluster analysis, Text mining, Knowledge domain and reasoning based approaches, decision making methods, pattern matching, search and optimization theory algorithms and semantic analysis are well-known big data analytics techniques and approaches in AI discipline [24–26]. Different AI based big data analysis techniques can be employed on these type of bid data to identify students at risk of failing by predicting their academic performance. AI based data analytics techniques can be applied to these datasets to automate the analytical model building to achieve the aim of predicting academic performance. These AI based predictive models can embed into an Educational DSS to support educational management to plan and offer support mechanisms that are beneficial and effective for struggling students to assist them in attaining their academic success goals.

In this research, we adopted an innovative research methodology to develop and evaluate a novel BDAS for accurately predicting the students at risk of failing in the early weeks of the semester by utilizing a trained model on the student LMS interaction dataset. This BDAS supports educators to focus more on teaching and research, instead of undertaking tedious and inefficient administrative duties which can be biased due to human intervention.

This study has three novelties. First, the innovative research methodology is grounded on the similarities of Design Science Research (DSR) and Design-based research (DBR) for developing and evaluating BDAS. DSR together with DBR are applied in educational artefact design for various technological interventions for enhancing learning flexibility and outcome. The DSR and DBR has been viewed to symbolise the designer mind and behaviour that are situated within the pragmatic philosophical tradition. DSR concerns on functioning artefacts while DBR does give importance to design novel artefacts applying technology-in-practice to educational settings. We anticipated that this methodological view suits our research to study the application of ML technologies for improving student learning. Second, the BDAS is based on LMS data to detect potential students who can fail earlier in the semester to enhance student learning with accurate and timely intervention. Third, an extended evaluation framework is used to rigorously evaluate

the BDAS based on simulation of real scenarios. The timely detection and measurement will improve the student progress which will result in increased retention and decreased attrition with a positive impact on the student, HE providers, and the economy.

The remainder of the paper is organized as follows. First, we review the educational environment, research methodology, BDAS, and evaluation framework to identify the gap to explore. After the background, the paper defines the Integrated DSR methodology. It also details the major components of the research including the hybrid methodology framework, artefact design, and evaluation framework. Subsequently, the study presents the results and contributions made. In the final section, the study summarized the study and suggests future directions.

Background and related work

Recently, AI has been adopted in the computing field extensively and effectively. The benefits and enhancement due of AI in the education sector have been highlighted in the literature. A few examples of the application of AI in the educational sector, but not limited to, are applications of data analytics, predicting student enrolments, a recommendation system for career pathway or resource management, adaptive tutoring, prediction of student readiness for employment, monitoring and predicting student academic performance or identifying struggling students. Table 1 presents a brief overview of the related previous works.

Existing studies does not focus on the LMS big data to predict academic performance earlier in the learning pathways. Most of studies has used data generated from transitional on-campus educational settings or completely online settings and not much studies studied data generated by student interaction with LMS in blended learning. Also, most of the existing research did not highlight the significance of identification of at-risk students in early stages of studies. There is a need to investigate a real-time automated analytical solution to identify student at risk of failing earlier in blended learning environment to timely offer strategies and remedial measures to keep the student academic progress on track. Furthermore, most of the related studies from research methodology and DSR artefact construction and evaluation are insufficient considering: that these studies did not use integrated DSR and DBR methodology to layout the study to design and develop an artefact; these studies Big data analytics approaches but do not employ DSR or DBR or integrated DSR paradigm; these studies did not evaluate the DSR artefacts according to their complexity. However, existing literature can be leveraged to extrapolate to achieve the objective of this study, thus, forming the foundation of this study.

Big data, LMS and big data analytics

Big data technologies can play a significant role in improving data processing, data storage, data analytics and visualization [27]. Big data creates significant impact on the transformation of learning process and adoption of relevant innovative technologies [13]. The overview of big data analytics in HE is illustrated in Fig. 2. LMS platforms are considered as major source of big data and is an essential application to plan, deliver, monitor, and assess learning process e.g., Moodle, Blackboard, Canvas, Forma LMS, OpenOLAT. Moodle and Blackboard are most popular LMS platforms. LMS platform

Table 1 A brief overview of related work

Early prediction of undergraduate Student's academic performance in completely online learning: a 5-year study [15]	Proposed a collection of AI models to predict student academic progress from LMS interaction data and student academic data like GPA and enrolment test data. The data consists of LMS log files, demographics, and academic achievement. No research methodology is identified
Predicting Students' Academic Performance Through Supervised Machine Learning [61]	Developed an AI based system to predict student performance from their demographical and LMS interaction data. The dataset comprises of demographical characteristics and LMS interaction data including gender, country, birthplace, view of the LMS content, quiz attempts, and assessment submissions. The nature of the dataset does not allow early prediction. The research methodology is not clear
Predicting Students' Academic Procrastination in Blended Learning Course Using Homework Submission Data [62])	Develop an algorithm to enhance students' academic progress by detecting struggling students through their homework submission behaviours e.g., no submission or late submission. The nature of the dataset does not allow enough time to offer timely interventions and support to enhance student academic performance. No research methodology is identified to construct the predictive model e.g., DSR or DBR
An Efficient Approach for Multiclass Student Performance Prediction based upon Machine Learning [52]	Predicted the students' performance by using four classification algorithms The same dataset is used in other studies as well but with different ML classifiers [63, 64]. The study used secondary school students, not HE and did not use of LMS data Used socio-economic attributes of students which do not allow timely identification of the at-risk student. The research approach is not based on the similarities of DSR and DBR principles
Design, development, and evaluation of a mobile learning application for computing education [65]	Applied DSR approach to developing mobile learning application for HE for better student learning. The research approach is only based on the DSR approach and not on DBR principles or similarities between DSR and DBR. No AI (DL or ML) models are used to predict student academic performance
Predicting Student Performance in Higher Educational Institutions Using Video Learning Analytics and Data Mining Techniques [66]	Created a model to predict student overall performance at the end of the semester by analysing student academic information and video interactions data. The model is trained and tested using was tested with eight classification algorithms. The research approach used is quantitative prediction methodology which is not based on the similarities of DSR and DBR principles. The study mentioned early stages, but it does not state a definitive timeframe within the semester to show whether there is enough time to offer support to enhance student performance

has three key purposes: (i) management of digital content material and student access record, (ii) management of assessments and student progress, (iii) management of student feedback and interaction [28].

LMS generates rich and huge volume of data which increases the need of innovative solutions to improve learning and education management. There is also an emerging requirement of LMS integrated tools to interpret and manipulate the data generated by LMS [28, 29].

