

Comparative Analysis of Quantum Kernel Methods and Deep Neural Networks for Heart Disease Prediction

Md Abdullah Enayet ^{1 *}, Shakila Sultana ², Mst Anupa Nasrin Khan ³, Tanaya Jakir ⁴, Masrufa Tasnim ⁵, Abidul Alam ⁶, ALAUDDIN ⁷ and Md Takowa Ullah ⁸

¹ Master of Science in Cybersecurity, Washington University of Science and Technology, USA.

² Master's in Accounting (M.A.), National University, Bangladesh.

³ Master of Science in Zoology (Entomology Focus), The University of Rajshahi, Bangladesh.

⁴ MS in Business Analytics, Trine University, USA.

⁵ Master of Business Administration, University of Dhaka, Dhaka, Bangladesh.

⁶ Master of Science in Information Technology, Washington University of Science and Technology, USA.

⁷ B.Sc. in Computer Science and Engineering, Department of Computer Science and Engineering, Daffodil International University, Ashulia, Dhaka, Bangladesh.

⁸ Diploma in Computer Science, Bogura Polytechnic Institute, Bogura, Bangladesh.

¹ menayet@student.wust.edu, ² sultanashakila6@gmail.com, ³ anupa516@hotmail.com, ⁴ tanayajakir@gmail.com, ⁵ mtreema@gmail.com, ⁶ aabidul@student.wust.edu, ⁷ alauddin23ala@gmail.com, ⁸ takowa830@gmail.com

ABSTRACT

Relative to the proposed methods for better clinical prediction on small tabular cohorts, quantum kernel methods are often proposed, but published comparisons rarely control input dimensionality, hyperparameter search budget or even the classical kernel baseline that a fidelity kernel directly replaces. We present a combined pre-registered style comparison of quantum kernel methods, deep neural networks, and five classical baseline models on three different cardiac cohorts (UCI datasets combined, $n=917$; Cleveland: $n=297$; Z-Alizadeh Sani: $n=303$), under an identical single protocol including 55 repeated stratified nested cross-validation, fold-fitted preprocessing and dimensionality reduction with consistent eight-dimensional input subspace across compared models, equal search time budget of 20 candidates per model. The quantum kernel did not outperform the radial-basis-function support vector machine (its direct classical analogue) in any of the three cohorts ($AUC=-0.006$, -0.014 , -0.014), and ℓ_2 -regularised logistic regression reached the highest mean ROC-AUC across 84 pairwise comparisons on ROC-AUC compared to other methods for each of two out of three cohorts, with 29 surviving Holm correction, and only one exceeding the standard deviation at fold-level. A sample-efficiency analysis finds the crossover where the network surpasses the quantum kernel at $n \approx 93$ and $n \approx 123$ patients for both UCI cohorts, lower than any clinically viable trial size. We also demonstrate that amplitude encoding - the best-performing quantum configuration whose fidelity kernel analytically reduces to a squared cosine similarity and is therefore not, functionally, a quantum feature map - actually yields exponentially concentrating operators directly measurable: off-diagonal Gram variance decreases from 1.06×10^{-2} for four qubits to 2.22×10^{-4} at ten (the second-order Pauli map yielding $MCC=0.000$ in one cohort). The collapse of the model is detected by threshold-dependent metrics long before the ROC-AUC. With batched state vector construction, Gram assembly is reduced from $O(n^2)$ circuit executions to $O(n) +$ dense product, and the complete 25-fold protocol is feasible. Our conclusions are that there is no tangible benefit of using quantum kernel methods for clinical tabular prediction at the cohort sizes and qubit budgets accessible today, and we highlight the diagnostics for any future such claim.

KEYWORDS: Quantum Machine Learning; Quantum Kernel Methods; Deep Neural Networks; Heart Disease Prediction; Exponential Concentration; Clinical Decision Support; Negative Results.

How to Cite: Md Abdullah Enayet , Shakila Sultana, Mst Anupa Nasrin Khan, Tanaya Jakir , Masrufa Tasnim, Abidul Alam , ALAUDDIN, Md Takowa Ullah, (2025) Comparative Analysis of Quantum Kernel Methods and Deep Neural Networks for Heart Disease Prediction, Vascular and Endovascular Review, Vol.8, No.10s, 394-409

INTRODUCTION

Over the last three decades, automated stratification of routinely collected clinical variables from referral populations has been a long-standing area for supervised learning, with coronary artery disease remaining the single largest cause of mortality globally. The canonical benchmark cohorts such as the Cleveland subset of the UCI Heart Disease collection[1] and Z-Alizadeh Sani cohort from Tehran[2] are small, on the order of a few hundred patients and low-dimensional, on the order of ten to fifty predictors. This very regime, in which statistical arguments for quantum kernel methods have been optimally made[3, 4], is one of limited data where a powerful kernel that cannot be efficiently estimated by classical means can extract structure less effectively captured by classical approaches.

That argument has been formalised. Huang et al. A quantum advantage in kernel learning occurs only if the target function fits into the geometry that a quantum kernel induces, and access to data greatly limits which problems can have a separation[6]. Kübler et al. We referred to [7] for interest in the inductive bias of quantum kernels and demonstrated that expressivity without alignment is a liability, not an asset. Making the most important implications, Thanasilp et al. An

example of this is [8], which showed that a wide class of fidelity kernels concentrates exponentially in the number of qubits: off diagonal entries of the Gram matrix converge to a constant as the register grows, and it can be shown that in this case, the kernel matrix approaches the identity leading to suboptimal classification performance, wherein our classifier can only memorise its training set.

Grounded in this theory, the applied literature on quantum machine learning for cardiac prediction has rapidly but largely optimistically grown. There is a pattern here that provides further help in weakening the conclusions it invites. We show that benchmarks are often carried out on a single train–test split rather than on a repeated cross-validated protocol; the raw feature set is provided to one model class and the raw features reduced to another class (or vice versa) so that any difference confounds model class with input dimensionality; hyperparameter search budgets are either not equalised or even specified; and comparisons often neglect to report performance of radial-basis-function support vector machine, which in fact represents the classical model which can be seen as a fidelity kernel operator [25]. In its absence, a reported victory over a neural network does not differentiate a quantum effect from the generic strength of kernel methods in small-sample regimes that classical kernels already have [12, 13].

It aims to purify those confounds and report on what it finds. A fixed evaluation protocol is used identically for quantum kernels, deep networks and five classical baselines over three cardiac cohorts. All compared methods are provided the same 8-dimensional subspace fitted on folds, every fold used in cross-validation and the number of hyperparameter candidates. We then pose three questions in that order: does the quantum kernel win; if not, at what sample size would it; and what attribute of the kernel drives the result.

1. Our contributions are as follows.
2. A confound-controlled comparison. Eight models on three cohorts <55 reimplemented with repeated stratified nested cross-validation using fold-fitted preprocessing, constant input subspaces, and identical search budget of 20 candidates per model reported with Wilcoxon signed-rank tests, Holm correction, effect sizes and Nemenyi critical-difference diagrams
3. Bootstrapped Crossover Sample-Efficiency Analysis Instead of qualitatively saying that the low-data advantage exists, we find the training-set size at which the deep finally outperforms quantum kernel, and provide a bootstrap confidence interval for this point.
4. A mechanistic diagnosis. We quantify kernel-target alignment, Gram eigenspectra, effective rank and off-diagonal variance demonstrating the exponential-concentration pathology is manifest at qubit counts down to eight and that it shows in threshold-dependent metrics long before ROC-AUC registers it.
5. Analytical caveat – reframing the top-line result. The quantum encoding which performed best, amplitude encoding, has a fidelity kernel whose closed form reduces to a positive-scaled squared cosine similarity, i.e. a classical kernel. This reduction is made explicit and we report the configuration as such.
6. A tractable construction. We create Gram matrices from batched statevectors, which converts On2 circuits to On + a single dense product, and we demonstrate that our scaling enables a 25-fold protocol over four qubit budgets on one workstation.

RELATED WORK

Quantum kernels

Schuld and Killoran[4], Havlíček et al. As a first step toward performance exploration, [3] framed feature maps into some Hilbert space and specified an induced kernel given by the fidelity between encoded states using parameterised quantum circuits. Subsequent work by Schuld[5] showed that such supervised quantum models are kernel methods in the classical sense, situating them firmly inside statistical learning theory and rendering the relevant question empirical over architectural: does the resulting kernel agree with the target?

