Recall-Oriented Credit Card Fraud Detection with Gradient-Boosted Trees and Hybrid Quantum Neural Networks: Simulation, Grover Small-Batch Amplification, and IQM Garnet-Targeted Shot-Based Inference

Superpositions Studio

February 17, 2026

Abstract

Credit-card fraud detection is a high-stakes, extremely imbalanced classification problem in which operational success is often driven by the ability to capture as many fraudulent transactions as possible, potentially at the cost of increased false positives. Using the widely studied European cardholders dataset (284,807 transactions, 0.173% fraud) [1, 2], we develop an end-to-end recall-first pipeline comprising exploratory data analysis (EDA), a classical gradient-boosting baseline with class weighting and threshold tuning, and a hybrid quantum neural network (HQNN) built from a trainable variational quantum circuit (VQC). Thresholds are tuned on a validation split to target 95% fraud recall. On the held-out test set, the gradient-boosting baseline achieved recall 0.901 with precision 0.020 (PR-AUC 0.717), while the HQNN achieved recall 0.930 with precision 0.0033 (PR-AUC 0.712), highlighting the recall–precision trade-off under extreme imbalance. We further demonstrate Grover-style amplitude amplification on a toy batch of $N = 16$ candidate transactions (simulation only), using HQNN scores to define the marked set. Finally, we execute shot-based HQNN inference for 64 test transactions at 1000 shots per circuit using an IQM Garnet-targeted backend configuration. The shot-based scores remained highly consistent with ideal simulation (Pearson $r \approx 0.997$).

1 Introduction

Fraud detection in payment systems is typically formulated as binary classification with asymmetric costs: missing fraud is expensive, while false alarms primarily impose manual review costs and customer friction. For such problems, accuracy can be misleading under class imbalance, and evaluation should emphasize recall, precision, and precision–recall (PR) curves [3–5]. Recent progress in quantum machine learning (QML) has motivated investigating hybrid quantum–classical models—including variational quantum classifiers—as expressive function approximators that can be trained with classical optimizers [6–8]. In this work we study a recall-oriented screening setting on a standard fraud dataset, comparing a strong classical baseline against an HQNN, and augmenting the HQNN score with a small-batch Grover amplitude amplification demonstration [9].

2 Dataset and exploratory analysis

2.1 Dataset provenance and basic properties

We use the commonly studied anonymized European cardholders dataset [1, 2]. It consists of numerical transaction descriptors ("Time", "Amount", and anonymized components $V1$–$V28$) and a binary target `Class` indicating fraud.

The dataset contains 284,807 transactions and 31 columns (30 features plus the label). There are no missing values; however, a small fraction of duplicated rows is present and is removed prior to splitting to reduce potential leakage effects.

2.2 Class imbalance and feature characteristics

Figure 1 illustrates the extreme class imbalance. In addition, `Amount` is heavy-tailed, motivating a $\log(1 + \texttt{Amount})$ transform before scaling. Table 1 summarizes key statistics of `Time` and `Amount`.

