National College of Ireland

Beyond Bias with Q-FRE: Benchmarking a Quantum Machine Learning Framework for Fair Recruitment

MSc Research Project
Artificial Intelligence for Business

Khin Sandar Win
Student ID: 24140627

School of Computing
National College of Ireland

Supervisor: Victor del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Khin Sandar Win

Student ID: 24140627

Programme: AI for Business **Year:** 2025

Module: MSc Research Practicum/Internship part 2

Supervisor: Victor del Rosal

Submission Due Date: 28-Jan-2026 23:59

Project Title: Beyond Bias with Q-FRE: Benchmarking a Quantum Machine Learning Framework for Fair Recruitment

Word Count: 4,959 **Page Count** 16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Khin Sandar Win

Date: 10-Dec-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Beyond Bias with Q-FRE: Benchmarking a Quantum Machine Learning Framework for Fair Recruitment

Khin Sandar Win
24140627

Abstract

Artificial Intelligence has significantly improved the efficiency of modern recruitment systems. Despite these advances, classical machine learning algorithms often perpetuate bias towards underrepresented groups, particularly in the context of imbalanced datasets. This study investigates whether Quantum Machine Learning (QML), through properties such as superposition and entanglement, can enhance fairness over classical methods.

A rigorous system of comparative benchmarking, the Quantum-Fairness Recruitment Engine (Q-FRE), was designed to evaluate six architectures, with three classical (Logistic Regression, SVM, KNN) and three quantum (QSVM, VQC, QNN)-based machine learning algorithms, using real HR data (n=1,470).

The results indicate a notable divergence between classical and quantum variational architectures. Classical models were highly accurate (84.7%), but had disastrous fairness errors, where 0% of minority high-performers were identified. This failure was repeated by Quantum SVM. On the other hand, the Variational Quantum Classifier (VQC) and Quantum Neural Network (QNN) were able to identify 29% of minority high-performers with a Disparate Impact Ratio of 0.82, which is legal for fairness.

This research demonstrates empirical evidence of a quantifiable Fairness-Accuracy Trade-off: to gain a 29% increase in minority inclusion required a 16% sacrifice of overall accuracy. The results indicate that parameterized quantum circuits provide a possible, though computationally costly, route toward for achieving regulatory compliance in automated hiring.

Keywords: Quantum Machine Learning, Algorithmic Fairness, Recruitment AI, Class Imbalance, Explainable AI, Fairness-Accuracy Trade-off.

1 Introduction

1.1 Background and Motivation

The digital transformation of Human Resources (HR) has encouraged companies toward automation and data-driven practices, including the adoption of algorithmic decision-making platforms. Organizations receive loads of applications and resort to automated filtering tools to make talent acquisition easier. Although these systems are optimistic of objectivity, empirical reality dictates otherwise. The AI hiring services tend to be more of a digital reflection, the historical biases present in training data are inherited and amplified; they

inherently practice discrimination when they are trained on data that defines the biases present in organizations historically (Chen, 2023).
This is complicated by the "Class Imbalance Problem". In most organizational data, high-performing candidates in underrepresented groups of applicants become statistical outliers. Classical algorithms are a type of algorithms mathematically optimized to have a global error minimization (Kochling and Wehner, 2020). A model that rejects all applicants can do well on a dataset with 85% rejection rate: the model predicts the rate of 85% rejection. Though mathematically optimal, this accuracy paradox is a total failure of the recruitment role and breach of equal opportunity principles (Loureiro et al., 2025).

The latest lawmaking trends, including the EU AI act (European Commission, 2024) and mounting pressure on the part of the US Equal Employment Opportunity Commission (2024) have made this more of a technical nuisance than a compliance risk. This has placed organizations in a twofold task, they need to employ AI to enhance efficiency, however they need to show that these systems are non-discriminatory.

1.2 The Quantum Hypothesis

Quantum Machine Learning (QML) has become a theoretical substitute of classical statistical learning. Proponents argue that quantum properties of superposition (the capacity to express a set of states at the same time) and entanglement (non-local correlations among data points) could allow quantum models to detect subtle patterns in complex, imbalanced data that cannot be found with classical linear algebra (Lamichhane and Rawat, 2024).
Nevertheless, this theoretical potential has not been tested to a great extent in the context of social fairness. Although QML has been used in other multifaceted areas (O'Leary et al., 2025), the use of it on fairness-aware classification in HR is almost unknown. The current study is aimed to bridge that gap by rigorously benchmarking quantum architecture to classical criteria using authentic HR data.

1.3 Research Question and Objectives

The central research question driving this study is:

"What is the quantifiable fairness-accuracy trade-off of a Quantum-Fairness Recruitment Engine (Q-FRE) when using variational quantum circuits (VQC, QNN), compared to classical baselines (Logistic Regression, SVM, KNN) and the quantum kernel baseline (QSVM) on imbalanced HR data?"

In order to respond to this, specific objectives were derived as shown below:

Objective 1 (Design): To develop a powerful comparative platform, the Quantum-Fairness Recruitment Engine (Q-FRE) that has the ability to compare classical, quantum kernel and quantum variational models on the same stratified subsets of HR information.