Big data is produced by users (e.g. educators, administrators, and students) interacting with LMS in different manners. For example, educators upload material to deliver digital course materials to their students and student access these for learning,

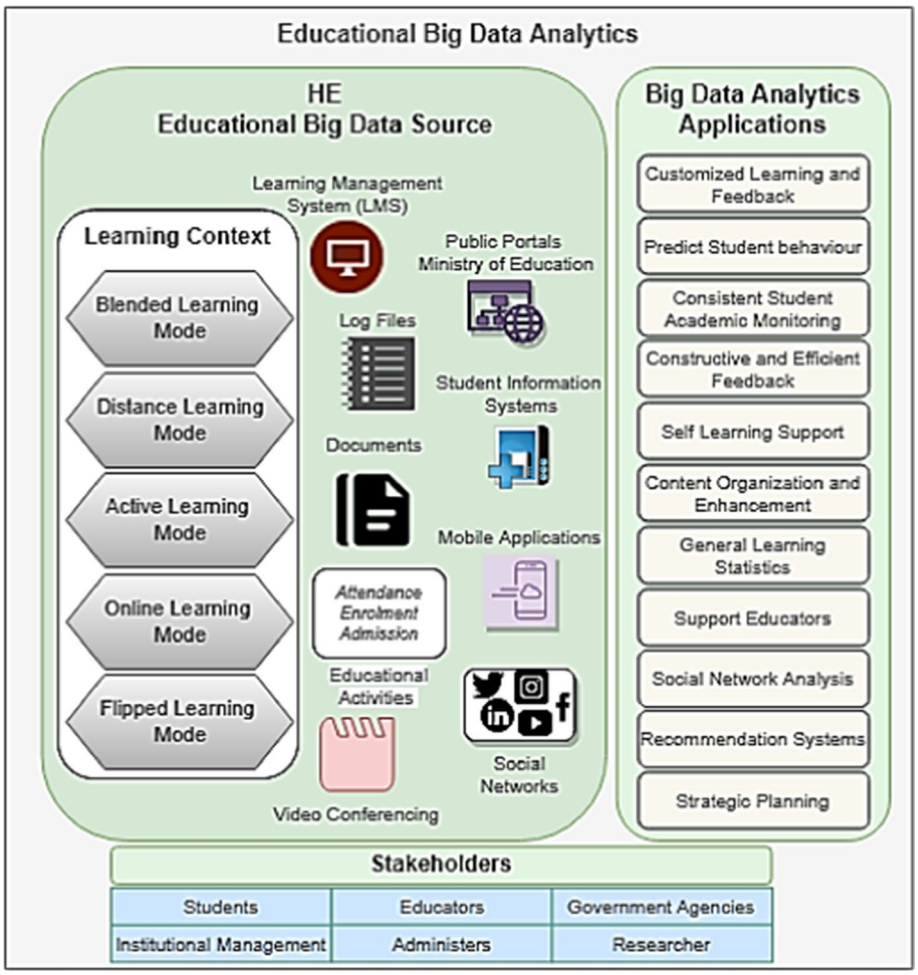

Fig. 2 Overview of big data analytics in HE

students attempt the LMS based tests related to a specific concept or students submits the assessment documents on LMS. Big data analytics applies set of analytical techniques to extract useful information and provide insight from big educational data related to students' learning behaviours, assessment scores, student learning styles, student logging in information, time spend on a task/module, assessment submission patterns, most visited page/content, completing a task or module or posting details about extracurricular activities [30–32].

Big data analytics allows to identify the real learning pattern of the students more accurately than the traditional practices. Big data analytics supports HE to make better and informed decision making based on the big data generated by LMS. It supports [28, 31, 33–35]:

- Customized and adaptive learning for better learning path
- Plagiarism detection in student submissions to improve academic integrity
- Student performance prediction for better course deliver planning
- Course Selection or Recommendation System

- Identification of students at risk based on their behaviour pattern to plan and delivery appropriate and timely interventions
- Dropout prediction
- Student participation and engagement measurement tracking to enhance learning experience
- Strategic planning to achieve HE goals

AI algorithms take all input data at once and process it to provide output, which is not possible in big data analytics due to its high velocity and huge volume. There are multiple approaches to solve this issue and apply AI algorithms on educational big data e.g., high-performing computing infrastructure, parallel processing approach and/or data processing platforms for data segmentation. In this study, data processing platform is suggested to deploy BDAS artefact [28, 31].Integrated Design Science Research Methodology.

Research methodology defines the guides and boundaries through which a study can be conducted ensuring its scientific value and significance. Researchers highlight research methodology as the most significant step to accomplish the purposes of the research. This study developed and used an innovative IS research methodology based on the similarities of two research approaches: DSR methodology from IS and Design based research (DBR) methodology. DBR is considered a DSR realization in the education sector to conduct research to develop and evaluate an BDAS as an IT and DSR artefact. DSR complements DBR and provides multi-paradigm perspectives to construct fundamental knowledge by researching social pragmatisms [36–38].

DSR approach suits the studies that will justify the research requirement and contribute to knowledge and development of the artefact [39]. For example, Miah et al. [40] have used the DSR framework to design a mobile based application for education; Carstensena and Bernhard [41] designed and improved teaching in the engineering education sector by utilizing the DSR methodology; Miah et al. [42] utilized DSR approach to extend mobile health information system; and Miah et al. [43] described development of the design of a DSS as method artefact. DBR methodology intends to achieve outcomes to improve student learning or enhanced understandings about teaching and learning or other educational phenomena [44]. The similarities among both methodologies are:

- Both are problem solving methodologies
- Both approaches design from a viable practical perspective
- Both approaches contribute to the knowledge based
- Both reflect on the nature of the theory
- Both produce the theoretical and practical artefact
- Both have an iterative cycle of design and rigorous evaluation

The study followed an integrated DSR methodology [45] consisting of five phases based on the similarities of DSR and DBR leveraging a variation of Peffer's DSR Methodology [39]. The five phases, as shown in Fig. 3, are: (1) Problem Identification; (2) Solution analysis; (3) Artefact Design and Development; (4) Evaluation; (5) Outcome Communication.