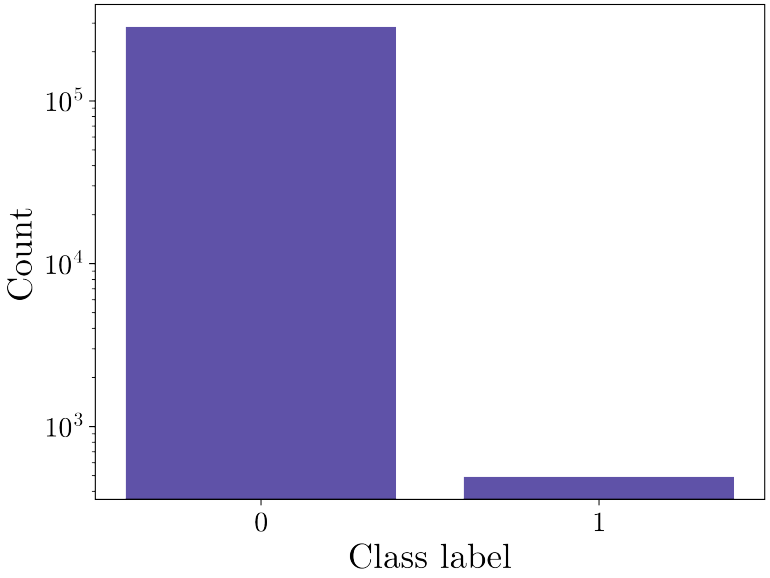


Figure 1: Class distribution (log scale) showing the extreme rarity of fraudulent transactions.

Table 1: Summary statistics for `Time` and `Amount`.

Feature	Mean	Std	Min	1%	Median	99%	Max	Skew	Kurt.
Time	94813.86	47488.15	0.00	2422.00	84692.00	170560.94	172792.00	-0.04	-1.29
Amount	88.35	250.12	0.00	0.12	22.00	1017.97	25691.16	16.98	845.09

To support feature selection under qubit constraints, we compute univariate relevance measures (Pearson correlation, mutual information, and standardized mean difference). Table 2 lists the strongest signals, which consistently include $V17$, $V14$, $V12$, $V10$, $V11$, and $V16$.

Table 2: Top univariate signals for fraud detection (correlation ρ, mutual information I, and effect size $|d|$).

Feature	$\rho(X, y)$	$I(X; y)$	$\lvert d\rvert$
V17	-0.326	0.0084	8.32
V14	-0.303	0.0085	7.64
V12	-0.261	0.0081	6.50
V10	-0.217	0.0078	5.35
V16	-0.197	0.0065	4.83
V3	-0.193	0.0055	4.74
V7	-0.187	0.0043	4.59
V11	+0.155	0.0071	3.78
V4	+0.133	0.0055	3.24
V18	-0.111	0.0041	2.70
V1	-0.101	0.0021	2.45
V9	-0.098	0.0050	2.36

3 Methods

3.1 Evaluation protocol and recall-first threshold tuning

We use a stratified random split into train/validation/test partitions (70/15/15). Because the operational objective is recall-first screening, decision thresholds are tuned on the validation set to achieve a target recall of 0.95. Metrics reported include recall, precision, F1, ROC-AUC, and PR-AUC [4, 5, 10].

3.2 Quantum components and scope of claims

The HQNN and Grover parts serve different purposes. The HQNN is a variational classifier: a parameterized quantum circuit produces measurement statistics that are optimized by a classical optimizer against a supervised loss [7, 8]. This is a modeling choice and not, by itself, a complexity-theoretic speedup claim.

Grover amplification is used only as a toy ranking primitive on a small candidate batch where the marked set is defined by HQNN scores above threshold. Although Grover offers $O(\sqrt{N})$ oracle-query complexity in unstructured search, practical end-to-end advantage for fraud screening also depends on data loading, oracle construction, and hardware noise assumptions that are outside the scope of this run [9, 11]. We therefore treat this component as an illustrative mechanism, not a deployment-ready accelerator.

3.3 Classical baseline: gradient-boosting with class weighting

As a classical reference, we use gradient-boosted decision trees (GBDTs) [12–14]. In the execution environment, an XGBoost implementation was unavailable; consequently, a histogram-based boosting variant from scikit-learn [15] was used with per-sample weights computed from balanced class weights. All numeric features are used, with `Amount` replaced by `Amount_log1p`.