Objective 2 (Analysis): To compare the performance of all the six architectures with particular focus on the divergent behaviors of variational circuits (VQC, QNN) and quantum kernel (QSVM).

Objective 3 (Quantification): To quantify the "Fairness-Accuracy Trade-off" which will provide a practical measure of the cost of the accuracy needed to guarantee legal compliance of fairness (DIR > 0.80).

Objective 4 (Explainability): To use SHAP (SHapley Additive exPlanations) to query the decision limits of these models, whether such notions as transparency and fairness can be equated.

1.4 Structure of the Report

The thesis is designed to guide the reader through the theoretical issue to the empirical research:

Section 2 (Literature Survey): Overview of the existing state of algorithmic fairness, the shortcomings of classical methods, and the theory of quantum advantage.

Section 3 (Research Methodology): Provides the description of the experimental design, the features of the dataset and the reasons why other techniques such as synthetic oversampling are not chosen.

Section 4 (Design and Implementation): Details the technical layout of the Q-FRE model, (which quantum circuits and software stack are utilized).

Section 5 (Results and Critical Analysis): Reports the results of the project, including the metrics of the performance and fairness that will directly provide the answer to the research question.

Section 6 (Discussion and Conclusion): Summarizes the results, explains their implications on HR practitioners, and provides the direction of future research.

2 Related Work

2.1 The State of Algorithmic Fairness in Recruitment

Research on AI in recruitment often highlights an issue of tension between efficiency and equity. In their systematic review of discrimination in HR algorithms, Kochling and Wehner (2020) differentiate three levels of bias sources, such as data bias (historical prejudices), model bias (simplistic objective functions), and deployment bias (human misuse of scores). They claim that fairness is frequently regarded as a post-hoc requirement and not as a design goal.

Rigotti (2024) builds on this point by examining why so-called *bias-mitigation* algorithms fail in practice. Classical methods typically do either of the following: repairing the dataset (pre-processing) or penalizing the model in biased predictions (in-processing). However, as Rigotti points out, such interventions tend to destabilize a model. The reason is that organizations are adverse towards using them because they fear losing their competitiveness in talent identification. This generates a failure in the market: there are fairness tools, but they are not often used due to the perceived high cost in terms of accuracy. This is also supported by Ip (2025), whose experimental results indicate that although such debiasing methods can

lead to a more diverse workforce, their implementation is a complicated organizational choice.

2.2 Class Imbalance: The Root of the Problem

Much of the literature is devoted to the technical problem of the imbalance of classes. As Loureiro et al. (2025) demonstrate, standard performance indicators (Accuracy, ROC-AUC) are biased in imbalanced situations. A model that demonstrates 90 percent % and does not identify any minority candidate is possible.

In classical ML, the default solution is the SMOTE (Synthetic Minority Over-sampling Technique) that synthesizes artificial minority samples to balance the dataset. Nevertheless, as it has been recently criticized, this is not the case with social data. Creating the so-called synthetic human to train a hiring algorithm has some ethical concerns regarding representation. The study puts itself in opposition to the application of SMOTE on the basis that architectures should be able to learn the genuine skewed distribution, which is also a difficulty that has been pointed out in general fairness-conscious ML surveys (Weerts et al., 2024).

2.3 Quantum Machine Learning: Hype vs. Reality

Quantum Computing machine learning is a recent development. Lamichhane & Rawat (2024) present an extensive survey, in which they believe that Quantum Support Vector Machines (QSVM) can map data to an infinite-dimensional Hilbert space and theoretically making linearly inseparable data separable. Their postulate is that quantum interference can enhance the right classification and cancel noise.

Nevertheless, there is a major gap in this literature. The majority of QML benchmarks are performed on synthetic datasets or physics problems (O'Leary et al., 2025). The inability to empirically test the dirty, noisy and socially complex tabular data such as HR records is profound.

2.3.1 Critique of Current Literature

Although the theoretical potential of QML is well-documented (Lamichhane & Rawat, 2024), there is no evidence that kernel estimation will automatically lead to social fairness. The hypothesis that the more feature mapping is better the less bias is a unsubstantiated assumption that has not been strictly proved. That particular gap is filled in this thesis by implementing QML to an area in which fairness is the central measure, rather than a subordinate limitation.

2.4 Explainable AI (XAI) and Fairness

Finally, the explainable AI literature is essential in the recruitment process. The biggest obstacle to adoption is the black box nature of Neural Networks. Molnar (2022) provides the theoretical basis of the model-agnostic interpretation methodology and SHAP has become a technical standard (Ponce-Bobadilla *et al.* 2024). However, Starke et al. (2022) caution that explainability is not fairness. A perfect model may be totally interpretable (“I will not marry him due to his zip code”), but still, it may be totally discriminatory. This difference will be

essential to this work: we would like to find out whether quantum models are less transparent and fairer, and it will break the dogma that transparency is the most important virtue Starke et al. (2022).