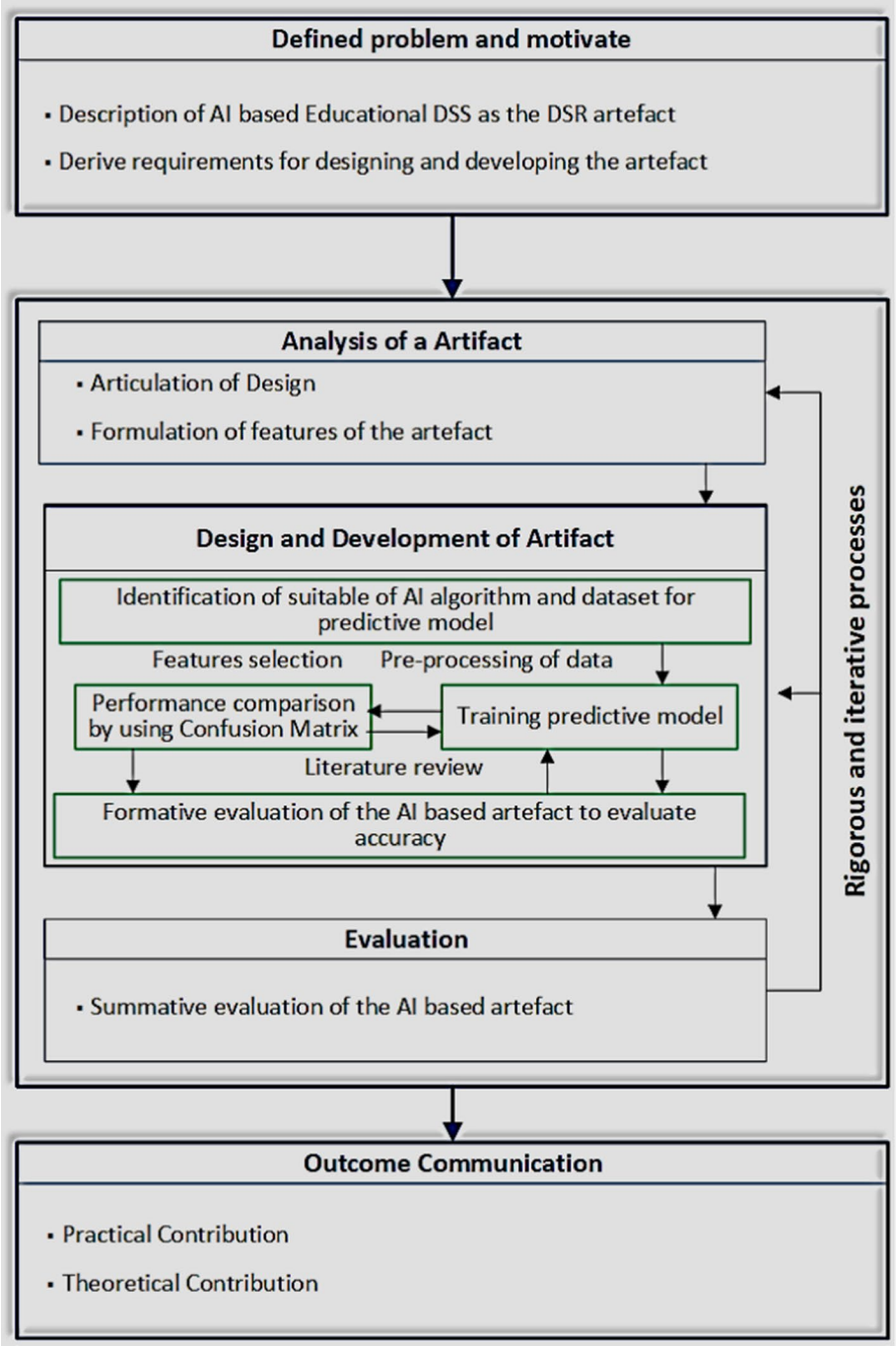

Fig. 3 Integrated DSR methodology

The study begins with a detailed problem description and analysis of existing studies to drive the design requirements and objective of designing an BDAS from the literature. This formulates the design principles of design and development of DSR artefact for a later phase by executing Systematic Literature Review and Meta Analysis. Next, the study evaluates the findings to establish design considerations for BDAS. In the third phase, BDAS as a DSR artefact is designed, developed, and evaluated formatively by

using AI data analysis techniques (ML and DL algorithms). In the final phases, the summative evaluation is carried out and the outcomes of the study are communicated as a contribution to the knowledge area.

Artefact description

This section focuses on the design process of the BDAS that addresses the identified problem of attrition related to student at risk of failure earlier in the semester. It provides the overview of the BDAS, details about the dataset utilized by the BDAS, training iterations of the BDAS to help to explain the structure and functionality of the DSR artefact i.e. BDAS.

Problem identification and objectives of the artefact

In the initial phases of our integrated DSR research methodology, an extensive systematic literature review and meta-analysis (SLRM) was conducted about the application of AI based technology in HE regarding student academic progress. The systematic literature review aims to understand the trends of application of AI based technology to a wide spectrum related to monitoring and predicting student academic performance and identify the different AI algorithms and process of development of AI models. The SLRM is conducted by using the PRISMA [46] framework with defining a search protocol incorporating inclusion and exclusion criteria and providing rich findings. The SLRM highlighted the phases, algorithms and evaluation metrics used in the studies. These algorithms and evaluation metrics form the foundation of the design and development of BDAS.

The objective of designing and developing the BDAS is to train and evaluate a predictive model with classified data to predict the student's academic progress. The predictive model must be sufficiently accurate to identify students who are at risk of failing. The prediction can assist educators to implement strategies to enhance student learning and improve their academic performance. BDAS can be integrated into coursework for timely and accurate identification of student academic progress, especially for the student at risk. This timely identification of students at risk supports earlier intervention to improve their academic performance. The generic computational model consists of Data collection, Data pre-processing, data analysis with algorithms and evaluation. This generic model is tailored for each iteration of the design and development phase for BDAS. Each iteration utilized various pre-processing techniques and different algorithms to achieve the objective of the BDAS. In case of educational big data, a large amount of real-time data is generated by LMS. The BDAS predictive model is trained on a set of historic LMS data of students' interaction with LMS as demonstrated in this study.. A distributed big data processing platform is used for collecting incoming big data and creating data segmentations e.g., Apache Kafka and Spark. These real-time big data small segments are fed to BDAS via pipelines to be classified to predict students academic performance for enhanced students academic progress and better decision making. A distributed big data processing platform is used for collecting incoming big data and creating data segmentations. These batches of big data are classified by the BDAS to identify students at risk and the ML models take all input data simultaneously to generate output, which is not possible in BDAS due to the massive volume and high

velocity of big data. There are various approaches to address this problem and apply AI algorithms to develop model on big educational data such as parallel processing techniques, high-performing computing infrastructure, data processing platforms for data partitioning. This study suggests adoption of data processing and handling platform for the BDA method architecture [28, 31]. However, this study primarily focuses on the design, development, and evaluation of the BDAS rather than the architectural environment of the BDAS. Figure 4 shows the process of design and development of DSR artefact as the BDAS.

