

People's Democratic Republic of Algeria

Ministry of Higher Education
and Scientific Research



Ferhat Abbas Setif-1 University

Faculty of Sciences

Department of Computer Science

Master's Thesis submitted in partial fulfillment of the Master's Degree

Option: Quantum Computing

TOPIC

Quantum System for Counterfactual Explanation Generation in Financial AI Models

Prepared by:

Ines KALKOUL

Supervised by:

Pr.Tarek NOUIOUA

Examination Committee (Jury):

Pr. Khadidja Harbouche

Pr. Safia Djemame Zazoua

Academic Year: 2025/2026

Abstract

Machine learning models are increasingly used in financial institutions to automate critical decisions such as credit scoring, fraud detection, and risk assessment. While these models achieve strong predictive performance, their black-box nature raises serious concerns regarding transparency and accountability, particularly in light of regulations such as GDPR Article 22 and MiFID II, which require that automated decisions affecting individuals be explainable and contestable.

Existing explainability methods such as LIME and SHAP identify which features influenced a decision but do not indicate what should be changed to obtain a different outcome. Counterfactual explanations address this gap by identifying the minimal modification to an applicant's profile that would flip a refused decision to an approved one. However, classical counterfactual generation methods rely on random or gradient-based search, which provides no guarantee of finding a minimal, realistic solution to this inherently combinatorial problem.

This thesis proposes a hybrid quantum-classical framework for counterfactual explanation generation in financial decision systems. A Gradient Boosting Machine serves as the black-box credit-scoring oracle. A Variational Quantum Circuit, distilled from this oracle, provides a parameter-efficient validity scorer. The Quantum Approximate Optimization Algorithm then searches the combinatorial space of candidate feature modifications, guided by a cost Hamiltonian that structurally encodes regulatory constraints, including GDPR-mandated attribute immutability.

Evaluated on the German Credit dataset under a stratified, unbiased protocol, the proposed approach achieves an 11.1 % improvement in proximity over DiCE, the strongest classical baseline, and a 48.2 % improvement over a non-quantum ablation, while maintaining 100 % structural regulatory compliance. The framework is further validated under simulated NISQ hardware noise, supporting its near-term practical feasibility.

Keywords: explainable AI, counterfactual explanations, quantum computing, QAOA, variational quantum circuit, GDPR.

Résumé

Les modèles d'apprentissage automatique sont de plus en plus utilisés par les institutions financières pour automatiser des décisions critiques telles que le scoring de crédit, la détection de fraude et l'évaluation des risques. Bien que ces modèles atteignent de bonnes performances prédictives, leur nature de boîte noire soulève de sérieuses préoccupations en matière de transparence et de responsabilité, notamment au regard de réglementations telles que l'article 22 du RGPD et la directive MiFID II, qui exigent que les décisions automatisées affectant les individus soient explicables et contestables.

Les méthodes d'explicabilité existantes telles que LIME et SHAP identifient les caractéristiques ayant influencé une décision, mais n'indiquent pas ce qui devrait être modifié pour obtenir un résultat différent. Les explications contre-factuelles répondent à cette limite en identifiant la modification minimale du profil d'un demandeur qui permettrait de transformer une décision de refus en décision d'approbation. Cependant, les méthodes classiques de génération de contre-factuels reposent sur une recherche aléatoire ou par descente de gradient, qui n'offre aucune garantie de trouver une solution minimale et réaliste à ce problème intrinsèquement combinatoire.

Cette thèse propose un cadre hybride quantique-classique pour la génération d'explications contre-factuelles dans les systèmes de décision financière. Une machine à gradient boosting sert d'oracle de scoring de crédit en boîte noire. Un circuit quantique variationnel, distillé à partir de cet oracle, fournit un évaluateur de validité économe en paramètres. L'algorithme QAOA explore ensuite l'espace combinatoire des modifications de caractéristiques candidates, guidé par un hamiltonien de coût qui encode structurellement les contraintes réglementaires, notamment l'immutabilité des attributs imposée par le RGPD.