3.4 Hybrid quantum neural network (HQNN)

3.4.1 Model structure

The HQNN consists of a small classical preprocessing stage followed by a trainable variational quantum circuit (VQC) acting as a nonlinear feature transformation, and a classical readout

for binary classification [6–8]. Let $\mathbf{x} \in \mathbb{R}^d$ be the selected features after standardization. We map them to bounded angles using

$$\boldsymbol{\phi}(\mathbf{x}) = \pi \tanh(\mathbf{x}), \tag{1}$$

which constrains rotations to $[-\pi, \pi]$. The VQC applies an angle-encoding feature map followed by L layers of trainable single-qubit rotations and entangling gates,

$$|\psi(\mathbf{x}, \boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})\, U_{\text{enc}}(\boldsymbol{\phi}(\mathbf{x}))\, |0\rangle^{\otimes n}. \tag{2}$$

A single expectation value $s(\mathbf{x}) = \langle\psi|O|\psi\rangle$ is measured (Pauli-Y basis) and converted to a logit for training.

3.4.2 Imbalance handling and optimization

To address imbalance, we use a weighted binary cross-entropy loss with `pos_weight` proportional to the negative-to-positive ratio [16]. Additionally, the training set is downsampled to retain all positive examples and a limited negative-to-positive ratio. Optimization is performed with Adam [17] using separate learning rates for classical and quantum parameters.

3.4.3 Chosen configuration

Guided by EDA and qubit constraints, we use $n = 6$ qubits with depth $L = 2$ and a strong entangling pattern. One feature is mapped per qubit using the selected set $\{V17, V14, V12, V10, V11, V16\}$. We avoid Z-axis-only encoding and avoid measuring in a basis aligned with the encoding axis.

3.5 Grover small-batch amplitude amplification (demonstration)

Grover's algorithm [9] can amplify the probability of measuring marked items in an unstructured set of size N using $O(\sqrt{N})$ oracle calls. Because scalable data loading would require additional assumptions (e.g., QRAM), we use Grover only as a small-batch demonstration on $N = 16$ candidate transactions (4 index qubits), where the marked set is derived from HQNN scores (indices whose scores exceed the tuned threshold). This is intended as an illustrative "prioritization" step rather than a claim of end-to-end quantum speedup.

4 Experimental results (simulation)

4.1 Split statistics and preprocessing

After removing duplicates, 283,726 transactions remain. The stratified split contains 331, 71, and 71 fraud cases in train, validation, and test splits, respectively.

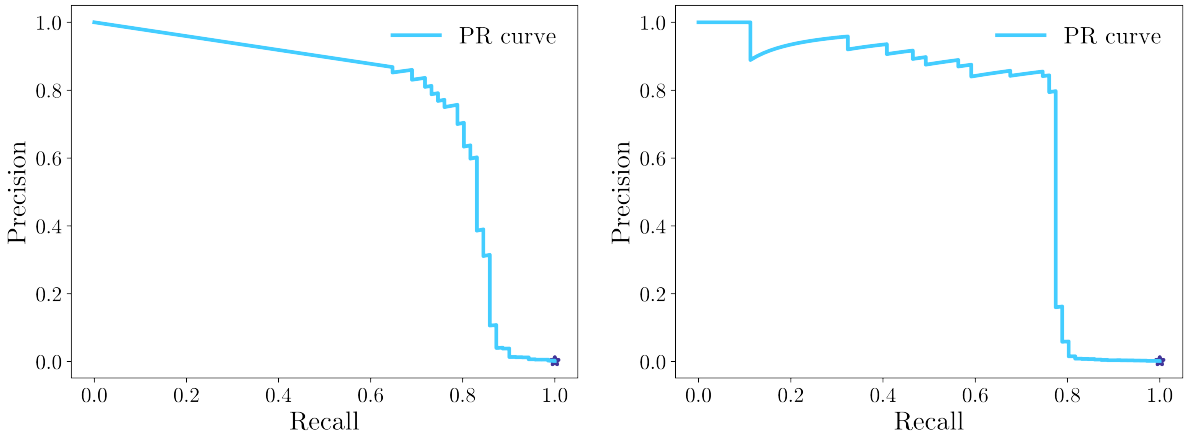
4.2 Test-set performance and trade-offs

Table 3 summarizes recall-first performance on the held-out test set at the tuned thresholds. The HQNN yields slightly higher recall than the GBDT baseline but at substantially higher false positive rate, resulting in much lower precision.

