

# A BiLSTM - DINA Enhanced Trajectory-driven Adaptive Decision Framework with Data Augmentation Using Longitudinal Academic Data

Swathieswari Mohanraj <sup>1</sup>, Shanmugavadivu P <sup>2</sup>

<sup>1</sup> Department of Computer science and Applications, Gandhigram Rural Institute (Deemed University), Gandhigram, Dindigul, Tamil nadu, India ,

<sup>2</sup> Department of Computer science and Applications, Gandhigram Rural Institute (Deemed University), Gandhigram, Dindigul, Tamil nadu, India ,

## Abstract

The application of learning analytics allows for the analysis of longitudinal data for learner-performance prediction. This study suggests a trajectory-based BiLSTM-DINA approach for predicting learner mastery from longitudinal academic performance in Mathematics, Science, and English. The data for this study involved 1,549 students who provided trajectory-based features in terms of achievement, progress, variability, and mastery of different subjects. The proposed approach was tested using the original dataset, SMOTE-balanced data, and GAN-augmented data. SMOTE enlarged the dataset to 1,550 data points, with one synthetic data point, while GAN created 5,000 synthetic data points, with 2,500 points per class.

The BiLSTM-DINA model was able to give accuracy values of 96.45% for the original dataset, 99.12% for SMOTE, and 95.80% for GAN. The SMOTE model had the best results, giving an F1-Score value of 0.9911, an AUC value of 0.9994, a MCC value of 0.9826, and Co-hen's Kappa value of 0.9824. The DINA model showed an attribute consistency index of 0.7664 for estimating subject level mastery. The SHAP analysis indicated Overall Mean, Mathematics Mean, Science Mean, and Learning Gain as the most influential features.

These results show that the effect of augmentation varied, such that SMOTE was seen to have the best configuration, whereas GAN had comparatively poor deep learning performance. It can therefore be seen that the integration of longitudinal prediction with cognitive diagnosis leads to the learner-mastery prediction as well as the subject mastery information. This is thus a foundation for the analysis of the learner performance and educational decisions in the future. However, the limitations of the results are that they only apply to the longitudinal academic data set considered in the study.

**Keywords:** Adaptive Learning; Learning Analytics; Educational Data Mining; BiLSTM; DINA; Explainable Artificial Intelligence; Longitudinal Learning Analytics; Student Mastery Prediction; Data Augmentation; Intelligent Tutoring Systems

## 1. Introduction

Learning analytics (LA) and educational data mining (EDM) have become significant research domains for learner behavior analysis, academic pattern recognition, and evidence-based decisions in education. Through these domains, educational establishments can make use of the data generated by learners for the purpose of assessing trends, behaviors, and developing personalized learning experiences [1,2,3]. With the quick growth of online learning platforms, intelligent tutoring systems, and digital assessment tools, there have been large amounts of educational data available for analysis regarding learner behaviors, activities, and academic progress [4]. Additionally, systematic review studies have noted the increased use of artificial intelligence technologies in higher education in terms of predicting,

personalizing, evaluating, and supporting learning [5]. In contrast to the traditional methods based solely on achieving results, longitudinal educational records allow researchers to study learning development, behavioral shifts, and knowledge acquisition over time [6].

### 1.1 Background and Research Motivation

Predicting student performance has slowly developed from traditional classification algorithms into modern prediction models that emphasize estimation of learner mastery, learning progress analysis, and detection of students who need additional help. It has been shown in recent EDM researches that the use of predictive analytics can be helpful in improving personalized learning support and decision making, in particular, by detecting at-risk learners and providing timely interventions [7,8]. However, many traditional machine learning methods use statically defined features which are manually extracted from the data and thus cannot capture sequential dependencies inherent in the longitudinal records [9].

The emergence of deep learning methodologies created an opportunity for building models which are capable of modelling complex nonlinear dependencies and sequential dependencies. The Long Short-Term Memory (LSTM) method was invented to solve the problem of learning long term dependencies of sequential data, while Bidirectional LSTM further extends this method and allows taking into account not only previous but also future states of sequences [10]. Some recent studies proved the efficiency of deep learning-based methods, including BiLSTM models, for student performance prediction and learning behaviour analysis.

However, in addition to high accuracy, educational decision support systems also require interpretable information about learner knowledge states, cognitive development, and the area in which the learner needs additional efforts. The cognitive diagnostic models can provide such information by predicting latent knowledge attributes related to the learner response. Among those models, the most popular is the DINA model [11].

Class imbalance, limited number of samples available, and unequal learner representation in the datasets can negatively affect the prediction performance in many educational datasets. The methods for data augmentation have thus been investigated in order to improve the quality of representation and predictive performance in case of limited and imbalanced datasets [12]. While SMOTE generates additional synthetic samples for minority classes, GANs give an opportunity to generate synthetic data distribution using deep generative model [13].

Moreover, with the growing involvement of artificial intelligence methods into educational decision-making, transparency and explainability of the decisions made became the key requirements. Methods for explaining AI, like SHAP, give the opportunity to researchers to reveal feature importance in AI-driven predictions [14].

Inspired by the above problems, this study introduces an adaptive decision-making framework based on BiLSTM-DINA trajectory, which uses longitudinal feature modeling, cognitive diagnosis, data augmentation, learner representation learning, clustering, and SHAP explainability for learner mastery prediction.

### 1.2 Problem Statement

Despite the progress made in student performance prediction, the current approaches to learner trajectory and cognitive mastery development modelling can be seen less emphasized in favor of predictive accuracy. It has been proved by previous researches that LA models tend to rely more on performance prediction, while deeper modeling of learner progress and knowledge states is rather problematic [15].

Sequential modeling methods like knowledge tracing and RNNs have proved the necessity of analyzing the temporal learning patterns by means of modeling changes of learner knowledge states [16,17]. Existing research on knowledge tracing and deep learning-based predictions are oriented mainly on prediction of future performance or modeling of changes of knowledge states of learners based on interaction sequences, while cognitive diagnostic models like DINA are oriented more on estimation of latent skill mastery of the learners from assessment responses [18]. That is why these two directions of research can be considered to investigate different sides of learning processes.

Models of cognitive diagnostics give important information about the mastery states and attribute-level knowledge estimation of learners through DINA and other Q-matrix based models [19]. However, the main feature of the models is that they are designed mostly around assessment response patterns, while the deep learning models like recurrent neural networks and BiLSTM architectures have proved to be good at modeling sequential education data and performance trends [20,21] and their prediction mechanisms require additional interpretation of the relation of learned representation and cognitive properties.

The efficiency of data augmentation methods like SMOTE and GANs in improving predictive models has been proved under conditions of insufficient data. However, the combination of data augmentation with longitudinal trajectory modeling and cognitive diagnosis needs further investigation.

Furthermore, recent studies prove the benefit of explainable artificial intelligence (XAI) approach in increasing transparency of the educational prediction models [22,23]. Existing educational studies that use XAI have focused mainly on the interpretation of prediction models and identification of important learning features. Similarly, the studies on data augmentation have focused mainly on improvement of robustness of the model under conditions of limited or imbalanced datasets [24]. These results show that the problem of prediction accuracy, interpretability and data augmentation have been investigated from different methodological points of view, which provides an opportunity to develop an integrated framework that takes into account longitudinal modeling, cognitive diagnosis, data augmentation and adaptive explanation.

That is why this paper tries to fill the gap by investigating the integrated adaptive learning framework that takes into account the above-mentioned aspects of learning process modeling.

### 1.3 Research Objectives and Research Questions

The study aims to:

Propose and test a BiLSTM–DINA model for predicting learner mastery from longitudinal academic achievement data.

Analyze the influence of SMOTE and GAN data augmentation and interpret learner mastery using DINA and SHAP analysis.

Accordingly, the study addresses the following research questions:

RQ1. How well can the proposed BiLSTM–DINA model predict learner mastery based on longitudinal performance data?

RQ2. What is the effect of SMOTE and GAN on prediction accuracy, and what is the contribution of DI-NA and SHAP to interpretation of learner mastery?

### 1.4 Major Contributions and Novelty

This study aims to introduce a unified trajectory-driven adaptive learning framework based on a single architecture to integrate the sequential modelling, cognitive diagnosis, data augmentation with SMOTE and GAN, latent representation learning with autoencoder, learner clustering, and explainability with SHAP. The framework is thoroughly validated with baseline comparisons, cross-validation, statistical testing, clustering evaluation, reconstruction analysis, synthetic data quality assessment, ablation studies. The proposed framework can be used for intelligent tutoring systems and LA, as it combines prediction, cognitive interpretation and adaptive learner support.

## 2. Related Work

AI in education has evolved from descriptive EDM to predictive models, sequential learners, cognitive diagnosis, data augmentation, and explainable AI (XAI). The literature related to the current research can therefore be examined along five dimensions: LA and learner performance prediction, longitudinal sequential modeling, cognitive diagnosis, data

augmentation, and XAI. This classification makes it possible to situate the suggested BiLSTM-DINA model in existing literature on educational-AI research.

## 2.1 Learning Analytics and Student Performance Prediction

EDM and LA aid in the analysis of academic, assessment, behavioral, and learning management data. According to Romero and Ventura, prediction, classification, clustering, recommendation, visualization, and decision support were the major applications of EDM and LA [25]. Some more recent studies that provide empirical evidence systematically reveal the extensive utilization of machine learning techniques like decision trees, random forest, Support Vector Machines (SVM), Neural Networks, and Ensemble learning to predict student performance [26].

Recent studies have also highlighted feature selection and adaptability of models. Malik et al. illustrated the significance of feature selection and classification techniques for educational performance prediction [27]. The significance of creating informative representations of the academic data can be seen through these studies. In the current study, longitudinal records of the academic data are represented through trajectory-based features, which capture cumulative performance and performance changes over time.

## 2.2 Longitudinal Learning and Sequential Deep Learning

Academic trajectories consist of longitudinal learner records with sequential observations that can give insight into the development of learners. Long Short-Term Memory (LSTM) models have been developed to capture long-term dependencies in sequences [28], while bidirectional recurrent models allow processing of information in both temporal directions [29]. This makes BiLSTM appropriate for modeling academic trajectories.

In educational applications, Deep Knowledge Tracing showed the capability of recurrent neural networks to model changes in the state of knowledge of learners through sequential interactions [30]. Further studies have provided evidence for the evolution of knowledge tracing via recurrent, attention-based, memory-based, and graph-based methods [31]. Kalita et al. further elaborated on the application of BiLSTM for student performance prediction along with SHAP-based interpretation [32].

The problem is that sequential modeling and cognitive diagnostic modeling have very different functions. The former predicts outcome based on observed sequences while the latter is about modeling mastery of particular attributes. For this reason, the current study tries to merge the two into one approach through sequential BiLSTM modeling and DINA-based cognitive diagnosis.

## 2.3 Cognitive Diagnostic Models

The Cognitive Diagnostic Models (CDMs) express the proficiency of the learner in terms of multiple skills or attributes. The DINA model is based on the conjunction of mastery in which all the required attributes contribute to the expected response [33]. In case of G-DINA models, the flexibility in modeling the association between attributes and item responses is improved [34].

Q-matrix plays an important role in cognitive diagnosis since it defines the connection between observable test results and underlying attributes. De la Torre and Chiu highlighted the significance of empirically validating the Q-matrix to improve the correspondence [35]. In the current study, the DINA model represents subject-level mastery states along with BiLSTM-based predictions. Such outputs can be considered as diagnostic representations derived from the model and dependent on chosen attributes, Q-matrix, and mastery criteria.

## 2.4 Data Augmentation in Educational Prediction

The Synthetic Minority Over-sampling Technique (SMOTE) creates synthetic minority-class examples by using interpolation among nearby examples and was initially devised for class imbalance [36]. In the context of educational prediction, SMOTE has demonstrated how augmentation outcomes depend on the setting of the model configuration [37]. Also, Švábenský et al. found that distinct augmentations can have distinct effects on the predictive outcome and stability [38].

In contrast, Generative Adversarial Networks (GANs) offer an alternative method that learns data distributions using adversarial training to create synthesized samples [39]. The generative methods have also been applied to multi-dimensional learning-performance data with consideration of the stability and fidelity of the synthesized samples [40].

As such, the current study compares SMOTE and GAN as experimental conditions relative to each other, instead of viewing data augmentation as an automatic generalization aid. Their impact is measured relative to the original-data condition as well in order to understand how different data representations affect predictive performance.

## 2.5 Explainable Artificial Intelligence in Education

Methods from the Explainable Artificial Intelligence (XAI) paradigm allow analysis of the contribution of each variable towards the model outcomes. The SHapley Additive exPlanations (SHAP) method offers the technique for additive feature attribution which can be applied for interpretation of model outcomes both globally and locally [41]. It is proven that educational tasks benefit from SHAP and similar techniques in analyzing influential academic variables [42].

In this current research, SHAP was employed as an auxiliary feature attribution method for the XGBoost algorithm to determine the trajectory-based features correlated with learner mastery prediction. This is why SHAP is not understood to be an explanation of the internal processes of the BiLSTM and DINA. Such distinction enables the explainability findings to support, but not to exaggerate, the predictions and cognitive diagnosis results.

The representative studies are summarized in Table 1.