ملخص

تُستخدم نماذج التعلم الآلي على نطاق واسع في المؤسسات المالية لأتمتة القرارات الحساسة مثل تقييم القروض، واكتشاف الاحتيال، وتقييم المخاطر. وعلى الرغم من أن هذه النماذج تحقق أداءً تنبؤيًا قويًا، فإن طبيعتها الغامضة، (الصندوق الأسود) تثير مخاوف جدية بشأن الشفافية والمساءلة، لا سيما في ظل تشريعات مثل المادة 22 من اللائحة العامة لحماية البيانات (GDPR) وتوجيه MiFID II، التي تتطلب أن تكون القرارات الآلية المؤثرة على الأفراد قابلة للتفسير والطعن.

تحدد طرق التفسير الحالية مثل LIME وSHAP السمات التي أثرت في القرار، لكنها لا تشير إلى ما ينبغي تغييره للحصول على نتيجة مختلفة. تعالج التفسيرات المضادة للواقع هذه الثغرة من خلال تحديد أدنى تعديل ممكن في ملف طالب القرض، يمكن أن يحوّل قرار الرفض إلى قرار بالموافقة. ومع ذلك، تعتمد الطرق الكلاسيكية لتوليد هذه التفسيرات على البحث العشوائي أو طرق التدرج، والتي لا تضمن إيجاد حل أدنى وواقعي لهذه المشكلة التركيبية بطبيعتها.

تقترح هذه الأطروحة إطارًا هجينًا كمي-كلاسيكي لتوليد التفسيرات المضادة للواقع في أنظمة القرار المالي. يُستخدم نموذج Gradient Boosting كنموذج صندوق أسود لتقييم القروض. وتوفر دائرة كمية تغايرية (VQC)، مُستخلصة من هذا النموذج، مُقيمًا للصلاحية فعالاً من حيث عدد المعاملات. ثم تقوم خوارزمية QAOA بالبحث في الفضاء التركيبي للتعديلات المحتملة للسمات، بالاستناد إلى هاميلتونانٍ تكلفة يُدمج بنيويًا القيود التنظيمية، بما هلب، بما في ذلك ثبات السمات المحمية التي تفرضها اللائحة العامة لحماية البيانات.

وبعد التقييم على مجموعة بيانات German Credit وفق بُروتوكول طبقي وغير منحاز، حقق النهج المقترح تحسنًا بنسبة 11.1%: في معيار التقارب مقارنة بـDiCE: وهو أقوى نموذج كلاسيكي مرجعي، وتحسنًا بنسبة 48.2% مقارنة بنسخة معدّلة بدون مكوّن كمي، مع الحفاظ على امتثال تنظيمي بنيوي بنسبة 100%. كما تم التحقق من صلاحية الإطار المقترح تحت محاكاة لتشويش الأجهزة الكمية الحالية (NISQ).

الكلمات المفتاحية: الذكاء الاصطناعي القابل للتفسير، التفسيرات المضادة للواقع، الحوسبة الكمية، QAOA، الدائرة الكمية التغايرية، اللائحة العامة لحماية البيانات.

