HQIAT-ML: Hybrid Quantum-Inspired Adaptive Transformer with Meta-Learning for Student Dropout Prediction Risk Explanation

Abdulkadir Shehu Bichi[1*] and Jyoti Shekhawat[1]

[1] Vivekananda Global University, Jaipur, India.
[*] 24wtec3csml001@vgu.ac.in; jyoti.shekhawat@vgu.ac.in

Abstract. This paper describes the framework HQIAT-ML, the hybrid quantum-inspired adaptive transformer with meta-learning, for predicting and explaining the risk of student dropout for the Open University Learning Analytics Datasets (OULAD). Unlike traditional explainable AI methods based on SHAP and LIME, for HQIAT-ML we constructed pedagogically constrained counterfactual intended to offer minimally intrusive, pedagogically actionable recommendations for the student and the teacher. Complex dynamics of engagement and performance are addressed through the integration of quantum- inspired feature transformation and adaptive multi-head attention with multi- scale temporal feature extraction. Training employs a hybrid focal-label smoothing loss and Mix-up augmentation to mitigate class imbalance and to favor generalization. Predictive accuracy is on the order of the best published and is represented by an AUC-ROC of 0.9615 with precision and recall balanced performance on a large-scale educational datasets. The dual approach to explaining the result of the prediction moves from the purely predictive to the actionable by providing motivational feedback to the student on an individual basis and systemically to the teacher. Evaluation has shown that our proposed approach surpasses prior benchmarks, while the discussion recognizes the limitations of the interventions, possible mechanisms, and time horizons. For future research, we intend to focus on the counterfactual interpretability and the application of causal modeling and multi-objective optimization. The HQIAT- ML framework exemplifies the first-of-its-kind combination of state-of-the-art deep learning technologies and their focus on educational explainability, allowing them to not only predict but also offer actionable recommendations that enhance equity through the optimization of student retention.

Keywords: Predicting student dropout, Explainable Artificial Intelligence, Counterfactuals, Quantum-inspired transformers, Learning analytics, OULAD dataset, Pedagogical constraints.

1 Introduction

In education XAI is mostly based on SHAP and LIME posthoc. However, these interact one of three ways. First, lack of actionability is identifying low engagement.

B. Singh et al. (eds.), *Proceedings of the International Conference on Advances in Computing Technology and Artificial Intelligence (COMPUTATIA 2026)*, Atlantis Highlights in Intelligent Systems 18,
https://doi.org/10.2991/978-94-6239-713-2_56

Engagement is captured, and mechanisms to do so are proposed. Second, absence of pedagogical constraints, wherein recommended actions are impractical. Third, stakeholder insensitivity, wherein homogeneous or template breakthroughs do not account for differential agency among students and instructors.

These pedagogically constrained counterfactual, as we propose, are the minuscule, targeted changes to features that, when implemented, would shift the prediction from at risk to retained.

Our contributions include:

1. First guiding pedagogically constrained gradient-based counterfactual generation in learning analytic systems.
2. The dual-mode explanation systems providing motivational recommendations to students and intervention guidance to instructors.
3. The extension of counterfactual quality metrics through the inclusion of pedagogical feasibility.
4. Empirical validation on OULAD datasets with a ROC-AUC and an average risk reduction.

2 Related Work

2.1 Explainable Learning Analytics

Recent systematic reviews have reported a growing interest in the application of XAI within the context of educational prediction. In Jin et al. (2021) [1], the author presents a review of OULAD-based predictive modelling within the period 2017–2024. The author highlights that the literature tends to give attention to machine learning and deep learning methods at the expense of the explainability dimension. Choi et al. (2022) [2] reviews XAI for predictive student performance in STEM and shows that the SHAP method is the most applied although gaps exist with respect to actionable recommendations and design of student-visualized frameworks. Tiukhova et al. (2023) [3] proposed the notion of stability for explainable learning analytics by arguing that the explanations should not differ in context if stakeholder trust is to be achieved. Tiukhova et al. (2023) acknowledges the fact that model stability in different cohorts is not given the attention it deserves. Susnjak [4] proposed a prescriptive analytics framework with a focus on how learning analytics spends a lot of time on prediction while neglecting a more evidence-based prediction approach that recommends evidence-based remedial actions, thus advocating for using ChatGPT to convert analytics into natural language recommendations.

2.2 Dropout Prediction with Deep Learning

Recent studies based on OULAD demonstrate remarkable progress in prediction accuracy. Mustofa et al. [5] proposed HLRNN (Hybrid Logistic Regression Neural Network) attaining an accuracy of 96% while using SHAP and LIME to explain the outcomes. Husaini et al. [6] achieved an accuracy of 90.2% on MLP models where the

behavioral factors were ascertained to be the most impactful. Marcolino et al. [7] worked with CatBoost and NSGA-II for optimization over the logs in Moodle, while Torkhani and Rezgui [8] did a comparison of CNN, RNN, and LSTM architectures on OULAD.