Tailnegative results on this line are now a well beanpole. A concentration result of Thanasilp et al. [8], Cerezo et al. [9] and Wang et al. Section 4. [10] developed barren-plateau and noise-induced-plateau phenomena that generally limit the scaling of circuit-based models in practice. These limitations have not yet seeped sufficiently into the applied literature, and empirical papers presenting quantum advantage on small clinical tables almost never report diagnostics that would separate a true kernel effect from an artifact of protocol.

Deep learning on clinical tables

There is a parallel work showing that deep networks are not automatically better on tabular data. 11The detailed proofs can be found in the extra material and are mainly based on some ideas from [12, 13]. Gorishniy et al. [13] wanted to assess how competitive gradient-boosted trees were with or without specialised tabular architectures, particularly at small sample sizes; It is worth noting that [14] proposed stronger deep baselines, but we arrive at similar conclusions. This means that a deep network is not automatically a worse opponent than this regime which makes classical kernel and linear baselines indispensable rather than ornamental.

Evaluation practice with small samples

Varoquaux[15] and Vabalas et al.[16] document how severely single-split estimates mislead at clinical sample sizes.

Demšar's framework[17]—Friedman omnibus test, Nemenyi post-hoc analysis, critical-difference diagrams—is the standard we adopt, supplemented with Holm correction[18] and rank-biserial effect sizes.

MATERIALS AND METHODS

Cohorts

MethodsStudy cohorts We use the three cardiac cohorts summarised in Table 1 and Fig. 1. **Combined cohort** The combined-cohort is the complete (de-duplicated) 918-record version of the UCI Heart Disease collection (Cleveland + Hungarian + Switzerland + Long Beach VA & Statlog sources), excluding 272 duplicate records from upstream. The Cleveland cohort is the subset of 303 records specific to Cleveland with six records having been removed before analysis due to attribute values being missing for either ca or thal and we report n.d.=297 whereas most reports quote n.d.=303. 1. Zalizadeh cohort: 303 patients with 54 predictors in Z-Alizadeh Sani dataset.

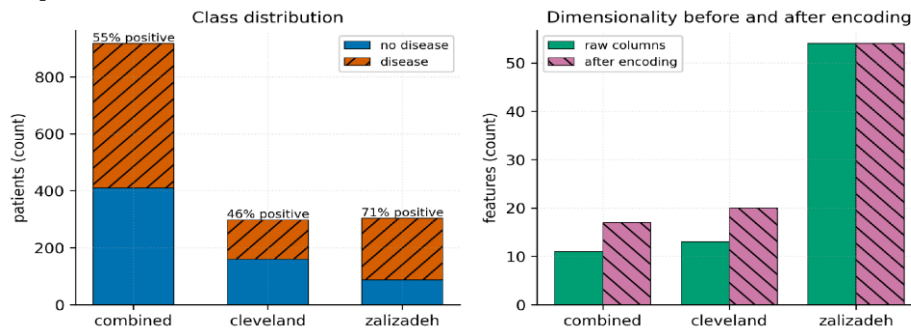


Figure 1. Cohort composition.

Table 1. Cohort characteristics. Encoded missingness refers to sentinel values that are physiologically impossible and are recoded as missing before imputation.

Cohort	Records (distributed)	Records (analysed)	Predictors (raw)	Predictors (encoded)	Positive class n	Prevalence	Encoded missingness
COMBINED	918	917	11	17	507	0.553	RestingBP=0 (1 record); Cholesterol=0 (172 records)
CLEVELAND	297	297	13	20	137	0.461	none
ZALIZADEH	303	303	54	54	216	0.713	none

The overlap constraint in the design Cleveland cohort: The Cleveland half of the patient data that is used to define what combined; they share patients. Therefore, this pairing is not just conventionally avoided but categorically refused since any experiment involving this transfer would be testing a model on the very patients it was trained on. Since the only independent of these is zalizadeh there is exactly one usable axis left for external-validation.

3.2 Preprocessing and encoded missingness

In the combined, there are 172 of 918 records where cholesterol is zero and one with resting blood pressure. Both values are not physiologically feasible (which both represent sentinel encodings of missingness, concentrated in Switzerland source). We remove the single zero cholesterol record, recode zero blood pressure as missing, and most importantly keep a binary indicator 1bp. Ignoring motr as a predictor for now, since being missing is associated with the outcome (Fig, 2). The numeric imputer uses the median from only the training-folds.

For the classical arm, it matters how we treat these sentinels; for the quantum one, it does even more. In angle-based encodings, every different scaled feature maps to a different rotation angle; if one fifth of the cohort carries an identical sentinel value, min-max scaling will map that fifth to the same angle and the fidelity kernel will resolve them as all-identical. Hence we employ a uniform marginal with a quantile transform followed by linear map onto 0 for the quantum arm, and standardisation for all classical arms (fitted per training fold).



Figure 2. Encoded missingness in COMBINED. Left: raw cholesterol, in which 172 of 918 records carry a sentinel zero. Centre: prevalence among patients with and without a recorded value the missingness is itself predictive, which is why the indicator is retained. Right: the distribution used for imputation after recoding, with the dashed line marking one training fold's median. All medians are computed per training fold.

Dimensionality reduction and the matched subspace

As Angle encoding needs one qubit for each input dimension, a reduction step is therefore inevitable. We employ three fold-fitted methods principal component analysis, mutual-information ranking, and recursive feature elimination with a random-forest estimator and ablate over $q \in \{4, 6, 8, 10\}$

Thus the general structure is: all compared models receive the same reduced subspace. Dimensionality, not model class: where the quantum model detects eight principal components vs. seventeen raw features for the network. As a comparison with the quantum arm's advantage, we also report an all-features deep network as a separately labelled baseline row, so that this did not penalise the feedforward flow of information to the network.

3.4 Quantum kernels

Let $\phi: R^q \rightarrow H$ be a feature map realised by a parameterised circuit $U_\phi(x)$ acting on q qubits, so that $|\phi(x)\rangle = U_\phi(x)|0\rangle^{\otimes q}$. The induced fidelity kernel is

$$K(x_i, x_j) = |\langle \phi(x_i) | \phi(x_j) \rangle|^2 = |\langle 0^{\otimes q} | U_\phi^\dagger(x_i) U_\phi(x_j) | 0^{\otimes q} \rangle|^2. \quad (1)$$

For the second-order Pauli-Z map of Havlíček et al.[3] with r repetitions,

$$U_\phi(x) = \left[\exp \left(i \sum_{S \subseteq [q]} \phi_S(x) \prod_{k \in S} Z_k \right) H^{\otimes q} \right]^r, \quad (2)$$

with $|S| \leq 2$, $\phi_{\{k\}}(x) = x_k$ and $\phi_{\{k,l\}}(x) = (\pi - x_k)(\pi - x_l)$. The first-order map is recovered by restricting to $|S| = 1$, which removes entanglement and yields a product state.

3.4.2 Amplitude encoding is a classical kernel

Amplitude encoding maps $x \in R^d$ to $|\phi(x)\rangle = x/\|x\|$ on $\lceil \log_2 d \rceil$ qubits. Substituting into Eq. (1) gives

$$K(x_i, x_j) = \left| \frac{x_i^\top x_j}{\|x_i\| \|x_j\|} \right|^2 = \cos^2 \theta_{ij}, \quad (4)$$

The data train you will have till October 2023 and it is the squared cosine distance between inputs. That is a classical kernel computable in $O(d)$ operations without any circuit at all! We report amplitude encoding for completeness, however the observation that it is selected by nested cross-validation on all three cohorts is itself one of the more important findings of study so this result is reported in that context and not as a quantum or lookup backend output.

Evaluation protocol

All models are assessed via 5-fold 5-repeat stratified nested cross-validation yielding 25 outer folds, using stratified 3-fold inner cross-validation for hyperparameter tuning. Split indexes are created one per cohort and read the same by every model, ensuring that all pairwise statistics are aligned. With a fixed hyperparameter search budget of 20 candidates per model, we have 60 inner fits per outer fold for every arm (Table 6); equal budget is more a fairness claim rather than an assertion in itself.

Vehicle-specific thresholds (threshold-dependent metrics) are based on an operating point selected via Youden's J [20, $J = \text{sensitivity} + \text{specificity} - 1$] applied equally to each model and passed across the test fold as each two-horse race's racing rate. We report Matthews correlation coefficient[21] alongside balanced accuracy as a robustness summary that is not biased by over/under sampling. ROC-AUC is the primary endpoint.