Table 1. Summary of Representative Studies and Their Relevance to the Present Study

| Study                              | Dataset Context                       | / Method                                           | Main Finding                                                                                                                              | Limitation / Unresolved Issue                                                                                                    | Relevance to Present Study                                                                                                                                                       |
|------------------------------------|---------------------------------------|----------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Romero &amp; Ventura (2020)</b> | EDM and LA literature                 | EDM/LA survey                                      | Established prediction, classification, clustering, recommendation, and visualisation, and decision support as major EDM/LA applications. | Provides a broad methodological overview rather than a specific integrated learner-trajectory and cognitive-diagnosis framework. | Establishes the broader learning-analytics and educational-data-mining foundation of the study.                                                                                  |
| <b>Chen et al. (2025)</b>          | Empirical studies in higher education | Systematic review of machine-learning applications | Demonstrated extensive application of machine learning to student performance, engagement, and related educational outcomes.              | The reviewed studies employ diverse outcomes, datasets, and modelling strategies.                                                | Provides evidence for the established role of machine learning in educational prediction and supports the need to position the present framework within this broader literature. |

|                           |                                              |                                                        |                                                                                                                                                         |                                                                                              |                                                                                                                                          |                                                                                                                                                                            |
|---------------------------|----------------------------------------------|--------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Malik et al. (2025)       | Multiple educational datasets                | Feature selection and classification models            | Demonstrated the relevance of systematic feature selection and classification strategies for educational performance prediction.                        | the of feature and classification rather than combined longitudinal and cognitive modelling. | Focuses primarily on feature selection and predictive classification rather than combined longitudinal and cognitive modelling.          | Supports the construction and selection of trajectory-based academic predictors.                                                                                           |
| Piech et al. (2015)       | Sequential student learning-interaction data | Deep Knowledge Tracing using recurrent neural networks | Demonstrated that recurrent deep learning can represent temporal patterns in learner interactions and estimate changing knowledge states.               |                                                                                              | Focuses on interaction-level knowledge tracing rather than subject-level mastery derived from longitudinal academic achievement records. | Provides a foundation for treating learner development as a sequence rather than as isolated observations.                                                                 |
| Abdelrahman et al. (2023) | Knowledge-tracing research literature        | Systematic survey of knowledge-tracing methods         | Documented the development of knowledge tracing from probabilistic approaches toward recurrent, attention-based, memory-based, and graph-based methods. |                                                                                              | Different sequential modelling approaches address different aspects of learner-state estimation.                                         | Positions BiLSTM within the broader methodological development of sequential learner modelling.                                                                            |
| Kalita et al. (2025)      | Student academic-performance data            | BiLSTM, machine-learning baselines, and SHAP           | Demonstrated the applicability of BiLSTM to student academic-performance prediction and the use of SHAP for feature-level interpretation.               |                                                                                              | Does not combine BiLSTM with DINA-based cognitive diagnosis or evaluate SMOTE and GAN as comparative augmentation conditions.            | Provides closely related evidence for sequential prediction with explainability and supports the motivation for extending predictive modelling toward cognitive diagnosis. |
| de la Torre (2008)        | Cognitive diagnostic assessment              | DINA                                                   | Established the DINA framework for representing cognitive mastery                                                                                       |                                                                                              | Interpretation depends on the selected cognitive attributes and their relationship to                                                    | Provides the methodological foundation for the DINA component.                                                                                                             |

|                                      |                                     |                                     |    |                                                                                                                              |                                                                                                             |                                                                                                                  |
|--------------------------------------|-------------------------------------|-------------------------------------|----|------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------|
|                                      |                                     |                                     |    | through a conjunctive attribute structure.                                                                                   | observed assessment outcomes.                                                                               |                                                                                                                  |
| <b>de la Torre (2011)</b>            | Cognitive diagnosis                 | G-DINA                              |    | Extended DINA by permitting more flexible relationships between attributes and item responses.                               | Greater model flexibility introduces additional specification and interpretation considerations.            | Provides context for the assumptions underlying the classical DINA formulation used in the present study.        |
| <b>de la Torre &amp; Chiu (2016)</b> | Cognitive diagnostic assessment     | Empirical Q-matrix validation       | Q- | Developed an empirical procedure for improving correspondence between assessment items and cognitive attributes.             | Q-matrix specification remains dependent on the assessment context and attribute structure.                 | Supports careful construction and validation of the subject-level Q-matrix.                                      |
| <b>Chawla et al. (2002)</b>          | Imbalanced classification datasets  | SMOTE                               |    | Demonstrated that interpolation between neighbouring minority-class observations can increase minority-class representation. | Developed primarily for class-imbalance correction rather than unrestricted generation of learner profiles. | Provides the methodological basis for evaluating SMOTE as a controlled augmentation condition.                   |
| <b>Švábenský et al. (2024)</b>       | Learning-analytics prediction tasks | Comparative augmentation strategies |    | Demonstrated that predictive effects vary across augmentation methods and modelling configurations.                          | Augmentation effects depend on the characteristics of the dataset and model.                                | Supports empirical comparison of augmentation conditions rather than assuming universal performance improvement. |
| <b>Goodfellow et al. (2014)</b>      | General generative modelling        | GAN                                 |    | Introduced adversarial learning for modelling data distributions and generating synthetic observations.                      | GAN performance can depend on training stability, architecture, and the fidelity of generated data.         | Provides the theoretical basis for the GAN-based augmentation condition.                                         |
| <b>Islam et al. (2025)</b>           | Student academic-                   | Machine learning with               |    | Demonstrated the use of XAI to identify influential academic factors associated                                              | Primarily considers conventional predictive models rather than                                              | Supports the use of feature-level explanation as a complementary                                                 |

|                             | performance<br>data              | SHAP<br>LIME                                    | and | with<br>performance<br>predictions. | student-<br>integrated<br>use of<br>predictive modelling,<br>synthetic<br>oversampling, and<br>explainability for<br>identifying students<br>requiring support.     | integrated<br>sequential cognitive<br>diagnosis.                                                         | component<br>educational<br>prediction.                                                                                                                                                  | of |
|-----------------------------|----------------------------------|-------------------------------------------------|-----|-------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| <b>Almazroei<br/>(2026)</b> | Seven<br>educational<br>datasets | Machine<br>learning,<br>SMOTE, SHAP<br>and LIME |     |                                     | Demonstrated an<br>integrated use of<br>predictive modelling,<br>synthetic<br>oversampling, and<br>explainability for<br>identifying students<br>requiring support. | Does not combine<br>longitudinal<br>BiLSTM<br>representations<br>with DINA-based<br>cognitive diagnosis. | Provides recent<br>evidence for<br>combining<br>prediction,<br>augmentation, and<br>explainability and<br>helps position the<br>integrated<br>analytical design of<br>the present study. |    |

As shown in Table 1 below, the literature has provided a strong basis for the methodologies for predictive learner modeling, sequential representation learning, cognitive diagnosis, data synthesis, and explainability; however, these individual components have largely been developed as separate analytical methods.

## 2.6 Literature Synthesis and Research Gap

The reviewed literature provides an alternative foundation for methodological approaches. EDM and LA provide frameworks that facilitate predictions and decision-making concerning student performance [25,26]. The sequential deep-learning literature emphasizes the importance of representing learners' observations in order [28–32]. The cognitive diagnostic literature provides mastery representations at the level of attributes using CDM and Q-matrix approaches [33–35]. Augmentation techniques reveal that SMOTE and generative methods can change the representation of training data, but it depends on the data and model configuration [36–40]. XAI works reveal the importance of feature level explanations for educational prediction [41,42].

This is because these approaches deal with various analytical tasks such as prediction, temporal modeling, cognition diagnosis, data generation, and model interpretation. In this current study, the mentioned functions are merged into one learner model framework without having to consider that individual methods have not been utilized before in educational research.

As such, the proposed framework incorporates longitudinal modeling through BiLSTM, cognitive diagnosis through DINA, data balancing through SMOTE and GAN, representation learning through autoencoder, learner clustering and supplementary SHAP analysis. The investigation explores these elements under both original, SMOTE balanced, and GAN augmented data scenarios to assess performance, robustness, cognitive interpretation, and feature level explanations.

The above position makes the research contribution an integrated and comparative analysis of existing longitudinal methods of predicting and interpreting learners' cognition, and not an unsubstantiated assertion that these techniques have not yet been studied.

## 3. Materials and Methods

This section introduces the proposed BiLSTM–DINA trajectory driven adaptive decision framework. The framework incorporates feature engineering along the timeline, SMOTE and GAN-based augmentation, bidirectional sequential modelling, cognitive diagnosis, latent representation learning, learner clustering, explainability and support in the form of rules.

### 3.1 Overall Framework Architecture

The proposed framework involves a series of analytical phases on the academic records of the student, where each phase is carried out longitudinally. The raw data is cleaned and processed to a set of trajectory-based features as shown in Fig. 1. The processed data are then tested for the original configuration, SMOTE-balanced configuration and GAN-augmented configuration.

The BiLSTM component is to foresee the learner level of mastery based on the longitudinal academic pattern, and the DINA component is to estimate the cognitive level of mastery of various subjects. An autoencoder learns a compact representation of the learners, K-Means produces groups of learners, and SHAP determines the features that impact the

action that the model takes. These outputs serve as a basis for designing personalized educational interventions, analytically.

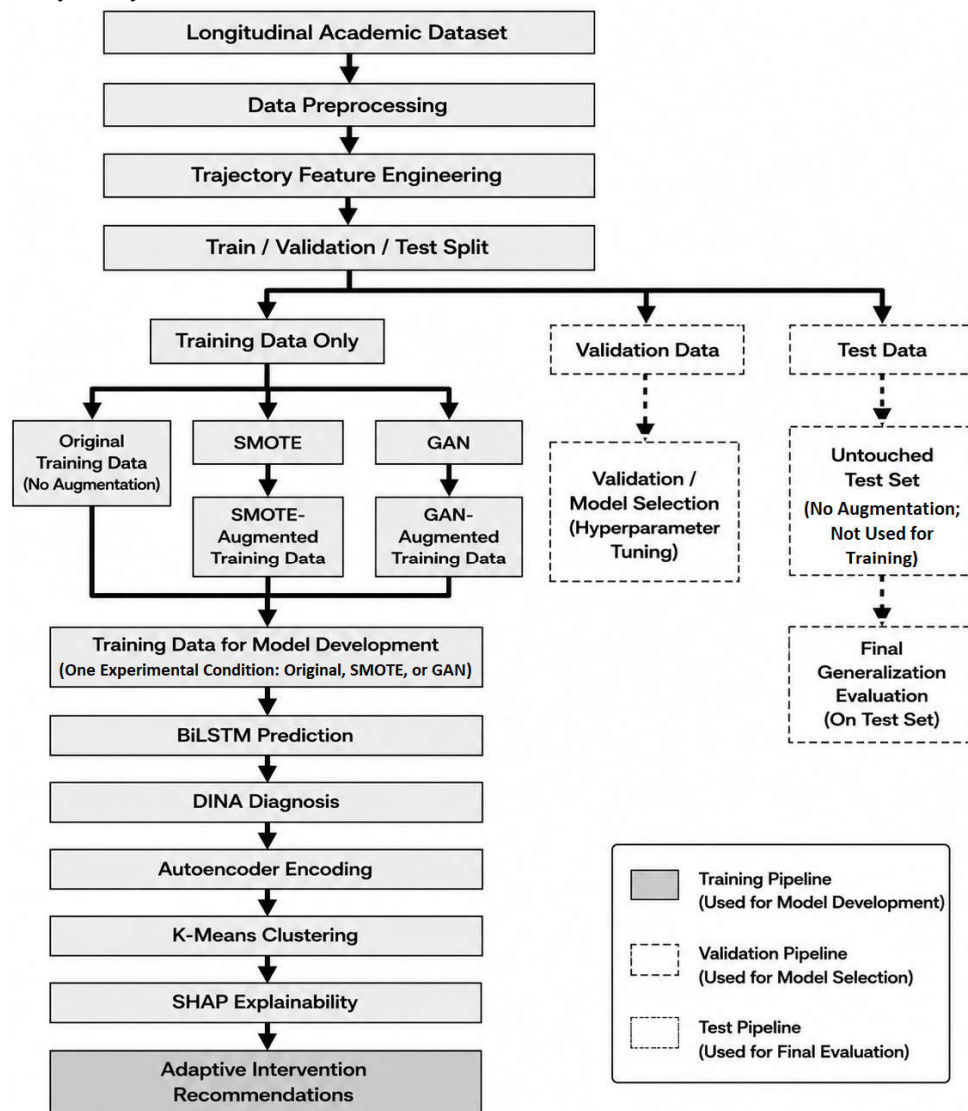


Fig. 1. Overall workflow of the proposed BiLSTM–DINA framework with alternative augmentation and independent test evaluation.

The architecture integrates predictive modelling with cognitive interpretation and segmentation of learners, thus offering a foundation for adaptive educational decision support other than classification.

### 3.2 Dataset Description

The framework was tested by applying it to the academic data of a longitudinal dataset of 1,549 students and 30 attributes. There were 22 numerical variables, eight categorical variables, and 21 academic score variables consisting of Mathematics, Science and English.

The first step involved preprocessing and feature engineering, which resulted in 13 trajectory-based predictors being retained. Mastery Status was chosen as the target variable; 775 learners were classified as mastered and 774 were classified as not mastered.

Characteristics of the dataset, such as attribute types, academic subjects, constructed attributes, target attribute, and distribution of classes, have been discussed in Table 2.

Table 2. Dataset characteristics

| Dataset characteristic | Value |
|------------------------|-------|
| Total students         | 1,549 |
| Total attributes       | 30    |

|                                 |                                    |
|---------------------------------|------------------------------------|
| <b>Numerical attributes</b>     | 22                                 |
| <b>Categorical attributes</b>   | 8                                  |
| <b>Academic subjects</b>        | Mathematics, Science, English      |
| <b>Academic score variables</b> | 21                                 |
| <b>Engineered features</b>      | 13                                 |
| <b>Target variable</b>          | Mastery Status                     |
| <b>Class distribution</b>       | Mastered = 775; Not Mastered = 774 |

The nearly equal class distribution provided a balanced basis for the original-data experiment.

### 3.3. Data Preprocessing

Data was enhanced and recoded for completeness during the preprocessing stage. Data was missing for science examination scores and was handled through mean substitution using scores from the relevant subject. There were eight categorical variables that were label encoded in order to change categorical data into numerical data.

Min–Max scaling was applied to all numerical variables to scale them into the range [0,1]:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where  $x$  is the original value of a feature;  $x_{min}$  is the minimum observed value of that feature;  $x_{max}$  is the maximum observed value of that feature; and  $x'$  is the normalized feature value after Min–Max scaling.

Afterwards, trajectory-based feature engineering was used for building predictive representations from longitudinal data. Preprocessing and trajectory-based feature engineering yielded an analytical matrix consisting of 1,549 students and 13 predictive features (1,549 × 13).

All the steps of the preprocessing process including handling of missing values, categorical encoding, feature scaling, trajectory-based feature engineering, and final validation are presented in Table 3.

Table 3. Missing-value treatment and preprocessing summary

| <b>Preprocessing step</b>              | <b>Method applied</b>                      | <b>Result</b>                                         |
|----------------------------------------|--------------------------------------------|-------------------------------------------------------|
| <b>Missing-value detection</b>         | Null-value analysis                        | Missing values found only in science scores           |
| <b>Missing-value treatment</b>         | Mean imputation                            | All missing values replaced                           |
| <b>Categorical encoding</b>            | Label encoding                             | Eight variables encoded                               |
| <b>Feature scaling</b>                 | Min–Max normalization                      | Values scaled to [1]                                  |
| <b>Feature engineering</b>             | Trajectory-based feature engineering       | Final analytical set contained 13 predictive features |
| <b>Final analytical feature matrix</b> | Post-preprocessing and feature engineering | 1,549 learners × 13 predictive features               |
| <b>Final validation</b>                | Missing-value check                        | 0 missing values                                      |

The preprocessing steps followed before generating each experiment data setup were the same, while SMOTE and GAN data augmentation were only done on the training dataset.

### 3.4. Trajectory-Based Feature Engineering

The longitudinal academic variables were transformed into 13 features representing achievement, progression, variability, and subject mastery.