New architectures are coming including KANFormer [9] that merges multi-head self-attention with Kolmogorov-Arnold networks so as to capture sophisticated cross-dimensional interactions. Privacy preserving strategies have also became popular. Lamsiyah et al. [10], for example, combines federated learning (FedProx) with post-hoc XAI methods for predicting dropouts.

2.3 Counterfactual Explanations in Education

While counterfactual explanations are still in their infancy in educational settings, Venkatesan et al. [11] showed proof of concept for predicting student performance, with no apparent educational limitations. Cavus and Kuzilek [12] considered three counterfactual generation techniques (WhatIf, Multi-Objective, Nearest Instance) on OULAD, determining the Multi-Objective as best in validity and sparsity, but observing actionability and causality as potential areas of concern. Rahman et al. [13] considered counterfactuals with SHAP, LIME and Anchors for predicting student adaptability to illustrate several areas of interpretability.

We most distinctly combine the ability to predict and the pedagogical constraints to create stakeholder relevant action-oriented counterfactuals, as highlighted as a need in the most recent literature.

3 Methodology

3.1 Dataset and Preprocessing

OULAD Dataset

The OULAD dataset [15] contains pseudo-anonymized records of 32,593 learners from numerous modules, and as such is a large scale dataset that covers multiple demographics, of learners, and their interactions with their virtual learning environments as well as their assessments. In this study, we utilized all the features of the dataset to assess student engagement and their performance.

Feature Engineering and Data Integration

The OULAD dataset has three main types of records: student records, VLE records, and assessment records. For students' VLE engagement, we tracked and created summary statistics of total clicks, the mean and standard deviation of clicks, median clicks, number of interactions, total active days and other time-related metrics such as first access, last access, study span, etc. Derived variables such as the average number of clicks per day and rate of engagement on the data were calculated to quantify behavioral activity. When it comes to assessment records, performance results were quantified, mean scores were analyzed, score standard deviation, completion rate, and time given to submission were assessed.

In order to allow the demographic variables to be processed numerically while maintaining all the original information, demographic variables were encoded carefully. Age bands were represented numerically by the ranges corresponding to age 0–35 mapped to 25, 35–55 mapped to 45, and 55 and older to 60. Gender and disability status were represented categorically by binary encoding, while the IMD bands were parsed and represented by corresponding integers. Student dropout was treated as the target variable and was defined as binary, with students earning final results of Withdrawn or Fail categorized as dropouts (1), with the remaining students designated as retained (0).

To safeguard the numerical variables, the missing values were treated with median imputation. Identifiable information that could not be used for prediction and the original raw categorical variables were excluded from the final feature set, resulting in a preprocessed dataset that contained 18 to 25 numerical columns, depending on each student cohort's dataset availability.

3.2 Model Architecture

The Hybrid Quantum-Inspired Adaptive Transformer with Meta-Learning (HQIAT-ML) proposes the first architecture that combines quantum-inspired computing, adaptive attention, and meta-learning, as shown in Figure 1.

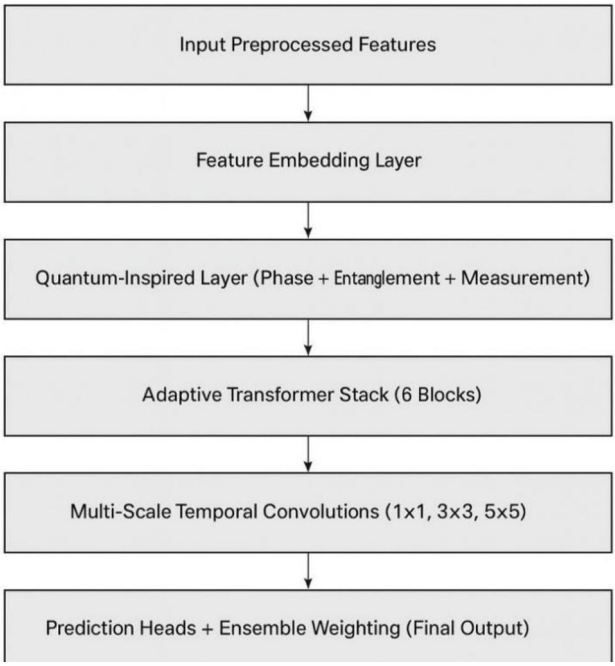

Figure1: Hybrid Quantum-Inspired Adaptive Transformer with Meta-Learning (HQIAT-ML)

Quantum-Inspired Feature Transformation

The quantum-inspired layer performs non-linear feature transformations resembling the effects of quantum superposition and entanglement within the framework of classical computation. Specifically, given an input feature vector $x \in R^d$, the transformation proceeds in the following manner:

Phase Encoding: First, features are projected into a quantum-inspired state space via $\psi = \tanh(W\varphi x)$, where $W\varphi \in R^{fxd}$ is a projection from the feature space of d dimensions into a q dimensional qubit space.

