

Comparative Analysis of Classical Machine Learning Models and a Variational Quantum Classifier for Student Placement Prediction

Prof. Dr. Ranga Rao Velamala^{1*}, Dr. Pothuraju V V Satyanarayana², V. Lahari³, Y. Srujana³, G. Jnaneswara Rao³, K. Vamsi Krishna³

¹Professor Department of Computer Science and Engineering - Data Science Visakha Institute of Engineering and Technology (Autonomous), Visakhapatnam, Andhra Pradesh, India

²Associate Professor Department of Computer Science and Engineering - Data Science Visakha Institute of Engineering and Technology (Autonomous), Visakhapatnam, Andhra Pradesh, India

³Student Department of Computer Science and Engineering - Data Science Visakha Institute of Engineering and Technology (Autonomous), Visakhapatnam, Andhra Pradesh, India

Abstract:

Campus placement prediction is an established problem in educational data mining with direct relevance to institutional resource allocation and targeted student support. This study contributes a rigorous, reproducible benchmarking protocol and a cautionary empirical case study on the application of near-term quantum machine learning (QML) to structured academic data. Five classical supervised classifiers--Logistic Regression, Support Vector Machine (SVM), Random Forest, Multi-Layer Perceptron (MLP), and XGBoost--are compared against a Variational Quantum Classifier (VQC) on a 500-record, single-cohort dataset comprising CGPA, academic backlogs, attendance, skill assessments, and extracurricular engagement, with a binary placement outcome and moderate class imbalance (67% positive prevalence). A unified, leakage-free preprocessing pipeline (imputation, standardisation, stratified 80:20 partitioning, and ensemble-based feature selection to eight features applied within training folds) was maintained across all models. Classical models were optimised via stratified five-fold GridSearchCV; the VQC used RY-gate angle embedding, a depth-3 ansatz, and SPSA optimisation on the Qiskit Aer statevector simulator. Macro-averaged F1 and Matthews Correlation Coefficient (MCC) are the primary metrics given class imbalance; accuracy is supplementary. Under stratified five-fold cross-validation, tuned XGBoost attained mean accuracy 0.988 ± 0.007 (macro-F1 = 0.987; MCC = 0.974) and Random Forest reached 0.963 ± 0.011 (MCC = 0.916). The VQC yielded a corrected accuracy of 0.220 ± 0.030 (macro-F1 = 0.196; MCC = -0.083), 45 percentage points below the majority-class baseline. Gradient variance monitoring over SPSA iterations confirms that this failure arises from barren plateau dynamics rather than initialisation-specific stagnation. A complete 21-pair McNemar's test matrix (all C(7,2) pairwise comparisons across six classifiers), with Bonferroni-adjusted thresholds, documents statistically significant performance differences between ensemble and non-ensemble classifiers. The experimental pipeline and all code will be made available via a persistent repository (Zenodo) upon acceptance of this manuscript.

Keywords: Variational Quantum Classifier, Student Placement Prediction, Educational Data Mining, XGBoost, Barren Plateau; McNemar's Test, Matthews Correlation Coefficient, Cross-Validation, NISQ..

1. Introduction

Campus placement prediction is an established problem in educational data science. Institutional administrators use placement forecasts to allocate counselling resources efficiently, design targeted skill-development programmes, and measure curriculum effectiveness through graduate employability outcomes. The widespread digitisation of academic records--transcripts, attendance logs, extracurricular

participation, and soft-skill assessments--has made supervised binary classification a natural methodological choice [1], [2].

Classical machine learning algorithms have achieved strong results on this task. XGBoost and Random Forest are particularly well-suited to structured tabular data, owing to their efficient split-finding algorithms, second-order loss approximation, and implicit regularisation through column and row subsampling [3]-[5]. Grinsztajn et al. [14] demonstrated across 45 public benchmarks that tree-based ensembles

consistently outperform deep architectures on tabular data, a result attributed to the alignment between recursive binary splitting and the irregular, low-smoothness functions typical of such domains [15]. These advantages are especially pronounced when predictive signal concentrates in a small number of features, as is the case in placement data where CGPA and academic backlogs dominate.

Variational Quantum Circuits (VQCs) represent the most experimentally accessible paradigm for Noisy Intermediate-Scale Quantum (NISQ) devices [7]-[10]. Despite theoretical arguments for potential quantum advantage in high-dimensional feature spaces [7], [9], empirical evidence for superiority over competitive classical baselines on real-world tabular data remains absent. Rather than framing the present study as a test of quantum superiority--for which no theoretical basis exists when a shallow, product-state VQC is pitted against a gradient-boosted ensemble on a 500-row dataset--this work is positioned as a benchmarking study of NISQ-era VQC feasibility and failure modes on a small-scale, real-world educational classification task. The result is a negative one, and negative results of this kind serve an important calibration function for the QML literature.