For learner  $i$ , the mean performance in subject  $s$  over  $T$  assessments was calculated as

$$\bar{x}_{i,s} = \frac{1}{T} \sum_{t=1}^T x_{i,s,t} \quad (2)$$

where  $\bar{x}_{i,s}$  is the mean performance of learner  $i$  in subject  $s$ ;  $T$  is the total number of assessments or time points considered for the subject;  $t$  is the assessment or temporal index, with  $t = 1, \dots, T$ ; and  $x_{i,s,t}$  is the observed score of learner  $i$  for subject  $s$  at time point  $t$ .

The overall mean performance across the three academic subjects was defined as

$$\bar{x}_{i,overall} = \frac{\bar{x}_{i,math} + \bar{x}_{i,science} + \bar{x}_{i,english}}{3} \quad (3)$$

where  $x_{i,overall}$  is the overall mean performance of learner  $i$ ;  $x_{i,math}$  is the mean Mathematics performance of learner  $i$ ;  $x_{i,science}$  is the mean Science performance of learner  $i$ ;  $x_{i,english}$  is the mean English performance of learner  $i$ ; and 3 represents the number of academic subjects considered.

Subject-wise learning gain was computed as

$$g_{i,s} = x_{i,s,T} - x_{i,s,1}, \quad (4)$$

where  $g_{i,s}$  is the learning gain of learner  $i$  in subject  $s$ ;  $x_{i,s,T}$  is the learner's score in subject  $s$  at the final assessment or time point  $T$ ;  $x_{i,s,1}$  is the learner's score in subject  $s$  at the initial assessment or time point 1;  $i$  denotes the learner;  $s$  denotes the academic subject; and  $T$  denotes the final assessment or time point.

The total gain was calculated by taking the average of all three subject gains. The performance variability was measured with the standard deviation of the learner's performance measurements:

$$\sigma_i = \sqrt{\frac{1}{M} \sum_{j=1}^M (x_{i,j} - \bar{x}_i)^2}. \quad (5)$$

where  $\sigma_i$  is the performance standard deviation of learner  $i$ ;  $M$  is the total number of performance observations available for learner  $i$ ;  $j$  is the index of an individual performance observation;  $x_{i,j}$  is the  $j$ -th performance observation of learner  $i$ ; and  $\bar{x}_i$  is the arithmetic mean of the  $M$  performance observations for learner  $i$ .

Subject mastery was determined using a data-adaptive median threshold based on the normalized mean performance of each subject:

$$\alpha_{i,s} = \begin{cases} 1, & \bar{x}_{i,s} \geq \tau_s \\ 0, & \bar{x}_{i,s} < \tau_s \end{cases} \quad (6)$$

where  $\alpha_{i,s}$  is the binary mastery indicator of learner  $i$  for subject  $s$ ;  $\bar{x}_{i,s}$  is the normalized mean performance of learner  $i$  in subject  $s$ ; and  $\tau_s$  is the subject-specific median of the normalized mean performance distribution.

A criterion based on the median was chosen as an adaptive data-based definition of mastery as it takes into consideration the performance distribution of the study sample and serves as a balanced starting point for further binary predictions and cognitive diagnoses.

Such engineered features served as the primary inputs for prediction, cognitive diagnoses, clustering, and explainability. The precise 13 trajectory-based features used in the final analytical matrix.

### 3.5. Hybrid Data Augmentation

Two data augmentation approaches were applied complementarily in order to investigate how well the suggested framework could be resilient against different types of data representations used in the training process. SMOTE was chosen due to its ability to create new synthetic samples using the interpolation between adjacent samples, which makes it very useful when dealing with class imbalances. GAN Augmentation was chosen as an alternative approach that could learn from the given representations. The objective of applying these two techniques was not to presume that augmented data would improve prediction in all cases, but rather to empirically compare the different ways in which augmentation impacts learning and representations. While SMOTE augmented neighborhood-based observations, GAN augmented the wider variations in the feature space.

There were 1,549 instances in the initial dataset, including 775 masters and 774 non-masters. The SMOTE technique helped increase the sample size to 1,550, thus creating one extra artificial instance of the data to achieve balance between the two categories. On the other hand, the class-wise GAN technique created 5,000 artificial instances, with 2,500 for each mastery category. The two approaches to data augmentation served as controlled experimental conditions to study the impact of different mechanisms of artificial data creation on prediction.

#### 3.5.1. SMOTE-Based Augmentation

SMOTE generated a synthetic learner observation by interpolating between an instance and one of its nearest neighbours:

$$x_{syn} = x_i + \lambda(x_j - x_i), \lambda \in [0,1]. \quad (7)$$

where  $x_{syn}$  is the generated synthetic learner observation;  $x_i$  is the selected learner observation from the minority class;  $x_j$  is a neighbouring learner observation from the same minority class; and  $\lambda$  is a random interpolation coefficient bounded between 0 and 1.

#### 3.5.2. GAN-Based Augmentation

The GAN consisted of a generator  $G$  and discriminator  $D$ . The generator transformed a latent vector into a synthetic learner profile:

$$x_{syn} = G(z). \quad (8)$$

where  $x_{syn}$  is the synthetic learner profile generated by the GAN;  $G$  is the generator network; and  $z$  is a latent random vector sampled from the predefined latent distribution  $p_z(z)$ .

The adversarial objective was expressed as

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}} [\log D(x)] + E_{z \sim p_z} [\log(1 - D(G(z)))]. \quad (9)$$

where  $G$  is the generator network;  $D$  is the discriminator network;  $V(G)$  is the adversarial value function;  $x$  is a real learner observation sampled from the real-data distribution  $p_{data}$ ;  $p_{data}$  is the probability distribution of real learner observations;  $z$  is a latent random vector sampled from the latent distribution  $p_z$ ;  $p_z$  is the probability distribution of the latent vectors;  $D(x)$  is the discriminator's estimated probability that  $x$  is a real observation;  $G(z)$  is the synthetic observation generated from  $z$ ; and  $E$  denotes mathematical expectation.

Reproducibility requires that the number of observations that are created through each of the augmentations is explicitly stated from the actual implementation. While the original configuration included the observed learner data, the SMOTE and GAN configurations involved the creation of synthetic observations based on their specific generation procedures. The datasets that resulted were then evaluated using the same process for predictive modeling and evaluation.

The final methodology therefore produced three configurations: the original dataset, the SMOTE-balanced dataset, and the GAN-augmented dataset.

### 3.6. Proposed BiLSTM–DINA Framework

The central predictive module fuses the temporal learning capability of BiLSTM with the cognitive diagnosis capability of DINA. The BiLSTM accounts for the longitudinal academic relationships and the DINA defines the observed performance to the mastery of Mathematics, Science, and English.

#### 3.6.1. BiLSTM Architecture

For an input sequence

$$X = \{x_1, x_2, \dots, x_T\}, \quad (10)$$

where  $X$  is the complete input sequence for a learner;  $x_t$  is the feature vector observed at time point  $t$ ;  $T$  is the total number of temporal observations in the sequence; and  $t$  is the temporal index ranging from 1 to  $T$ .

The forward hidden states are computed as

$$\vec{h}_t = LSTM_f(x_t, \vec{h}_{t-1}) \quad (11)$$

where  $\vec{h}_t$  is the forward hidden state at time point  $t$ ;  $LSTM_f$  is the forward-direction LSTM function;  $x_t$  is the input feature vector at time point  $t$ ; and  $\vec{h}_{t-1}$  is the previous forward hidden state.

The backward hidden states are computed as

$$h_t^{\leftarrow} = LSTM_b(x_t, h_{t+1}^{\leftarrow}). \quad (12)$$

where  $h_t^{\leftarrow}$  is the backward hidden state at time point  $t$ ;  $LSTM_b$  is the backward-direction LSTM function;  $x_t$  is the input feature vector at time point  $t$ ; and  $h_{t+1}^{\leftarrow}$  is the subsequent backward hidden state used when processing the sequence in reverse order.

The bidirectional representation is obtained as

$$h_t = [\vec{h}_t; h_t^{\leftarrow}], \quad (13)$$

where  $h_t$  is the combined bidirectional hidden representation at time point  $t$ ;  $\vec{h}_t$  is the forward hidden state;  $h_t^{\leftarrow}$  is the backward hidden state; and  $[\cdot]$  denotes vector concatenation.

The predicted mastery probability is calculated as

$$y = \sigma(Wh + b) \quad (14)$$

where  $y$  is the predicted probability of the learner belonging to the mastery class;  $\sigma(\cdot)$  is the sigmoid activation function;  $W$  is the weight matrix of the output classification layer;  $h$  is the final or aggregated bidirectional hidden representation supplied to the classification layer; and  $b$  is the bias term of the classification layer.

The extracted bidirectional representations are fed into the classification layer for calculating the probabilities of mastery of the learners. The complete list of hyperparameters is described in Section 4.

#### 3.6.2. DINA Cognitive Integration

For learner  $i$ , the cognitive mastery profile was represented as

$$\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK}), \quad (15)$$

where  $\alpha_i$  is the cognitive mastery profile of learner  $i$ ;  $\alpha_{ik}$  is the binary mastery status of learner  $i$  for cognitive skill  $k$ ;  $K$  is the total number of cognitive skills; and  $k$  is the cognitive-skill index ranging from 1 to  $K$ .

The binary mastery indicator was defined as

$$\alpha_{ik} = \{1, \text{if skill } k \text{ is mastered by learner } i, 0, \text{otherwise.}\} \quad (16)$$

where  $\alpha_{ik}$  is the binary mastery indicator for learner  $i$  and cognitive skill  $k$ ;  $i$  denotes the learner;  $k$  denotes the cognitive skill; 1 indicates mastery of the skill; and 0 indicates non-mastery.

The ideal DINA response was defined as

$$\eta_i = \prod_{k=1}^K \alpha_{ik}^{q_k}, \quad (17)$$

where  $\eta_i$  is the ideal response or ideal mastery state of learner  $i$ ;  $K$  is the total number of cognitive skills;  $k$  is the cognitive-skill index;  $\alpha_{ik}$  is the mastery status of learner  $i$  for skill  $k$ ; and  $q_k$  is the corresponding binary Q-matrix requirement for skill  $k$ , where  $q_k = 1$  indicates that the skill is required for the assessed subject or item and  $q_k = 0$  indicates that the skill is not required.

This integration associates predicted learner performance with interpretable subject-level mastery states.

### 3.6.3. Loss Function and Training Strategy

Binary Cross-Entropy was used for mastery classification:

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]. \quad (18)$$

where  $L$  is the binary cross-entropy loss;  $N$  is the total number of training observations;  $i$  is the learner index;  $y_i$  is the true binary mastery label for learner  $i$ ;  $\hat{y}_i$  is the predicted mastery probability for learner  $i$ ; and  $\log$  denotes the natural logarithm.

The optimization scheme was developed to ensure stability, avoid over-fitting, and provide consistent training under all conditions. The full set-up for the training procedure is described in section 4.

### 3.7. Latent Representation Learning

Compact nonlinear representations of learner trajectories were created using an autoencoder. The encoder transforms the input feature vector into a compact latent representation:

$$z = f(W_e x + b_e), \quad (19)$$

where  $z$  is the latent representation produced by the encoder;  $f(\cdot)$  is the encoder activation function;  $W_e$  is the encoder weight matrix;  $x$  is the input feature vector of the learner; and  $b_e$  is the encoder bias vector.

The decoder reconstructs the original feature vector as

$$\hat{x} = g(W_d z + b_d). \quad (20)$$

where  $\hat{x}$  is the reconstructed feature vector;  $g(\cdot)$  is the decoder activation function;  $W_d$  is the decoder weight matrix;  $z$  is the latent representation generated by the encoder; and  $b_d$  is the decoder bias vector.

The reconstruction objective was defined using mean squared error (MSE):

$$MSE = \frac{1}{N} \sum_{i=1}^N \|x_i - \hat{x}_i\|^2. \quad (21)$$

where  $MSE$  is the mean squared reconstruction error;  $N$  is the number of learner observations;  $i$  is the learner index;  $x_i$  is the original feature vector of learner  $i$ ;  $\hat{x}_i$  is the reconstructed feature vector of learner  $i$ ;  $\|\cdot\|_2$  denotes the Euclidean norm; and the superscript 2 denotes the squared Euclidean distance.

The learned latent representation is able to preserve the important features of learner progress while minimizing redundancy in the original feature space. These representations were then used for learner clustering.

### 3.8. Learner Clustering and Adaptive Decision Framework

The compressed instances of the learner representations were grouped based on their academic and cognitive properties. Mappings were created to rule-based intervention recommendations.

#### 3.8.1. K-Means Clustering

K-Means was used to partition the latent representations by minimizing the within-cluster sum of squared distances:

$$J = \sum_{k=1}^K \sum_{z_i \in C_k} \|z_i - \mu_k\|_2^2 \quad (22)$$

where  $J$  is the K-Means objective function;  $K$  is the number of learner clusters;  $k$  is the cluster index;  $C_k$  is the set of learner representations assigned to cluster  $k$ ;  $z_i$  is the latent representation of learner  $i$ ;  $\mu_k$  is the centroid of cluster  $k$ ;  $\|\cdot\|_2$  denotes the Euclidean norm; and the superscript 2 denotes the squared Euclidean distance.

The best number of learner clusters was selected using the Elbow Method. The quality of the clusters was further examined using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score.

#### 3.8.2. Personalized Intervention Strategy

The four groups of learners were considered 'high performing', 'moderate', 'low performing' and 'at risk'. The relationship between the rule-based decision layer and the types of resources available was found to be as follows: High-performing learners were linked to advanced materials and enrichment activities, moderate learners to guided practice and periodic feedback, low-performing learners to remedial resources, and at-risk learners to more frequent monitoring and individualized instructional support. The above mappings are intended for interventions and are not the result of interventions that have been tested in a real educational context.

### 3.9. Model Explainability

For better interpretability in the prediction of learner mastery, SHAP (Shapley Additive Explanations) was performed on a selected XGBoost model. SHAP analysis did not consider the BiLSTM or DINA models directly; instead, XGBoost

was chosen as a surrogate decision tree for the analysis of the impact of the same set of trajectory features for predicting mastery. The rationale behind this is to ensure that the attribution of features reported here corresponds directly to the model under SHAP analysis.

For the explainability analysis, the XGBoost classifier was trained using the GAN-augmented configuration, with an 80:20 stratified train-test split. The TreeExplainer implementation of SHAP was then used to estimate feature-level contributions on the held-out test observations. For feature  $j$ , the SHAP contribution was calculated using the Shapley-value formulation:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{j\}) - f(S)] \quad (23)$$

where  $\phi_j$  is the SHAP contribution of feature  $j$ ;  $F$  is the complete set of input features;  $j$  is the feature being explained;  $S$  is a subset of features that excludes feature  $j$ ;  $|S|$  is the number of features in subset  $S$ ;  $|F|$  is the total number of input features;  $f(S)$  is the model output when only the features in subset  $S$  are considered;  $f(S \cup \{j\})$  is the model output when feature  $j$  is added to subset  $S$ ; and  $!$  denotes the factorial operation.