Artefact design and development

An AI based DSR artefact is a complex artefact and designed according to the requirements and objectives identified in previous phases. Design approaches developed around contextual knowledge and general practices lead to enhanced artefact design [47]. This study has used two sets of iterations to design and develop the BDAS as a predictive model based on existing approaches in literature: ML based predictive model; DL based predictive model. In this phase, we apply ML and DL algorithms to design and develop ML based and DL based predictive models as DSR artefacts to identify potential students at risk of failing accurately from a dataset based on student LMS interaction. This iterative approach in this phase provides continuous improvement of the construction of DSR artefact by evaluating various performance metrics by using the confusion matrix in each iteration. These performance metrics of different AI algorithms in each iteration are compared to select the best predictive model.

BDAS as a DSR artefact is constructed by a series of tasks consisting of Data collection, Data pre-processing, Data analysis with AI algorithms, Evaluation and successful decision marking [13, 48]. All these tasks are tailored to develop and evaluate ML and DL based predictive models. The workflow of training an AI based artefact is illustrated in Fig. 5.

This study has sourced a freely available dataset the UCI (University of California, Irvine) ML repository [49] comprising 230,318 instances of students' activities and interactions with LMS to train the predictive model. The dataset consists of 14 features

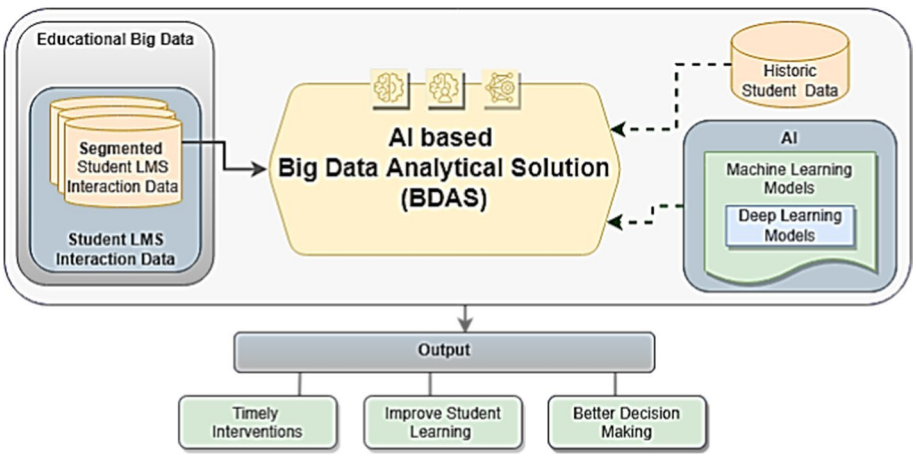

Fig. 4 Overview of BDAS as a DSR artefact

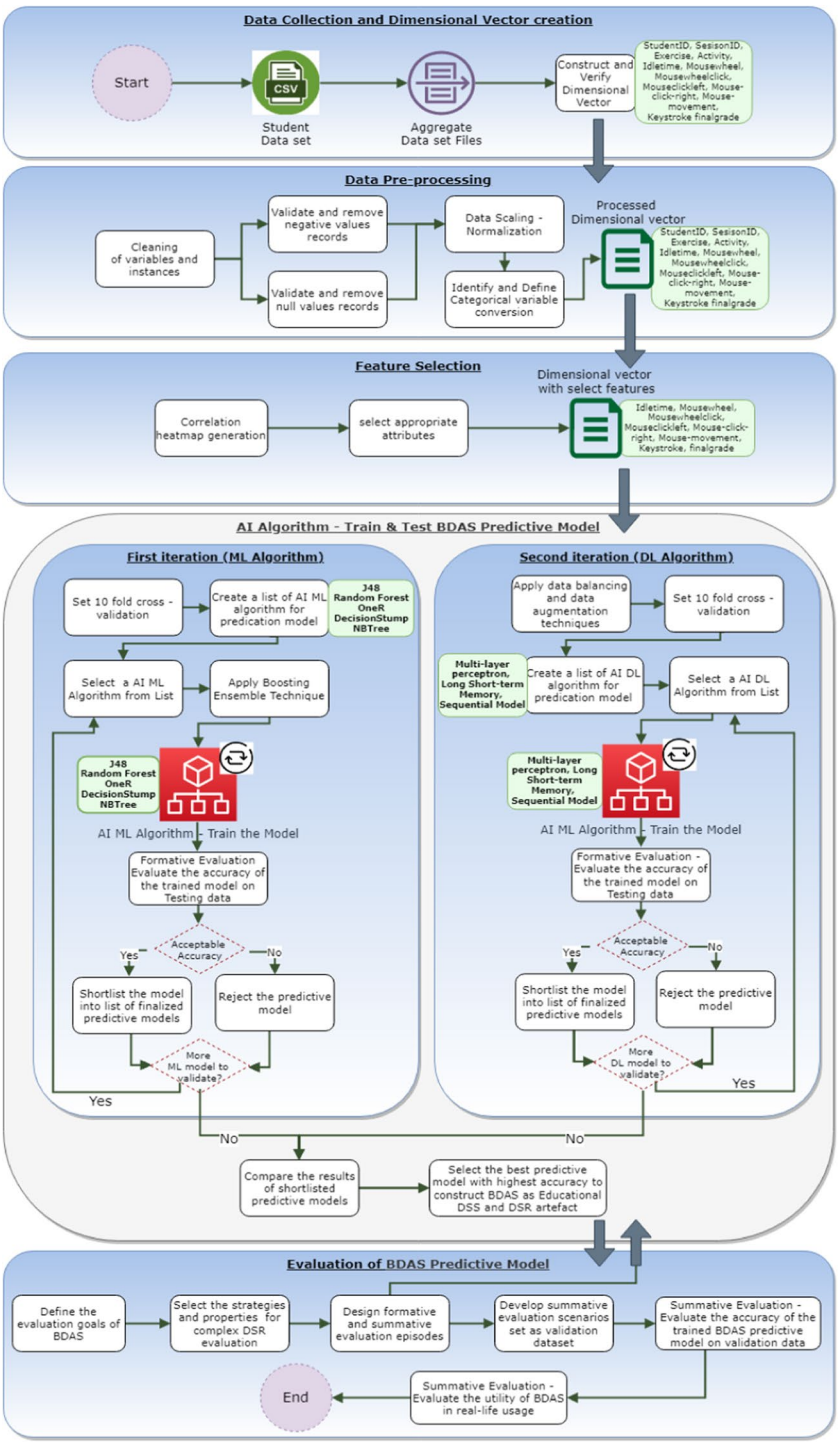

Fig. 5 Workflow of the rigorous and iterative phase of integrated DSR methodology to design, develop and trained BDAS as a DSR artefact