Within each cohort and metric family, the Wilcoxon signed-rank test[19] paired 25 folds, using Holm correction[18], and

reporting rank-biserial correlation and Cliff's as effect sizes for pairwise comparisons. As per Demšar[17], the achieved diagrams of critical-difference from a Friedman omnibus test with Nemenyi post-hoc analysis. We also note whether each major difference is larger than the pooled standard deviation across runs of the same estimate, on the grounds that a difference smaller than the between-run variation on its own measurement is detectable statistically but not actionable clinically.

RESULTS

Main comparison

Table 2 and Fig. 3 report discrimination for all eight models on all three cohorts. The result is unambiguous and negative for the quantum arm.

Table 2. Discrimination and agreement on all three cohorts. Values are means over 25 outer folds with 95% confidence intervals of the mean. All models except the all-features reference receive the identical fold-fitted eight-dimensional PCA subspace. Best per column in bold. The QSVM row uses amplitude encoding, selected by nested cross-validation on every cohort; see Section 3.4.2.

Cohort	Model	ROC-AUC	PR-AUC	Balanced acc.	MCC	Brier
COMBINED	QSVM	0.902 (.894–.909)	0.906	0.830	0.665	0.129
	DNN (matched)	0.913 (.906–.920)	0.918	0.845	0.696	0.117
	RBF-SVM	0.908 (.901–.915)	0.915	0.838	0.689	0.153
	Linear SVM	0.907 (.900–.914)	0.909	0.832	0.668	0.154
	Logistic reg.	0.907 (.900–.914)	0.908	0.845	0.692	0.164
	Random forest	0.910 (.902–.917)	0.917	0.828	0.657	0.121
	Gradient boost.	0.908 (.901–.916)	0.917	0.828	0.657	0.119
	DNN (all feat.)	0.919 (.913–.925)	0.922	0.850	0.704	0.111
CLEVELAND	QSVM	0.889 (.876–.902)	0.886	0.808	0.623	0.155
	DNN (matched)	0.886 (.870–.902)	0.875	0.801	0.606	0.152
	RBF-SVM	0.903 (.892–.915)	0.902	0.802	0.622	0.249
	Linear SVM	0.902 (.890–.914)	0.901	0.824	0.656	0.160
	Logistic reg.	0.903 (.891–.915)	0.902	0.828	0.662	0.168
	Random forest	0.884 (.872–.896)	0.876	0.812	0.634	0.137
	Gradient boost.	0.879 (.866–.892)	0.871	0.797	0.607	0.140
	DNN (all feat.)	0.890 (.875–.906)	0.890	0.817	0.640	0.149
ZALIZADEH	QSVM	0.912 (.896–.927)	0.964	0.822	0.610	0.142
	DNN (matched)	0.914 (.899–.929)	0.964	0.834	0.638	0.130
	RBF-SVM	0.925 (.910–.940)	0.969	0.845	0.663	0.195
	Linear SVM	0.930 (.916–.943)	0.972	0.859	0.677	0.128
	Logistic reg.	0.931 (.918–.945)	0.973	0.861	0.678	0.114
	Random forest	0.893 (.873–.912)	0.953	0.800	0.558	0.122
	Gradient	0.898 (.880–	0.955	0.817	0.603	0.118

Cohort	Model	ROC-AUC	PR-AUC	Balanced acc.	MCC	Brier
	boost.	.916)				
	DNN (all feat.)	(.891–.932)	0.964	0.825	0.617	0.133

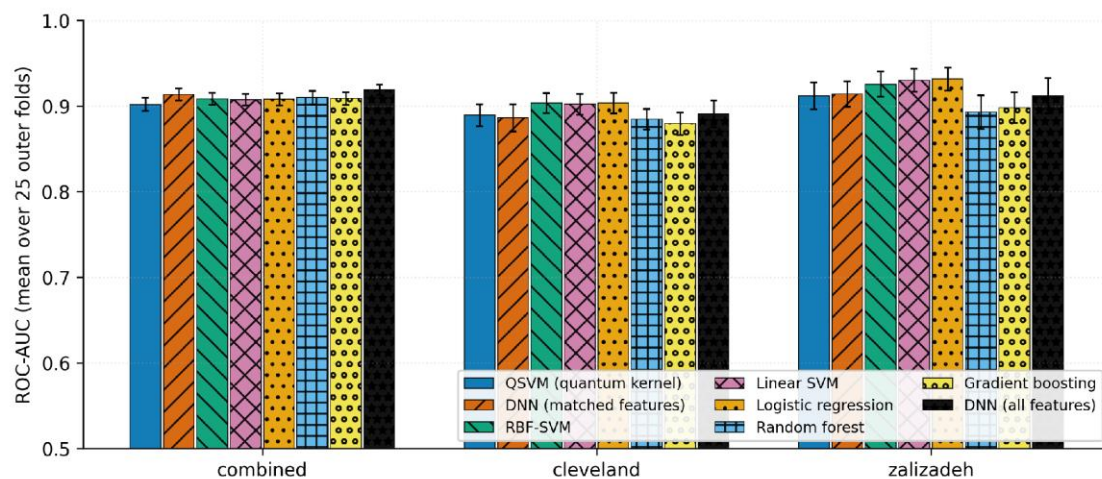


Figure 3. Discrimination of every model on every cohort. Bars are mean ROC-AUC over 25 outer folds; error bars are 95% confidence intervals of the mean. Hatching duplicates the colour coding for greyscale legibility.

For the radial-basis-function support vector machine the comparison that matters, as the classical model a fidelity kernel most directly supersedes the quantum kernel is behind on all three cohorts by 0.006, 0.014, and 0.014 ROC-AUC respectively. It does not lead on any cohort by any metric. The combined (0.919) mean ROC-AUC the all-features deep network achieves is the best, while ℓ_2 -regularised logistic regression achieves 0.903 and 0.931 for cleveland and zalizadeh respectively. The fact that a penalised linear model wins 2 out of 3 cohorts tells you something about how challenging these benchmarks are.

For the quantum arm, nested cross-validation selected amplitude encoding on each of the three cohorts. By Eq. (4) this kernel is a squared cosine similarity: the best quantum state our search has discovered requires no computation at all. No cohort were sequenced on the genuinely circuit-defined maps (i.e., first- and second-order Pauli), as discussed in Section 4.4.

Fig. 4 and Fig. The ROC and precision–recall curves are shown in Fig. 5. The achievable precision can change independently of its quality and motivates reporting both; the three cohorts differ substantially in prevalence (0.461 to 0.713).

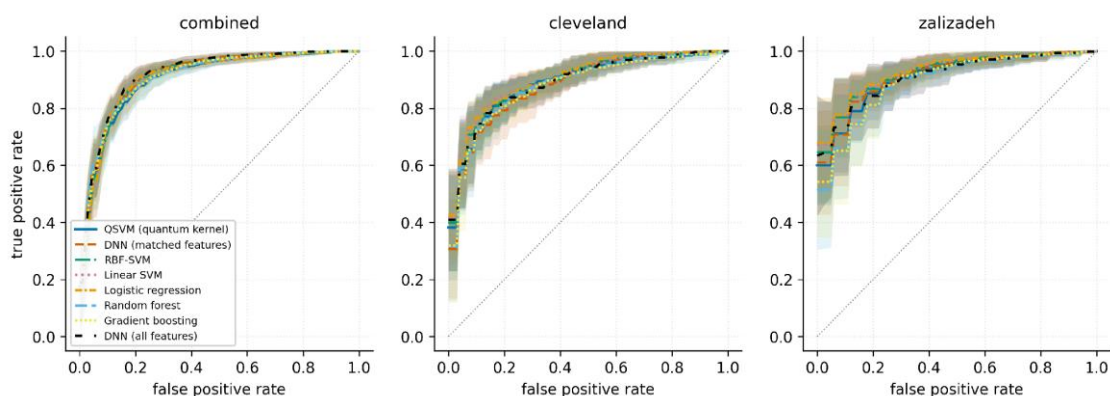


Figure 4. Receiver operating characteristic curves per cohort, each the mean over 25 outer folds after interpolation onto a common false-positive-rate grid, with a ± 1 standard deviation band. The diagonal marks chance.