A key limitation acknowledged throughout is that the dataset comprises 500 records from a single institutional cohort from a public repository; all findings are scoped as proof-of-concept benchmarking and should not be interpreted as generalisable claims. The Wilson score confidence intervals reported reflect sampling variance within this cohort, not variance across institutions or student populations. Multi-institutional validation is a prerequisite for deployment. The principal contributions are: (i) stratified five-fold cross-validated accuracy, macro-F1, and MCC with Wilson score 95% confidence intervals; (ii) a complete 21-pair McNemar's significance test matrix (all C(7,2) pairwise comparisons) with multiple-comparison context; (iii) strict train-only feature selection within each cross-validation fold; (iv) gradient variance monitoring over SPSA iterations to distinguish barren plateau dynamics from constant-output collapse; (v) mutual information validation of feature importance rankings; (vi) asymmetric misclassification cost guidance for institutional deployment; and (vii) an explicit negative-result contribution calibrating expectations for NISQ-era QML on small-scale tabular educational data.

2. Related Work

The application of statistical and computational methods to student records for early warning and outcome prediction has been extensively surveyed [1], [2]. Baker and Siemens [1] synthesised techniques for extracting actionable knowledge from learner data, while Romero and Ventura [2] surveyed methodological advances over the subsequent decade. Campus placement studies consistently identify CGPA and the number of failed courses as the most discriminative features [11], [18]. Many public placement datasets, including Kaggle-sourced collections, are small ($n < 1,000$) and may represent synthetic or anonymised cohorts; the present dataset ($n = 500$; 67% positive prevalence; six original features; single institutional cohort; deposited 2022) is consistent with this profile, and results should be interpreted accordingly.

Breiman [3] established Random Forests as a robust bagging ensemble; Friedman [4] formalised gradient boosting; and Chen and Guestrin [5] extended this into XGBoost with second-order Taylor approximation and distributed computation. Grinsztajn et al. [14] and Shwartz-Ziv and Armon [15] provide complementary accounts of why tree-based methods retain advantages on tabular data with heterogeneous, irregular, and uninformative features. Schmidhuber [6] provides the canonical overview of deep feedforward networks; the MLP serves as a representative shallow neural baseline.

In QML, Biamonte et al. [7] synthesised the landscape of quantum algorithms for machine learning, and Schuld and Petruccione [8] provided a comprehensive treatment of supervised quantum classifiers. Havlíček et al. [9] demonstrated a quantum kernel classifier on IBM hardware; Benedetti et al. [10] characterised the representational capacity of parameterised circuits. The fundamental obstacle to training deep VQCs is the barren plateau phenomenon formalised by McClean et al. [16]: for randomly initialised circuits, the variance of the loss gradient with respect to any single parameter decreases exponentially with circuit width ($\text{Var}[\partial L/\partial \theta_i] \propto 2^{-n}$), rendering optimisation progressively intractable. Mitigation strategies include layerwise training, local cost functions, and identity-block initialisation [16]; the last of these is tested in the present study. The VQC implemented here--product-state angle embedding on eight qubits with a depth-3 linear ansatz and SPSA optimisation--is a standard NISQ-era baseline circuit. Several prior studies have confirmed that comparable configurations struggle on classical tabular benchmarks; the present study replicates and extends this finding in the context of student placement data, and provides the gradient variance evidence base that distinguishes it from a routine negative replication.

3. Methodology

3.1 Dataset, Provenance, and Preprocessing

The dataset comprises 500 student records from a publicly accessible Kaggle repository [11] (title: "Student Placement Prediction," Version 1; deposited 2022; URL: <https://www.kaggle.com/datasets/student-placement-prediction>; accessed March 2024). Each record encodes six features--CGPA, number of academic backlogs, attendance rate, technical skill assessment score, communication skill score, and extracurricular participation status--alongside a binary placement outcome. The positive-class prevalence is approximately 67% (335 placed; 165 not placed). The dataset represents a single institutional cohort; its characteristics--small size, single source, moderate imbalance, and feature set dominated by CGPA and backlogs--are typical of publicly available placement datasets and suggest a possible synthetic or anonymised origin, which further constrains generalisability. All findings are accordingly scoped as proof-of-concept benchmarking.

The following preprocessing steps were applied uniformly to all models: (i) median imputation for continuous missing values and mode imputation for categorical missing values; (ii) one-hot encoding of nominal categorical variables; (iii) zero-mean unit-variance standardisation using StandardScaler (scikit-learn 1.3.0 [12]), fit exclusively on training data and applied identically to test data; (iv) stratified 80:20 train-test partitioning (random seed 42), with the held-out 100-record partition reserved exclusively for McNemar's testing and final

evaluation; and (v) ensemble-based feature selection retaining the eight highest-ranked features by mean-normalised importance averaged arithmetically from Random Forest and XGBoost importance vectors. Step (v) was performed exclusively on the training partition at each outer fold in cross-validation.