including time-series based features i.e., Session number, Student number, Exercise number, Activity name abbreviation, Start time of the activity, End time of the activity, Idle time during activity, Mouse wheel movement count, Mouse wheel click count,

count of Mouse left click, count of Mouse right click, Mouse movement count, count of Keystroke and final marks as given in the following Table 2.

The dataset is pre-processed and normalized, and features are selected by correlational analysis to build a dimensional vector including categorised features. The dataset consists of multiple comma-separated value (csv) documents containing data regarding sessions and students. An additional csv document contains the final marks of each student who attended the session at the end of the semester shows the result of the student. During the data pre-processing, negative, empty or null values are eliminated from the dataset. The dimensional vector is built by aggregating each feature for each student and merging them with the total final marks for the students. The final marks of the students are converted into classification categorical variables i.e. "Pass" or "Fail". Appropriate features are selected from the dataset from the 13 features and 1 categorical variable by using Correlation heat map to identify a +ve or −ve correlation with the final result (final total) as depicted in Fig. 5. For instance, the heat map shows that when keystroke has +ve correlation with final result i.e. when "keystroke" has high value then there is a higher probability that the final result (final total) will be higher value as well. This transformed dataset is then used to train the predictive model by using ML and DL algorithms to detect students at risk of failing.

Model improvement using multiple iterations aligns with the continuous improvement target of the artefact of our integrated DSR methodology. Each improvement iteration is executed to boost the predictive classification accuracy of the model and attain best suited model to develop BDAS. In the first iteration, ML model is trained using multiple ML algorithms and improved by tuning the classifiers with an ensemble technique Adaptive Boosting (AdaBoost). In the second improvement iteration, the dataset is balanced by applying data augmentation techniques such as Synthetic Minority Oversampling Technique (SMOTE) and different DL algorithms are applied to create the ML predictive model with improved prediction accuracy. In each model improvement iteration, different ML and DL techniques are used which derived from the literature review

Table 2 Features of the dataset used in the study

	Dataset features	Brief description
1	session_Id	Number based value for the session number
2	student_Id	Number based value for the student number
3	exercise	Number based value for the exercise number in a certain session
4	activity	Text based value for the abbreviation of activity categories
5	start_time	Date based value for the start date and time of an activity
6	end_time	Date based value for the end date and time of an activity
7	idle_time	Number based value for idle time during the activity
8	mouse_wheel	Number based value for count of the mouse wheel
9	mouse_wheel_click	Number based value for count of mouse wheel clicks
10	mouse_click_left	Number based value for count of mouse left clicks
11	mouse_click_right	Number based value for count of mouse right clicks
12	keystroke	Number based value for total count of keystrokes in a certain activity
13	mouse_movement	Number based value for the distance covered by the mouse movements
14	final_total	Number based value for the final marks of the student at the end of the semester

and analysis of the existing related works. The study has selected Decision tree classifiers as extensive existing work [50, 51] reveals that Decision tree based predictive models are simpler and exhibits better performance on educational data. Further, numerous studies have used ensemble techniques to develop predictive models to forecast the academic performance of the students [52–54]. In addition, MLP is selected as it is widely used to develop classification prediction modelling in the literature [55].

In the first iteration, five tree based ML supervised algorithms (J48, Random Forest, OneR, Decision Stump, NBTree,) are used to train and evaluate the predictive model. These tree based algorithms use a series of if–then decisions to generate highly accurate, easily interpretable predictions, to identify potential students at risk of failing. A booster ensemble technique is applied to the transformed dataset to further fine-tune it. The predictive model is trained and tested by using k-fold cross validation on the training and testing data using the above five ML supervised algorithm iteratively. In the final step, performance metrics are compared for all the predictive models based on five ML algorithms to select the most accurate predictive model to construct BDAS. In the real-time implementation of the BDAS, a data processing framework, e.g., Apache spark, will be used to receive and segment the real-time big data stream from LMS and decomposes the large data into small batches to be processed and classified by the BDAS predictive model.

In the second iteration of continuous improvement of the AI based artefact, two different data pre-processing techniques are used to modify the class distribution and augment the dataset to resolve the implications of an imbalance dataset. DL algorithms are made up of neural networks with several layers of differentiable nonlinear nodes. Three DL algorithms Long Short-term Memory (LSTM), Multi-layer perceptron (MLP) and Sequential Model (SM), are applied to train the augmented dataset which demonstrated higher classification accuracy of the prediction model and reduces false prediction. The higher classification accuracy and reduced false prediction mean a low instance of incorrectly not identifying students who are not at-risk, therefore addressing the objective of the general description of the BDAS as a DSR artefact.

Artefact evaluation

The evaluation phases focus on whether the developed artefact has achieved the purpose it is designed for and it is a vital phase of a study in the DSR domain. The evaluation of the developed artefact within its context is a vital component of the evaluation strategy [56]. In this study, BDAS as the artefact is evaluated by an innovative DSR evaluation framework to evaluate the utility, efficacy, and effectiveness [57, 58] of the artefact with hybrid evaluation requirements by using the Confusion matrix, given in Table 3. In addition, to train, test and evaluate an AI based predictive model the original dataset in sectioned into three sections i.e., Training dataset, Testing dataset and Validation dataset. The predictive model is trained and testing on the training dataset and testing dataset respectively during the construction of the predictive model. The trained predictive model is evaluated to define a generalize predictive model by using the validation dataset.