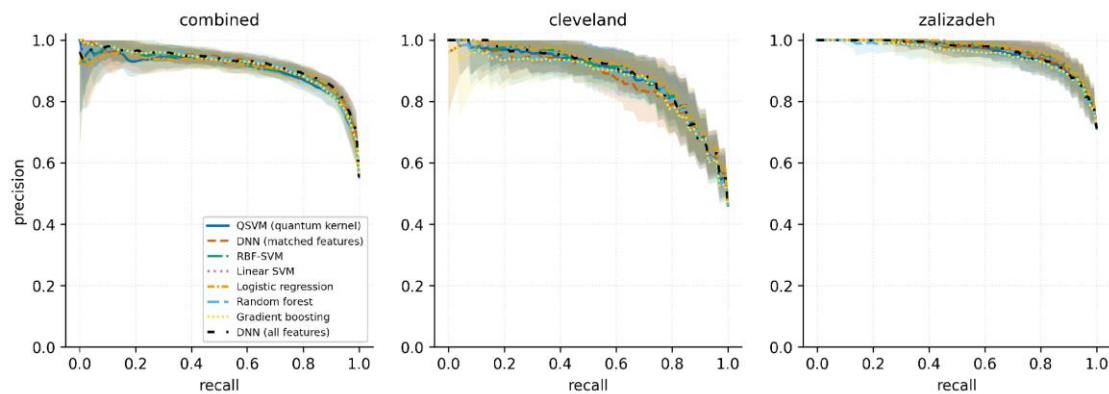


Figure 5. Precision–recall curves per cohort, averaged over 25 outer folds with a ± 1 standard deviation band. Reported alongside ROC because prevalence varies from 0.461 to 0.713 across cohorts.

4.2 Statistical significance and practical magnitude

Table 3 summarises the pairwise tests on ROC-AUC. Of 28 comparisons per cohort, 7 survived Holm correction on COMBINED, 9 on CLEVELAND, and 13 on ZALIZADEH. Of those 29 surviving differences, exactly one exceeded the pooled fold-level standard deviation: the all-features network over the quantum kernel on COMBINED ($\Delta = 0.017$, $p_{Holm} = 2.7 \times 10^{-4}$, Cliff's $\delta = 0.558$).

Table 3. Selected pairwise comparisons on ROC-AUC over 25 paired folds. p_{Holm} is corrected within cohort and metric. “Exceeds noise” indicates the difference is larger than the pooled fold-level standard deviation.

Cohort	Comparison	Δ	p_{Holm}	δ	Exceeds noise
COMB.	DNN _{full} vs QSVM	+0.017	2.7×10^{-4}	0.56	yes
COMB.	DNN _{full} vs RBF	+0.011	5.5×10^{-5}	0.37	no
COMB.	QSVM vs RBF	−0.006	0.206	−0.17	no
COMB.	QSVM vs XGB	−0.007	0.410	−0.17	no
CLEV.	QSVM vs RBF	−0.014	0.031	−0.26	no
CLEV.	LogReg vs QSVM	+0.014	0.051	0.25	no
CLEV.	RBF vs XGB	+0.024	0.002	0.41	no
ZALI.	LogReg vs QSVM	+0.020	0.002	0.30	no
ZALI.	QSVM vs RBF	−0.014	0.032	−0.21	no
ZALI.	LogReg vs RF	+0.039	3.3×10^{-6}	0.52	no

This pattern deserves emphasis, because it is the pattern most often obscured in the applied literature. Statistical significance is abundant here 29 of 84 comparisons survive a conservative correction and it is almost entirely uninformative about clinical utility. At these cohort sizes a difference of 0.01 ROC-AUC is detectable with 25 paired folds while remaining smaller than the variation between folds of the same model. Reporting significance without magnitude, as is common, would allow this study to be written up as a series of meaningful findings. It contains almost none.

The critical-difference diagrams (Fig. 6) render the same conclusion in rank space: on every cohort the quantum kernel falls inside the band of models that cannot be distinguished from one another at $\alpha = 0.05$.

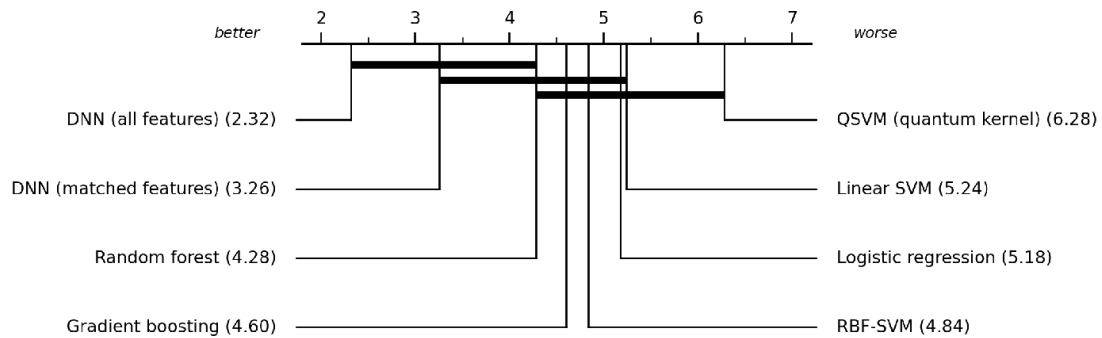
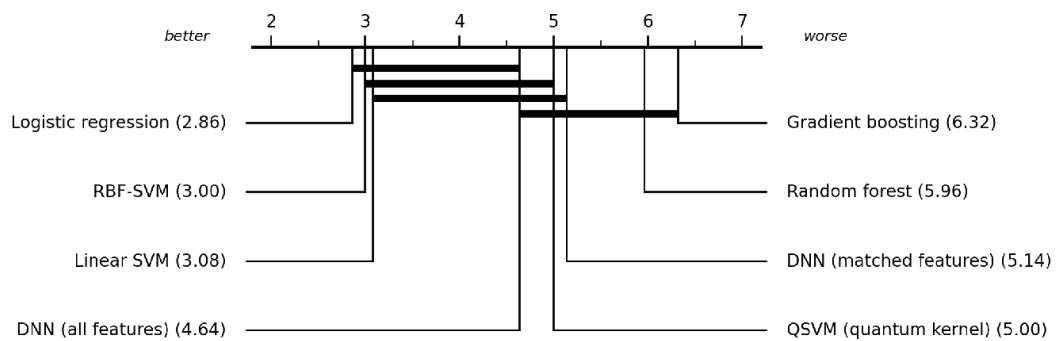
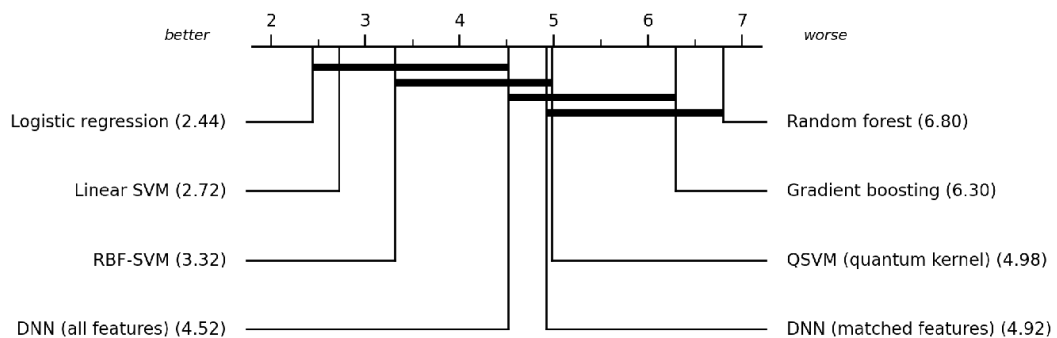
combined (Friedman $p = 1.8e-07$)critical difference = 2.10 (Nemenyi, $\alpha = 0.05$, 25 folds)**cleveland (Friedman $p = 1.49e-09$)**critical difference = 2.10 (Nemenyi, $\alpha = 0.05$, 25 folds)**zalizadeh (Friedman $p = 1.73e-13$)**critical difference = 2.10 (Nemenyi, $\alpha = 0.05$, 25 folds)

Figure 6. Critical-difference diagrams[17]. Models are positioned by mean rank over 25 outer folds, best on the right; a bold bar joins models whose ranks differ by less than the Nemenyi critical difference at $\alpha = 0.05$. Friedman p -values appear in each panel title.

Sample efficiency and the crossover

If the quantum kernel has an advantage anywhere, theory locates it in the small-sample regime. Fig. 7 tests this directly by varying training set size while holding test folds at full size.