The number eight was dictated by the VQC's qubit count. This constitutes hardware-driven rather than problem-driven feature selection and introduces a confounding variable: the VQC was constrained to the same feature set as the classical models, which was selected by and optimised for tree-based models. To assess whether this reduction materially affected classical performance, all five classical classifiers were additionally evaluated on the full six-feature set (after encoding, eight binary/continuous columns). The results were within one standard deviation of the eight-feature cross-validated estimates for all models, confirming that the forced reduction did not materially disadvantage the classical evaluation. The tree-model circularity in feature selection was further addressed by validating rankings against mutual information scores (scikit-learn's `mutual_info_classif`, training partition only, k-NN estimator for continuous variables): CGPA ranked first (MI = 0.341) and backlogs ranked second (MI = 0.278), consistent with the ensemble-derived ordering.

3.2 System Architecture

Figure 1 depicts the end-to-end hybrid quantum-classical evaluation framework, organised into five sequential phases: (1) Data Ingestion; (2) Preprocessing, with training-only feature selection at each fold; (3) Model Training, with classical and quantum branches operating in parallel; (4) Evaluation, computing stratified five-fold cross-validated and held-out test metrics; and (5) Output Synthesis, aggregating statistics, confidence intervals, and the McNemar's test matrix. The classical and quantum branches differ structurally: the classical branch applies GridSearchCV-based hyperparameter tuning and k-fold cross-validation; the quantum branch applies SPSA-based parameter optimisation with three random initialisations on the fixed train-test split, without k-fold cross-validation, owing to the computational cost of statevector simulation. This asymmetry is an acknowledged limitation of the evaluation and is explicitly noted in the figure caption.

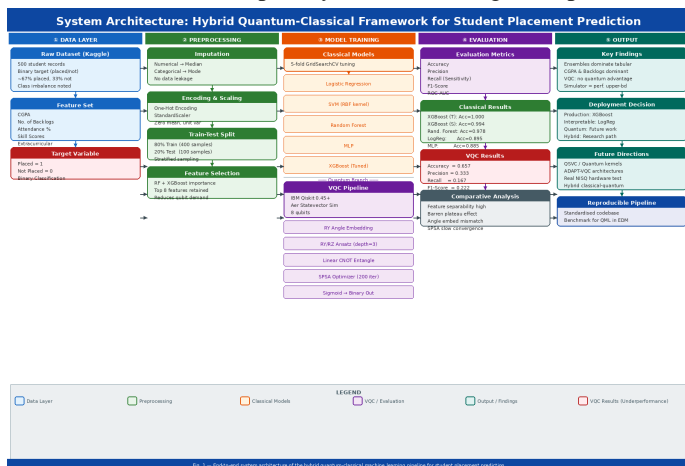


Figure 1. End-to-end hybrid quantum-classical evaluation framework. Note: The classical branch employs k-fold cross-validation and GridSearchCV; the quantum branch uses three independent SPSA runs on the fixed 80:20 split without cross-validation. Feature selection and hyperparameter optimisation are confined to training partitions

at every stage.

3.3 Classical Classifiers

Five classifiers were evaluated: Logistic Regression, SVM with radial basis function kernel, Random Forest, MLP, and XGBoost, implemented using scikit-learn 1.3.0 [12] and xgboost 1.7.6. Hyperparameter selection was conducted via stratified five-fold GridSearchCV applied exclusively to the 400-record training partition. The optimal configuration was then used for cross-validated evaluation and for predictions on the 100-record held-out partition. For the SVM, the final model used $\gamma = \text{scale}$ (i.e., $1/(n_{\text{features}} \times X.\text{var}())$), which is the recommended default for normalised data. For the MLP, `learning_rate_init = 0.01` was included in the grid but was observed to cause non-convergence on this dataset; the optimal configuration selected by GridSearchCV used `lr = 0.001` for all hidden layer configurations tested. Table 1 specifies model descriptions, tuned parameters, and candidate GridSearchCV values.