The efficacy and effectiveness of the BDAS have evaluated whether the artefact provides the desired output or not i.e., the high classification accuracy. The BDAS as DSR

Table 3 The Confusion matrix to evaluate the performance of the BDAS predictive model

	Predicted values		
	Positive	**Negative**	
Actual values			
Positive	True positive (TP)	False negative (FN)	Recall/sensitivity $\frac{TP}{(TP+FN)}$
Negative	False positive (FP)	True negative (TN)	
	Precision $\frac{TP}{(TP+FP)}$		Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

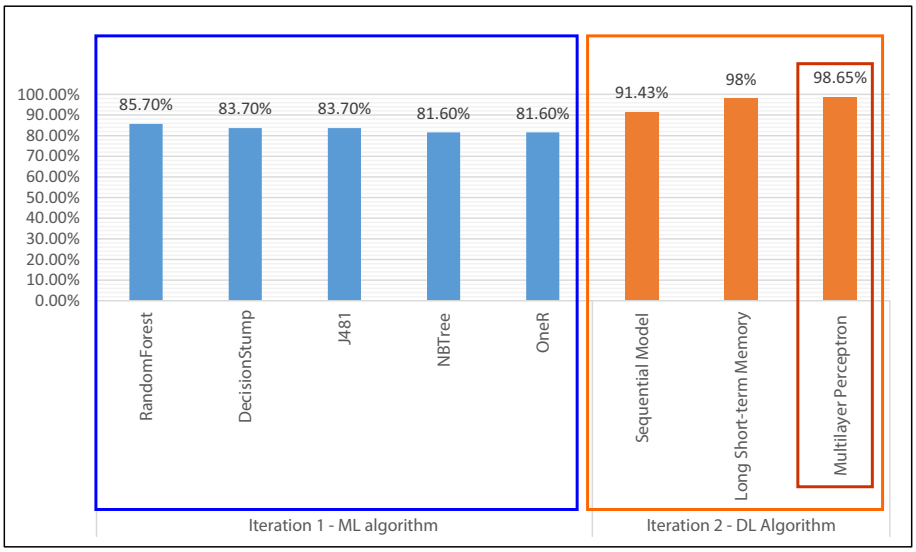

Fig. 6 Comparison of formative evaluations of predictive models by using the Confusion matrix

artefact is evaluated by an innovative evaluation framework, which extends Venable's [59] Framework for Evaluation in Design Science (FEDS) and composed a series of formative and summative evaluation episodes. The innovative evaluation framework has extended the 2 and 4 steps of FEDS which are: (1) Define the evaluation goal(s), (2) Select the strategy, (3) Establish the properties to evaluate, and (4) Design and Develop the evaluation episodes. For fourth step, there is not much guidance available on how to plan and execute formative or summative evaluation episodes. In this innovative evaluation framework, the steps of each evaluation episodes are structured according to the phases of the IT-dominant BIE (Building, Intervention, and Evaluation) schema of the Action Design Research (ADR) [60]. The innovative evaluation framework emphasises on executing a formative evaluation in the very beginning of the study to evaluate the significance of the artefact. Later formative evaluation episodes as interim evaluations are executed to improve the artefact during the design and development phase. The formative evaluation episodes are executed using the training and testing dataset (as explained above). The comparison of classification accuracy from the formative evaluation episodes is presented in Fig. 6. The comparison clearly demonstrates that predictive model accuracy has been improved during the iterative design and development phase and MLP outperformed other models with an accuracy of 98.65% (see Table 3 below).

The summative evaluation episodes highlight the outcome and impact of the implemented artefact in a context, thus performed towards the completion of the study. One of the summative episodes was performed to evaluate the effectiveness and efficacy of the predictive model by accurately identifying the students at risk early in the semester. Validation dataset is used to execute the terminal evaluation episode to evaluate the effectiveness the BDAS predictive model and generate a generalise the BDAS predictive model. The second and final summative episode, an ex-post evaluation, to evaluate the utility of real users with live unseen data is left for future work.

Discussion and conclusion

The study outlined an integration of two research methodologies DSR and DBR based on key similarities between them to design, construct and evaluate a new DSRartefact called BDAS.. The methodological view forms an appropriate research paradigm for designing, developing, and evaluating the BDAS artefact that can be implemented to enhance academic performance with timely intervention strategies for those who are at risk of failing and to support better decision making.

Several technological opportunities like learning analytics are emerging due to the big data from LMS in HE. The objective of BDAS artefact complements existing practices to support educators to discover the potential students at a very early risk in the semester and contact students to take remedial actions and mitigate the risk of dropping out. This paper presents the steps to design and develop an AI based BDAS by using integrated DSR methodology and rigorously evaluate to improve the accuracy of BDAS identifying the students at risk. The big data analytics approach contributes to the knowledge area as it utilized multiple AI techniques to improve the accuracy of predictive model i.e., performing correlations between LMS attributes to select attributes, tuning of classifier algorithm parameters, augmenting the dataset and applied both ML and DL algorithms to select best performing predictive model to construct BDAS artefact.

In a broader sense, our solution design research aimed to promote studies of predictive artefact design that have potentials to advance technology-based innovations in other aspects in education sector [61, 62]. Extension of the studies to design predictive artefact provide enormous opportunities for creation of new practical knowledge, although it is recommended that exploration of design research methodology such as design science [63, 64] can be of paramount integral study-task. Studies in future would enable advancement in designing more with innovations in other problem domains, such as for healthcare information management [65–67] and supply chain [68] for delivering predictive outcome.

This paper presents the two phases to design and develop predictive model to improve identification accuracy. This AI based BDAS can be an alarming system for educators to provide appropriate support by taking necessary steps to improve students academic progress. Our BDAS approach fills the gap of using data generated by student interaction with LMS in blended learning and automated process almost real-time and an early detection of student at risk of failing in blended learning environment, which is beneficial from both academic and administrative perspectives. In addition, in this paper, a great focus is given to evaluate the AI based BDAS by executing numerous formative and summative evaluation episodes. The innovative evaluation framework provides well

designed phases including evaluation episode plans to guide future researchers about evaluating hybrid artefact like BDAS. The AI based BDAS as Educational DSS would be useful for students and educators from different HE providers (e.g., Massive open online course (MOOC), universities, Non-University Higher Education (NUHE) not to derail their learning pathway.

High performing computational infrastructure and interoperability of educational big data are required for practical deployment of BDAS in educational system. In the future, we will work on the full implementation of the BDAS and integration of the BDAS into the LMS of the students to evaluate the efficiency and utility in the real-time use of the BDAS by students and educators as clients. The extension will enhance the details about how the BDAS might support decision-making about which strategies to use for students identified at risk.