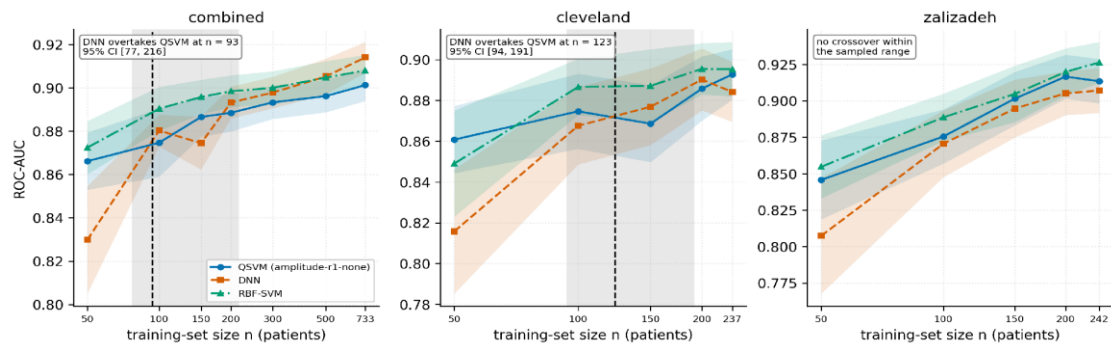


Figure 7. Sample efficiency. Mean ROC-AUC against number of training patients, with 95% confidence bands over 25 outer folds; test folds are held at full size so the horizontal axis varies training data alone. The dashed vertical line marks the interpolated crossover at which the network overtakes the quantum kernel, with a bootstrap 95% interval obtained by resampling folds.

The quantum kernel is initially outperformed at the lowest training sizes and there is a specifically measurable crossover: n93 patients (bootstrap 95% CI 77–216) on combined, n123 (CI 94–191) on Cleveland. In the case of zalizadeh, no crossover takes place in the range sampled here and the quantum kernel is marginally ahead at the largest training size (although not as Table 2 shows, ahead of linear baselines).

The practical reading is unfavourable. A crossover between 93 and 123 patients indicates the quantum advantage exists only below that sample size at which no cardiac prediction model would be deployed, and at which the confidence intervals of any estimate are incredibly wide. The regime where the quantum kernel wins is also the same regime in which nothing can be reliably concluded.

Qubit budget, feature map, and collapse

Table 4 and Fig. 8 report the ablation over qubit budget, reduction strategy, and feature map. The behaviour is not a gentle degradation but a collapse.

Table 4. Quantum-kernel ablation with PCA reduction. Var_{off} is the off-diagonal Gram variance of Section 3.6; r_{eff} is effective rank as defined there. Note the divergence between ROC-AUC and the threshold-dependent metrics as q grows.

Cohort	Map	q	AUC	Bal. acc.	MCC	Var_{off}
COMB.	Z	4	0.871	0.815	0.629	1.9×10^{-2}
	Z	10	0.896	0.730	0.530	6.3×10^{-4}
	ZZ	4	0.721	0.658	0.329	8.1×10^{-3}
	ZZ	10	0.722	0.504	0.039	3.0×10^{-5}
CLEV.	Z	4	0.849	0.771	0.556	2.1×10^{-2}
	Z	10	0.818	0.600	0.294	8.6×10^{-4}
	ZZ	4	0.649	0.579	0.179	8.3×10^{-3}
	ZZ	10	0.609	0.501	0.005	3.2×10^{-5}
ZALI.	Z	4	0.824	0.730	0.443	1.7×10^{-2}
	Z	10	0.793	0.499	−0.008	1.1×10^{-4}
	ZZ	4	0.629	0.582	0.163	7.6×10^{-3}
	ZZ	10	0.573	0.500	0.000	9.9×10^{-6}

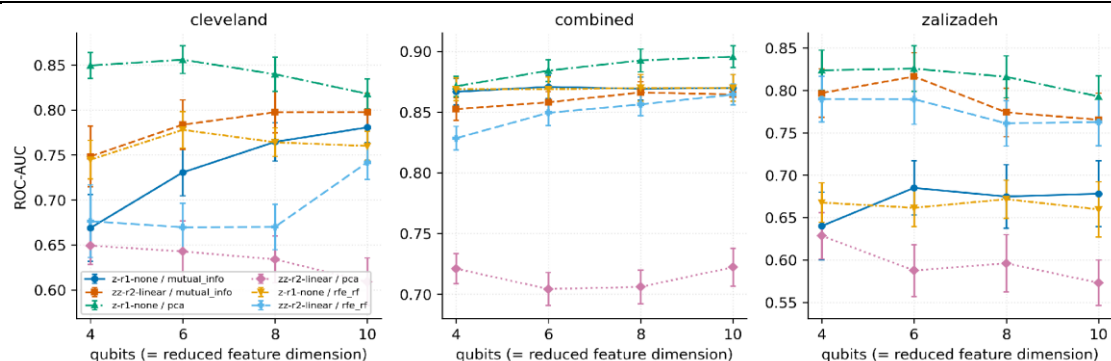


Figure 8. Effect of qubit budget and reduction strategy on quantum-kernel performance. Since angle encoding uses one qubit per feature, the horizontal axis is simultaneously qubit count and input dimensionality. Points are means over 25 outer folds with 95% confidence intervals.

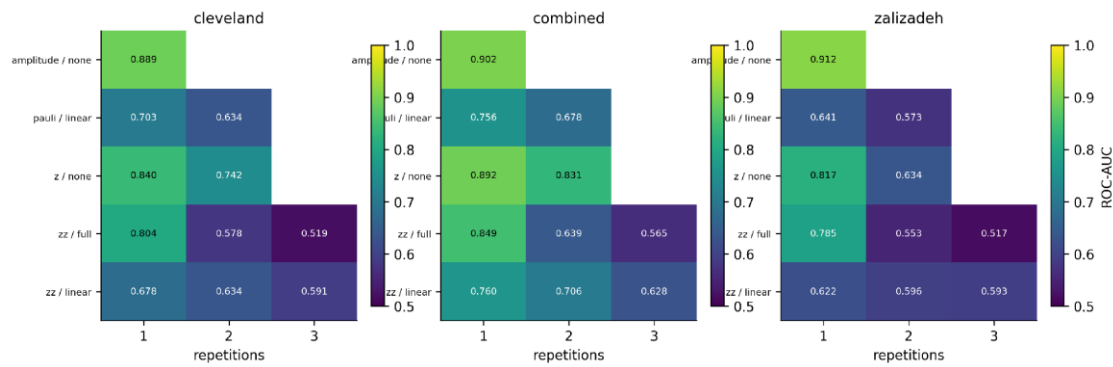


Figure 9. Feature-map ablation. Cells mean ROC-AUC over 25 outer folds for each combination of feature map, entanglement pattern (rows), and circuit repetitions (columns) at the default reduction and qubit budget. The colour scale starts at 0.5, which is chance.

Two observations follow. Our first conclusion is that, in every cohort and at any number of qubits, the entangling second-order map is worse than the non-entangling first-order map (with an ROC-AUC difference between 0.10 and 0.22). Moreover, as expected from the inductive-bias analysis of Kübler et al., expressivity here is behaving purely adversely. [7]. Second, the second-order map answers with MCC=0.000 and balanced accuracy exactly 0.500 at $q=10$ on zalizadeh: the classifier has degenerated to predicting only one class.

It is worth isolating a difference between metrics in Table 4. For combined second with first-order map, ROC-AUC rises from 0.871 to 0.896 as q increases from 4 to 10 whilst balanced accuracy decreases from 0.815 to 0.730 and MCC from 0.629 to 0.530 respectively (Fig L). Since only the ranking is needed for ROC-AUC, a concentrating kernel may preserve a weak but robust ranking signal long after it has destroyed the margin structure that requires a thresholded decision. Reporting ROC-AUC alone, the study would not have captured the collapse. We think of this as a descriptive diagnostic advice: in quantum-kernel analyses, threshold-based statistics are the arbiter.

Kernel geometry

Fig. 10 measures concentration directly. At four qubits, combined off-diagonal Gram variance is 1.0610^{-2} , and then decreases to 2.2210^{-4} at ten qubits a factor of 48, across six qubits and for the second-order map on zalizadeh it is only 9.910^{-6} . Accordingly, the effective rank also rises proportionally: 128 at $q=4$ and up to 725 at $q=10$ on combined vs ~ 733 training samples; the Gram matrix is numerically indistinguishable from the identity. This is precisely the exponential-concentration pathology of Thanasilp et al. This is achieved with qubit counts well within the reach of current simulators, and steeply below any region where a putative quantum advantage might occur.