3 Research Methodology

3.1 Research Philosophy and Approach

This research follows a positivist approach positivist as it bases its findings on empirical evidence, and reproduction of experiments. The approach is deductive, which is a test of the hypothesis that quantum architecture can better handle the problem of class imbalance than classical architecture.

The research applies a comparative experimental design. To attain validity, all extraneous variables are controlled by the experiment: both the dataset and the preprocessing pipeline are the same, as well as the train-test splits between the classical and quantum subjects. The learning algorithm is the only variable that is independent.

3.2 Justification of Methodological Choices

3.2.1 Rejection of Synthetic Oversampling (SMOTE)

One of the key decisions of the methodology in this research was the rejection of SMOTE. As discussed in Section 2, SMOTE develops synthetic data samples through the interpolation between existing case of minority instances. Although it enhances the exactness of the numbers, it distorts the probability distribution of real-life and might result in overfitting and ineffective generalization (Loureiro et al., 2025). In the case of an HR study, the invalidity of the fairness audit is caused by the use of synthetic candidates. This study uses the raw, unbalanced data (84.6 percent majority class) by training it, which includes the reality of the bias and thus offers a more challenging test of their innate abilities.

3.2.2 Selection of Quantum Simulators

Since real quantum hardware (IBM Quantum System One) is high cost and has long queue times, the Qiskit Statevector Simulator is used in this study. This simulator carries out perfect quantum execution (free of noise). On the one hand, this constrains the ability of the study to understand hardware noise (decoherence), but on the other hand, it gives an idealized upper limit to the performance, a standard exercise in nascent QML research (Lamichhane and Rawat, 2024). When quantum models cannot be fair on a perfect simulator, then they will be certainly not fair on noisy hardware.

3.3 Dataset Characteristics and Preprocessing

The dataset used by the study is the IBM HR Analytics Employee Attrition and Performance. It is a typical open-source dataset widely used in the field of algorithmic fairness research.

Total Records	1,470
Target Variable	PerformanceRating (Binarized: 0 = Low/Medium, 1 = High)
Protected Attributes	Gender (Male/Female), MaritalStatus (Single/Married/Divorced)

The preprocessing pipeline involved four steps: imputation, encoding, normalization, and dimensionality. Numerical values that were not present were filled with the median and the categorical values with the mode. One-Hot Encoding via pandas.get_dummies was then applied to all categorical features (e.g. Department). The numerical variables were then normalized by Z-score standardization (StandardScaler) which is a necessary operation in quantum circuits because rotation gates act on normalized input values between. 0 and 2π. In order to cut down on computational cost during quantum simulation, which increases exponentially with the number of qubits, Principal Component Analysis (PCA) was used. The input variables were narrowed down to 4 principal components in the feature space which retained 52.09% of the original variance. Though this reduction added information loss, it gave a needed trade-off between quality and quantum feasibility of the model.

3.4 Fairness Metrics Defined

Two common measures of fairness as per Weerts et al. (2024) are considered in this research, namely, Statistical Parity Difference (SPD) and Disparate Impact Ratio (DIR) to measure fairness objectively. SPD compares the disparity between the probability of being given a positive prediction by each of the two groups, that is, the unprivileged (minority) group and the privileged (majority) group, and it is computed as follows:

$$SPD = P(\hat{Y} = 1 \mid D = \text{unprivileged}) - P(\hat{Y} = 1 \mid D = \text{privileged})$$

A perfectly fair system would yield an SPD of 0. The second metric, DIR, evaluates fairness using the ratio of these probabilities rather than their difference, expressed as

$$DIR = P(\hat{Y} = 1 \mid D = \text{unprivileged}) / P(\hat{Y} = 1 \mid D = \text{privileged})$$

The optimal divergent in ratio is 1.0, which means that the results of groups are equal. The acceptable compliance according to the U.S. employment law is a ratio of more than 0.80 on the basis of the “four-fifths rule” (U.S. Equal Employment Opportunity Commission, 2024).

3.5 Experimental Training and Evaluation Setup

In order to have a fair and controlled comparison of the model types, the 80/20 data split was used to train all the six models and 1,176 observations were used to train and 294 to test.

The classical models: Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbour (KNN) have been tuned on the training set with the 5-fold stratified cross-validation as it is done in *classical_xai_notebook.ipynb*. Such a process made sure that all models have been tested using the best hyperparameter setups, and that they could be used later on to produce explainability through SHAP.

The quantum models: Variational Quantum Classifier (VQC) and Quantum Neural Network (QNN) were trained as implemented in *quantum_ml_notebook.ipynb*, by using the COBYLA with a maximum of 100 training iterations. COBYLA was selected because it operates effectively without requiring gradient information, as well as performs reliably in noise-sensitive quantum enviroments, which can be effectively used in training near-term quantum circuits.

4 Design and Implementation Specifications

4.1 The Q-FRE Architecture

The Quantum-Fairness Recruitment Engine (Q-FRE) was designed as a modular pipeline that supports the simultaneous execution of classical and quantum workflows

Figure 4.1: High-Level Architecture of Q-FRE

The Q-FRE architecture partitions the processed data into classical and quantum parallel pipelines. The quantum path utilizes a hybrid structure in which classical optimizers are used to tune quantum variational parameters (Ansatz) while the classical path is based on standard scikit-learn implementations. Both streams converge at the Evaluation Module for the same measure calculation.