The sum of the absolute SHAP values was calculated for all observations not in the training data set in order to get a global ranking of features by importance. Larger mean absolute SHAP values mean that the feature has more impact on the prediction from the XGBoost model. Positive SHAP values contribute positively to the prediction while negative SHAP values contribute negatively. Positive SHAP values push the prediction towards positive mastery class and vice versa. This produced insights about how trajectory features, such as cumulative achievement, learning gains, knowledge of the specific subjects, and variability in performance influenced prediction of mastery in learners. Feature-level explanations, therefore, offer an explanation that complements the predictive and cognitive aspects of the proposed BiLSTM-DINA approach and do not necessarily explain the inner workings of the model itself.

#### 4. Experimental Design

This section introduces the experimental settings for testing the proposed framework. It explains the environment used to implement, the parameters involved in hyperparameter tuning, comparison models used and the evaluation criteria for the results, as well as the used protocol for validations and its consistency and repeatability.

##### 4.1 Experimental Environment

The proposed framework has been implemented in Google Colab environment using Python programming language with GPU acceleration. Embedded in TensorFlow 2.x, the BiLSTM network and Autoencoder were implemented using the Keras API, while Scikit-learn was used for preprocessing, baseline machine learning models, clustering algorithms, cross-validation, and performance evaluation. Gradient boosting experiments were conducted using XGBoost, and the numerical computation, data manipulation, and visualization were handled by NumPy, Pandas, and Matplotlib, respectively. All experiments were conducted using the same software environment, thus ensuring consistency and reproducibility.

##### 4.2 Hyperparameter Configuration

The identical hyper-parameter settings were kept throughout all three experiments in order to make sure that the variations in the performance are attributable to the differences in data configurations and not in the model's learning capacity or hyper-parameters themselves. Hyper-parameter values were chosen a priori as experimental settings and not as globally optimal ones. The chosen configuration represents a moderate-capacity recurrent neural network suitable for the dimensionality of the learner dataset and uses dropout regularization in order to reduce overreliance on individual hidden representations. For stable adaptive gradient optimization, the Adam optimizer was used with a learning rate of 0.001, while the batch size of 32 represents a compromise between gradient stability and efficient computation. For evaluating the model's behavior on data not included in estimating parameters, the validation portion of 20% was kept the same consistently. The same parameters were kept for all augmentation conditions to retain consistency in experimentation. The entire configuration used for the proposed framework is summarized in Table 4.

Table 4. Hyperparameter configuration of the proposed framework

| Component | Hyperparameter          | Value   |
|-----------|-------------------------|---------|
| BiLSTM    | Hidden Units            | 64      |
|           | Number of BiLSTM Layers | 1       |
|           | Dense Layer Units       | 32      |
|           | Dropout Rate            | 0.30    |
|           | Activation Function     | ReLU    |
|           | Output Activation       | Sigmoid |
|           | Optimizer               | Adam    |
|           | Learning Rate           | 0.001   |

|                            |                    |                             |
|----------------------------|--------------------|-----------------------------|
| <b>Autoencoder</b>         | Batch Size         | 32                          |
|                            | Epochs             | 30                          |
|                            | Validation Split   | 20%                         |
|                            | Encoder Units      | 8                           |
|                            | Latent Dimension   | 4                           |
|                            | Decoder Units      | 8                           |
|                            | Optimizer          | Adam                        |
|                            | Loss Function      | MSE                         |
|                            | Epochs             | 50                          |
| <b>Random Forest</b>       | Batch Size         | 32                          |
|                            | Number of Trees    | 200                         |
| <b>SVM</b>                 | Kernel             | Radial Basis Function (RBF) |
| <b>Logistic Regression</b> | Maximum Iterations | 1000                        |
| <b>XGBoost</b>             | Evaluation Metric  | Log Loss                    |
| <b>K-Means</b>             | Number of Clusters | 4                           |
| <b>Cross Validation</b>    | Strategy           | Stratified 5-Fold           |

The selected hyperparameter values were kept constant throughout all experiments to ensure a fair comparison among the evaluated models.

The hidden-layer size of the BiLSTM is 64 to have enough representation capability to model nonlinear relations among the features of the trajectories without making the model too complex for the number of observations. A dense layer of size 32 has been used as a compact non-linear transformation before classification while the dropout ratio of 0.30 has been used as a regularizer. The choice of the sigmoid function as the activation function in the output layer was justified by the nature of the binary goal of mastery. The number of epochs of 30 during training was held constant among all configurations in order to ensure similar computation settings, whereas the 50-epoch budget for autoencoder was used in order to have enough iterations for reconstruction training. The choice of the 8-4-8 architecture of the autoencoder was made to generate compact four-dimensional latent space with a symmetric path for reconstruction.

#### 4.3 Baseline Models

To test the predictive effectiveness of the proposed framework, a comparison was made with representative traditional machine learning and deep learning models in the same experimental conditions. The comparison included Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost, LSTM and GRU.

#### 4.4 Performance Evaluation Metrics

A number of evaluation criteria were used to measure the predictive power, quality of clustering, ability to reconstruct, cognitive consistency, and significance. The criteria used in this study are outlined in Table 5.

Table 5. Performance evaluation metrics

| Category              | Evaluation Metric                      | Purpose                                    |
|-----------------------|----------------------------------------|--------------------------------------------|
| <b>Classification</b> | Accuracy                               | Overall prediction performance             |
|                       | Precision                              | Positive prediction accuracy               |
|                       | Recall (Sensitivity)                   | Detection of positive learner mastery      |
|                       | F1-score                               | Balance between Precision and Recall       |
|                       | Balanced Accuracy                      | Performance under balanced/imbalanced data |
|                       | AUC                                    | Classification discrimination capability   |
|                       | Root Mean Square Error (RMSE)          | Prediction error measurement               |
|                       | Matthews Correlation Coefficient (MCC) | Balanced binary classification evaluation  |
|                       | Cohen's Kappa                          | Agreement beyond random chance             |
| <b>Clustering</b>     | Silhouette Score                       | Cluster cohesion and separation            |

|                               |                              |                                           |
|-------------------------------|------------------------------|-------------------------------------------|
|                               | Davies–Bouldin Index         | Cluster compactness                       |
|                               | Calinski–Harabasz Score      | Cluster separation quality                |
| <b>Autoencoder</b>            | MSE                          | Reconstruction quality                    |
| <b>Synthetic Data</b>         | Kolmogorov–Smirnov (KS) Test | Real vs synthetic distribution similarity |
| <b>Statistical Validation</b> | Paired t-test                | Comparison between two models             |
|                               | One-way ANOVA                | Comparison among multiple models          |
| <b>Cognitive Diagnosis</b>    | ACI                          | Validation of learner mastery estimation  |

Using multiple evaluation measures allows the predictive, diagnostic, representational, clustering, and statistical properties of the framework to be examined from complementary perspectives.

RMSE was utilized as an additional error metric in predicting the binary mastery state and was calculated based on the final binary predictions as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (24)$$

where  $N$  is the number of evaluated observations,  $y_i$  is the observed binary mastery label for learner  $i$ , and  $\hat{y}_i$  is the corresponding predicted binary label.

Because both  $y_i$  and  $\hat{y}_i$  take binary values of 0 or 1, the squared error corresponds to the misclassification indicator. Consequently, for the hard binary predictions used in this study, RMSE is mathematically related to accuracy as  $RMSE = \sqrt{1 - Accuracy}$ . RMSE is therefore reported as a complementary error measure rather than as an independent measure of classification discrimination.

#### 4.5 Experimental Protocol and Generalization Assessment

The three experiments were conducted using the original dataset, SMOTE balanced dataset, and GAN generated dataset. The same methodology was used in all the experiments to provide a fair comparison between the proposed framework and baseline models. This was achieved by analysis of the model's performance based on the evaluation metrics discussed in the previous subsection.

### 5. Results

In this section, the results that have been found in an empirical manner on the basis of the original dataset, the SMOTE balanced dataset, and the GAN augmented dataset are mentioned. The methodology starts with validating the DINA model-derived cognitive structure and goes through performance prediction, cross-validation, statistical significance analysis, learning of latent representations, clustering, and explainability.

#### 5.1. Exploratory Data Analysis

The experiments conducted start by validating the cognitive diagnostic component that is part of the proposed methodology. The reliability of the estimation of subject level mastery was validated using the Q-matrix and the ACI before proceeding to the evaluation of the prediction performance.

The DINA model used the three-skill Q-matrix for the academic subjects in the data set as shown in Table 6.

Table 6. Q-matrix used in the DINA model

| Academic Subject   | Mathematics Skill | Science Skill | English Skill |
|--------------------|-------------------|---------------|---------------|
| <b>Mathematics</b> | 1                 | 0             | 0             |
| <b>Science</b>     | 0                 | 1             | 0             |
| <b>English</b>     | 0                 | 0             | 1             |

Each academic subject has a one-to-one correspondence with its relevant cognitive ability according to the Q-matrix. The math scores would affect mathematics proficiency, science scores science proficiency, and English scores English proficiency. This gives a clear basis for the cognitive diagnosis at the subject level.

The ACI was used to determine how consistent the DINA-based mastery construct is. Table 7 shows the results of the cognitive validation process.

Table 7. DINA validation summary

| Metric                            | Value                   |
|-----------------------------------|-------------------------|
| <b>Number of Cognitive Skills</b> | 3                       |
| <b>Q-Matrix Skill Count</b>       | 3                       |
| <b>Mastery Threshold</b>          | Subject-specific median |

|                       |                                        |
|-----------------------|----------------------------------------|
| <b>ACI</b>            | 0.7664                                 |
| <b>Interpretation</b> | Moderate-to-good cognitive consistency |

The ACI value of 0.7664 shows that there is an acceptable level of correspondence between the observed academic achievement and the estimated cognitive mastery states. This finding justifies the inclusion of the three-skills DINA-based diagnostic representation as an integral part of the proposed framework.

## 5.2. Comparative Predictive Performance Across Experimental Configurations

The predictions of the tested models were analyzed in terms of the original, SMOTE, and GAN-based setups by the means of evaluation metrics. The analysis encompasses five classic ML models (Logistic Regression, Decision Tree, Random Forest, SVM, and XGBoost) as well as three DL models (LSTM, GRU, and the proposed BiLSTM-DINA). The findings are reported in Table 8.

Table 8. Comparative predictive performance under original, SMOTE and GAN configurations

| <b>Dataset / Model Configuration</b>   | <b>Accuracy</b> | <b>Balance d Accuracy</b> | <b>Precision</b> | <b>Recall</b> | <b>F1-score</b> | <b>AUC</b> | <b>RMSE</b> | <b>MCC</b> | <b>Cohen's Kappa</b> |
|----------------------------------------|-----------------|---------------------------|------------------|---------------|-----------------|------------|-------------|------------|----------------------|
| <b>Original</b> Logistic Regression    | 0.9387          | —                         | 0.9359           | 0.9419        | 0.9389          | 0.9937     | 0.2476      | —          | —                    |
| <b>Original</b> Decision Tree          | 1.0000          | —                         | 1.0000           | 1.0000        | 1.0000          | 1.0000     | 0.0000      | —          | —                    |
| <b>Original</b> Random Forest          | 0.9968          | —                         | 0.9936           | 1.0000        | 0.9968          | 1.0000     | 0.0568      | —          | —                    |
| <b>Original</b> SVM                    | 0.9452          | —                         | 0.9539           | 0.9355        | 0.9446          | 0.9934     | 0.2342      | —          | —                    |
| <b>Original</b> XGBoost                | 1.0000          | —                         | 1.0000           | 1.0000        | 1.0000          | 1.0000     | 0.0000      | —          | —                    |
| <b>Original</b> LSTM                   | 0.9709          | 0.8444                    | 0.9725           | 0.9965        | 0.9843          | 0.9949     | 0.1704      | 0.7960     | 0.7848               |
| <b>Original</b> GRU                    | 0.9645          | 0.8234                    | 0.9691           | 0.9930        | 0.9809          | 0.9869     | 0.1884      | 0.7475     | 0.7369               |
| <b>Original</b> BiLSTM-DINA (Proposed) | 0.9645          | 0.7885                    | 0.9627           | 1.0000        | 0.9810          | 0.9942     | 0.1884      | 0.7453     | 0.7142               |
| <b>SMOTE</b> Logistic Regression       | 0.9290          | —                         | 0.9182           | 0.9419        | 0.9299          | 0.9926     | 0.2664      | —          | —                    |
| <b>SMOTE</b> Decision Tree             | 1.0000          | —                         | 1.0000           | 1.0000        | 1.0000          | 1.0000     | 0.0000      | —          | —                    |
| <b>SMOTE</b> Random Forest             | 1.0000          | —                         | 1.0000           | 1.0000        | 1.0000          | 1.0000     | 0.0000      | —          | —                    |
| <b>SMOTE</b> SVM                       | 0.9387          | —                         | 0.9416           | 0.9355        | 0.9385          | 0.9922     | 0.2476      | —          | —                    |
| <b>SMOTE</b> XGBoost                   | 1.0000          | —                         | 1.0000           | 1.0000        | 1.0000          | 1.0000     | 0.0000      | —          | —                    |
| <b>SMOTE</b> LSTM                      | 0.9895          | 0.9894                    | 1.0000           | 0.9789        | 0.9893          | 0.9992     | 0.1027      | 0.9791     | 0.9789               |
| <b>SMOTE</b> GRU                       | 0.9772          | 0.9771                    | 1.0000           | 0.9542        | 0.9766          | 0.9963     | 0.1512      | 0.9553     | 0.9543               |
| <b>SMOTE</b> BiLSTM-DINA (Proposed)    | 0.9912          | 0.9912                    | 1.0000           | 0.9824        | 0.9911          | 0.9994     | 0.0937      | 0.9826     | 0.9824               |
| <b>GAN</b> Logistic Regression         | 0.8750          | —                         | 0.8698           | 0.8820        | 0.8759          | 0.9408     | 0.3536      | —          | —                    |
| <b>GAN</b> Decision Tree               | 0.9670          | —                         | 0.9587           | 0.9760        | 0.9673          | 0.9670     | 0.1817      | —          | —                    |
| <b>GAN</b> Random Forest               | 0.9970          | —                         | 0.9960           | 0.9980        | 0.9970          | 0.99995    | 0.0548      | —          | —                    |
| <b>GAN</b> SVM                         | 0.9970          | —                         | 0.9940           | 1.0000        | 0.9970          | 0.99997    | 0.0548      | —          | —                    |
| <b>GAN</b> XGBoost                     | 0.9950          | —                         | 0.9940           | 0.9960        | 0.9950          | 0.99994    | 0.0707      | —          | —                    |
| <b>GAN</b> LSTM                        | 0.9450          | 0.9450                    | 1.0000           | 0.8900        | 0.9418          | 0.9929     | 0.2345      | 0.8954     | 0.8900               |
| <b>GAN</b> GRU                         | 0.9580          | 0.9580                    | 0.9935           | 0.9220        | 0.9564          | 0.9928     | 0.2049      | 0.9184     | 0.9160               |
| <b>GAN</b> BiLSTM-DINA (Proposed)      | 0.9580          | 0.9580                    | 0.9935           | 0.9220        | 0.9564          | 0.9934     | 0.2049      | 0.9184     | 0.9160               |

Note: Bold values indicate the best value within each dataset/configuration for the corresponding metric. For RMSE, the lowest value is considered best. “—” indicates that the corresponding metric was not reported for the traditional machine-learning models in the original experimental results.