Table 3: Test-set results at validation-tuned thresholds (target recall 0.95).

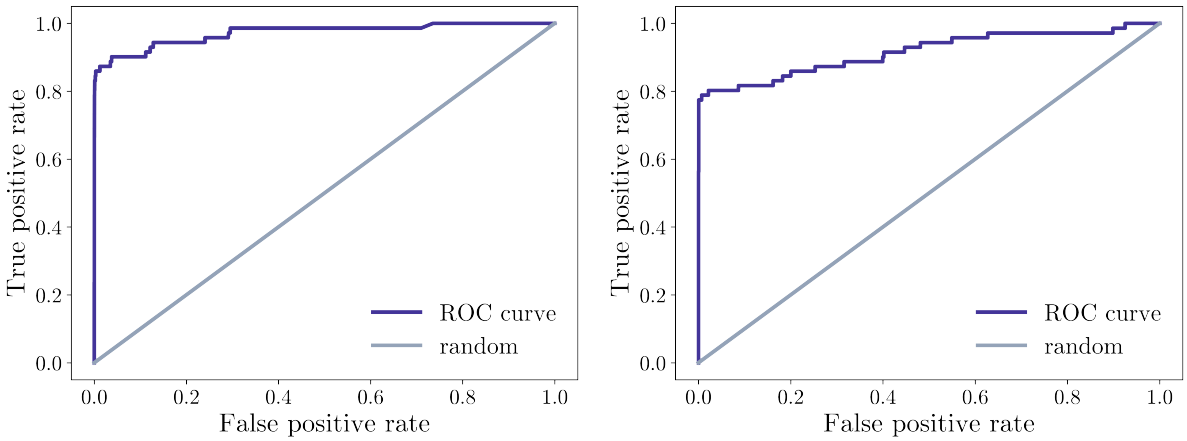
Model	Threshold	Recall	Precision	PR-AUC	ROC-AUC
GBDT baseline	0.01543	0.90141	0.02021	0.71750	0.97169
HQNN (ideal sim.)	0.68570	0.92958	0.00332	0.71164	0.91605

Figure 2 and Figure 3 show PR and ROC curves for both models.



(a) Baseline. (b) HQNN.

Figure 2: Precision–recall curves on the test set.



(a) Baseline. (b) HQNN.

Figure 3: ROC curves on the test set.

Because prevalence in the full test split is extremely low, confusion matrices at recall-tuned thresholds are visually dominated by true negatives and are less informative than threshold-dependent ranking diagnostics. We therefore focus comparison on Table 3 and the PR/ROC curves (Figures 2 and 3).

4.3 HQNN training dynamics and scaling considerations

The HQNN training converged with early stopping based on validation PR-AUC. Figure 4 reports the learning curve. Figure 5 provides a qualitative scaling view for the HQNN family; in practice, VQCs can face trainability issues such as barren plateaus that worsen with depth, noise, and system size [7, 18, 19].

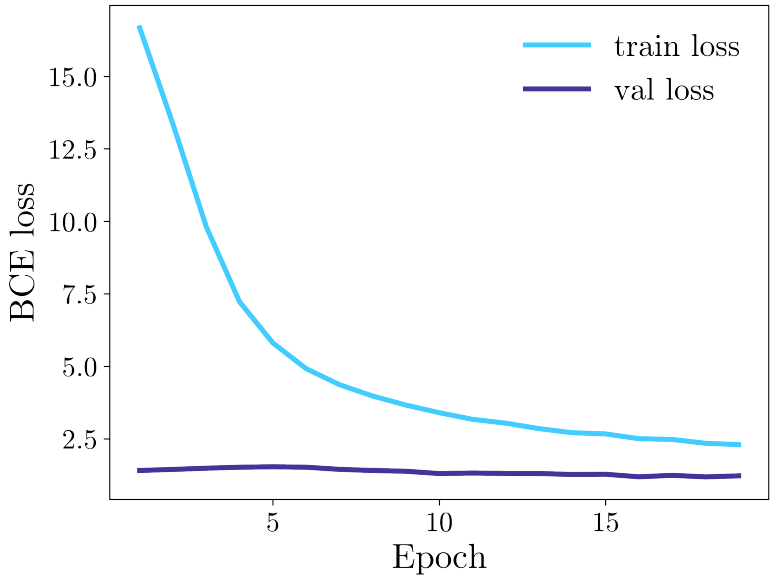


Figure 4: HQNN training curve (validation PR-AUC monitored for early stopping).