Abbreviations

ADR	Action Design Research
AdaBoost	Adaptive Boosting
AI	Artificial Intelligence
BDAS	Big Data Analytical Solution
BIE	Building, Intervention, and Evaluation
csv	Comma-separated value
DBR	Design-based research
DEST	Department of Education, Science, and Training
DL	Deep Learning
DSR	Design Science Research
DSS	Decision Support System
FEDS	Framework for Evaluation in Design Science
HE	Higher Education
IS	Information System
LMS	Learning Management System
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multi-layer perceptron
MOOC	Massive open online course
NUHE	Non-University Higher Education
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SLRM	Systematic literature review and meta-analysis
SM	Sequential Model
SMOTE	Synthetic Minority Oversampling Technique

Acknowledgements
Not applicable.

Author contributions
All authors contributed equally. Both authors read and approved final manuscript.

Funding
Not applicable.

Availability of data and materials
The data used in the study is downloaded from public data repository and is available publicly.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 17 September 2022 Accepted: 3 October 2023
Published online: 13 October 2023

References

1. Cherastidtham I, Norton A. University attrition: what helps and what hinders university completion? Grattan Institute; 2018.
2. Martín MD, Jansen L, Beckmann EA. Understanding the Problem Student attrition and retention in university Language & Culture programs in Australia. The Doubters' Dilemma. Exploring student attrition and retention in university language and culture programs. ANU Press; 2016. p. 1–30.
3. TEQSA. Tertiary education quality and standards agency (TEQSA)'s risk assessment framework. Australian Government; 2019. Contract No.: 2.3.
4. Ulriksen L, Madsen LM, Holmegaard HT. What do we know about explanations for drop out/opt out among young people from STM higher education programmes? Stud Sci Educ. 2010;46(2):209–44.
5. Sarra A, Fontanella L, Di Zio S. Identifying students at risk of academic failure within the educational data mining framework. Soc Indicat Res. 2019;146:41–60.
6. Ferguson H. Parliment of Australia; 2021. Cited 2022. https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/FlagPost/2021/November/HELP-2020-21.
7. Miah SJ, Solomonides I, Gammack J. A design-based research approach for developing data-focussed business curricula. Educ Inf Technol. 2020;25:553–81.
8. Miah SJ, Solomonides I. Design requirements of a modern business master's degree course: perspectives of industry practitioners. Educ Inf Technol. 2021;26:763–81.
9. Panel HES. Final report—improving retention, completion and success in higher education. Department of Education and Training (DEST); 2017. Contract No.: ISBN: 978-1-76051-156-2.
10. Institute TV. Student attrition report comprehensive analysis and recommendations. Victoria University; 2013.
11. Aljohani O. A comprehensive review of the major studies and theoretical models of student retention in higher education. High Educ Stud. 2016;6:1–18.
12. Beer C, Lawson C. The problem of student attrition in higher education: an alternative perspective. J Furth High Educ. 2017;41(6):773–84.
13. Miah SJ, Miah M, Shen J. Editorial note: learning management systems and big data technologies for higher education. Educ Inf Technol. 2020;25:725–30.
14. Plak S, Cornelisz I, Meeter M, van Klaveren C. Early warning systems for more effective student counselling in higher education: evidence from a Dutch field experiment. High Educ Q. 2022;76(1):131–52.
15. Bravo-Agapito J, Romero SJ, Pamplona S. Early prediction of undergraduate Student's academic performance in completely online learning: a five-year study. Comput Hum Behav. 2021;115: 106595.
16. Miah SJ. An ontology based design environment for rural decision support. Griffith University; 2008.
17. Sarker K, Deraman A, Hasan R, Abbas A. Ontological practice for big data management. Int J Comput Dig Syst. 2019;8:265–73.
18. Fahd K, Parvin S, Souza-Daw Ad, editors. A Framework for Real-time Sentiment Analysis of Big Data Generated by Social Media Platforms. In: 2021 31st international telecommunication networks and applications conference (ITNAC); 2021 24–26 Nov. 2021.
19. Hasan R, Palaniappan S, Mahmood S, Abass A, Sarker K. Dataset of students' performance using student information system, moodle and the mobile application "eDify." Data. 2021;6:1–10.
20. Kumar P, editor Big Data Analytics: An Emerging Technology. In: 2021 8th international conference on computing for sustainable global development (INDIACom); 2021 17–19 March; 2021.
21. Chunzi S, Xuanren W, Ling L, editors. The application of big data analytics in online foreign language learning among college students: empirical research on monitoring the learning outcomes and predicting final grades. In: 2020 2nd international conference on machine learning, big data and business intelligence (MLBDBI); 2020 23–25 Oct; 2020.
22. Babiceanu RF, Seker R. Big Data and virtualization for manufacturing cyber-physical systems: a survey of the current status and future outlook. Comput Ind. 2016;81:128–37.
23. Sun X, Fu Y, Zheng W, Huang Y, Li Y. Big educational data analytics, prediction and recommendation: a survey. J Circ Syst Comput. 2022;31:2230007.
24. Sekeroglu B, Abiyev R, Ilhan A, Arslan M, Idoko J. Systematic literature review on machine learning and student performance prediction: critical gaps and possible remedies. Appl Sci. 2021;11:10907.
25. Rahmani A, Azhir E, Ali S, Mohammadi M, Ahmed O, Ghafour M, et al. Artificial intelligence approaches and mechanisms for big data analytics: a systematic study. PeerJ Comput Sci. 2021;7: e488.
26. Begum A, Fatima F, Haneef R. Big data and advanced analytics: helping teachers develop research informed practice; 2019, In book: Explorations in Technology Education Research, 594–601, https://doi.org/10.1007/978-3-030-11890-7_57.
27. Otoo-Arthur D, van Zyl T. A scalable heterogeneous big data framework for e-learning systems. IEEE; 2020. p. 1–15.
28. Ang L-M, Ge F, Seng K. Big educational data & analytics: survey, architecture and challenges. IEEE Access. 2020;8:116392–414.
29. Cantabella M, Martínez-España R, Ayuso B, Yáñez J, Muñoz A. Analysis of student behavior in learning management systems through a big data framework. Fut Gen Comput Syst. 2018;90:262–72.
30. Otoo-Arthur D, van Zyl T. A systematic review on big data analytics frameworks for higher education—tools and algorithms; 2019., Proceedings of the 2nd International Conference on E-Business, Information Management and Computer Science, 15, 1–9, https://doi.org/10.1145/3377817.3377836 p. 1–9
31. Elatia S, Ipperciel D. Learning analytics and education data mining in higher education. IGI Global; 2021. p. 108–26.