#### 5.2.1 Generalization and Overfitting Assessment

The existence of perfect or nearly perfect classification performance for certain conventional machine learning algorithms should not be interpreted as proof of their superiority in prediction. Using the default setting, Decision Tree and XGBoost had test set accuracy of 1.0000, while using the SMOTE setting, Decision Tree, Random Forest, and XGBoost had 1.0000. These outcomes reveal that the constructed predictor space made it easy for these models to separate classes within the dataset. Nevertheless, having a perfect test set accuracy is not enough to ensure that there was no overfitting.

Thus, generalization ability of models was tested with the help of two types of evidences which could be provided by the experiment setup: stratified fivefold cross-validation and independent test set evaluation. With the initial setting of model parameters, Random Forest and XGBoost yielded mean cross validation accuracies of 0.9987 and 0.9994 and both obtained independent-test accuracy of 1.0000. In the SMOTE scenario, the cross-validation accuracies were consistent at 0.9987 and 0.9994 for Random Forest and XGBoost, respectively. Both models achieved an independent-test accuracy of 1.0000 in this case. Such a similarity in cross-validation and independent-test accuracies suggests that the high accuracy of the model was not restricted to one specific partition.

Furthermore, the same diagnosis suggests that the observed behavior was configuration dependent. In terms of GAN configuration, the Random Forest, SVM, and XGBoost provided independent test accuracies of 0.9970, 0.9970, and 0.9950, respectively, corresponding to five-fold cross-validation values of 0.9928, 0.9972, and 0.9914. This drop compared to the original and SMOTE configurations proves that the impressive results obtained by the baseline classifiers were representation dependent and not uniformly good in all experimental setups.

In this case, the 1.0000 performance of Decision Tree, Random Forest, and XGBoost is considered as experimental findings based on the current data and experiment methodology, and not as proof of their absolutely perfect prediction ability. Even though explicit history of losses during training was not saved for the conventional baseline models, hence the inability to compare losses for training and validation, the concordance of cross-validation and independent test performances serves as another check of generalization ability. In such a way, the results do not show absence of overfitting; rather, they show that the good performing baseline models preserved their prediction behavior on different data subsets.

#### 5.2.2 Original Configuration

The reported RMSE values are aligned with the corresponding accuracy values since RMSE was computed on the basis of the final binary predictions. In the case of the suggested BiLSTM-DINA model, the accuracy of 0.9645 in the original dataset is equivalent to RMSE equal to about 0.1884. The same is true for SMOTE accuracy of 0.9912 that corresponds to RMSE value of about 0.0937 and GAN accuracy of 0.9580 that is equivalent to RMSE value of about 0.2049. Hence, the relationship between accuracy and RMSE is expected under the current binary hard-prediction paradigm and does not show any inconsistencies between the two metrics. These findings mean that the developed feature space allows separating information for some traditional classifiers.

However, the presence of such perfect results does not imply that they represent perfectly generalized results. Since the mastery performance measure has been extracted from the academic performance parameters, some traditional algorithms might detect very deterministic dependencies between the designed predictors and the target. For that reason, the results are interpreted based on the validation/cross-validation results instead of just the test results.

While the best result from among the deep learning models was obtained by LSTM with accuracy of 0.9709, the proposed BiLSTM-DINA gave an accuracy of 0.9645 with a recall of 1.0000. This implies that the proposed model was able to classify all instances of mastery correctly in the test set used for evaluation but had a lower balanced accuracy of 0.7885 compared to LSTM.

The confusion matrix for the developed BiLSTM-DINA model is shown in Fig. 2. The model correctly classified 15 samples as negatives, 284 as positives, misclassified 11 samples as positive, and had no false negative predictions. Therefore, the perfect recall was achieved due to lack of false negative predictions, whereas 11 false positive predictions lowered precision.

The classification accuracy of the BiLSTM-DINA model can be further illustrated using the confusion matrix in Fig. 2.

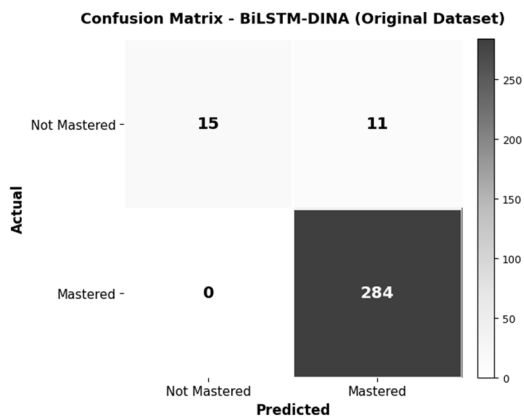


Fig. 2. Confusion matrix of the proposed BiLSTM–DINA model on the original dataset.

From the confusion matrix, there were 15 true negatives, 284 true positives, 11 false positives, and 0 false negatives. Consequently, the classifier accurately classified all positive cases of mastery, which produced a recall of 1.0000. False positives of 11 made the precision low, contributing to the low balanced accuracy compared to LSTM.

ROC curve for this is shown in Fig. 3.

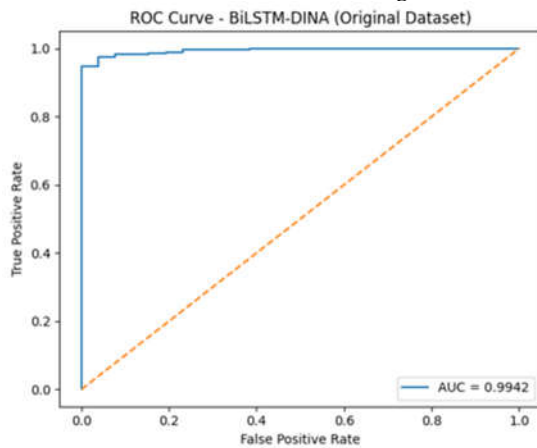


Fig. 3. ROC curve of the proposed BiLSTM–DINA model on the original dataset.

ROC analysis yielded an AUC of 0.9942, denoting good discrimination capability between the mastered and non-mastered classes in all classification cut-offs.

Autoencoder training behavior under original configuration is illustrated in Fig. 4.

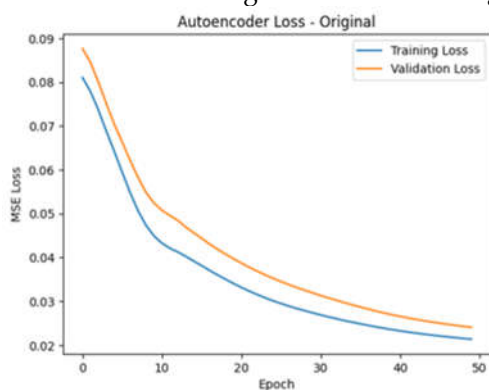


Fig. 4. Autoencoder training and validation loss on the original dataset.

The training and validation losses quickly decreased in the early epochs and then started to converge. The common pattern of convergence shows that there is stable reconstruction without divergence between the training and validation losses.

### 5.2.3 SMOTE Configuration

The BiLSTM-DINA model performed best in deep learning when configured in SMOTE settings. In this case, the model scored accuracy and balanced accuracy of 0.9912, F1 score of 0.9911, AUC of 0.9994, MCC of 0.9826, and Cohen's Kappa of 0.9824. The lowest RMSE was also scored by the model among the three deep learning models (0.0937).

The accuracy achieved by LSTM was 0.9895, whereas GRU reached 0.9772. Therefore, the introduced BiLSTM-DINA gave the highest performance among all evaluated deep learning architectures in the SMOTE setting.

Similarly, the conventional classifiers performed quite exceptionally well for the SMOTE setting. Decision Tree, Random Forest, and XGBoost delivered accuracy, precision, recall, F1-score, and AUC of 1.0000, while their RMSE was 0.0000. These results imply that there is significant class separability in SMOTE. Nevertheless, the flawless classification is cautiously interpreted due to the fact that augmenting data synthetically may change the geometry of the feature space. In this way, the results are reviewed together with the validation and cross-validation findings to check if the obtained performance still holds true for various partitions of the data set.

Confusion matrix of the BiLSTM-DINA model with SMOTE is depicted in Fig. 5.

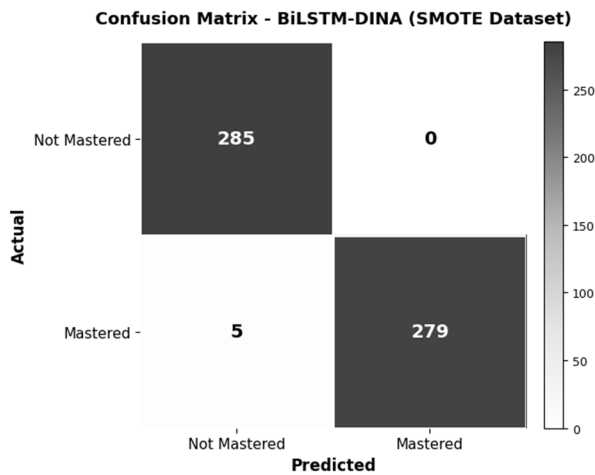


Fig. 5. Confusion matrix of the proposed BiLSTM–DINA model on the SMOTE dataset.

The model identified 285 instances of non-mastered learners and 279 instances of mastered learners with an error rate of five false negatives and zero false positives. Precision was 1.0000, recall was 0.9824, and accuracy was 0.9912, which correspond with the results shown in Table 8.

The ROC curve for the SMOTE configuration is illustrated in Fig. 6.

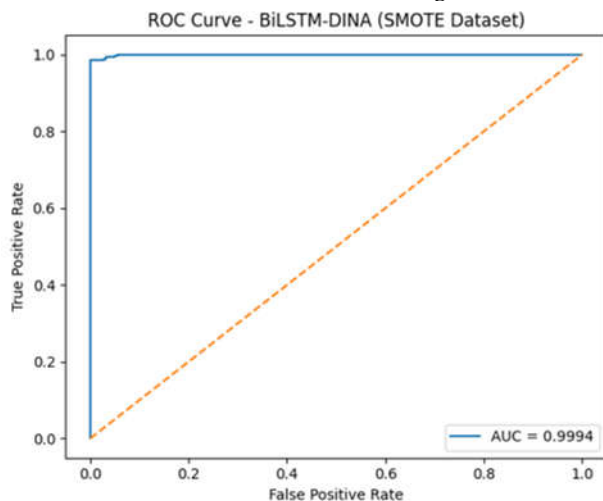


Fig. 6. ROC curve of the proposed BiLSTM–DINA model on the SMOTE dataset.

An area under the curve of 0.9994 was obtained from the ROC curve, showing high discriminative capacity for the classes and marking the highest AUC obtained from the deep learning architecture framework amongst the three architectures.

Autoencoder Training and Validation Loss Plot is shown in Fig. 7.

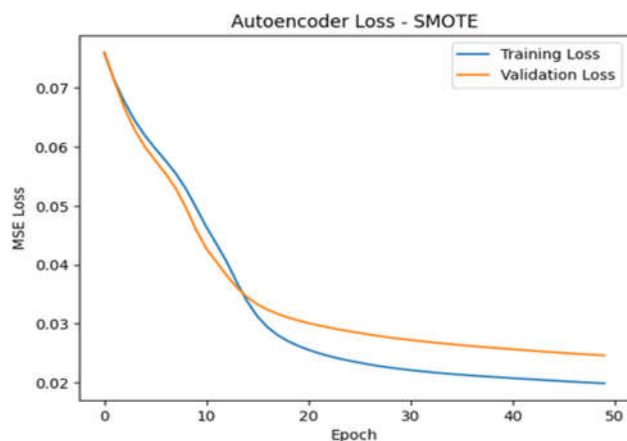


Fig. 7. Autoencoder training and validation loss on the SMOTE dataset.

The losses for training and validation reduced gradually and converged during training, showing that representation learning was stable with no significant difference between the two loss curves.

#### 5.2.4 GAN-Augmented Configuration

GAN enhanced model showed different results compared to SMOTE. In terms of machine learning models, the best results were shown by the Random Forest and SVM with the highest accuracy of 0.9970. The SVM model also had the highest recall value of 1.0000 and AUC of 0.99997. XGBoost showed excellent results with the accuracy of 0.9950 and AUC of 0.99994.

Among the deep-learning models, GRU and the proposed BiLSTM-DINA achieved identical accuracy (0.9580), balanced accuracy (0.9580), precision (0.9935), recall (0.9220), F1-score (0.9564), RMSE (0.2049), MCC (0.9184), and Cohen's Kappa (0.9160). However, the proposed BiLSTM-DINA achieved a slightly higher AUC of 0.9934 compared with 0.9928 for GRU, indicating marginally stronger ranking-based discrimination.

Confusion Matrix of the suggested BiLSTM-DINA architecture for the GAN configuration is shown in Fig. 8.

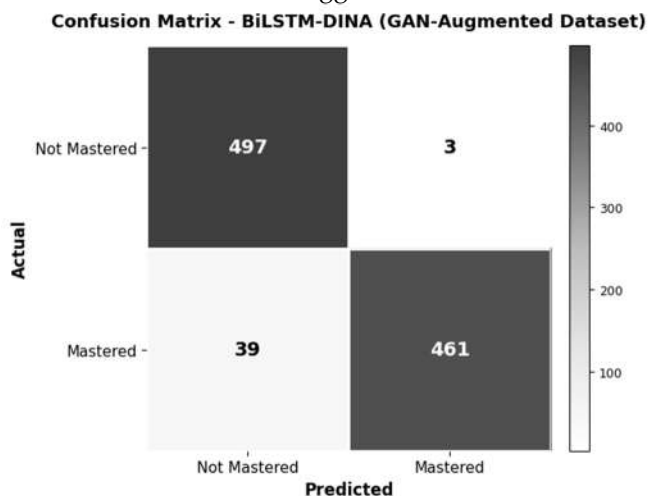


Fig. 8. Confusion matrix of the proposed BiLSTM-DINA model on the GAN-augmented dataset.