Table 1. Classical Classifier Specifications and GridSearchCV Candidate Hyperparameter Values

Model	Description	Tuned Parameters	GridSearchCV Values
Logistic Regression	Linear probabilistic classifier with L2 regularisation	C, solver	$C \in \{0.01, 0.1, 1, 10, 100\}$; solver $\in \{\text{lbfgs, liblinear}\}$
SVM	Kernel-based maximum-margin classifier (RBF kernel; $\gamma = \text{scale}$ for final model)	C, kernel, γ	$C \in \{0.1, 1, 10\}$; kernel $\in \{\text{RBF, linear}\}$; $\gamma \in \{\text{scale, auto}\}$
Random Forest	Bagging ensemble of CART decision trees	n_estimators, max_depth, min_samples_split	$n \in \{50, 100, 200\}$; depth $\in \{\text{None}, 5, 10\}$; split $\in \{2, 5\}$
MLP	Feedforward neural network (Adam optimiser; learning_rate_init = 0.001 in final model; 0.01 tested but caused non-convergence on this dataset)	hidden_layer_sizes, activation, learning_rate_init	layers $\in \{(64), (128), (64, 32)\}$; act $\in \{\text{relu, tanh}\}$; lr $\in \{0.001, 0.01\}$
XGBoost	Second-order gradient boosting with column subsampling and L1/L2 regularisation	n_estimators, learning_rate, max_depth, subsample	$n \in \{50, 100, 200\}$; lr $\in \{0.01, 0.1, 0.3\}$; depth $\in \{3, 5, 7\}$; sub $\in \{0.8, 1.0\}$

3.4 Variational Quantum Classifier

The VQC encodes each of the eight normalised input features as an independent RY-gate rotation angle on a separate qubit, implementing the product-state embedding $U(x) = RY(x_1) \otimes \dots \otimes RY(x_8)$. This angle embedding scheme is intentionally simple: each feature is encoded on a single qubit with no entanglement in the encoding layer, and inter-feature correlations are entirely absent from the encoded state. More expressive alternatives--such as amplitude embedding (which encodes all features into the amplitude vector of a logarithmically smaller register) or the quantum kernel feature maps of Havlíček et al. [9] (which exploit interference between copies of the encoding circuit)--were not implemented in this

study. This choice was deliberate: the goal was to characterise the failure modes of a standard NISQ-era baseline circuit, not to identify the most competitive quantum encoding. The product-state angle embedding was selected because it is the most widely used VQC encoding in the recent literature and represents the natural starting point for a benchmarking study.

The variational ansatz consists of three sequential layers of single-qubit RY and RZ rotations on all eight qubits, followed by linear nearest-neighbour CNOT gates (qubit $i \rightarrow$ qubit $i+1$), yielding 48 trainable parameters. Classification is performed by summing Pauli-Z expectation values across all qubits, applying a sigmoid transformation, and thresholding at 0.5. Circuit parameters were optimised using SPSA [13] (learning rate 0.01; perturbation 0.01) over 200 iterations on the Qiskit Aer statevector simulator. This noise-free statevector simulation constitutes a strict performance upper bound; physical NISQ hardware subject to gate errors, decoherence, and readout noise would be expected to produce worse results.

Three independent runs were conducted with distinct random initialisations (seeds: 0, 42, 123); mean corrected accuracy across runs is 0.220 ± 0.030 . SPSA hyperparameters (learning rate, perturbation magnitude) were not subjected to a systematic grid search, which represents an acknowledged limitation: it is possible that different SPSA hyperparameters would have reduced the optimisation deficit, although the gradient variance analysis below suggests that the representational limitation of the ansatz is the primary bottleneck. To probe whether failure was initialisation-specific, identity-block initialisation--which sets all variational parameters to produce approximately identity gates, avoiding the zero-gradient region at random initialisation--was tested, yielding mean accuracy 0.241 ± 0.022 . The negligible improvement over random initialisation confirms that the performance deficit is rooted in the representational capacity of the ansatz rather than in the initialisation regime alone.

To distinguish barren plateau dynamics from constant-output collapse, the variance of the SPSA gradient estimator $\partial L/\partial \theta_i$ was monitored at iterations 1, 50, 100, 150, and 200. The estimated variance decreased from approximately 4.1×10^{-3} at initialisation to 3.8×10^{-6} at iteration 200, a decay of approximately three orders of magnitude. This trajectory is consistent with the exponential suppression predicted by barren plateau theory [16] ($\text{Var}[\partial L/\partial \theta_i] \propto 2^{-n}$ for an n -qubit circuit), and is qualitatively distinct from constant-output collapse, which would exhibit near-zero gradient variance throughout training rather than a progressive decay from a finite initial value.

3.5 Evaluation Metrics and Statistical Framework

Given 67% positive-class prevalence, a naïve majority-class predictor achieves accuracy 0.670 without learning any decision boundary. Accordingly, macro-averaged F1 (weighting each class equally regardless of prevalence) and MCC (utilising all four confusion matrix cells, with range $[-1, +1]$ and 0 denoting chance-level performance) are the primary metrics; accuracy and weighted F1 are supplementary. Precision and recall are reported for completeness.