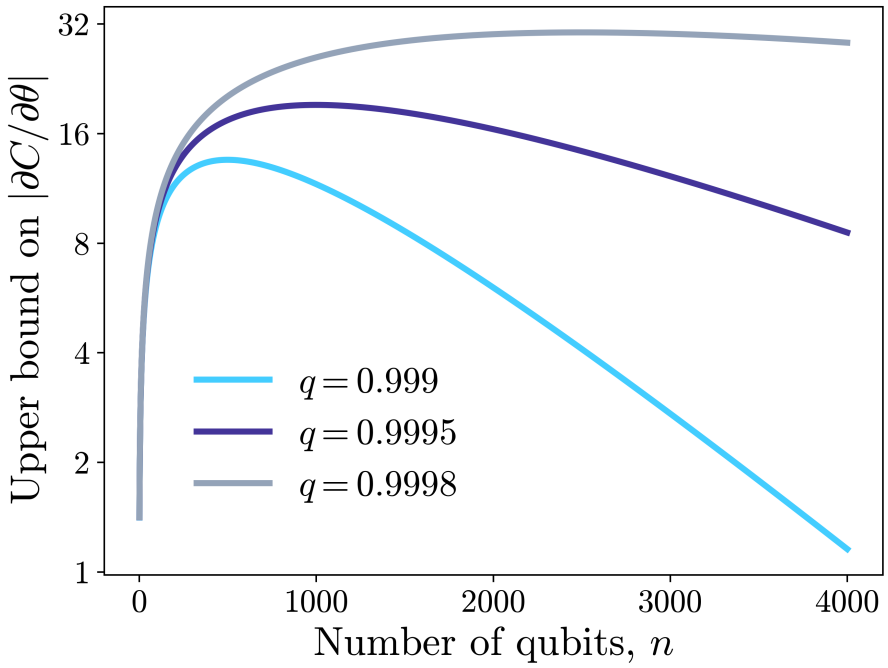


Figure 5: Illustrative scaling considerations for HQNN/VQC approaches (qualitative).

4.4 Grover demonstration results

For a toy batch of $N = 16$ candidate transactions (simulation), one Grover iteration was applied with $M = 6$ marked indices determined by HQNN scores. The resulting measurement distribution concentrates probability mass on the marked indices (Figure 6), consistent with amplitude amplification.