Contents

Part I: Theoretical Foundations & State of the Art **12**

1 Introduction . . . 13

2 Financial Regulations and the Need for XAI . . . 14

2.1 GDPR Article 22 — Right to Explanation . . . 14

2.2 MiFID II — Financial Decision Transparency . . . 14

2.3 Basel III — Risk Model Auditability . . . 14

3 Counterfactual Explanations . . . 15

3.1 Definition and Core Idea . . . 15

3.2 The Four Quality Criteria . . . 15

4 Related Work . . . 16

4.1 Classical XAI Methods and Their Limits . . . 16

4.2 The Shared Root Cause and the Quantum Opportunity . . . 19

4.3 XAI in the Financial Domain . . . 20

4.4 Quantum Approaches to Finance . . . 20

5 Conclusion . . . 22

Chapter 2: Background on Classical and Quantum Machine Learning **23**

1 Introduction . . . 23

2 Classical Deep Learning . . . 24

2.1 Artificial Neural Networks . . . 24

2.2 Multi-Layer Perceptron . . . 24

2.3 Activation Functions . . . 25

2.4 Loss Functions and Training Objective . . . 26

2.5 Optimisation: Adam . . . 28

3 Quantum Machine Learning Foundations . . . 29

3.1 Quantum Data Encoding . . . 29

3.2 Parameterised Quantum Circuits . . . 32

3.3 Variational Quantum Classifier . . . 34

3.4 Variational Quantum Eigensolver . . . 37

3.5 Quantum Approximate Optimisation Algorithm . . . 37
3.6 Comparison of VQE and QAOA . . . 38
3.7 Parameter-Shift Rule for Gradient Computation . . . 39
4 The Barren Plateau Problem . . . 40
5 Quantum Tools: Qiskit . . . 40
6 Types of Quantum Computers and Simulation Environments . . . 41
6.1 The NISQ Era: Concepts and Limitations . . . 42
6.2 Qiskit Aer: Classical Simulation of Quantum Circuits . . . 43
6.3 IBM Quantum Processors . . . 44
6.4 Comparison of Execution Environments . . . 45
6.5 The Quantum Simulation Boundary . . . 46
7 Conclusion . . . 47

Part II: Conception & Implementation 48

1 Dataset, Experimental Protocol and Black-Box Model 49
1 Introduction . . . 49
2 Dataset . . . 50
2.1 Source and Acquisition . . . 50
2.2 Feature Space . . . 51
2.3 Class Distribution and Imbalance Analysis . . . 52
3 Experimental Protocol . . . 52
3.1 Preprocessing Pipeline . . . 52
3.2 Train/Test Split . . . 53
3.3 Evaluation Metrics and Validation . . . 53
3.4 Baseline Model Specification . . . 54
3.5 Implementation Overview . . . 55
3.6 Feature Sensitivity Analysis . . . 56
4 Conclusion . . . 56

2 Hybrid MLP–VQC Model 58
1 Introduction . . . 58
2 Scientific Framing and Evaluation Criteria . . . 58
2.1 The Correct Scientific Question . . . 58
3 Hybrid Architecture . . . 59
3.1 Design Rationale . . . 59
3.2 Three-Phase Construction . . . 60
3.3 Full Architecture Specification . . . 60
3.4 Backbone Design Choices . . . 60
3.5 Angle Encoding Strategy . . . 61

3.6 Variational Ansatz . . . 62
3.7 Measurement Scheme . . . 62
3.8 Gradient Computation: Parameter-Shift Rule . . . 62
4 Training Protocol . . . 63
4.1 Phase 1: Classical Backbone Pre-training . . . 63
4.2 Phase 2: Backbone Freezing and Embedding Precomputation . . . 65
4.3 Phase 3: VQC Training on Frozen Embeddings . . . 66
5 Experimental Evaluation . . . 67
5.1 Performance Comparison . . . 67
5.2 Per-Class Performance . . . 68
5.3 Statistical Significance of the AUC Gap . . . 68
5.4 Quantum Contribution Analysis . . . 68
6 Discussion . . . 69
6.1 Why the VQC Has Higher Accuracy but Lower AUC than the Classical MLP . . . 69
6.2 Why the Separation Matters Scientifically . . . 69
6.3 Role in the Downstream Pipeline . . . 70
7 Conclusion . . . 70
1 Introduction . . . 71
2 Problem Formulation . . . 71
2.1 Counterfactual Explanation . . . 71
2.2 QUBO Encoding of Feature Selection . . . 72
2.3 Evaluation Metrics . . . 73
3 Hybrid QAOA–COBYLA Pipeline . . . 73
3.1 Phase 1 — QAOA Feature Mask Selection . . . 73
3.2 Phase 2 — COBYLA Magnitude Refinement . . . 74
3.3 Feasibility Verification . . . 75
3.4 NISQ Noise Model . . . 76
4 Discussion . . . 76
5 Limitations and Future Work . . . 80
6 Conclusion . . . 80