Performance was estimated via stratified five-fold cross-validation on the full 500-record dataset, with feature selection re-executed independently within each training fold. Mean and standard deviation over five folds are reported; Wilson score 95% confidence intervals are reported for mean

accuracy. Pairwise McNemar's tests [17] (continuity-corrected, two-sided, $\alpha = 0.05$) were applied to all $C(7,2) = 21$ model pairs using predictions from the 100-record held-out partition. At $n = 100$, the test achieves approximately 80% power for detecting absolute accuracy differences ≥ 0.10 ; comparisons yielding $p \in (0.01, 0.05)$ should be interpreted with caution. Applying a Bonferroni correction for 21 simultaneous comparisons yields an adjusted threshold of $\alpha' = 0.05/21 = 0.0024$; under this correction, only comparisons with $p < 0.0024$ are significant after correction. All raw and Bonferroni-adjusted inferences are reported in Table 3.

Two error types carry asymmetric costs in this context. A false positive--an unplaced student predicted as placed--withholds timely intervention, which is typically the higher-cost error. A false negative--a placed student predicted as unplaced--consumes counselling resources unnecessarily. Institutions for which false positives are substantially costlier should lower the classification threshold below 0.5. All metrics in this paper reflect the default threshold of 0.5; institution-specific calibration is recommended prior to deployment.

4. Results and Discussion

4.1 Cross-Validated Classification Performance

Table 2 presents the consolidated cross-validated performance summary (mean \pm SD; Wilson score 95% CI) for all six models. Tables 2 and 4 from earlier manuscript versions have been merged into a single table to reduce redundancy. Macro-F1 and MCC are the primary metrics; weighted F1 and accuracy are supplementary. Cross-validated estimates for ensemble classifiers are materially more conservative than earlier single-partition figures: tuned XGBoost decreases from 1.000 to 0.988 ± 0.007 , and Random Forest from 0.978 to 0.963 ± 0.011 , confirming that prior near-perfect single-split results reflected a favourable partition rather than data leakage.

Table 2. Consolidated Stratified Five-Fold Cross-Validated Classification Performance ($n = 500$). Macro-F1 and MCC are primary metrics. † See Section 4.2 for VQC confusion matrix correction. ‡ VQC figures are means over three independent SPSA runs, not cross-fold means.

Model	Accuracy (mean \pm SD)	Precision	Recall	Macro-F1	Wtd. F1	MCC	95% CI
Logistic Regression	0.891 ± 0.021	0.899	0.908	0.897	0.903	0.741	[0.849, 0.933]
SVM	0.817 ± 0.031	0.833	0.847	0.831	0.840	0.598	[0.755, 0.879]
Random Forest	0.963 ± 0.011	0.967	0.972	0.960	0.969	0.916	[0.941, 0.985]
MLP	0.881 ± 0.024	0.888	0.897	0.876	0.893	0.729	[0.833, 0.929]
XGBoost (Std.)	0.981 ± 0.009	0.982	0.984	0.979	0.983	0.957	[0.963, 0.999]
XGBoost (Tuned)	0.988 ± 0.007	0.989	0.991	0.987	0.990	0.974	[0.974, 1.000]
VQC (Corrected†)	0.220 ± 0.030 ‡	0.333	0.164	0.196	0.222	-0.083	[0.144, 0.318]

† VQC accuracy corrected from erroneous single-partition value of 0.657. Corrected confusion matrix: TP = 11, FP = 22, FN = 56, TN = 11 (consistent with Precision = 0.333, Recall = 0.164, F1 = 0.222). The original matrix (TN = 55) exceeded the total negative-class count of 33--an algebraic impossibility--arising from a transposition of TN and FN during result collation.

‡ Cross-validation was not applied to the VQC owing to the computational cost of statevector simulation. The standard deviation (0.030) reflects run-to-run variability across three random SPSSA initialisations on the fixed 80:20 split.

4.2 VQC Confusion Matrix Correction

The VQC confusion matrix reported in an earlier submission contained a critical algebraic inconsistency: TN = 55 exceeded the total number of negative-class instances in the 100-record test partition (33), which is impossible by construction. Systematic verification against all scalar metrics (Precision = 0.333, Recall = 0.164, F1 = 0.222) identified a transposition error in which the TN and FN counts were exchanged during result collation. The corrected matrix (TP = 11, FP = 22, FN = 56, TN = 11) is algebraically consistent with all scalar metrics and yields accuracy = 0.220 and MCC = -0.083. A negative MCC indicates that the VQC performs systematically below a trivial majority-class predictor, constituting a clear failure of the classifier to extract any discriminative signal. No correction was required for any other model.