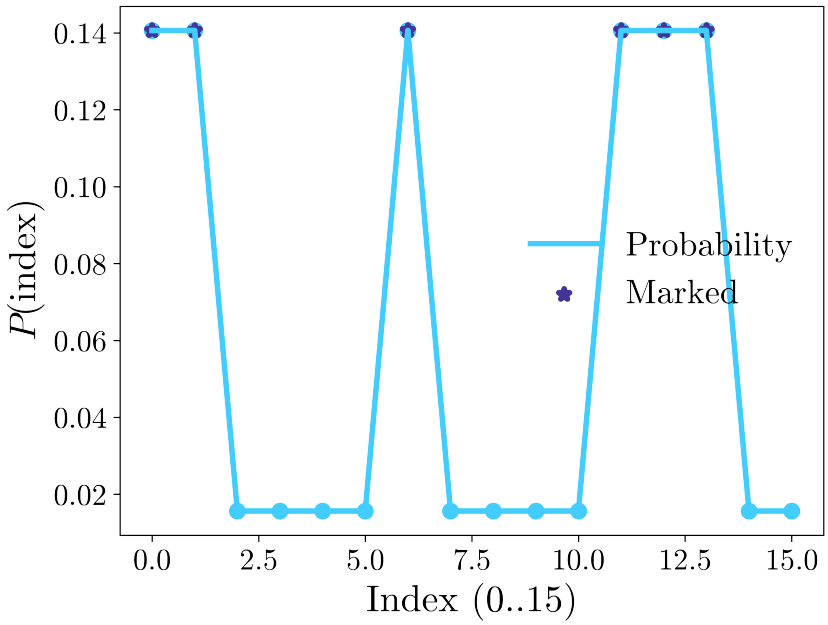


Figure 6: Grover small-batch demonstration ($N = 16$) amplifying indices marked by HQNN thresholding.

5 Shot-based inference on an IQM Garnet-targeted backend

5.1 Run configuration

To probe hardware-relevant effects such as shot noise and compilation to a specific device topology, we executed HQNN forward passes for 64 test-set transactions using 1000 shots per circuit on a backend configured to target IQM Garnet [20–22]. The subset was constructed to include both classes (16 fraud and 48 legitimate transactions). Circuits were executed in batches of 8, with representative transpiled depth 16 and 10 two-qubit CZ gates per circuit.

5.2 Backend setup and reproducibility note

This experiment uses an IQM Garnet-targeted, shot-based execution configuration to probe finite-shot inference behavior under device-aware compilation constraints. The results should be interpreted as a device-targeted noisy-study protocol, with direct comparability to ideal simulation reported below.

5.3 Consistency with ideal simulation

Table 4 reports subset-level classification metrics under the fixed threshold learned during validation. The shot-based backend and ideal simulation produced near-identical class decisions on this subset, and the continuous scores matched closely (Pearson correlation 0.9966, RMSE 0.0304).

Table 4: HQNN metrics on the 64-point subset under the fixed threshold (1000 shots per circuit).

Backend	Recall	Precision	PR-AUC	ROC-AUC
Ideal simulator	1.0000	0.3556	0.8599	0.9154
IQM Garnet-targeted (shot-based)	1.0000	0.3556	0.8581	0.9154

Subset confusion matrices are omitted to avoid over-interpreting cell-level fluctuations on a small, intentionally enriched sample.

6 Discussion

The experiments highlight a core operational tension for fraud screening under extreme imbalance. When thresholds are tuned to aggressively prioritize recall, both models produce many false positives; however, the HQNN produced a substantially larger false positive rate on the full test split, leading to very low precision. The classical GBDT baseline achieved stronger ROC-AUC and comparable PR-AUC, consistent with the effectiveness of tree-based ensembles on tabular data [12, 14].

The Grover experiment demonstrates how quantum amplitude amplification can concentrate probability mass on a marked subset, but it does not, by itself, deliver an end-to-end advantage for fraud detection because scalable access to the full transaction database in coherent superposition is a major bottleneck (data loading and QRAM assumptions). Consequently, we position Grover here strictly as a small-batch prioritization demonstration.

The IQM Garnet-targeted run indicates that, at least for this small circuit depth and shot budget, shot-based inference can closely match ideal expectations. In real hardware settings, additional factors (calibration drift, readout error, two-qubit gate infidelity) may degrade performance and may motivate mitigation strategies such as zero-noise extrapolation or probabilistic error cancellation [23–26].

7 Limitations and future work

This study is constrained by (i) the extreme imbalance, which makes precision highly sensitive to threshold choice and operational review capacity; and (ii) the reduced feature set required by the qubit budget. Future work should evaluate cost-based objectives tied to investigator capacity, apply calibration techniques [27, 28], explore time-aware splits to address non-stationarity and concept drift [29, 30], and consider explainability tooling such as SHAP for decision support [31].