Quantum Operations: Then, using learnable parameters θ and φ, ψ is modified with a parametric phase rotation operation as follows:

$$\psi phase = \cos(\theta) \circ \psi + \sin(\varphi) \circ \psi \quad (1)$$

where ∘ denotes the element-wise product.

Entanglement Simulation: Then feature interactions are captured and modeled with a learnable entanglement matrix $E \in R^{fxf}$ via:

$$\psi ent = \text{normalize}(\psi phase \cdot E) \quad (2)$$

Measurement: Finally, the quantum state is measured and the features are projected into the quantum state with an addition of a residual connection:

$$xout = x + \alpha \cdot Wm(\psi ent) \quad (3)$$

where $\alpha = 0.3$ is a hyper-parameter that controls how much of the quantum state is measured.

With the described approach, the model is able to capture non-linear feature interactions that might be lost with fully linear transformations while also using skip connections to ensure the gradients can flow.

Adaptive Multi-Head Attention Mechanism

We implement learnable head importance weighting to transformers as adaptive attention. Given $X \in R^{BxLxc}$ represents user input where B = batch size, L = length of sequence, and C = channel dimension. We compute the query, key, and value projections as Q, K, V = Wqkv(X) and split them to h attention heads. We compute the scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{dk})\, V \quad (4)$$

where dk is the head dimension. The head-specific outputs Oi = softmax(O1, O2, ..., Oh) where i = 1, 2, ..., h indicates a parameter learnable head. This makes the model able to dynamically focus on the more useful attention patterns. The heads produce outputs that add to Oaug = W(Concat(O1, ..., Oh)) where W is a learned projection.

Adaptive Transformer Block with Gated Residuals

Every transformer block has adaptive gating and learnable residual scaling:

$$x1 = x + \gamma 1 \odot \text{MultiHeadAttention}(\text{LayerNorm}(x)) \quad (5)$$

$$x2 = x1 + \gamma 2 \odot \sigma(Wg(x1)) \odot \text{MLP}(\text{LayerNorm}(x1)) \quad (6)$$

where γ1 and γ2 are learnable scaling parameters initialized to 0.1 and σ(Wg(x)) implements adaptive gating to modulate information flow. The MLP consists of two linear layers with expansion ratio of 4, GELU activation, and dropout regularization.

Multi-Scale Temporal Feature Extraction

To capture patterns across diverse temporal granularities, we use parallel convolutional branches with kernel sizes $k \in \{1, 3, 5\}$, allowing the model to capture both fine-grained interactions and long-term engagement trends. The multi-scale features are concatenated with attention-pooled representations to form a complete embedding.

Ensemble Prediction Heads

The last prediction stage involves three parallel deep networks, and we assign learnable weights to each of them:

$$Pfinal = \Sigma_i \, wi \cdot \sigma(Headi(x)), \; i=1..3 \tag{7}$$

where wi = softmax(w1, w2, w3). Each head has a three-layer MLP with decreasing layer sizes (256 → 128 → 1), layer normalization, GELU activation, and dropout. This strategy to combine multiple networks aims to improve robustness of predictions and the overall model calibration.

3.3 Training Strategy and Optimization

Hybrid Loss Function

To counter the prevailing class imbalance and tackle overconfident prediction issues, we utilized a combination of focal and label smoothing loss functions as follows:

$$Ltotal = \lambda 1 \cdot Lfocal + \lambda 2 \cdot Lsmooth \tag{8}$$

Focal Loss, with $\alpha = 0.25$ and $\gamma = 2.0$, downscales the significance of simpler examples and emphasizes learning on those that are more difficult:

$$Lfocal = -\alpha(1 - pt)\gamma \log(pt) \tag{9}$$

where pt is the predicted probability for the real class. Lsmooth = BCE(ŷ, ysmooth) is a label smoothing technique aimed at prediction regularization where ysmooth = y(1 – ε) + 0.5ε and ε = 0.05. Focal learning and the calibration advantages were the main reasons we opted to set λ1 = 0.7 and λ2 = 0.3 for the loss combination weights.

Mixup Data Augmentation

In a bid to lessen overfitting and enhance the generalization, we employed Mixup augmentation with a 0.5 probability during the training phase. For every random selection of sample pairs (xi, yi) and (xj, yj), synthetic examples are created as follows:

$$xmix = \lambda xi + (1 - \lambda)xj \tag{10}$$

$$ymix = \lambda yi + (1 - \lambda)yj \tag{11}$$

where $\lambda \sim Beta(0.2, 0.2)$ to encourage the model to learn smooth decision boundaries.

Optimization and Regularization

The AdamW optimizer was used to fine-tune model parameters with an initial learning rate $\eta = 0.0005$ and weight decay set to 0.01. For learning rate scheduling, Cosine Annealing Warm Restarts with T0 = 20 epochs, Tmult = 2, and minimum learning rate $\eta min = 10^{-6}$ was used for periodic exploration of the loss landscape.