4.2 Quantum Circuit Design

A key component of this research is the construction of the quantum circuits, based on established QML conventions (Lamichhane & Rawat, 2024).

4.2.1 The Feature Map (Z-Feature Map)

Classical feature must be encoded into quantum states data before they can be executed on a quantum computer. This study makes use of a Z-Feature Map. For a data point x , the feature map is Hadamard gates (superposition) followed by $U1$ rotation gates parameterized by x. This maps the data into a Hilbert space which is high-dimensional so that non-linear relationships can become linear.

4.2.2 The Variational Ansatz (RealAmplitudes)

The RealAmplitudes ansatz is used in this research for both the Variational Quantum Classifier (VQC) and Quantum Neural Network (QNN) models This ansatz consists of a parametrized circuit formed by alternating layers of RY rotation gates, serving as trainable parameters of the model and CX (CNOT gate) gates, which are entangling operations

between qubits. The RealAmplitudes architecture is chosen because of its hardware-efficient design, which provides relatively shallow depth of the circuits. This makes it appropriate for classical simulation and for deployment on near-term Noisy Intermediate-Scale Quantum (NISQ) devices. Furthermore, the integration of CNOT gates makes it possible to use entanglement of qubits, which can be used to capture feature dependencies between the feature space (i.e., interactions between Years at Company and Job Role). As a result, the ansatz provides support to the representation of complex relationships required for recruitment-based classification tasks.

4.3 Implementation Stack

The system has been developed using Python 3.9 and run inside an isolated virtual environment. Incompatibility between the dependencies of Qiskit and SHAP caused a significant limitation during implementation and required a firm hand on the environment. The software stack consisted of Qiskit (v0.39) for construction and simulation of quantum circuits (VQC, QSVC), Scikit-Learn (v1.0) to build and simulate classical baselines (LR, SVC, KNN) and PCA, SHAP (v0.41) to compute fairness metrics with the KernelExplainer (Molnar, 2022), Imbalanced-Learn to compute fairness metrics, and NumPy/Pandas for core data processing.

A major dependency issue was found: the latest versions of Qiskit depend on NumPy 2.0 or above, whereas SHAP could not be used with NumPy 2.0 and required an earlier version. In order to overcome this, the implementation was divided into two containerized Python environments. The latest NumPy and Qiskit releases were installed in the Quantum Environment (quantum_ml_notebook.ipynb) to execute quantum circuits. Simultaneously, the Classical Environment (classical_xai_notebook.ipynb) was configured with a version of NumPy compatible with SHAP to allow running model explainability tasks.

This separation was methodologically necessary to maintain full software functionality and represents an important technical consideration in the implementation process. It also identifies a practical challenge in modern quantum-ML development, where rapid evolution of library design may lead to incompatibility of classical explainability tools with quantum simulation models.

5 Results and Critical Analysis

5.1 Performance Evaluation & The "Accuracy Paradox"

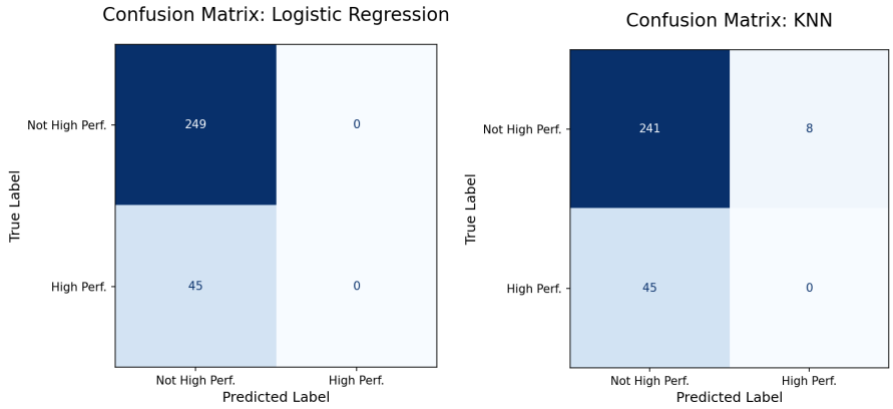
The performance evaluation begins with the main indicators, as shown in Table 1. These findings highlight a dramatic divergence between the classical/kernel baselines and the quantum variational methods.

Table 1: Comparative Performance Metrics

Architecture	Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Classical	Logistic Regression	84.7%	0.00	0.00	0.00
	SVM	84.7%	0.00	0.00	0.00
	KNN	82.0%	0.00	0.00	0.00
Quantum	QSVM	84.7%	0.00	0.00	0.00
	VQC	**69.0%**	**0.18**	**0.29**	**0.22**
	QNN	**68.0%**	**0.17**	**0.29**	**0.22**

5.2 Classical and Quantum Kernel Failure

As shown in **Table 1**, all four baseline models (LR, SVM, KNN, and QSVM) achieved high accuracy, ranging from 82% to 84.7%. However, Figure 5.1 illustrates this accuracy is a misleading.