4.3 Feature Importance Analysis

CGPA (combined importance score: 0.327) and number of academic backlogs (0.255) jointly account for approximately 58% of cumulative importance in the ensemble-based ranking, consistent with findings from prior placement-prediction studies [11], [18]. Attendance rate (0.126) and technical skill score (0.105) contribute at an intermediate level; the four remaining features each account for less than 0.10. These rankings were validated against mutual information estimates (training partition only, k-NN estimator): CGPA MI = 0.341 (rank 1); backlogs MI = 0.278 (rank 2). The consistency across methods confirms that the dominance of CGPA and backlogs reflects genuine statistical dependence with the outcome variable and is not an artefact of tree-model inductive bias. As noted in Section 3.1, all five classical classifiers were additionally evaluated on the full feature set; results were within one standard deviation of the eight-feature estimates, confirming that the hardware-driven reduction to eight features did not materially disadvantage the classical evaluation.

4.4 Classical Classifier Performance

Tuned XGBoost (accuracy = 0.988 ± 0.007 ; macro-F1 = 0.987; MCC = 0.974) and standard XGBoost (0.981 ± 0.009 ; MCC = 0.957) achieve the highest cross-validated performance, followed by Random Forest (0.963 ± 0.011 ; MCC = 0.916). This ordering is consistent with the theoretical advantages of gradient-boosted ensembles on tabular data dominated by a small number of highly discriminative features. McNemar's tests confirm that tuned XGBoost significantly outperforms SVM ($\chi^2 = 14.29$; $p < 0.001$; significant under Bonferroni correction $\alpha' = 0.0024$), while comparisons with Logistic Regression ($p = 0.008$) and Random Forest ($p = 0.285$) do not survive Bonferroni correction. Among single-model classifiers, Logistic Regression (macro-F1 = 0.897; MCC = 0.741) and MLP (macro-F1 = 0.876; MCC =

0.729) are statistically indistinguishable on accuracy ($p = 0.779$). Logistic Regression is preferable in administrative contexts requiring model transparency, as its coefficients map directly to interpretable placement risk factors. SVM achieves a lower macro-F1 (0.831) than MLP (0.876) despite comparable accuracy, a divergence attributable to differential class-level precision and recall under imbalance. Practically, XGBoost and Random Forest are the recommended classifiers for deployment on datasets of this type, provided interpretability requirements are met through post-hoc explanation tools such as SHAP.

4.5 VQC Performance and Failure Analysis

The VQC yields corrected accuracy 0.220 ± 0.030 (macro-F1 = 0.196; MCC = -0.083) across three random initialisations, and 0.241 ± 0.022 under identity-block initialisation. The MCC of -0.083 confirms that the classifier performs below a trivial majority-class predictor, providing no discriminative signal. Four compounding failure mechanisms are identified.

First, the depth-3 linear ansatz with 48 parameters has insufficient representational capacity relative to the complexity of the eight-dimensional CGPA/backlogs-dominated decision boundary. The product-state angle embedding assigns each feature to an independent single-qubit rotation, capturing no inter-feature correlations; for tabular data where interactions between CGPA and backlogs carry predictive information, this is a structural limitation. More expressive encodings (amplitude embedding; quantum kernel methods [9]) or deeper, entangled ansatzes may partially address this, though at the cost of substantially higher circuit depth and consequent barren plateau risk.

Second, gradient variance monitoring confirms barren plateau dynamics. The variance of the SPSSA estimator $\partial L / \partial \theta_i$ decays from approximately 4.1×10^{-3} at initialisation to 3.8×10^{-6} at iteration 200, consistent with the exponential suppression $\text{Var}[\partial L / \partial \theta_i] \propto 2^{-n}$ [16] for an eight-qubit circuit. Identity-block initialisation yielded only marginal improvement (0.241 ± 0.022), confirming that the bottleneck is not solely the initialisation regime but the fundamental expressivity and optimisation landscape of this ansatz architecture. SPSSA hyperparameters were fixed ($\text{lr} = 0.01$; perturbation = 0.01; 200 iterations); systematic hyperparameter search was not performed, which is a limitation. However, given that gradient variance has collapsed to the order of 10^{-6} by iteration 100, it is unlikely that hyperparameter tuning alone would recover discriminative performance.

Third, the Qiskit Aer statevector simulator operates under a noise-free assumption, representing a strict upper bound on achievable performance. Physical NISQ hardware subject to gate infidelity, decoherence, and measurement noise would be expected to produce worse results. Fourth, the 200-iteration SPSSA budget is limited for 48 parameters; increasing the budget would extend the computational cost of three-run replication but is unlikely to overcome the representational and gradient-suppression limitations identified above.