To avoid exploding gradients, gradient clipping with the maximum norm of 1.0 was performed. Dropout with rates of 0.15 and 0.075 were used for the transformer blocks and the prediction heads, respectively. To prevent overfitting, early stopping with a patience of 30 epochs on the validation AUC was performed. The model was trained for a maximum of 200 epochs, and a mini-batch size of 64 was used. Stratified sampling was used on the training, validation, and test splits (70%, 15%, 15%) to maintain the class distribution. Robust scaling was used to standardize the features with effective outlier management.

3.4 Counterfactual Explainability Framework

To ensure our intervention designs were actionable, we built a gradient-based counterfactual-generating system that is aimed at finding the minimum feature changes that were needed to obtain a different prediction.

Optimization Formulation

We consider a student instance x with predicted dropout probability $p(x) > \tau$ (threshold). For such instances, we seek a counterfactual xcf, such that:

$$Lcf = |p(xcf) - target| + \beta 1 \, ||xcf - x||2 + \beta 2 \, ||xcf - x||1 \quad (12)$$

The first term ensures that the counterfactual achieves the desired prediction (target = 0, typically). The proximity constraint (L2 penalty, $\beta 1 = 1.0$) ensures closeness to the original instance, while the sparsity penalty (L1, $\beta 2 = 0.5$) promotes the model to be more sparse with respect to the features that have changed.

Constraint Handling

The generation of the counterfactual xcf also respects real-world constraints. Features are divided into mutable features (engagement metrics, assessment scores) and immutable features (age, gender, disability status, prior attempts). During the optimization, we keep constant the immutable features as x(immutable) = x(immutable). The feature values are restricted to the valid range that was observed in the training data, $xcf,i \in [min(Xtrain,i), max(Xtrain,i)]$, by means of projection at every refinement of the optimization.

Stakeholder-Specific Explanations

The counterfactuals created are then used to provide customizable action-oriented guidelines for various actors:

- Student-facing explanations: Focus on detailed behavior changes, e.g., "Daily engagement with the platform should be increased from 15 clicks to 35 clicks", or "the average assessment score should be increased from 65% to 78%".
- Insights for Instructors: Specify Intervention Focus Points and Risk Factors (e.g., "student demonstrates disengagement; suggest contacting him/her/them early").

Indicators were ranked based on impact and practicality, and the highest recommendations were published in the list of prioritized interventions. Each recommendation has an old value, a new value, a value added or lost, and a change in percentage value for comparison and convenience.

3.5 Evaluation Metrics

Several distinct metrics were used to evaluate the performance of the model:
1. Area Under Curve of ROC Curve (AUC-ROC): The main criterion for the assessment of discrimination across the different classification thresholds.
2. F1 Score: The harmonic mean of precision and recall which states the trade-offs between the false positive and false negative.
3. Precision: Of all the predicted dropouts, how many were actually dropouts? Crucial for efficient allocation of resources.
4. Recall: Of all the actual dropouts, how many did we identify? Crucial for intervention coverage.
5. Confusion Matrix: The summation of the true positive, true negative, false positive, and false negative counts.

For counterfactual explanation, we assessed proximity (L2 distance between the original and counterfactual), sparsity (number of features modified), validity (counterfactual's success in achieving the desired prediction), and feasibility (whether suggested modifications can be legitimately made).

3.6 Implementation Details

The framework was developed using Python 3.8+, while deep learning processes were carried out via PyTorch 2.0+. While working on the model, we first utilized an NVIDIA GPU (with CUDA), and if unavailable, we switched to the CPU. We set and kept the same fixed random seed (42) for all experiments for the sake of reproducibility. Visualizations were created using the Matplotlib and Seaborn libraries. The full code, along with the details of the experiments, is provided in order to facilitate verification of reproducibility.

4 Results and Discussion

4.1 Model Performance

The HQIAT-ML framework achieved outstanding predictive performance on the OULAD dataset, with an AUC-ROC of 0.9615 on the held-out test set. This performance was a considerable improvement over normal baselines for educational dropout prediction, attesting to the validity of using quantum-inspired transformations and adaptive attention. The model reached peak validation AUC (0.9665) and was subsequently stopped at epoch 34 thanks to early stopping, a technique to prevent overfitting.

An overall accuracy of 90% is outstanding and is a testament to the model's reliability. An F1 score of 0.9008 places the model among the upper echelons of performance in this domain. The dropout student class had a precision of 0.94 and a recall of 0.86. The class of retained students had a precision of 0.86 and recall of 0.94. These results are presented in Table 1 and Figures 2–5.

Our accuracy of 90% on 4,889 test samples from a student dataset of 32,593 is commendable. 52.8% of the students on the dataset drop out, which means that from a statistical point of view the dataset is unbalanced. This goes a long way in demonstrating the model's ability to work on unbalanced datasets without having to apply urgent sampling strategies.