The confusion matrix had 497 true negatives, 461 true positives, three false positives, and 39 false negatives. The accuracy attained by the model as a result was 0.9580, while 39 false negative results gave a recall of 0.9220. Unlike the original model and the SMOTE method, the more number of false negatives implies that variability was added to the classification of learners who have mastered the skill using GAN augmentation.

Figure 9 shows the ROC curve of the GAN-augmented method.

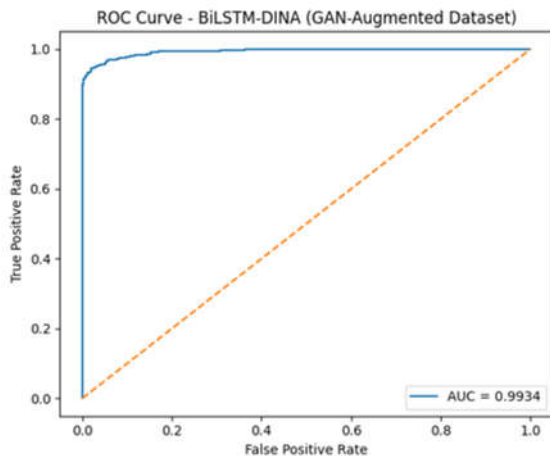


Fig. 9. ROC curve of the proposed BiLSTM–DINA model on the GAN-augmented dataset.

The model obtained an AUC score of 0.9934. This shows that there is good separation between the two classes of mastery despite the added variance through GAN based augmentation.

Autoencoder learning behavior for GAN configuration is shown in Fig. 10.

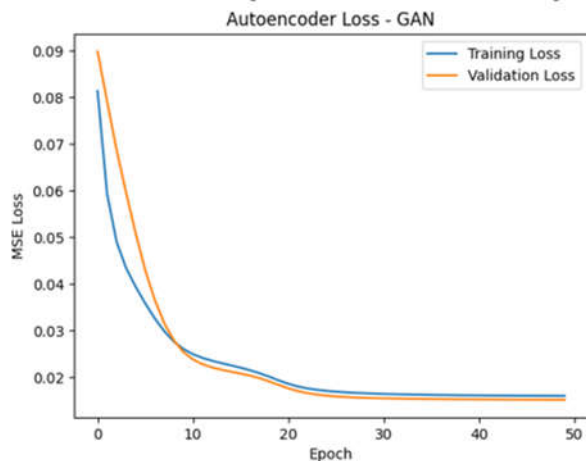


Fig. 10. Autoencoder training and validation loss on the GAN-augmented dataset.

The reconstruction error reduced systematically through the training process while the training and validation trends were quite steady. This suggests that there was consistent learning of the latent representation through the GAN-enabled architecture.

Training and validation reconstruction loss curves offer complimentary evidence on the issue of Autoencoder convergence. Autoencoder training and validation reconstruction loss curves were considered in order to find out about convergence of reconstruction and possibility of overfitting. In the Original setting, there was a tendency of the validation reconstruction loss being a little bit higher than the training one when the Autoencoder converged. This was expected since the Autoencoder parameters are estimated based on the training set observations and not based on the validation set observations. Thus, an increase in validation reconstruction error compared to training reconstruction error is a natural result.

It should be noted that the marginally higher validation loss does not necessarily indicate overfitting in itself. The appropriate diagnostic is the pattern exhibited by the two loss curves during the training process. In the current experimental setup, both training and validation losses have reduced and converged without showing any persistent divergence. This implies that the Autoencoder did learn a stable latent space representation, rather than fitting the specificities of the training data at all times. The behavior of training and validation loss for Original, SMOTE, and GAN setups are shown in Figures 4, 7, and 10 respectively.

The same methodology was used for interpreting the augmentation architectures. There is no doubt that the validation loss can stay slightly higher than the training loss because validation data are not seen during parameter tuning; but there is no divergence of the two plots, which does not signify reconstruction overfitting in any significant way. Reconstruction quality is thus interpreted alongside independent indicators of reconstruction performance and clustering evaluation. The corresponding behavior of the model with SMOTE and GAN augmentation is presented in Figs. 7 and 10, respectively. The above curves are taken into consideration along with the independent test

reconstruction outcome and not in isolation from each other, hence giving within-training evidence of convergence as well as out-of-sample evidence of reconstruction.

Hence, the train-validation loss divergence is the generalization gap that is expected between training observation optimization and validation observation, whereas the convergence of both gives evidence of stable Autoencoder training.

### 5.2.5 Overall Comparative Interpretation

Overall, the results demonstrate that the effect of data augmentation was configuration- and model-dependent rather than uniformly beneficial. The SMOTE configuration produced the strongest performance for the proposed BiLSTM-DINA model, achieving 0.9912 accuracy, 0.9912 balanced accuracy, 0.9911 F1-score, 0.9994 AUC, 0.9826 MCC, and 0.9824 Cohen's Kappa. This improvement indicates that the neighbourhood-based synthetic representation used by SMOTE provided a favourable training configuration for the proposed model. On the other hand, the GAN architecture yielded poor classification performance of the deep-learning classifier, with the accuracy and AUC of the model being 0.9580 and 0.9934, respectively. It means that an increased variability in synthetic learners' representations may not lead to better classification accuracy when the distribution is different from that of the target classification problem."

Thus, the use of augmentation in this research serves as an experiment for testing and comparing models and not as an assumption to improve the performance of the machine learning model. The comparison of the results achieved using SMOTE and GAN allows us to identify whether the predicted behavior is consistent across different methods of generating synthetic data. The results obtained show that the method of augmentation has an impact on the behavior of the machine learning model.

### 5.5. Cross-Validation Results

Stratified 5-fold cross-validation analysis was conducted on the selected traditional machine learning algorithms to determine their consistency in different data partitions. The mean accuracy, standard deviation, and 95% confidence interval from the repeated experiment results are presented in Table 9.

Table 9. Cross-validation performance

| Dataset  | Model               | Mean Accuracy | Standard Deviation | 95% Confidence Interval |
|----------|---------------------|---------------|--------------------|-------------------------|
| Original | Logistic Regression | 0.9206        | 0.0120             | 0.9101 – 0.9311         |
| Original | Random Forest       | 0.9987        | 0.0026             | 0.9964 – 1.0010         |
| Original | SVM                 | 0.9290        | 0.0106             | 0.9197 – 0.9383         |
| Original | XGBoost             | 0.9994        | 0.0013             | 0.9982 – 1.0005         |
| SMOTE    | Logistic Regression | 0.9194        | 0.0102             | 0.9104 – 0.9283         |
| SMOTE    | Random Forest       | 0.9987        | 0.0016             | 0.9973 – 1.0001         |
| SMOTE    | SVM                 | 0.9258        | 0.0087             | 0.9182 – 0.9334         |
| SMOTE    | XGBoost             | 0.9994        | 0.0013             | 0.9982 – 1.0005         |
| GAN      | Logistic Regression | 0.8790        | 0.0086             | 0.8715 – 0.8865         |
| GAN      | Random Forest       | 0.9928        | 0.0027             | 0.9904 – 0.9952         |
| GAN      | SVM                 | 0.9972        | 0.0015             | 0.9959 – 0.9985         |
| GAN      | XGBoost             | 0.9914        | 0.0026             | 0.9891 – 0.9937         |

The findings from cross-validation show that the models exhibit consistency in their predictive accuracy regardless of the experiment setup. XGBoost produced the highest accuracy in the Original (0.9994) data set, whereas SVM exhibited its strength in the GAN-augmented data set (0.9972). Random Forest maintained consistency in terms of accuracy in all the data sets.

Good cross-validation performance of the traditional classifiers is another indication that the flawless or nearly flawless testing outcomes were not limited to one fold only. However, it must be stated that cross-validation consistency alone cannot be taken as an indicator of lack of data-specific effects. Thus, good generalization within the data can be inferred from the outcomes; generalization to other datasets needs to be externally confirmed.

### 5.6. Statistical Significance Analysis

Hypothesis testing for statistical significance was carried out to test the differences found between the various models. Comparison of the models was done using the paired t-test while the differences between all the models was tested using One Way ANOVA. The results are shown in Table 10.

Table 10. Statistical test results

| Dataset  | t-test Comparison        | t-statistic | p-value | ANOVA F-value | ANOVA p-value          |
|----------|--------------------------|-------------|---------|---------------|------------------------|
| Original | XGBoost vs Random Forest | 0.4082      | 0.7040  | 111.8327      | $6.03 \times 10^{-11}$ |
| SMOTE    | XGBoost vs Random Forest | 0.5345      | 0.6213  | 170.8212      | $2.33 \times 10^{-12}$ |

|            |                          |         |        |          |                        |
|------------|--------------------------|---------|--------|----------|------------------------|
| <b>GAN</b> | XGBoost vs Random Forest | -1.2472 | 0.2804 | 585.7794 | $1.46 \times 10^{-16}$ |
|------------|--------------------------|---------|--------|----------|------------------------|

In case of paired t-test, the p-values were 0.7040, 0.6213, and 0.2804 for Original, SMOTE, and GAN data sets respectively, which means there was no statistically significant difference between XGBoost and Random Forest. However, in case of One-way ANOVA, all tests were statistically significant ( $p < 0.001$ ).

#### 5.7. Autoencoder Reconstruction and Clustering Performance

Representation Learning was analyzed through the behavior of the reconstruction loss during training as well as reconstruction performance on an independent test set. The Autoencoder was trained for latent representation learning, using MSE as the objective function for reconstructing the data. Training and validation curves of the reconstruction loss were considered for analysis of convergence and over-fitting, and also the MSE and RMSE of the reconstruction were computed on the independent test set. The clustering effectiveness of the latent representations generated was further tested through the use of the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score.

Reconstruction effectiveness of the Autoencoder for Original, SMOTE, and GAN methods has been presented in Table 11 with the respective MSE and RMSE figures.

Table 11. Autoencoder reconstruction performance

| <b>Dataset</b>  | <b>Autoencoder MSE</b> | <b>Autoencoder RMSE</b> |
|-----------------|------------------------|-------------------------|
| <b>Original</b> | 0.0218                 | 0.1477                  |
| <b>SMOTE</b>    | 0.0207                 | 0.1439                  |
| <b>GAN</b>      | 0.0158                 | 0.1257                  |

The Autoencoder achieved reconstruction MSE values of 0.0218, 0.0207, and 0.0158 for the Original, SMOTE, and GAN configurations, respectively. The corresponding RMSE values were 0.1477, 0.1439, and 0.1257. The GAN configuration produced the lowest reconstruction error, indicating comparatively lower reconstruction loss under the GAN-augmented configuration.

Quality of learner grouping after latent features extraction was tested through three clustering validation metrics. The obtained results are shown in Table 12.

Table 12. Clustering validation results

| <b>Dataset</b>  | <b>Clustering Method</b> | <b>Silhouette Score</b> | <b>Davies-Bouldin Index</b> | <b>Calinski-Harabasz Score</b> |
|-----------------|--------------------------|-------------------------|-----------------------------|--------------------------------|
| <b>Original</b> | K-Means                  | 0.7026                  | 0.4723                      | 4427.08                        |
| <b>Original</b> | DBSCAN                   | 0.8959                  | 0.5784                      | 33772.62                       |
| <b>Original</b> | Hierarchical             | 0.7028                  | 0.4722                      | 4425.88                        |
| <b>SMOTE</b>    | K-Means                  | 0.6828                  | 0.5196                      | 4033.77                        |
| <b>SMOTE</b>    | DBSCAN                   | 0.9454                  | 0.0797                      | 212432.73                      |
| <b>SMOTE</b>    | Hierarchical             | 0.6060                  | 0.6433                      | 2835.98                        |
| <b>GAN</b>      | K-Means                  | 0.2837                  | 1.2163                      | 3551.57                        |
| <b>GAN</b>      | DBSCAN                   | -0.5170                 | 1.3300                      | 12.42                          |
| <b>GAN</b>      | Hierarchical             | 0.2191                  | 1.5176                      | 2985.54                        |

Clustering efficiency differed among the methods tested. The best clustering performance was recorded for DBSCAN on the Original (Silhouette = 0.8959) and SMOTE (Silhouette = 0.9454; Davies-Bouldin Index = 0.0797) datasets. For the GAN data, the overall clustering efficiency was poor, but K-Means had the best Silhouette Score (0.2837).

#### 5.8. Explainability Analysis

SHAP-based explainability was used to examine the contribution of the engineered trajectory features to learner-mastery prediction. Importantly, the SHAP analysis was performed on the XGBoost classifier trained using the GAN-augmented configuration rather than directly on the BiLSTM or DINA components. Therefore, the results should be interpreted as a global feature-attribution analysis of the representative tree-based predictive model, providing complementary evidence about which trajectory characteristics are most influential in the mastery-prediction task. Global explanations of the model are provided in Figures 11 and 12.

Table 13. SHAP-based feature importance ranking

| <b>Rank</b> | <b>Feature</b>   | <b>Importance Score</b> |
|-------------|------------------|-------------------------|
| <b>1</b>    | Overall Mean     | 2.5540                  |
| <b>2</b>    | Mathematics Mean | 1.6858                  |
| <b>3</b>    | Science Mean     | 1.1434                  |
| <b>4</b>    | Learning Gain    | 1.0801                  |
| <b>5</b>    | Science Gain     | 0.9630                  |

|    |                            |        |
|----|----------------------------|--------|
| 6  | Science Mastery            | 0.8728 |
| 7  | Mathematics Mastery        | 0.8631 |
| 8  | English Mean               | 0.8517 |
| 9  | English Mastery            | 0.8007 |
| 10 | Performance Stability      | 0.7012 |
| 11 | Mathematics Gain           | 0.5851 |
| 12 | English Gain               | 0.5688 |
| 13 | Overall Standard Deviation | 0.3014 |

As shown in Table 13, Overall Mean had the highest mean absolute SHAP importance (2.5540), followed by Mathematics Mean (1.6858), Science Mean (1.1434), Learning Gain (1.0801), and Science Gain (0.9630). Overall Standard Deviation had the lowest importance score (0.3014). This ranking indicates that cumulative achievement and learning-progression variables contributed more strongly to the predictive output than the overall variability measure under the evaluated configuration.

The global SHAP Summary Plot in Figure 11 shows the distribution of contributions at the level of each feature for all the learners considered, whereas the SHAP Mean Absolute Importance Ranking in Figure 12 is the equivalent to the above graph but expressed as the mean absolute value. Both charts provide complementary information about the feature's contribution and global importance, respectively.