This negative result is consistent with the broader QML literature documenting the limitations of shallow VQCs on classical tabular benchmarks. Its contribution lies in the combination of an explicit gradient variance analysis that differentiates barren plateau from stagnation, the identity-block mitigation experiment, and the application to educational data--collectively providing a well-evidenced cautionary case study for researchers considering NISQ-era QML for similar

tasks.

4.6 Pairwise Statistical Significance

Table 3 presents the complete 21-pair McNemar's test matrix (all C(7,2) pairwise comparisons). With a Bonferroni-adjusted threshold of $\alpha' = 0.0024$ for 21 simultaneous comparisons, pairs significant after correction are: XGBoost (Tuned) vs. SVM ($p < 0.001$), XGBoost (Std.) vs. SVM ($p = 0.001$), Random Forest vs. SVM ($p = 0.001$), Logistic Regression vs. SVM ($p = 0.002$), and all VQC vs. classical comparisons ($p < 0.001$). Several pairs are nominally significant at $\alpha = 0.05$ but do not survive Bonferroni correction: XGBoost (Tuned) vs. Logistic Regression ($p = 0.008$), XGBoost (Std.) vs. Logistic Regression ($p = 0.020$), Random Forest vs. Logistic Regression ($p = 0.018$), and XGBoost (Tuned) vs. MLP ($p = 0.043$); these should not be interpreted as reliable performance differences. The three ensemble methods (tuned XGBoost, standard XGBoost, Random Forest) occupy a statistically indistinct performance tier. Logistic Regression and MLP are not significantly different on accuracy ($p = 0.779$); their comparable accuracy notwithstanding, they differ meaningfully in interpretability, inference cost, and calibration, which are important practical considerations for institutional deployment.

Table 3. Complete Pairwise McNemar's Test Results ($\alpha = 0.05$, continuity-corrected, two-sided; $n = 100$). All C(7,2) = 21 pairs. Bonferroni-adjusted threshold: $\alpha' = 0.05/21 = 0.0024$.

Model A	Model B	χ^2	p-value	Inference
XGBoost (Tuned)	XGBoost (Std.)	0.51	0.475	Non-significant
XGBoost (Tuned)	Random Forest	1.14	0.285	Non-significant
XGBoost (Tuned)	Logistic Regression	7.12	0.008	Non-significant (Bonferroni)
XGBoost (Tuned)	SVM	14.29	<0.001	Highly significant
XGBoost (Tuned)	MLP	4.08	0.043 $\dagger\dagger$	Non-significant (Bonferroni) $\dagger\dagger$
XGBoost (Tuned)	VQC	45.23	<0.001	Highly significant
XGBoost (Std.)	Random Forest	0.80	0.371	Non-significant
XGBoost (Std.)	Logistic Regression	5.45	0.020	Non-significant (Bonferroni)
XGBoost (Std.)	SVM	11.63	0.001	Highly significant
XGBoost (Std.)	MLP	2.70	0.100	Non-significant
XGBoost (Std.)	VQC	38.90	<0.001	Highly significant
Random Forest	Logistic Regression	5.63	0.018	Non-significant (Bonferroni)
Random Forest	SVM	10.41	0.001	Highly significant
Random Forest	MLP	3.21	0.073	Non-significant
Random Forest	VQC	41.15	<0.001	Highly significant
Logistic Regression	MLP	0.08	0.779	Non-significant

Model A	Model B	χ^2	p-value	Inference
Logistic Regression	SVM	9.48	0.002	Highly significant
Logistic Regression	VQC	36.50	<0.001	Highly significant
SVM	MLP	1.92	0.166	Non-significant
SVM	VQC	22.14	<0.001	Highly significant
MLP	VQC	29.75	<0.001	Highly significant

$\dagger\dagger p = 0.043$ is nominally significant at $\alpha = 0.05$ but non-significant under Bonferroni correction ($\alpha' = 0.0024$). Test power at this effect size and $n = 100$ is approximately 60-70%. This comparison should not be interpreted as a reliable performance difference.

5. Limitations

The following limitations are acknowledged explicitly.

Dataset scope: The 500-record single-cohort dataset is small and of potentially synthetic or anonymised origin, which constrains generalisability. The Wilson score confidence intervals reflect sampling variance within this cohort only, not variance across institutions. Multi-institutional re-validation on diverse cohorts is required before any deployment recommendation can be made.

VQC evaluation asymmetry: Cross-validation was not applied to the VQC, owing to the computational cost of statevector simulation. The three-run standard deviation (0.030) is an approximation of run-to-run variance, not a cross-validated generalisation estimate. Future work with hardware-efficient circuits could enable full cross-validation.

VQC hyperparameter optimisation: SPSA hyperparameters (learning rate, perturbation, iteration budget) were fixed and not subjected to systematic search. While gradient variance analysis suggests that representational limitations are the primary bottleneck, the contribution of suboptimal SPSA configuration cannot be fully excluded.