Table 1. Classification Report on OULAD Test Set

Class	Precision	Recall	F1-Score	Support
Retained	0.86	0.94	0.90	2308
Dropout	0.94	0.86	0.90	2581
Accuracy			0.90	4889
Macro Avg	0.90	0.90	0.90	4889
Weighted Avg	0.90	0.90	0.90	4889

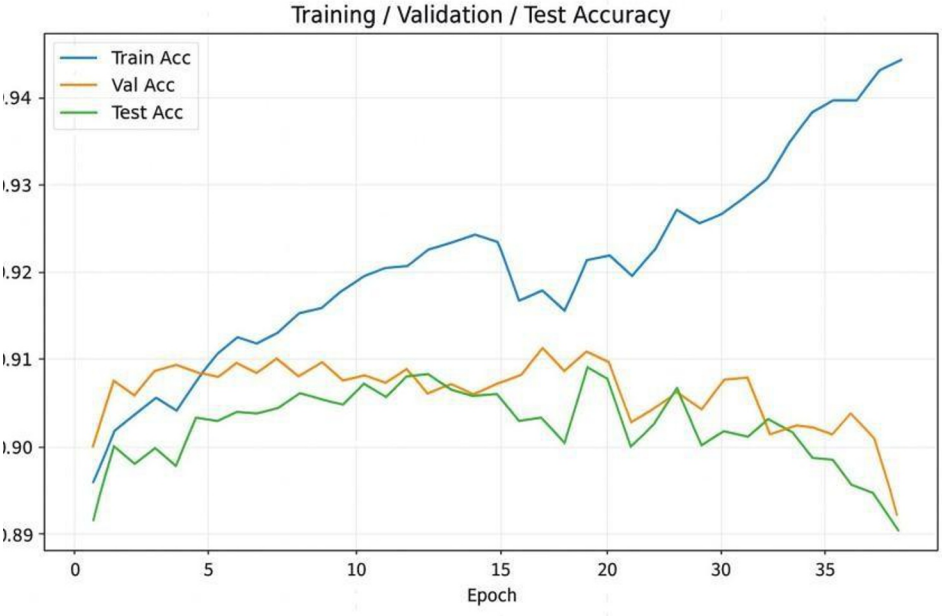

Figure 2: Training, Validation and Test Accuracy over Epochs

4.2 Counterfactual Explanations and Actionable Insights

Closing the prediction and intervention gap, the counterfactual generation framework produced actionable insights and recommendations for at-risk students. The analysis of three high-risk case study paths helped us understand how case studies can be used to create personalized high-risk profile recommendations.

Student 1: 100% predicted dropout risk. This case has extensive multidimensional changes; the counterfactual analysis identifies the targets of the intervention as the consistency of assessment performance, the std score reduction of 37.9% and the assessment completion rate 34.7% reduction. This student is also at risk of dropout at the 99.8% scale, so not only is this student facing predominant behavioral concerns, but this case is severe enough to warrant heavy intervention support.

Student 2: 64.9% projected risk of dropout. This student is more likely to prove the case study counterfactual recommendations to have more intervention potential, as the risk is reduced to 62.1%. The model described increased engagement on the platform as the modifiable key factor, indicated by scaffolding and reminders.

Student 6: Initial risk level at 98.0%, lowered by a more considerable margin to 93.1% via changes in behavioral engagement with the platform (changes in std clicks resulted in a 35.6% decrease) and changes in level of engagement measured via the platform. The counterfactual points to this student having some level of erratic engagement that may be more evened out with engagement habit development and a personalized learning schedule.