The dominance of the Overall Mean feature implies that the cumulative academic achievement feature is the one with the highest predictive power among the engineered features. The dominance of the Mathematics Mean and Science Mean features implies that the subject-specific academic achievement feature made a significant contribution to the discrimination of mastery. In addition, the Learning Gain feature dominance implies that the change in performance over time made a significant contribution to discrimination of mastery. On the other hand, the lack of overall standard deviation's predictive dominance implies that overall performance variance has a relatively low contribution to the predicted output in the experiment.

The global SHAP summary plot is shown in Fig. 11.

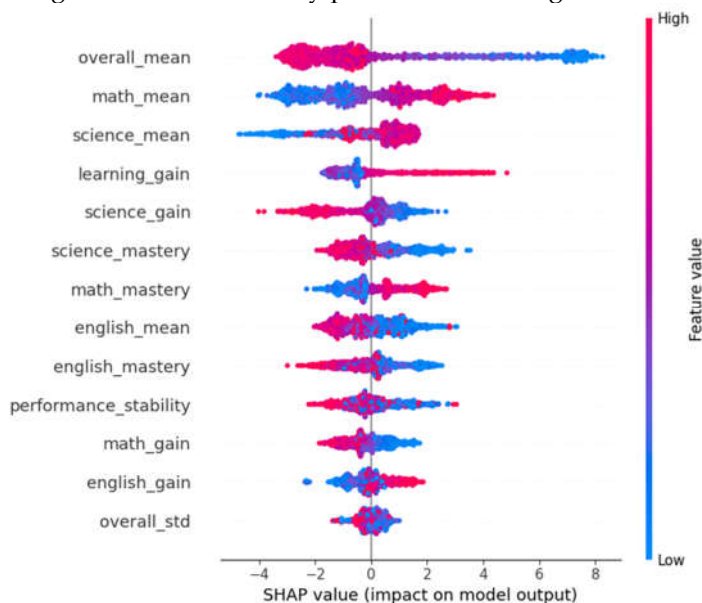


Fig. 11. SHAP summary plot

SHAP summary plot shows the importance of individual trajectory-based features in terms of their overall impact on the prediction model for all learners. Features appearing near the top of the graph have more importance in making the prediction. The importance of features in ranking form is shown in Fig. 12.

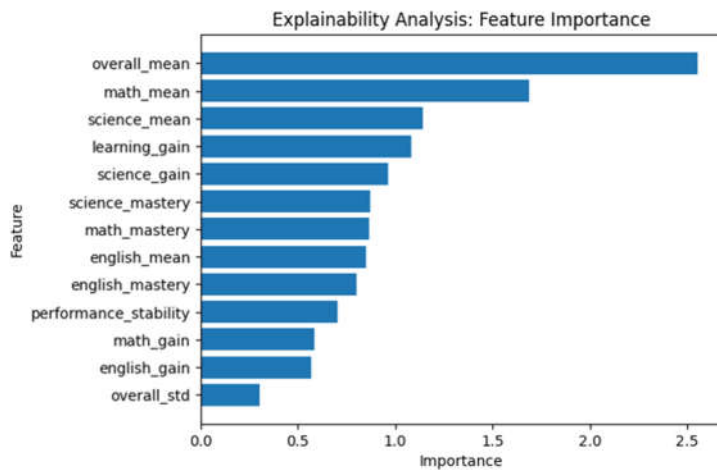


Fig. 12. SHAP feature importance bar chart

The bar graph offers a numeric measure of feature importance and strengthens the ordering indicated in Table 13. Overall Mean makes the most important contribution to prediction, while Mathematics Mean and Science Mean follow next; however, Overall Standard Deviation contributes the least amount to prediction.

#### 5.9. Synthetic Data Quality Assessment

The similarity between the original and synthetic data was examined via statistical distributions of the parameters. Mean differences and Kolmogorov-Smirnov (KS) statistics were computed for the selected trajectory-based characteristics. The findings are shown in Table 14.

Table 14. Synthetic data quality assessment

| Feature                           | Real Mean | SMOTE Mean | GAN Mean | SMOTE Mean Difference | GAN Mean Difference | SMOTE KS Statistic | GAN KS Statistic |
|-----------------------------------|-----------|------------|----------|-----------------------|---------------------|--------------------|------------------|
| <b>Math Mean</b>                  | 0.5770    | 0.5769     | 0.4654   | 0.0001                | 0.1117              | 0.0006             | 0.3524           |
| <b>Science Mean</b>               | 0.6592    | 0.6592     | 0.5922   | 0.0000                | 0.0670              | 0.0004             | 0.2356           |
| <b>English Mean</b>               | 0.6490    | 0.6489     | 0.4510   | 0.0001                | 0.1980              | 0.0005             | 0.4595           |
| <b>Overall Mean</b>               | 0.6946    | 0.6945     | 0.5840   | 0.0001                | 0.1105              | 0.0006             | 0.2611           |
| <b>Overall Standard Deviation</b> | 0.4124    | 0.4123     | 0.4734   | 0.0000                | 0.0610              | 0.0005             | 0.3397           |
| <b>Learning Gain</b>              | 0.5144    | 0.5142     | 0.4913   | 0.0001                | 0.0230              | 0.0006             | 0.1798           |
| <b>Performance Stability</b>      | 0.2872    | 0.2872     | 0.5143   | 0.0000                | 0.2271              | 0.0005             | 0.4852           |

The samples produced by SMOTE were very similar to the original samples with the near-zero mean difference and KS statistics of about 0.0005. By comparison, the GAN-generated dataset had larger deviations especially for English Mean, Performance Stability and Mathematics Mean, suggesting higher variability in the synthetic learner profiles. To better illustrate these distributional difference, feature-wise comparisons are provided between Original and GAN datasets in the following figures: (Fig.) 13–17.

Figure 13 compares the Mathematics Mean distributions and illustrates the shift between the original and GAN-generated samples.

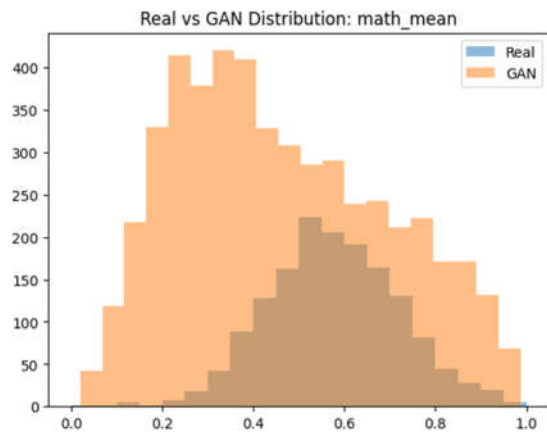


Fig. 13. Real versus GAN distribution of Mathematics Mean

The distribution shows the difference between the original Mathematics Mean and the generated samples from the GAN and the associated KS statistic is reported in Table 14, summarising the shift.

Figure 14 presents the Science Mean distributions, showing the difference in distributional shape between the original and GAN-generated data.

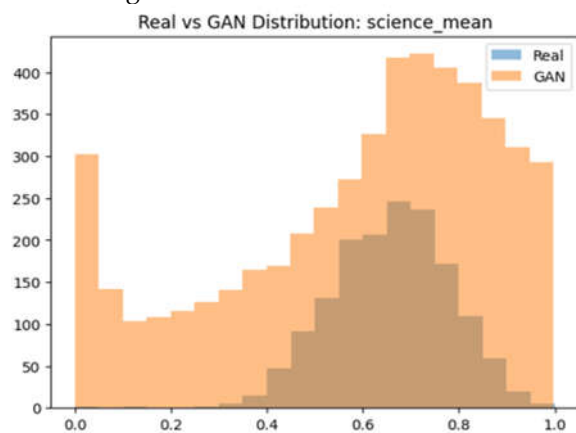


Fig. 14. Real versus GAN distribution of Science Mean

The distribution of the Science Mean samples has noticeable differences between the original and the generated sets, as noted in Table 14.

Figure 15 illustrates the greater separation between the original and GAN-generated English Mean distributions.

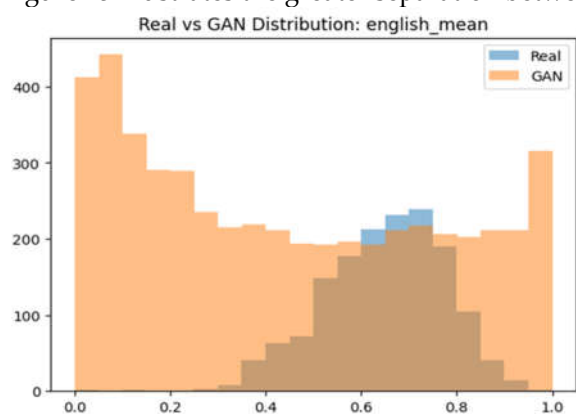


Fig. 15. Real versus GAN distribution of English Mean

The English Mean distribution shows an increased separation between the original and GAN samples, in line with the higher Mean difference and KS statistic shown in Table 14.

Figure 16 shows the distributional shift in Overall Mean between the original and GAN-generated datasets.

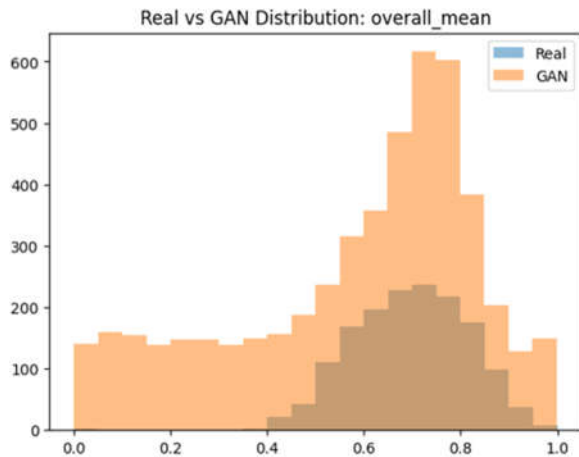


Fig. 16. Real versus GAN distribution of Overall Mean

The Overall Mean distribution is recognizable, with a shift between the real distribution and the GAN-generated distribution, as evident in the mean difference of 0.1105 and KS statistic of 0.2611.

Figure 17 highlights the comparatively largest distributional deviation for Performance Stability between the original and GAN-generated data.

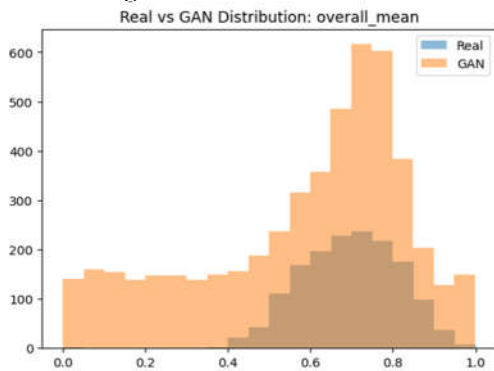


Fig. 17. Real versus GAN distribution of Performance Stability

Performance Stability exhibits the largest deviation between the real and GAN-generated distributions, as reflected by the highest mean difference and KS statistic.

#### 5.10. Ablation Study

In order to assess the impact of each of the framework elements, an ablation study was carried out through stepwise elimination of GAN Augmentation, DINA Incorporation, and Trajectory-based Feature Engineering. This is presented in Table 15.

Table 15. Ablation study results

| Configuration                                  | Accuracy | Precision | Recall | F1-score | AUC    | RMSE   |
|------------------------------------------------|----------|-----------|--------|----------|--------|--------|
| <b>Full Model (GAN + Trajectory + DINA)</b>    | 0.9800   | 0.9878    | 0.9720 | 0.9798   | 0.9972 | 0.1414 |
| <b>Without GAN (SMOTE + Trajectory + DINA)</b> | 0.9419   | 0.9477    | 0.9355 | 0.9416   | 0.9929 | 0.2410 |
| <b>Without GAN/SMOTE (Original Dataset)</b>    | 0.9645   | 0.9337    | 1.0000 | 0.9657   | 0.9986 | 0.1884 |
| <b>Without DINA</b>                            | 0.9710   | 0.9563    | 0.9871 | 0.9714   | 0.9973 | 0.1704 |
| <b>Without Trajectory Features</b>             | 0.8613   | 0.7828    | 1.0000 | 0.8782   | 0.9945 | 0.3724 |

Among all the tested ablation settings, the entire framework performed best with an accuracy of 98.00%. Removing trajectory-based features caused the largest decline in performance, whereas excluding DINA produced only a modest reduction. Replacing GAN with SMOTE also reduced overall predictive performance.

## 6. Discussion

The findings are discussed according to the two research questions addressing the effectiveness of the proposed BiLSTM-DINA framework and the effects of data augmentation and interpretability components. The discussion is based on the three experimental configurations evaluated in the study.

### 6.1. Discussion of RQ1: Effectiveness of the Proposed BiLSTM-DINA Framework

The results demonstrate that the proposed BiLSTM–DINA framework provided strong learner-mastery prediction across the three experimental configurations. The model achieved an accuracy of 96.45% on the Original dataset, 99.12% with SMOTE, and 95.80% with GAN. The SMOTE configuration produced the strongest performance, with an F1-score of 0.9911, AUC of 0.9994, MCC of 0.9826, and Cohen's Kappa of 0.9824.

Moreover, the results show that trajectory-based feature representation is very crucial. Longitudinal academic data was represented as features which captured academic success, progress, variation, and subject proficiency. It can be seen that deletion of trajectory-based features caused more decrease in performance compared to the deletion of DINA. This shows that trajectory-based representation played a major role in the prediction part while DINA provided an additional diagnosis aspect.

The DINA component provided subject-level mastery information for Mathematics, Science, and English. The obtained Attribute Consistency Index (ACI) of 0.7664 indicates moderate-to-good consistency for the specified three-skill cognitive structure. The DINA output therefore complements the overall mastery prediction by providing subject-level mastery representations within the defined Q-matrix and mastery criterion.

For contextual comparison, the performance of the proposed framework is presented alongside representative studies in Table 16. Because these studies use different datasets, modelling procedures, and experimental conditions, the comparison is descriptive rather than a direct test of superiority.