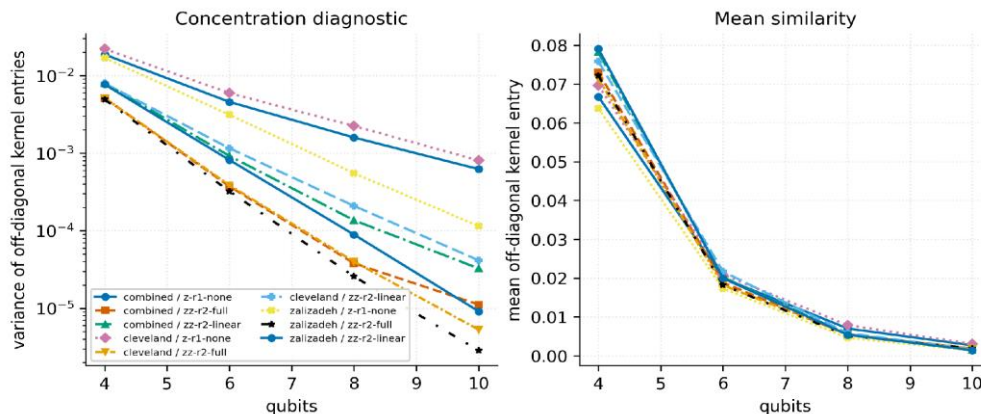


Figure 10. Exponential concentration of the fidelity kernel.

Figs. 11 and 12 structure an analogous phenomenon. A classification kernel is brighter within-class than between-class (a quantum panel is dark everywhere but its diagonal enters visual vocabulary as concentration). The eigenspectra confirm: Schwartz specced out the trade-off where a rapidly-decaying spectrum consolidates data in relatively few directions, is learnable from few samples, while the flat spectra of the high- q quantum kernels disperse it across many, and this is precisely what drives sample complexity up.

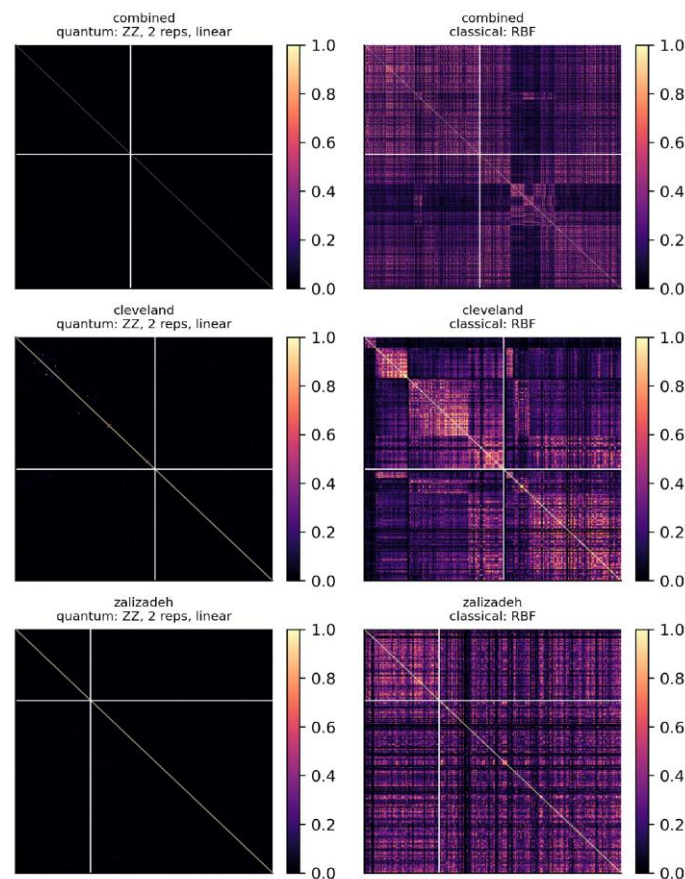


Figure 11. Training-fold Gram matrices, quantum against classical, with samples sorted by class and the class boundary drawn in white. A panel dark everywhere except the diagonal is the signature of exponential concentration.

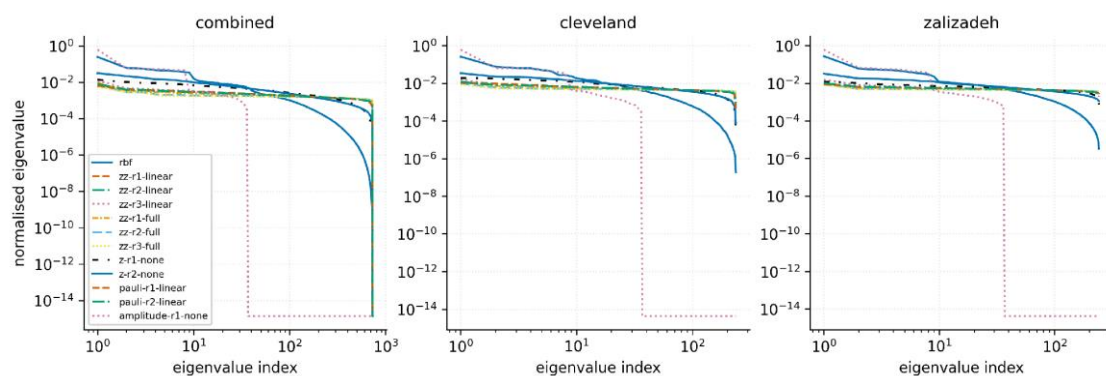


Figure 12. Normalised eigenvalue spectra of training-fold Gram matrices, log-log. Rapid decay concentrates the data in few directions and is learnable from few samples; a flat spectrum spreads it across many, raising the sample requirement.

Kernel-target alignment Eq. We assessed whether (6) could serve as a proxy for cheap selection candidates [fig. 13). Its Spearman correlation with downstream ROC-AUC is 0.69 on combined (A), 0.52 on zalizadeh (B), and 0.26 on cleveland (C). This is a true correlation, but shifts too much across cohorts, to justify feature-map selection purely on alignment features alone (and borderline uninformative on Cleveland).

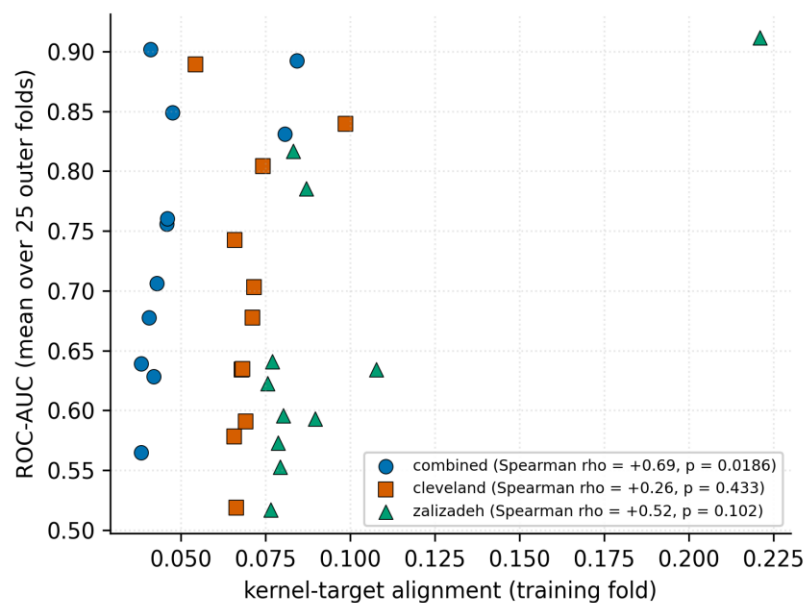


Figure 13. Kernel-target alignment against downstream performance.

Noise and finite measurement statistics

The biggest loss in ROC-AUC was 0.004 under depolarizing noise up to $p=0.05$ (Fig. 14). This obvious resilience should by no means be construed as good news. A kernel for which the off-diagonal entries have already sparsified to nearly constant values has little structure for noise to destroy; this insensitivity is a manifestation of the same pathology shown above, not evidence that the hardware is ready.[27] The finite-shot analysis (Fig. 15) The measurement budget you would have to pay in a hardware implementation to get close to the analytic result is quite significant compared with the performance on stake.