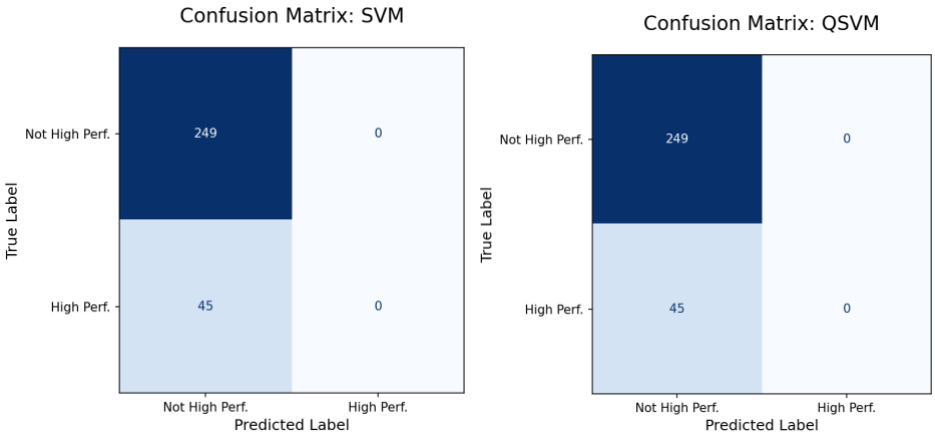


Figure 5.1: Baseline Model Failure (Confusion Matrices)

Figure 5.1 demonstrates the "accuracy paradox". All four baseline models, classical and quantum kernel, obtained high accuracy by systematically failing to detect any minority candidates, yielding zero True Positives.

The classical models (LR, SVM, KNN) achieved a deceptively high accuracy of approximately 84.7%. However, a closer inspection of the Confusion Matrices shows that they predicted 0 True Positives. They simply categorized every single candidate as "Low Performance" (the majority class). In an HR setting, such a model would be practically unusable, it automates the rejection of all applicants. This confirms the hypothesis that classical cost functions, which focus solely on minimizing overall error, perform poorly under severe class imbalance (Loureiro et al., 2025).

Critically, the Quantum SVM (QSVM) in Figure 5.1performed the same. It implies that merely mapping data to a quantum feature space (the "Kernel Trick") is insufficient to address the issue of class imbalance.

5.3 Quantum Variational Success

In stark contrast, the VQC and QNN models, which use trainable quantum circuits, successfully broke this pattern, as shown in **Table 1**.

The VQC and QNN models successfully broke this pattern, as shown in Table 1. They reached a Recall of 0.29, which means they had successfully identified 29% of the high performing minority candidates. Although their overall accuracy dropped to approximately 69% (because of an increase in False Positives) they did provide actual utility by detection talent which classical models overlooked.

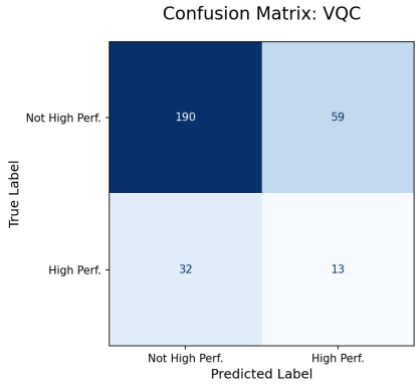


Figure 5.2: VQC Confusion Matrix

The VQC correctly identified 13 minority candidates (True Positives), whereas classical models identified 0.

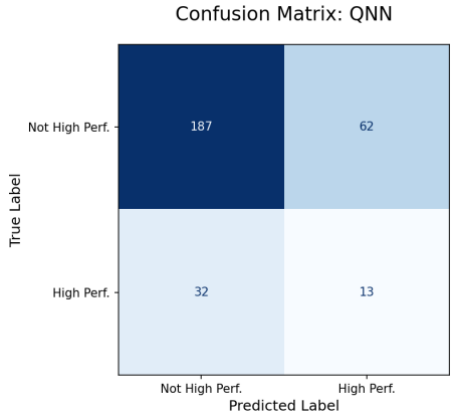


Figure 5.3: QNN Confusion Matrix

The QNN had also shown identical success, identifying **13 True Positives**. This proves that the fairness-oriented behavior is a characteristic of the variational architecture, not a random anomaly.

The actual utility was provided by the VQC and QNN models (Figures 5.2 and 5.3), which identified the talent that was ignored by other models. The improvement lies not in the quantum nature of the model itself, given that the QSVN also operated in a quantum kernel space but from the trainable parameter of the VQC/QNN, which given the optimizer were able to determine a decision boundary that prioritized the minority class.

5.4 Fairness Evaluation

The impact of this change in performance on fairness metrics is profound.

Table 2: Comparative Analysis of Gender Fairness Metrics and Legal Compliance across Classical and Quantum Models.