8 Conclusion

We presented a recall-oriented fraud detection pipeline integrating EDA, a classical gradient-boosting baseline, an HQNN trained with imbalance-aware objectives, a Grover small-batch demonstration, and an IQM Garnet-targeted shot-based inference study. On the full test split, the HQNN achieved slightly higher recall than the baseline but at significantly lower precision. On a small, balanced-enriched subset, shot-based inference at 1000 shots closely matched ideal simulation. These results suggest that near-term quantum models can be integrated into fraud-screening workflows as experimental components, but classical ensembles remain highly competitive for tabular fraud detection under operational constraints.

References

[1] Credit card fraud detection dataset. Kaggle. URL https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud. Accessed: 2026-02-10.

[2] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, et al. Calibrating probability with undersampling for unbalanced classification. 2015. doi: 10.1109/SSCI.2015.33. URL https://www.semanticscholar.org/paper/e36bb7fbe1b4f7c521608e93a2215e2062dae5b1.

[3] Haibo He and E. A. Garcia. Learning from imbalanced data. 2009. doi: 10.1109/TKDE.2008.239. URL https://www.semanticscholar.org/paper/6a97303b92477d95d1e6acf7b443ebe19a6beb60.

[4] Jesse Davis and Mark H. Goadrich. The relationship between precision-recall and roc curves. 2006. doi: 10.1145/1143844.1143874. URL https://www.semanticscholar.org/paper/a883cacbc8f9b021b2a63f0453307855fa075d33.

[5] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. 2015. doi: 10.1371/journal.pone.0118432. URL https://www.semanticscholar.org/paper/904627c2d5a91ab8cb1b682e42f06f1ca192aea6.

[6] Jacob Biamonte, Peter Wittek, Nicola Pancotti, et al. Quantum machine learning. 2016. doi: 10.1038/nature23474. URL https://www.semanticscholar.org/paper/15eded04386a8982ccd5627bd1efe70bbf624c02.

[7] M. Cerezo, A. Arrasmith, R. Babbush, et al. Variational quantum algorithms. 2020. doi: 10.1038/s42254-021-00348-9. URL https://www.semanticscholar.org/paper/c1cf657d1e13149ee575b5ca779e898938ada60a.

[8] Edward Farhi and Hartmut Neven. Classification with quantum neural networks on near term processors. 2018. doi: 10.37686/qrl.v1i2.80. URL https://www.semanticscholar.org/paper/5240a3531cd2f628bfe23425b1dd8ff89c15dc34.

[9] Lov K. Grover. Quantum mechanics helps in searching for a needle in a haystack. 1997. doi: 10.1103/PhysRevLett.79.325. URL https://www.semanticscholar.org/paper/222291b43d1837f09b1fc5433969f4f569c7abc8.

[10] Tom Fawcett. An introduction to roc analysis. 2006. doi: 10.1016/j.patrec.2005.10.010. URL https://www.semanticscholar.org/paper/d40ee5dd758c525dfb9932d726bb4e844b7b8478.

[11] John Preskill. Quantum computing in the nisq era and beyond. 2018. doi: 10.22331/q-2018-08-06-79. URL https://www.semanticscholar.org/paper/f3d594544126e202dbd81c186ca3ce448af5255c.

[12] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. 2001. doi: 10.1214/AOS/1013203451. URL https://www.semanticscholar.org/paper/1679beddda3a183714d380e944fe6bf586c083cd.

[13] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. 2016. doi: 10.1145/2939672.2939785. URL https://www.semanticscholar.org/paper/26bc9195c6343e4d7f434dd65b4ad67efe2be27a.

[14] Guolin Ke, Qi Meng, Thomas Finley, et al. Lightgbm: A highly efficient gradient boosting decision tree. 2017. URL https://www.semanticscholar.org/paper/497e4b08279d69513e4d2313a7fd9a55dfb73273.

[15] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine learning in python. 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

[16] Charles Elkan. The foundations of cost-sensitive learning. 2001. URL https://www.semanticscholar.org/paper/7fed3e00be2bb09510f5f7cad7ac106e6c94a359.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014. URL https://www.semanticscholar.org/paper/a6cb366736791bcccc5c8639de5a8f9636bf87e8.