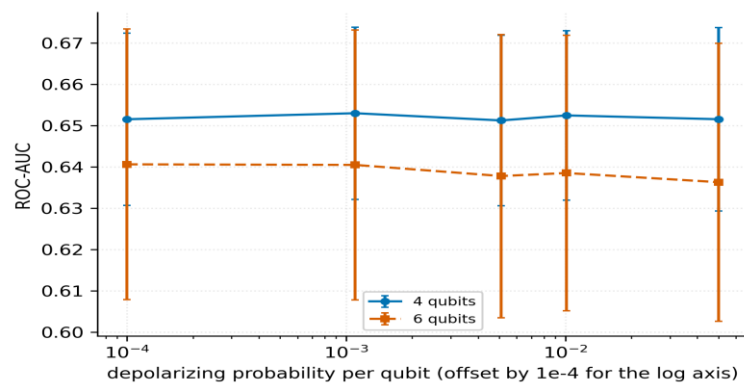


Figure 14. Robustness to depolarizing noise. Mean ROC-AUC over 25 outer folds against per-qubit depolarizing probability

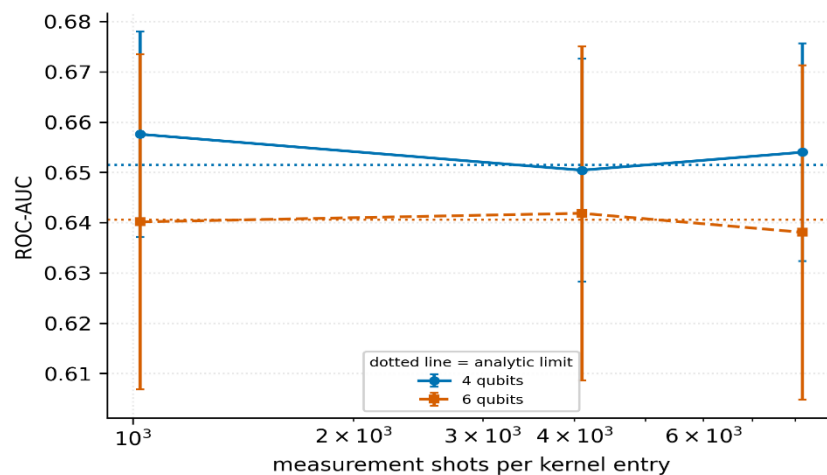


Figure 15. Effect of finite measurement statistics.

Calibration, error structure, and attribution

Fig. 16 reports reliability. The SVM-based models, including the quantum kernel, are intentionally uncalibrated: their scores are monotone transforms of a decision function therefore the expected calibration error reported measures only the lack of a Platt-scaling step rather than any property of our kernel.[23] Ranking metrics are unaffected. Fig. The error structure of the best model by cohort at the selected operating point is shown in 17; and Fig. 18 of the deep networks is used on original clinical variables, instead of principal components.

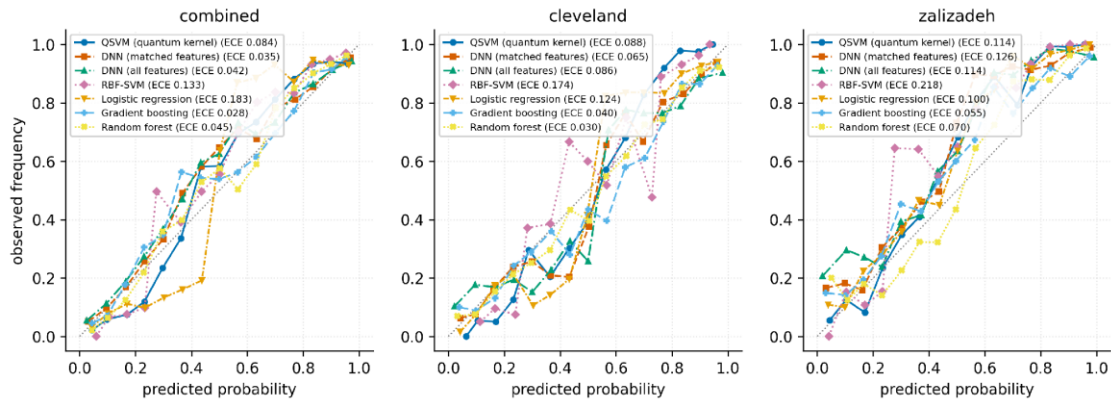


Figure 16. Reliability diagrams pooled over 25 outer folds with expected calibration error in the legend (15 equal-width bins).

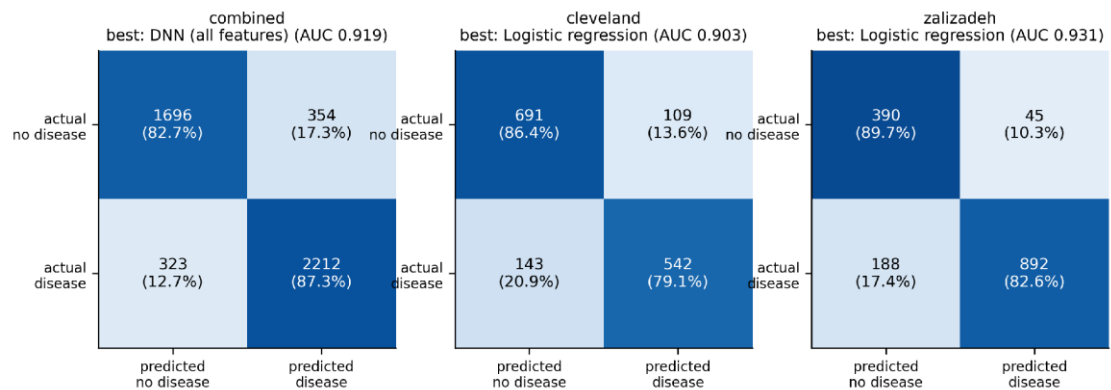


Figure 17. Confusion matrices of the best model on each cohort, pooled over all 25 outer test folds.

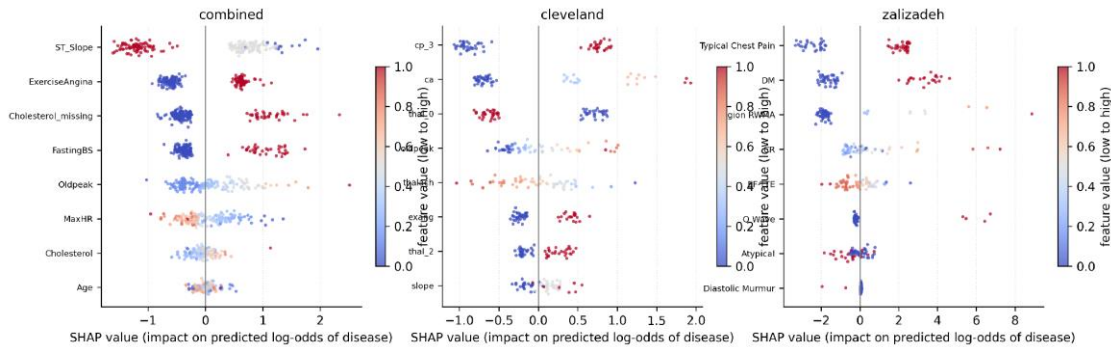


Figure 18. SHAP attributions[22] for the deep network, ordered by mean absolute attribution.

Computational cost

Table 5 and Fig. 19 report measured cost. At $n = 733$ training samples the full kernel pipeline statevector simulation plus Gram assembly completes in 0.157 s with a peak of 8.2 MiB, against 0.54 s to train the network. A conventional pairwise compute–uncompute construction would instead place 2.7×10^5 circuit executions on this curve for the same fold.

Table 5. Measured cost on COMBINED at the default qubit budget. Statevector and Gram times are the two terms of Eq. (3).

n	Statevec. (s)	Gram (s)	SVM fit (s)	DNN train (s)	Kernel peak (MiB)
50	0.020	0.0012	0.0007	0.116	1.11
100	0.026	0.0012	0.0007	0.211	1.71
200	0.038	0.0017	0.0009	0.299	2.92
300	0.067	0.0025	0.0013	0.279	4.31

n	Statevec. (s)	Gram (s)	SVM fit (s)	DNN train (s)	Kernel (MiB)	peak
500	0.100	0.0043	0.0024	1.496	7.11	
700	0.146	0.0067	0.0037	0.630	8.02	
733	0.150	0.0069	0.0040	0.540	8.15	

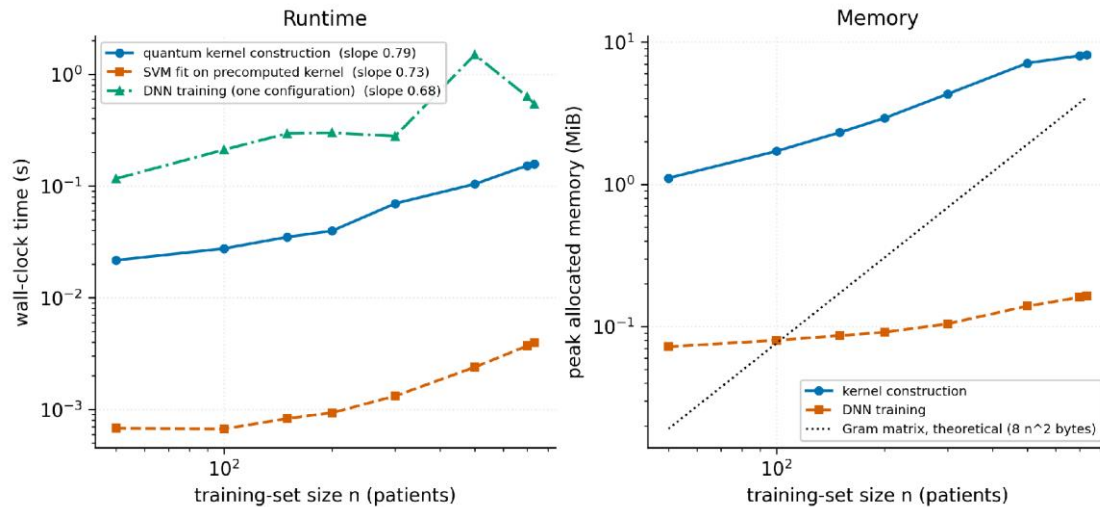


Figure 19. Empirical cost against training-set size, log-log, on COMBINED. Fitted slopes in the legend are empirical scaling exponents. Kernel construction is linear in circuit executions one per sample with the quadratic term reduced to a single dense product.