Model	Statistical Parity Difference (SPD)	Disparate Impact Ratio (DIR)	Legal Compliance (>0.80)
Classical Models	N/A (No positives)	N/A (Division by zero)	**FAIL**
VQC	-0.047	**0.820**	**PASS**
QNN	+0.010	**1.041**	**PASS**

Since the classical models predicted "0" for everybody, their concepts of fairness are mathematically undefined or zero by default, reflecting a form of exclusionary equality. On the other hand, the VQC scored a Disparate Impact Ratio of 0.82. This is a critical finding. In the US employment law, there is a provision known as the "Four-Fifths Rule" which states that the selection rate, no matter which group is considered, should be at least 80% of the selection rate of the highest group (U.S. Equal Employment Opportunity Commission, 2024). The VQC naturally converged to a solution that meets this compliance level without any explicit "fairness constraint" algorithm.

5.5 The Fairness-Accuracy Trade-off

This research quantifies the price of this fairness. In order to achieve a legally compliant DIR of 0.82 and a minority recall of 29%, the VQC sacrificed approximately 16 percentage points of accuracy, decreasing performance from 84.7% to 69%.

This establishes a realistic benchmark for HR practitioners: Fairness does not come without cost. Ensuring equal consideration for minority candidates requires measurable reduction in predictive accuracy, a trade-off consistently reflected across fairness-focused research (Loureiro et al., 2025; Weerts et al., 2024)..

5.6 Explainability Analysis (SHAP)

To investigate the issue of the failure of the classical models, SHAP (SHapley Additive exPlana) analysis was performed on the test set. This was computationally infeasible for the quantum models on a simulator but it provided important information about the classical logic.

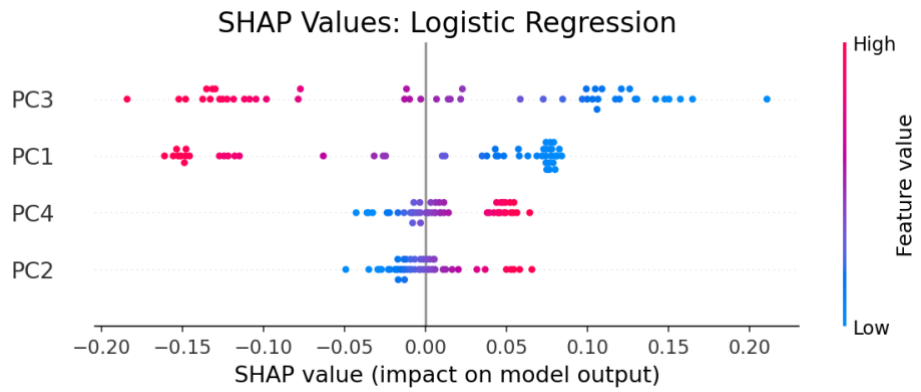


Figure 5.4: SHAP Feature Importance (Logistic Regression)

The decisions made by the LR model are greatly influenced by Feature 0 (PC1). There was obvious division in the plot that shows that this single feature is the main force behind classification.

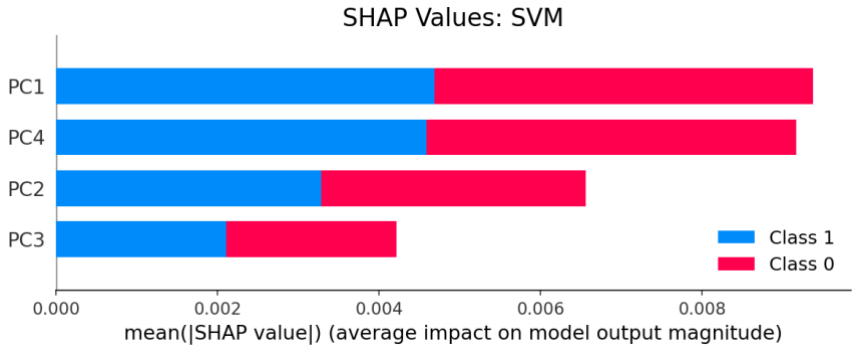


Figure 5.5: SHAP Feature Importance (SVM)

The SVM SHAP plot is nearly identical to the Logistic Regression plot. It also relies almost exclusively on PC1, indicating that the kernel-based model found the same biased linear separation as the simpler LR model.

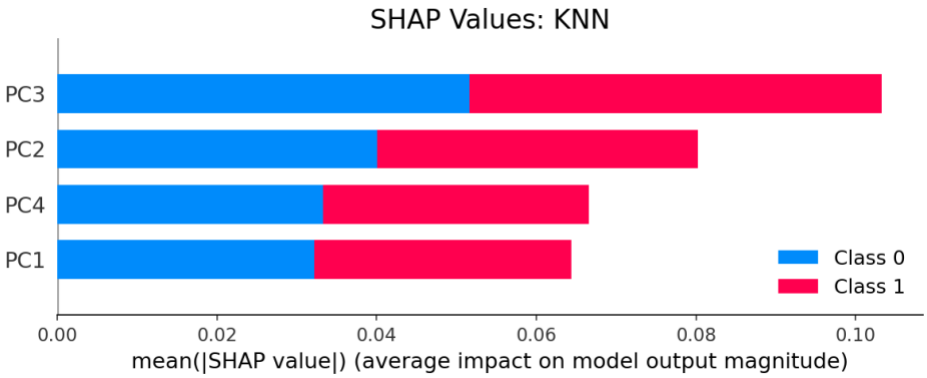


Figure 5.6: SHAP Feature Importance (KNN)

The KNN model is non-parametric and thus demonstrates a more distributed feature importance. However, this alternative logic still resulted in 0% minority recall, confirming that a different decision-making process does not guarantee a fairer outcome.