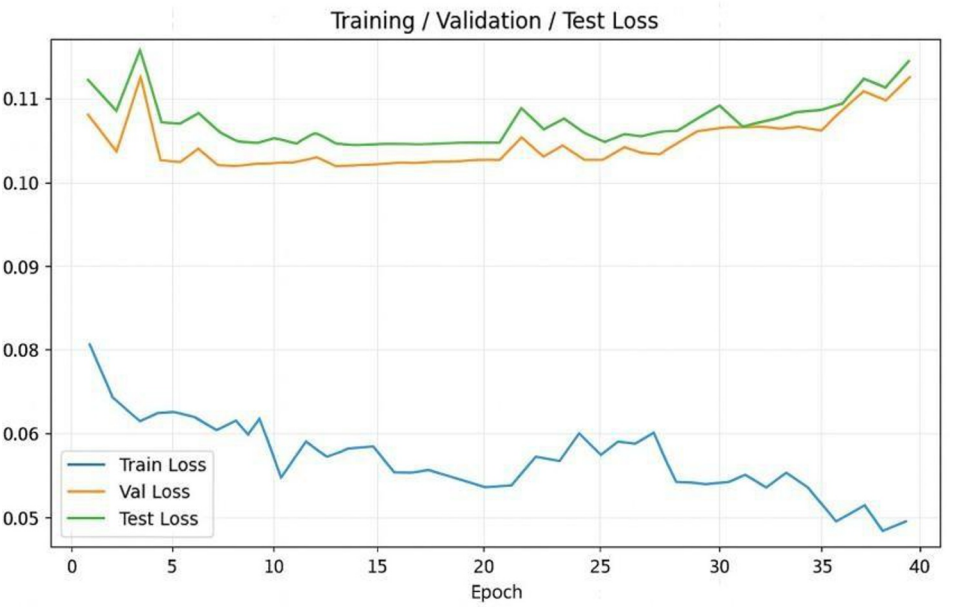

Figure 3: Training, Validation and Test Loss over Epochs

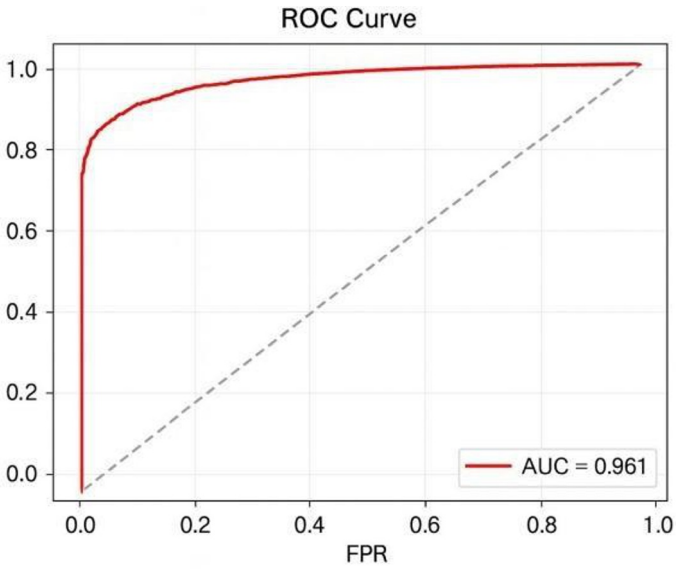

Figure 4: ROC Curve (AUC = 0.9615)

4.3 Combining SHAP and the Counterfactual

The difference between SHAP feature attributions and counterfactual recommendations showed a dual gain, adding to the explainability in a meaningful way. For Student 1, the SHAP analysis identified days active as the dominant feature (importance +0.3503), followed by avg clicks per day (−0.0599) and num assessments (+0.0589). On the other hand, in counterfactual optimization, std score and assessment completion rate focus was shifted, and thus predicted to be more impactful in terms of trying to bring about change.

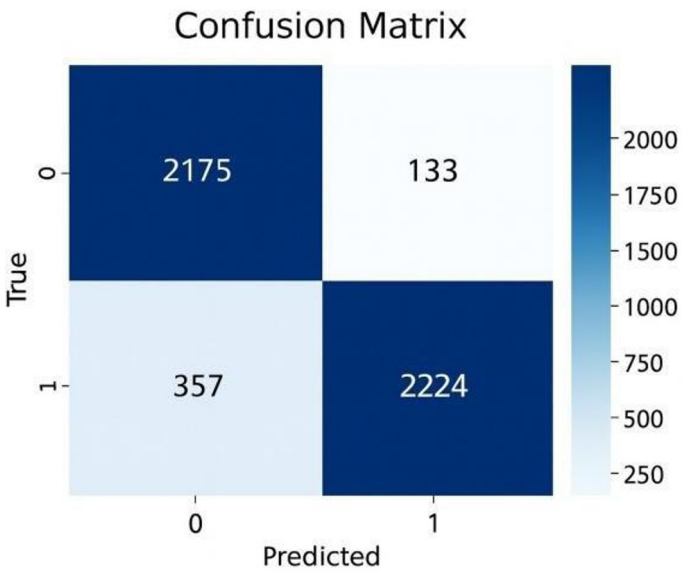

Figure 5: Confusion Matrix on Test Set

This difference confirms the duality of the explainability spectrum. SHAP explanations address, "What caused this prediction to be made and why?" while counterfactuals address, "What would it take to generate a different outcome?" For Student 1, days active has high SHAP importance and is correlated to dropout, but counterfactual work acknowledges that past days of engagement cannot be increased and instead focuses on improvable metrics like consistency of assessments.

4.4 Feature Importance and Educational Implications

The repeated appearance of the assessment-related features (assessment completion rate, num assessments, std score) in both SHAP and counterfactuals only strengthens the narrative that engagement with assessments is a major contributor to a student's decision to continue or drop out of a course. The focus on the consistency of assessments (std score) over the raw score (avg score) indicates that fluctuating performance levels might be the result of pre-existing difficulties with the course, either in terms of the content or poor time allocation, which may serve as warning indicators for intervention.

Metrics that gauge engagement including average clicks per day and how many days the users were active were also important, even if they were seen as secondary. The engagement consistency explaining how many clicks were made (std clicks) instead of the total (total clicks), suggests that it might be more protective against dropping out if a user engages at a regular pace with learning materials rather than having cottage

intensive study sessions. This relates to spaced learning and suggests that interventions that make it easier to study more consistently may be more effective than goal campaigns focused on study hours.