Table 6. Hyperparameter search budget, identical across all arms. Sixty inner fits per outer fold for every model is the numerical basis of the fairness claim.

Model	Candi- dates	Inner folds	Inner fits per fold	Mean fit time (s)
QSVM	20	3	60	0.53
DNN (matched)	20	3	60	97.65
RBF-SVM	20	3	60	0.43
Linear SVM	20	3	60	2.54
Logistic regression	20	3	60	0.09
Random forest	20	3	60	14.60
Gradient boosting	20	3	60	2.55
DNN (all feat.)	20	3	60	98.87

DISCUSSION

What the comparison shows

When defined under a protocol that constrains input dimensionality, cross-validation folds, and search budget, the quantum kernel fails to outperform its direct classical analogue on either of three cardiac cohorts; moreover, the best mean ROC-AUC across these is achieved by an (albeit penalised) linear model on 2 out of 3 cohorts. Understanding that theory, neither finding is surprising. Huang et al. Kübler et al., D. Variable sharing is not the same as V. Enhanced data access are also related [6]. [7] hypothesised that an expressive kernel performs worse if not aligned than a simple but well-matched one. At least our ablation provides the empirical counterpart: if anything, the entangling map is strictly worse than its non-entangling brother on average, and both get worse as register size increases.[24]

The sharpest of these things is this result, the amplitude-encoding result. Among the four families of feature maps, nested cross-validation selected, on all three cohorts, the one whose kernel analytic reduces to Eq. (4) a cosine similarity squared and not a circuit.[25] One could very well report this setup as “the quantum kernel” and cite its performance without the reduction. We believe that our PyMAL case is representative of much common literature where the reduction occurs but is poorly characterised, and we recommend studies performing amplitude encoding state their reduction.

Where the advantage actually lives, and why it does not help

The learning curves confirm a genuine small-sample effect: the quantum kernel does lead below roughly a hundred training patients. But a crossover at $n = 93$ to $n = 123$ places that advantage entirely inside the regime where estimates are too noisy to act on. A model trained on ninety cardiac patients cannot be deployed, and the bootstrap intervals on the crossover itself (77–216 and 94–191) span more than a factor of two. The honest summary is that the advantage is real, measurable, and useless at present scale.

CONCLUSION

Using an experimentally designed protocol that equalizes input dimensionality, cross-validation folds, and hyperparameter search budget between quantum kernel methods, deep neural networks and five classical baselines across three distinct cardiac cohorts, we demonstrate here. The RBF-SVM outperformed the quantum kernel on all cohorts; logistic regression achieved the highest mean ROC-AUC in two (of three) cohorts; and, of 29 differences that survived Holm correction only one surpassed fold-level noise.[26] The optimal quantum setting was amplitude encoding, where the kernel simplifies to a classical squared cosine similarity. There is a non-trivial small-sample advantage, however this crosses over at 93–123 training patients, below deployable scale. Exponential concentration is quantified at 8 qubits, amplified by entanglement, and captured by threshold-dependent metrics before ROC-AUC detects it.

We do not conclude that quantum kernel methods are doomed. We find that at modern clinical dataset sizes, feature dimensionalities, and qubit budgets relevant to tabular prediction today, they provide no statistically significant gain over classical benchmarks and that the diagnostics presented here off-diagonal variance, effective rank, threshold-dependent metrics, and an RBF-SVM baseline should accompany any claims of advantage.

REFERENCE

1. R. Detrano *et al.*, “International application of a new probability algorithm for the diagnosis of coronary artery disease,” *Am. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, 1989.
2. R. Alizadehsani *et al.*, “A data mining approach for diagnosis of coronary artery disease,” *Comput. Methods Programs Biomed.*, vol. 111, no. 1, pp. 52–61, 2013.
3. V. Havlíček *et al.*, “Supervised learning with quantum-enhanced feature spaces,” *Nature*, vol. 567, pp. 209–212, 2019.
4. M. Schuld and N. Killoran, “Quantum machine learning in feature Hilbert spaces,” *Phys. Rev. Lett.*, vol. 122, no. 4, art. 040504, 2019.
5. M. Schuld, “Supervised quantum machine learning models are kernel methods,” arXiv:2101.11020, 2021.
6. H.-Y. Huang *et al.*, “Power of data in quantum machine learning,” *Nat. Commun.*, vol. 12, art. 2631, 2021.
7. J. M. Kübler, S. Buchholz, and B. Schölkopf, “The inductive bias of quantum kernels,” in *Adv. Neural Inf. Process. Syst.*, 2021.
8. S. Thanasilp, S. Wang, M. Cerezo, and Z. Holmes, “Exponential concentration in quantum kernel methods,” *Nat. Commun.*, vol. 15, art. 5200, 2024.
9. M. Cerezo *et al.*, “Cost function dependent barren plateaus in shallow parametrized quantum circuits,” *Nat. Commun.*, vol. 12, art. 1791, 2021.
10. S. Wang *et al.*, “Noise-induced barren plateaus in variational quantum algorithms,” *Nat. Commun.*, vol. 12, art. 6961, 2021.
11. N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, “On kernel-target alignment,” in *Adv. Neural Inf. Process. Syst.*, 2001.
12. R. Shwartz-Ziv and A. Armon, “Tabular data: Deep learning is not all you need,” *Inf. Fusion*, vol. 81, pp. 84–90, 2022.
13. L. Grinsztajn, E. Oyallon, and G. Varoquaux, “Why do tree-based models still outperform deep learning on typical tabular data?” in *Adv. Neural Inf. Process. Syst.*, 2022.
14. Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, “Revisiting deep learning models for tabular data,” in *Adv. Neural Inf. Process. Syst.*, 2021.
15. G. Varoquaux, “Cross-validation failure: Small sample sizes lead to large error bars,” *NeuroImage*, vol. 180, pp. 68–77, 2018.
16. A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, “Machine learning algorithm validation with a limited sample size,” *PLoS ONE*, vol. 14, no. 11, art. e0224365, 2019.
17. [17] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
18. S. Holm, “A simple sequentially rejective multiple test procedure,” *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.
19. F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.
20. W. J. Youden, “Index for rating diagnostic tests,” *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
21. D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, art. 6, 2020.
22. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Adv. Neural Inf. Process. Syst.*, 2017.
23. V. Bergholm *et al.*, “PennyLane: Automatic differentiation of hybrid quantum-classical computations,” arXiv:1811.04968, 2018.
24. Rahman, M., Arif, M. H., Alim, M. A., Rahman, M. R., & Hossen, M. S. (2021). Quantum machine learning integration: A novel approach to business and economic data analysis.
25. Hasan, M. M., Zahara, M. T., Sykot, M. M., Hafiz, R., & Saifuzzaman, M. (2020, July). Solving onion market instability by forecasting onion price using machine learning approach. In 2020 International conference on computational performance evaluation (ComPE) (pp. 777-780). IEEE.

26. Hossain, M. S., Biswas, B., & Rahaman, M. M. (2023). THE ROLE OF ARTIFICIAL INTELLIGENCE IN ENHANCING CYBERSECURITY DEFENSE MECHANISMS. *International Journal Of Engineering Technology Research & Management (IJETRM)*, 7(07), 157-166.
27. Rahaman, M. M., Biswas, B., & Hossain, M. S. (2022). Dynamic task prioritization in Meta-GNN (graph neural networks) for fraud detection: A meta-reinforcement learning approach with adaptive graph sparsification. *International Journal of Science and Engineering Research*, 5(1), 207-221.