32. Anshari M, Alas Y, Yunus N, Sabtu N, Hamid M. Online learning: trends, issues, and challenges in the big data era. J E-Learn Knowl Soc. 2016;12:121–34.
33. Sharma A, Dhaka A, Nandal A, Swastik K, Kumari S. Big data analysis: basic review on techniques. 2021, In book: Advancing the Power of Learning Analytics and Big Data in Education. p. 208–33, https://doi.org/10.4018/978-1-7998-7103-3.ch010
34. Ashaari MA, Dara Singh K, Abbasi G, Amran A, Cabanillas F. Big data analytics capability for improved performance of higher education institutions in the Era of IR 4.0: a multi-analytical SEM & ANN perspective. Technol Forecast Soc Change. 2021;173: 121119.
35. Şahin M, Yurdugül H. Educational data mining and learning analytics: past. Present and Future. 2020;9:121–31.
36. Singh H, Miah S. Smart education literature: a theoretical analysis. Educ Inf Technol. 2020;25:3299–328.
37. Genemo H, Miah S, McAndrew A. A design science research methodology for developing a computer-aided assessment approach using method marking concept. Educ Inf Technol. 2015;21:1769–84.
38. Miah SJ, Gammack J. Ensemble artifact design for context sensitive decision support. Aust J Inf Syst. 2014;18(2):5–20.
39. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. J Manag Inf Syst. 2007;24(3):45–77.
40. Singh H, Miah SJ. Design of a mobile-based learning management system for incorporating employment demands: case context of an Australian University. Educ Inf Technol. 2018;24(2):995–1014.
41. Carstensen A-K, Bernhard J. Design science research—a powerful tool for improving methods in engineering education research. Eur J Eng Educ. 2019;44(1–2):85–102.
42. Miah SJ, Gammack J, Hasan N. Extending the framework for mobile health information systems research: a content analysis. Inf Syst. 2017;69:1–24.
43. Miah SJ, Kerr D, Hellens L. A collective artefact design of decision support systems: design science research perspective. Inf Technol People. 2014;27:259–79.
44. Anderson T, Shattuck J. Design-based research. Educ Res. 2012;41:16–25.
45. Fahd K, Miah SJ, Ahmed K, Venkatraman S, Miao Y. Integrating design science research and design based research frameworks for developing education support systems. Educ Inf Technol. 2021;26:4027–48.
46. Page M, McKenzie J, Bossuyt P, Boutron I, Hoffmann T, Mulrow C, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372: n71.
47. Miah SJ, Gammack JG, McKay J. A metadesign theory for tailorable decision support. J Assoc Inf Syst. 2019;20:570–603.
48. Janssen M, Voort H, Wahyudi A. Factors influencing big data decision-making quality. J Bus Res. 2016;70:338–45.
49. Vahdat M, Oneto L, Anguita D, Funk M, Rauterberg M. A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator. Springer International Publishing; 352–366, 2015, accessed on 5 Oct, 2023, https://doi.org/10.1007/978-3-319-24258-3_26
50. Thai-Nghe N, Janecek P, Haddawy P. A comparative analysis of techniques for predicting academic performance. IEEE; 2007. p. T2G7.
51. Rokach L. Ensemble-based classifiers. Artif Intell Rev. 2010;33:1–39.
52. Jain A, Solanki S, editors. An efficient approach for multiclass student performance prediction based upon machine learning. In: 2019 international conference on communication and electronics systems (ICCES); 2019 17–19 July; 2019, https://doi.org/10.1109/ICCES45898.2019.9002038
53. Xu J, Moon K, Schaar M. A machine learning approach for tracking and predicting student performance in degree programs. IEEE J Select Top Signal Process. 2017;11:742–53.
54. Tenpipat W, Akkarajitsakul K, editors. Student dropout prediction: a KMUTT case study. In: 2020 1st international conference on big data analytics and practices (IBDAP); 2020 25–26 Sept; 2020.
55. Altay O, Varol AE. A novel hybrid multilayer perceptron neural network with improved grey wolf optimizer. Neural Comput Appl. 2023;35(1):529–56.
56. Miah SJ, Debuse J, Kerr D. A development-oriented IS evaluation approach: case demonstration for DSS. Aust J Inf Syst. 2012. https://doi.org/10.3127/ajis.v17i2.694.
57. Hevner AR, March ST, Park J, Ram S, et al. Design Science in Information Systems Research. Manag Inf Syst Quart. 2004;28:75.
58. Venable J. The role of theory and theorising in design science research. In: First international conference on design science research in information systems and technology; 2006.
59. Venable J, Pries-Heje J, Baskerville R. FEDS: a framework for evaluation in design science research. Eur J Inf Syst. 2016;25(1):77–89.
60. Sein M, Henfridsson O, Purao S, Rossi M, Lindgren R. Action design research. MIS Q. 2011;35:37–56.
61. Miah SJ, Samsudin AZH. EDRMS for Academic Records Management: A Design Study in a Malaysian University, Educ Inf Technol. 2016;22:1895–1910
62. Muhammad JS, Isa AM, Samsudin, AZH, Miah SJ. Critical factors for implementing effective information governance in Nigerian universities: A case study investigation, Educ Inf Technol. 2020;25:5565–5580
63. Miah SJ, McGrath GM, Kerr D (2016). Design science research for decision support systems development: recent publication trends in the premier IS journals, Australas J Inf Syst. 20, https://doi.org/10.3127/ajis.v20i0.1482
64. Ali MS, Miah SJ. Identifying Organizational Factors for Successful Business Intelligence Implementation, International J Bus Intell Res. 2017;9(2). https://doi.org/10.4018/IJBIR.2018070103
65. Miah SJ, Hasan N, Gammack JG. A Methodological Requirement for designing Healthcare Analytics Solution: A Literature Analysis, J Health Inform. 2019;26(4):2300–2314
66. Miah SJ, Hasan N, Gammack JG. A Follow-up Decision Support Artifact for Public Healthcare: A design research perspective, Healthc Inform Res. 2021;25(4):313–323
67. Hasan N, Bao Y, Miah SJ. Exploring the impact of ICT usage among indigenous people and their quality of Life: operationalizing Sen's capability approach, Inf Technol Dev. 2021;28(2):230–250

68. de Vass T, Shee H, Miah SJ (2018). Internet of Things for improving Supply Chain Performance: A Qualitative study of Australian retailers, 29th Australasian Conference on Information Systems- ACIS 2018, Sydney, Australia, https://aisel.aisnet.org/acis2018/90

Publisher's Note