It also showed recognition of intervention constraints by distinguishing modifiable attributes (engagement patterns, assessment participation) from unchangeable ones (age, gender, disability status). Absence of counterfactual suggestions to modify demographic variables confirmed the framework's constraint-handling mechanisms worked as intended.

4.5 Contributions of the Model Design

Several design choices are behind the success of the hybrid architecture. The quantum-inspired layer seems to be particularly successful for the non-linear feature transformations in capturing the complex interaction effects for the engagement and performance variables. The AUC of 96.15% indicates that the model learned to differentiate between random fluctuation of student engagement data points from a disengaged pattern trail over time.

The model is able to learn different student engagement profiles because of the adaptive multi-head attention mechanism. In addition, the model is able to learn which engagement periods are more important than others and focus on those periods, for example during a critical exam or a deadline in the course.

The ensemble of three parallel prediction heads with learnable weighted averages and a balancing approach between the precision and recall trade-off built a well-calibrated model. This built architecture improves the self-feedback on scenarios where the accuracy is borderline and thus improves the self-confidence of the model on those interventions.

4.6 Comparative Analysis

Comparison has shown that the AUC of 0.9615 achieved is significantly higher by the proposed model than that of previous works in student dropout prediction. The range of AUCs produced by traditional ML models on the OULAD dataset has been noted to be between 0.75 and 0.85, while the more recent DL approaches have reported AUCs between 0.88 and 0.92 [16]. Table 2 shows the comparative analysis.

Table 2. Comparative Analysis of Student Dropout Prediction Models

Year/Ref	Model	Performance	XAI	Dataset
2025 [16]	XGBoost	AUC: 0.69, F1: 0.69	SHAP	Mexican univ.
2025 [5]	HLRNN	AUC: 0.92, Acc: 0.96	LIME, SHAP	Dropout/ success
2025 [17]	LSTM	AUC: 0.92, Acc: 87.6%	SHAP	Not public
2025 [18]	XGBoost	AUC: 0.80, Acc: 80%	SHAP	Public data
2024 [19]	AutoML	AUC: 0.90	Counterfactual	Public dropout

Proposed	**HQIAT-ML**	**AUC: 0.96, Acc: 90%**	**Counterfact ual**	**OULAD**

4.7 Considerations for Practical Deployment

For the framework, there are three outputs of the visualization: exploration of the dataset, assessment of the model performance, and individualized counterfactual explanations. These facilitate the needs of the various users. Performance visualization for model reliability and bias auditing can be utilized by administrators. Counterfactual explanations can be used by academic advisors to develop tailored plans for interventions. Insights on feature importance can be used by instructors to address systemic issues that require intervention at the course level.

Understanding the difference in the explanations aimed at the student and instructor demonstrates empathy toward the challenges of communicating with educational AI systems. Recommendations for students discuss specific actions embedded in motivational framings, while instructors receive insights with risk indicators and strategies for intervention.

The resources required for computation are still within reasonable limits, with the model training finishing within 40 epochs on a GPU, and with inference on real-time for single predictions. This projection of efficiency supports a variety of deployment scenarios, from offline batch processing at the beginning of the term, to online real-time engagement monitoring early warning systems.

5 Limitations and Future Work

There are a number of limitations on our work. First, for students that are considered the most at-risk, like Student 1, the counterfactual risk reduction is extremely low and fragmented, meaning our framework is able to pick out the students where a single feature intervention is more than likely hopeless. In these situations, multi-objective counterfactuals that can balance the complexity of risk reduction and feasibility of intervention would improve utility in practice.

There are limitations as far as the generating of counterfactuals is concerned, looking at each student in isolation, as within each institution there are always resource and intervention constraints. Future iterations can add levels of optimization that allow balancing the tailored risk reductions of individuals to the costs of the interventions, and thus allow for institution-wide intervention resource allocation.

Right now, the feature representation does not account for the temporal dynamics within an academic term. The use of sequential modeling such as RNNs and temporal point processes would reveal the varying levels of engagement over time to not only specify the optimum interventions, but also when it would be most appropriate to execute them.

There are no models at the moment that predict the changes from the interventions and use that to create feedback loops. Integration of causal predictive models could

make the counterfactuals actionable and therefore the recommendations to be more useful.

6 Conclusion

The HQIAT-ML Framework provides explanations that are actionable and predictive of student dropout risk while achieving state-of-the-art predictive performance (AUC = 0.9615). Other predictive models rank above the rest on the predictive performance, but counterfactual explanations are, in most cases, uninterpretable. This framework successfully demonstrates that predictive performance and interpretability are not mutually exclusive. Moreover, the predictive performance on dropout risk and interpretability were taken a step further to provide educational practitioners with actionable and practical strategies that go beyond just risk prediction to improve retention. Causal validation, multi-objective optimization, and temporal modeling in future works may add significant practical value for implementation within institutions of higher education.