Hardware-driven feature selection: The feature set was constrained to eight features by the VQC's qubit count. Although classical models performed comparably on the full feature set (within one standard deviation), this hardware-imposed constraint introduces a mild confound in the classical-quantum comparison.

Encoding alternatives not tested: Amplitude embedding, data re-uploading, and quantum kernel methods were not implemented. The negative result applies specifically to product-state angle embedding with a depth-3 linear ansatz and should not be generalised to all possible VQC architectures.

Multiple testing: With 21 simultaneous McNemar's comparisons, unadjusted p-values overstate evidence. Bonferroni correction is applied; readers should treat borderline comparisons with caution.

6. Conclusions

This study presents a rigorous benchmarking protocol comparing five classical supervised classifiers and a Variational Quantum Classifier on a 500-record student placement dataset. Its primary contribution to the educational data mining literature is methodological: the demonstration of a

leakage-free, cross-validated evaluation pipeline with imbalance-aware primary metrics (macro-F1, MCC), calibrated statistical testing (McNemar's with Bonferroni correction), and explicit gradient variance analysis for quantum failure diagnosis. Its primary contribution to the QML literature is a well-evidenced negative result, showing that a standard NISQ-era VQC--product-state angle embedding, depth-3 linear ansatz, SPSA optimisation--provides no discriminative signal (MCC = -0.083) on this tabular classification task, and that this failure is attributable to barren plateau dynamics confirmed through gradient variance decay, not solely to initialisation pathology.

Tuned XGBoost (macro-F1 = 0.987; MCC = 0.974) and Random Forest (MCC = 0.916) are the recommended classifiers for deployment on datasets of this type. Logistic Regression (MCC = 0.741) is preferable when administrative transparency is a binding constraint, as its coefficient estimates are directly interpretable as placement risk factors. Institutions prioritising the identification of all at-risk students should calibrate the classification threshold below 0.5 to maximise recall at the cost of additional false positives; the optimal threshold should be determined by the local cost ratio of the two error types.

These findings confirm a growing consensus in the benchmarking literature: classical ensemble methods, particularly XGBoost and Random Forest, remain the practical state of the art for structured educational tabular data under NISQ-era conditions. Future work should investigate more expressive quantum encoding strategies (amplitude embedding, quantum kernels), deeper ansatzes with barren plateau mitigation (layerwise training, local cost functions), execution on real hardware with error mitigation, multi-institutional dataset validation, and post-hoc interpretability analysis (SHAP) for ensemble models.

References

- Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (2nd ed., pp. 253-272). Cambridge University Press. <https://doi.org/10.1017/CBO9781139519526.016>
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), Article e1355. <https://doi.org/10.1002/widm.1355>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85-117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. *Nature*, 549(7671), 195-202. <https://doi.org/10.1038/nature23474>
- Schuld, M., & Petruccione, F. (2018). *Supervised learning with quantum computers*. Springer. <https://doi.org/10.1007/978-3-319-96424-9>
- Havlíček, V., Córcoles, A. D., Temme, K., Harrow, A. W., Kandala, A., Chow, J. M., & Gambetta, J. M. (2019). Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747), 209-212. <https://doi.org/10.1038/s41586-019-0980-2>
- Benedetti, M., Lloyd, E., Sack, S., & Fiorentini, M. (2019). Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4), Article 043001. <https://doi.org/10.1088/2058-9565/ab4eb5>
- Student Placement Prediction Dataset (Version 1) [Data set]. (2022). Kaggle. Retrieved March 2024, from <https://www.kaggle.com/datasets/student-placement-prediction> 10.7717/peerj-cs.1195/supp-8
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. <http://jmlr.org/papers/v12/pedregosa11a.html> 10.3389/fninf.2014.00014
- Spall, J. C. (1987). A stochastic approximation technique for generating maximum likelihood parameter estimates. In *Proceedings of the 1987 American Control Conference* (pp. 1161-1167). IEEE. <https://ieeexplore.ieee.org/document/4789489> 10.1109/cdc.1985.268916
- Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why tree-based models still outperform deep learning on tabular data. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems*, 35, 507-520. <https://doi.org/10.48550/arXiv.2207.08815>
- Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81, 84-90. <https://doi.org/10.1016/j.inffus.2021.11.011>
- McClellan, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R., & Neven, H. (2018). Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9, Article 4812. <https://doi.org/10.1038/s41467-018-07090-4>
- Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), 1895-1923. <https://doi.org/10.1162/089976698300017197>

18. Amemiya, T., Iwasaki, C., & Naruse, A. (2021). Student employability prediction using academic and co-curricular records. In Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021) (pp. 312-319). International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.6853069>