Table 16. Comparison of the proposed framework with representative studies.

| Study                        | Model         | Dataset                       | Accuracy       | Key Characteristics                                                                    |
|------------------------------|---------------|-------------------------------|----------------|----------------------------------------------------------------------------------------|
| <b>Ahajjam et al. (2022)</b> | ANN           | Student Performance           | 94.20%         | No cognitive diagnosis                                                                 |
| <b>Kalita et al. (2025)</b>  | BiLSTM + SHAP | Longitudinal Academic Data    | 97.50%         | No cognitive diagnostic integration                                                    |
| <b>Jiao et al. (2022)</b>    | Deep Learning | Online Learning Dataset       | 96.30%         | No data augmentation                                                                   |
| <b>Malik et al. (2025)</b>   | XGBoost       | Educational Dataset           | 97.10%         | No trajectory learning                                                                 |
| <b>Present Study</b>         | BiLSTM–DINA   | Longitudinal Academic Dataset | 99.12% (SMOTE) | Trajectory-based prediction with cognitive diagnosis and comparative data augmentation |

This comparative study serves to provide contextual information regarding the performance of prediction in the current study as well as its incorporation of longitudinal trajectory modeling and cognitive diagnosis. The framework presented here includes a prediction approach based on BiLSTM along with a mastery level assessment for subjects based on DINA, while the experiment also explores other possibilities for data augmentation. However, the comparative study needs to be viewed cautiously due to the differences between datasets and settings in various studies.

The internal validation results further support the consistency of the conventional predictive models across data partitions. For example, XGBoost achieved a mean five-fold accuracy of 0.9994 under both Original and SMOTE configurations, while Random Forest achieved 0.9987 in both configurations. Under GAN, XGBoost achieved 0.9914 cross-validation accuracy and 0.9950 independent-test accuracy. These results indicate strong within-dataset predictive consistency while also demonstrating sensitivity to the training-data configuration.

Overall, RQ1 is addressed by demonstrating that the proposed BiLSTM–DINA framework can provide strong learner-mastery prediction from the evaluated longitudinal academic data, while DINA supplies complementary subject-level mastery information.

## 6.2 Discussion of RQ2: Effect of Data Augmentation and Interpretability

The augmentation experiments demonstrate that the effect of synthetic data depends on the augmentation method. The Original dataset contained 1,549 learner records. SMOTE increased the dataset to 1,550 observations, adding one synthetic observation, whereas GAN generated 5,000 synthetic learner profiles, with 2,500 profiles for each class. These

configurations were evaluated to examine the effect of alternative training-data representations rather than to assume that augmentation would necessarily improve performance.

SMOTE produced the strongest performance for the proposed BiLSTM–DINA model. Accuracy increased from 96.45% in the Original configuration to 99.12% with SMOTE, whereas the GAN configuration produced 95.80% accuracy. The SMOTE configuration also achieved the highest F1-score, AUC, MCC, and Cohen's Kappa among the three configurations.

The GAN configuration nevertheless retained strong discrimination, with an AUC of 0.9934. Its lower classification accuracy occurred alongside greater distributional differences between the GAN-generated and original learner profiles. The synthetic-data analysis showed that SMOTE produced near-zero mean differences from the original data, whereas GAN generated larger deviations for several trajectory-based features.

These results suggest that larger synthesized data is not always associated with better prediction performance. In the current research, the smaller-sized SMOTE synthesis turned out to be more suitable for BiLSTM-DINA compared to the much larger sized GAN synthesis. Consequently, the influence of data augmentation should be considered along with the prediction performance.

DINA helps interpret because it provides both overall mastery estimate and subject-level mastery information. ACI equals 0.7664 which is consistency obtained using the three-skill DINA model. Interpretation is always dependent on the Q matrix and mastery criteria used in the study.

SHAP gives a separate interpretation for each individual feature. It was used for interpreting the XGBoost classifier, not for BiLSTM and DINA themselves. The analysis found that Overall Mean, Mathematics Mean, Science Mean, and Learning Gain were the most influential trajectory features. Overall Mean was the feature with the highest SHAP importance, Mathematics Mean, Science Mean, and Learning Gain followed.

Therefore, DINA and SHAP give complementary interpretations. While DINA describes the subject mastery, SHAP reveals the trajectory features that are responsible for XGBoost prediction. This way the analysis goes further than just classification without suggesting that SHAP interprets how BiLSTM and DINA work.

### 6.3 Overall Interpretation

Consequently, the overall results suggest that the proposed framework considers the combination of longitudinal predictions, cognitive diagnoses, comparative augmentations, and complementary feature level interpretations. The best performance of BiLSTM–DINA was observed using SMOTE, whereas GAN resulted in lower predictions' performance and higher distribution deviation from original learners' data.

At the same time, it should be noted that each of the framework components contributes unequally. Trajectory based features have influenced the prediction performance, whereas DINA provided subject-based mastery knowledge. Finally, SHAP showed the contribution of academic traits to the prediction of XGBoost.

The results therefore support the framework as an analytical approach for learner-mastery prediction and cognitive interpretation within the evaluated dataset. They provide a basis for future adaptive educational decision-support applications, but the present study does not experimentally evaluate real-time intervention or learner outcomes following intervention.

### 6.4 Study Limitations

There are two main limitations of the research. First, the framework has been tested using only one longitudinal dataset for academics in Mathematics, Science, and English. As a result, the conclusions drawn from the research do not provide any generalization to other educational institutes, learners, or subjects.

Second, the DINA model was tested by applying a fairly straightforward three-skill Q-matrix approach. It gives an indication of subject level mastery but does not indicate the relationship between the cognitive processes within different components of knowledge.

### 6.5 Future Research Directions

Future studies could assess the framework through larger sample sizes and datasets from multiple institutions to determine its behavior in different educational contexts. Also, other types of cognitive diagnostic models and Q-matrices could be explored to capture fine-grained learner knowledge.

Possible future research directions include transformer- and graph-based sequential models, learning analytics data collected in real time, and privacy-preserving methods such as federated learning. This way, the proposed framework could be extended to cover more diverse educational contexts while overcoming the current study limitations.

### 7. Conclusion

This study developed and evaluated a trajectory-based BiLSTM–DINA framework for predicting learner mastery from longitudinal academic performance. The framework integrated trajectory-based feature engineering, BiLSTM

prediction, DINA cognitive diagnosis, data augmentation, and explainability using SHAP. The analysis was conducted using longitudinal academic data from 1,549 learners across Mathematics, Science, and English.

The BiLSTM–DINA framework achieved 96.45% accuracy on the original dataset, 99.12% with SMOTE, and 95.80% with GAN. The SMOTE configuration produced the strongest results, including an F1-score of 0.9911, AUC of 0.9994, MCC of 0.9826, and Cohen's Kappa of 0.9824. The DINA component achieved an ACI of 0.7664, providing subject-level mastery information. SHAP analysis further identified Overall Mean, Mathematics Mean, Science Mean, and Learning Gain as the leading features associated with the predictive output.

The findings regarding the augmentation reveal that synthetic data did not yield a consistent improvement for prediction. The SMOTE configuration turned out to be the most effective one, whereas the GAN configuration was less predictive. Hence, it is important to assess augmentation based on its impact on the prediction problem.

In summary, the research outcomes illustrate that longitudinal modeling and cognitive diagnosis can be integrated to provide predictions about the mastery of learners and mastery information about subjects. This gives rise to an appropriate analytical framework in the context of learning analytics as well as future applications in adaptive education decision support systems. Nonetheless, the current research utilized only one longitudinal academic data set and a fairly simple three-skill DINA Q-matrix. Thus, it is necessary to consider the current findings within the scope of the current data set.

## REFERENCES

- [1] Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018b). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98–110. <https://doi.org/10.1016/j.chb.2018.07.027>
- [2] Baker, R. S. (2014). Educational Data Mining: An advance for intelligent systems in education. *IEEE Intelligent Systems*, 29(3), 78–82. <https://doi.org/10.1109/mis.2014.42>
- [3] Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & education*, 104, 18-33. <https://doi.org/10.1016/j.compedu.2016.10.001>
- [4] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, 10(3). <https://doi.org/10.1002/widm.1355>
- [5] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators?. *International journal of educational technology in higher education*, 16(1), 39. <https://doi.org/10.1186/s41239-019-0171-0>
- [6] Matcha, W., Gašević, D., Uzir, N. A. A., Jovanović, J., & Pardo, A. (2019, March). Analytics of learning strategies: Associations with academic performance and feedback. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 461-470). <https://doi.org/10.1145/3303772.3303787>
- [7] Jiao, P., Ouyang, F., Zhang, Q., & Alavi, A. H. (2022). Artificial intelligence-enabled prediction model of student academic performance in online engineering education. *Artificial Intelligence Review*, 55(8), 6321–6344. <https://doi.org/10.1007/s10462-022-10155-y>
- [8] Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., ... & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *Ieee Access*, 9, 7519-7539. <https://doi.org/10.1109/ACCESS.2021.3049446>
- [9] Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2016). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. *IEEE Transactions on Learning Technologies*, 10(1), 17-29. <https://doi.org/10.1109/TLT.2016.2616312>
- [10] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [11] De La Torre, J., & Chiu, C. (2015). A general method of Empirical Q-matrix validation. *Psychometrika*, 81(2), 253–273. <https://doi.org/10.1007/s11336-015-9467-8>
- [12] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1), 1-48. <https://doi.org/10.1186/s40537-019-0197-0>
- [13] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>

- [14] Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and education: artificial intelligence*, 3, 100074. <https://doi.org/10.1016/j.caeai.2022.100074>
- [15] Villegas-Espinoza, A. E., & Necochea-Chamorro, J. I. (2025). Using Deep Learning in Student Performance Prediction: A Systematic review. *TEM Journal*, 2472–2482. <https://doi.org/10.18421/tem143-51>
- [16] Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., & Hu, G. (2019). Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1), 100-115. <https://doi.org/10.1109/TKDE.2019.2924374>
- [17] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in neural information processing systems*. Association for Computing Machinery, 201-204. <https://doi.org/10.48550/arXiv.1506.05908>
- [18] De La Torre, J. (2008a). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115–130. <https://doi.org/10.3102/1076998607309474>
- [19] De La Torre, J. (2008b). An empirically based method of Q-Matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45(4), 343–362. <https://doi.org/10.1111/j.1745-3984.2008.00069.x>
- [20] Schuster, M., & Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681. <https://doi.org/10.1109/78.650093>
- [21] Kalita, E., Alfarwan, A. M., Aouifi, H. E., Kukkar, A., Hussain, S., Ali, T., & Gaftandzhieva, S. (2025). Predicting student academic performance using Bi-LSTM: a deep learning framework with SHAP-based interpretability and statistical validation. *Frontiers in Education*, 10. <https://doi.org/10.3389/feduc.2025.1581247>
- [22] Singh, T., Dutta, S., Vyas, S., & Rocha, Á. (2024). Explainable AI for education: Recent trends and challenges. In *Information systems engineering and management*. <https://doi.org/10.1007/978-3-031-72410-7>
- [23] Lu, Y., Rajendra, A. S., Zhang, J., & Zhao, T. (2026). Explainable machine learning for student performance prediction. *AI In Education*, 2(2), 17. <https://doi.org/10.3390/aieduc2020017>
- [24] Nikolenko, S., I. (2019). Synthetic data for deep learning. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.1909.11512>
- [25] Romero, C., & Ventura, S. (2010). Educational Data Mining: A review of the state of the art. *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)*, 40(6), 601–618. <https://doi.org/10.1109/tsmcc.2010.2053532>
- [26] Romero, C., & Ventura, S. (2012). Data mining in education. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, 3(1), 12–27. <https://doi.org/10.1002/widm.1075>
- [27] Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, 37, 13–49. <https://doi.org/10.1016/j.tele.2019.01.007>
- [28] Gašević, D., Dawson, S., & Siemens, G. (2014). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64–71. <https://doi.org/10.1007/s11528-014-0822-x>
- [29] Malik, S., Patro, S. G. K., Mahanty, C., Hegde, R., Naveed, Q. N., Lasisi, A., Buradi, A., Emma, A. F., & Kraiem, N. (2025). Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution. *Scientific Reports*, 15(1), 8738. <https://doi.org/10.1038/s41598-025-92324-x>
- [30] Xiong, Z., Li, H., Liu, Z., Chen, Z., Zhou, H., Rong, W., & Ouyang, Y. (2024). A review of Data mining in Personalized Education: Current trends and future Prospects. *Frontiers of Digital Education*, 1(1), 26–50. <https://doi.org/10.1007/s44366-024-0019-6>
- [31] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6), 602–610. <https://doi.org/10.1016/j.neunet.2005.06.042>
- [32] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- [33] Ahajjam, T., Moutaib, M., Aissa, H., Azrour, M., Farhaoui, Y., & Fattah, M. (2022). Predicting students' final performance using artificial neural networks. *Big Data Mining and Analytics*, 5(4), 294–301. <https://doi.org/10.26599/bdma.2021.9020030>
- [34] Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student Engagement Predictions in an e-Learning System and their impact on student course assessment scores. *Computational Intelligence and Neuroscience*, 2018, 1–21. <https://doi.org/10.1155/2018/6347186>

- [35] Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., Yin, Y., ... & Wang, S. (2020, April). Neural cognitive diagnosis for intelligent education systems. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 04, pp. 6153-6161). <https://doi.org/10.1609/aaai.v34i04.6080>
- [36] Ifenthaler, D., & Yau, J. Y. (2018). Utilising learning analytics for study success in higher education. *ASCILITE Publications*, 406–411. <https://doi.org/10.14742/apubs.2018.1958>
- [37] Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, 2, 100020. <https://doi.org/10.1016/j.caeai.2021.100020>
- [38] Matcha, W., Uzir, N. A. A., Gašević, D., & Pardo, A. (2019). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE transactions on learning technologies*, 13(2), 226-245. <https://doi.org/10.1109/TLT.2019.2916802>
- [39] Chen, X., Zou, D., Xie, H., & Wang, F. L. (2021). Past, present, and future of smart learning: a topic-based bibliometric analysis. *International Journal of Educational Technology in Higher Education*, 18(1), 2. <https://doi.org/10.1186/s41239-020-00239-6>
- [40] Susnjak, T. (2022). A Prescriptive Learning Analytics Framework: Beyond Predictive Modelling and onto Explainable AI with Prescriptive Analytics and ChatGPT. *arXiv* (Cornell University). <https://doi.org/10.48550/arXiv.2208.14582>
- [41] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786), 504–507. <https://doi.org/10.1126/science.1127647>
- [42] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- [43] Boujmiraz, S., Darhmaoui, H., & Maliani, A. D. E. (2026). Predicting student performance: A comprehensive review of machine learning, deep learning, and explainable AI approaches. *Computers and Education Artificial Intelligence*, 10, 100548. <https://doi.org/10.1016/j.caeai.2026.100548>
- [44] González-Pérez, L. I., García-Peñalvo, F. J., & Argüelles-Cruz, A. J. (2025). Data-Driven Learning Analytics and Artificial Intelligence in Higher Education: A Systematic review. *IEEE Revista Iberoamericana De Tecnologías Del Aprendizaje*, 20, 440–451. <https://doi.org/10.1109/rita.2025.3615512>
- [45] Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20(1). <https://doi.org/10.1186/s41239-023-00392-8>