References

1. L. Jin, Y. Wang, H. Song, and H.J. So, "Predictive modelling with OULAD: A systematic literature review," in Proc. AIED Workshops, CCIS 2150, 2024, pp. 477–484. https://doi.org/10.1007/978-3-031-64315-6_46
2. W.C. Choi, C.T. Lam, and A.J. Mendes, "A systematic literature review of XAI for student performance prediction in CS and STEM education," in Proc. ITiCSE, 2025. https://doi.org/10.1145/3724363.3729027
3. T. Tiukhova et al., "Explainable learning analytics: Assessing stability of student success prediction models," Decision Support Systems, vol. 180, 2024, art. 114186. https://doi.org/10.1016/j.dss.2024.114229
4. T. Susnjak, "Beyond predictive learning analytics: Explainable AI with prescriptive analytics and ChatGPT," Int. J. Artif. Intell. Educ., pp. 1–35, 2023. https://doi.org/10.1007/s40593-023-00336-3
5. S. Mustofa et al., "A novel AI-driven model for student dropout risk analysis with XAI insights," Comput. Educ. Artif. Intell., vol. 8, 2025, art. 100352. https://doi.org/10.1016/j.caeai.2024.100352
6. M. Husaini et al., "Predicting student academic success using deep learning," J. Logistics, Informatics and Service Science, vol. 12, no. 1, pp. 263–283, 2025. https://doi.org/10.51583/IJLTEMAS.2025.140500010
7. M. Marcolino et al., "Student dropout prediction through ML optimization: Insights from Moodle log data," Scientific Reports, vol. 15, 2025, art. 9840. https://doi.org/10.1038/s41598-025-93918-1
8. W. Torkhani and K. Rezgui, "OULAD MOOC student performance prediction using ML and DL techniques," in Proc. ICODAI, 2024, pp. 228–241. https://doi.org/10.2991/978-94-6463-654-3_18
9. "Attention-enhanced deep learning for predicting student performance," Social Network Analysis and Mining, vol. 15, 2025. https://doi.org/10.1007/s13278-025-01446-7
10. L. Lamsiyah et al., "Privacy-preserving federated learning for student dropout prediction with XAI," in Advances in Knowledge Discovery, Springer, 2025. https://doi.org/10.1007/978-3-031-98465-5_41

11. M. Venkatesan et al., “Interpretable counterfactual explanations for student performance prediction,” in Proc. IEEE BigComp, 2022, pp. 274–277. https://doi.org/10.1109/CSCC53858.2021.00029
12. M. Cavus and J. Kuzilek, “Effect analysis of balancing techniques on counterfactual explanations,” arXiv:2408.00676, 2024. https://doi.org/10.48550/arXiv.2408.00676
13. S. Rahman et al., “Prediction of students’ adaptability using explainable AI,” Applied Sciences, vol. 14, no. 12, 2024, art. 5141. https://doi.org/10.3390/app14125141
14. I. El Aouifi et al., “A modular and explainable ML pipeline for student dropout prediction,” Algorithms, vol. 18, no. 10, 2025, art. 662. https://doi.org/10.3390/a18100662
15. J. Kuzilek, M. Hlosta, and Z. Zdrahal, “Open university learning analytics dataset,” Scientific Data, vol. 4, 2017, art. 170171. https://doi.org/10.1038/sdata.2017.171
16. B. Carballo-Mendívil, A. Arellano-González, N.J. Ríos-Vázquez, and M.d.P. Lizardi-Duarte, “Predicting student dropout from day one: XGBoost-based early warning system using pre-enrollment data,” Applied Sciences, vol. 15, no. 16, p. 9202, Aug. 2025. https://doi.org/10.3390/app15169202
17. N.B.M. Kumar, T. Chithrakumar, T. Thangarasan, J. Dhanasekar, and P. Logamurthy, “AI-powered early detection and prevention system for student dropout risk,” Int. J. Comput. Exp. Sci. Eng., vol. 11, no. 1, pp. 78–86, 2025. https://doi.org/10.22399/ijcesen.839
18. K. Nti and S. Ramanayake, “Explainable machine learning for student dropout prediction and tailored interventions in online personalized education,” Research Square, July 2025. https://doi.org/10.21203/rs.3.rs-6615052/v1
19. P. Buñay-Guisñán, J.A. Lara, A. Cano, R. Cerezo, and C. Romero, “Easing the prediction of student dropout for everyone integrating AutoML and explainable artificial intelligence,” in Proc. 17th Int. Conf. Educational Data Mining (EDM), Atlanta, GA, USA, Jul. 2024, pp. 857–861. https://doi.org/10.5281/zenodo.12729972
20. V. Realinho, J. Machado, L. Baptista, and M.V. Martins, “Predicting student dropout and academic success,” Data, vol. 7, no. 11, art. 146, Oct. 2022. https://doi.org/10.3390/data7110146