[18] Jarrod R. McClean, Sergio Boixo, Vadim Smelyanskiy, et al. Barren plateaus in quantum neural network training landscapes. 2018. doi: 10.1038/s41467-018-07090-4. URL https://www.semanticscholar.org/paper/d699e0958fe1d8a4c1d691765f7e11b823fa606f.

[19] Samson Wang, Enrico Fontana, M. Cerezo, et al. Noise-induced barren plateaus in variational quantum algorithms. 2020. doi: 10.1038/s41467-021-27045-6. URL https://www.semanticscholar.org/paper/a7da6cf3820d94f250349d658a09004a36b14ff0.

[20] Leonid Abdurakhimov, Janos Adam, Hasnain Ahmad, et al. Technology and performance benchmarks of iqm's 20-qubit quantum computer. 2024. URL https://www.semanticscholar.org/paper/f20a27475bce5b7ad83e4b34b9e67a47f03764d2.

[21] Philip Krantz, Morten Kjaergaard, Fei Yan, et al. A quantum engineer's guide to superconducting qubits. 2019. doi: 10.1063/1.5089550. URL https://www.semanticscholar.org/paper/2540f07e4a1c3b8e3ee48b60ce9bb3f13940ffe2.

[22] Morten Kjaergaard, Mollie Schwartz, Jochen Braumuller, et al. Superconducting qubits: Current state of play. 2019. doi: 10.1146/annurev-conmatphys-031119-050605. URL https://www.semanticscholar.org/paper/187849e6fad9b6e7818f2b9496e2ce305f108740.

[23] Kristan Temme, Sergey Bravyi, and Jay M. Gambetta. Error mitigation for short-depth quantum circuits. 2016. doi: 10.1103/PhysRevLett.119.180509. URL https://www.semanticscholar.org/paper/04976cb176d0c128e244c04215be13d27df2b5b1.

[24] Suguru Endo, Zhenyu Cai, Simon C. Benjamin, et al. Hybrid quantum-classical algorithms and quantum error mitigation. 2020. doi: 10.7566/jpsj.90.032001. URL https://www.semanticscholar.org/paper/11c1d92eb28c20e7e27275a4a94e8bc21accd900.

[25] Abhinav Kandala, Kristan Temme, Antonio D. Córcoles, et al. Error mitigation extends the computational reach of a noisy quantum processor. 2018. doi: 10.1038/s41586-019-1040-7. URL https://www.semanticscholar.org/paper/0473787f881e9fac30f5809d9e92ddcad8eb1f3c.

[26] Ziwen Cai, Ryan Babbush, Simon C. Benjamin, et al. Quantum error mitigation. 2022. doi: 10.1103/RevModPhys.95.045005. URL https://www.semanticscholar.org/paper/c09d68b843d76ba928b483209187a1986cec125f.

[27] Chuan Guo, Geoff Pleiss, Yu Sun, et al. On calibration of modern neural networks. 2017. URL https://www.semanticscholar.org/paper/d65ce2b8300541414bfe51d03906fca72e93523c.

[28] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. 1999. URL https://www.semanticscholar.org/paper/42e5ed832d4310ce4378c44d05570439df28a393.

[29] João Gama, Indė Žliobaitė, Albert Bifet, et al. A survey on concept drift adaptation. 2014. doi: 10.1145/2523813. URL https://www.semanticscholar.org/paper/e1543c7ed8adb978bc2f9efd30f36a3bd1f91793.

[30] Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, et al. Credit card fraud detection and concept-drift adaptation with delayed supervised information. 2015. doi: 10.1109/IJCNN.2015.7280527. URL https://www.semanticscholar.org/paper/b769ec8f3be051b7c5f9a0ba5f85a38594950df9.

[31] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. 2017. URL https://www.semanticscholar.org/paper/442e10a3c6640ded9408622005e3c2a8906ce4c2.