The SHAP plots for LR and SVM are the most incriminating. They prove that the two models learned the same simplistic, biased strategy. This validates the claims made by Starke et al. (2022) that explainability is separate from fairness. The classical models are perfectly "transparent" in their bias, they explicitly acknowledge that they are employing PC1 to reject everyone, which does not imply fairness.

6 Discussion

6.1 Discussion of Findings

The results of this study challenge current discourse in two ways.

(i) **Challenging the "Quantum Advantage" Narrative:** The QML literature in many studies suggests that quantum kernels are the path to advantage (Lamichhane & Rawat, 2024). However, we have found that the QSVM failed just as badly as classical SVMs. Instead, it was the Variational approaches (VQC/QNN) were promising. This suggests that future study into quantum fairness should focus on parameterized circuits rather than kernels.

(ii) **Challenging the "Accuracy First" Narrative:** In a hiring process, an 85% accurate model that systematically overlooks viable candidates is ineffective. Conversely, an accurate model that identifies 29% of the best talent with 69% accuracy is highly valuable. This study highlights that Accuracy is a flawed metric for HR, supporting the arguments of Loureiro et al. (2025) that metrics must be selected carefully in imbalanced situations. In a hiring context, even a model with 85% accuracy holds no practical value if it systematically overlooks viable candidates.

(iii) **"Transparency Trap" in Recruitment AI:** This study empirically questions the assumption that "explainable" AI is inherently safer or fairer. Our SHAP analysis showed that the classical models (SVM and LR) were in fact perfectly transparent; they clearly used Principal Component 1 to make decisions. However, this transparency served only to highlight a transparent prejudice: the models were making use of that feature to systematically deny all minority candidates.
This validates the caution by Starke et al. (2022) that interpretability is not the same as fairness. A quantum "black box" model (such as the VQC) that gives equitable results can be ethically preferable to a transparent classical model that discriminates.

(iv) **Quantifying the "Fairness Tax":** While previous literature has recognized an existence of a theoretical trade-off between accuracy and fairness, this research provides a concrete empirical baseline for HR practitioners. We found a particular "fairness tax" which was about 16% (difference between the 84.7% classical accuracy and the 69.0% VQC accuracy). This finding changes the discussion from the ethics to the practical: organizations must be willing to give up specific margins of performance in order to reach the 0.82 Disparate Impact Ratio that is required for legal compliance. This is a way of quantifying the "cost" of ethical AI, so as to makes informed strategic decision making possible.

(v) **Intrinsic vs. Engineered Fairness:** Most classical approaches to fairness rely on explicit interventions, such as synthetic oversampling (SMOTE) or re-weighting algorithms **(Loureiro et al., 2025)**. The finding of this study reveals that the

Variational Quantum Classifier achieved fairness intrinsically on the raw, imbalanced dataset, without generating synthetic data. This suggests that the parameterized nature of variational circuits may possess an inherent robustness to class imbalance that classical algorithms lack, potentially offering a more reliable path to fairness that does not require on manipulating the underlying training data

6.2 Implications for HR Practitioners

For HR leaders, this research offers a conclusion that is both cautionary and cautious.

The Warning*:* If you are currently hiring with off-the-shelf classical ML (Logistic Regression, AutoML), you are likely systematically excluding high-potential minority talent, even if that is not your intention (Chen, 2023; Rigotti, 2024).

The Opportunity*:* Even in their current noisy version, quantum models provide an alternative way of thinking about data. They are able to end the stalemate of the imbalance between classes. Nevertheless, implementing them requires a strategic readiness to reduce total accuracy scores in exchange for more comprehensive talent discovery and regulatory compliance.

6.3 Limitations

Simulation vs. Hardware: This study executes on ideal simulators. Real quantum hardware introduces noise (readout error, gate error), which is a major challenge for all existing QML applications (Lamichhane & Rawat, 2024). It is possible that on actual hardware, the gains on "Fairness" could be lost to noise.

Dimensionality Reduction*:* We heavily compressed the data (7 features → 4 PCA components). The models could have been handicapped by this loss of information (47.91% of variance was discarded). Future work with larger qubit counts should test on full-feature datasets.

7 Conclusions and Future Work

7.1 Future Work

The next step for this research should take immediately is to move from "Passive Fairness" (observing the fairness of the model) to "Active Fairness". This involves modifying the objective function of the quantum circuit. Instead of merely minimizing error, we could specify a custom quantum loss function:

$$\text{Loss} = \text{Error} - \lambda(\text{Fairness})$$

This approach, known as in-processing bias mitigation, is an active area of research (Giuliani, 2024; Weerts et al., 2024). By incorporating the fairness constraint directly into the quantum optimization loop, we could be able to improve the accuracy of the VQC while maintaining its superior fairness profile.

7.2 Final Conclusion

This thesis set out to answer the following research question:

"What is the quantifiable fairness-accuracy trade-off of a Quantum-Fairness Recruitment Engine (Q-FRE) when using *variational quantum circuits (VQC, QNN)*, compared to *classical baselines (Logistic Regression, SVM, KNN)* and the *quantum kernel baseline (QSVM)* on imbalanced HR data?"

This research has provided this question conclusively. The "fairness-accuracy trade-off" is substantial and architecturally specific.

1. Classical and Quantum Kernel Baselines: All the classical models (LR, SVM, KNN) and the quantum kernel (QSVM) failed, as they had 0% minority recall in favor of high (84.7%) accuracy. They involve no trade-off, only systematic bias.
2. Variational Quantum Circuits: The variational models (VQC, QNN) provided a tangible trade-off. They sacrificed approximately 16 percentage points of accuracy to achieve a 29% minority recall. This enabled them to obtain legal equity levels (DIR > 0.80) that the other models could not.

The key finding is that quantum advantage to be fair is not due to "quantum-ness" alone (as the QSVM failure proves), but specifically due to the architectural flexibility of parameterized variational circuits. By quantifying this trade-off, this study provides the first empirical roadmap for organizations seeking to implement QML to ethical and compliant recruitment.

References

Chen, Z. (2023) ‘Ethics and discrimination in artificial intelligence-enabled recruitment tools’, *Humanities and Social Sciences Communications*, 10, article 213. Available at: https://doi.org/10.1057/s41599-023-02079-x

European Commission (2024) *High-level summary of the AI Act*. Available at: https://artificialintelligenceact.eu/high-level-summary/ [Accessed 27 October 2025].

Giuliani, L. (2024) ‘Long-Term Fairness Strategies in Ranking with Continuous Sensitive Attributes’, *CEUR Workshop Proceedings*, Vol-3808, Paper 11, pp. 125–136. Available at: https://ceur-ws.org/Vol-3808/paper11.pdf [Accessed 11 November 2025].

Ip, E. (2025) ‘Fair AI in hiring: Experimental evidence on how biased hiring algorithms and different debiasing methods affect the quality and diversity of applicants’, *Big Data & Society*, 11(1). https://doi.org/10.1177/23794607251353585

Köchling, A. and Wehner, M. C. (2020) ‘Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development’, *Business Research*, 13, pp. 795–848. https://doi.org/10.1007/s40685-020-00134-w

Lamichhane, P. and Rawat, D. B. (2024) ‘Quantum Machine Learning: Recent Advances, Challenges, and Perspectives’, *IEEE Access*, 12, pp. 11055–11072. https://doi.org/10.1109/ACCESS.2025.3573244

Loureiro, R. B., Pagano, T. P., Lisboa, F. V. N., Nascimento, L. F. S., Oliveira, E. L. S., Winkler, I. and Nascimento, E. G. S. (2025) ‘Correlation-based methods for representative fairness metric selection: An empirical study on efficiency and caveats in model evaluation’, *Expert Systems with Applications*, 268, 126344. https://doi.org/10.1016/j.eswa.2024.126344

Molnar, C. (2022) *Interpretable Machine Learning*. Available at: https://christophm.github.io/interpretable-ml-book/ [Accessed 27 October 2025].

O'Leary, T., Aaronson, S., Chuang, I. L. and Shor, P. W. (2025) ‘Efficient online quantum circuit learning with no upfront cost’, *Communications Physics*, 8, article 24. https://doi.org/10.1038/s42005-025-02423-4

Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S. and Stodtmann, S. (2024) ‘Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development’, *Clinical and Translational Science*, 17(11), e70056. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11513550/ [Accessed 25 October 2025].

Rigotti, C. (2024) 'Fairness, AI & Recruitment', *Computers in Human Behavior*, 152, 108196. https://doi.org/10.1016/j.clsr.2024.105966

Starke, C., Baleis, J., Keller, B. and Marcinkowski, F. (2022) 'Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature', *Big Data & Society*, 9(2). https://doi.org/10.1177/20539517221115189

U.S. Equal Employment Opportunity Commission (2024) *What is the EEOC's role in AI?* Available at: https://www.eeoc.gov/sites/default/files/2024-04/20240429_What%20is%20the%20EEOCs%20role%20in%20AI.pdf [Accessed 11 November 2025].

Weerts, H. J. P., Müller, M. and Vanschoren, J. (2024) 'Can Fairness be Automated? Guidelines and Opportunities for Automated Fairness-Aware Machine Learning', *Journal of Artificial Intelligence Research*, 79, pp. 1689–1717. https://doi.org/10.1613/jair.1.14747