Quantum Cognition Machine Learning for Forecasting Chromosomal Instability

Giuseppe Di Caro[1,2,*], Vahagn Kirakosyan[1,*], Alexander G. Abanov[1,3], Luca Candelori[1,4], Nadine Hartmann[2], Ernest T. Lam[2], Kharen Musaelian[1], Ryan Samson[1], Martin T. Wells[1,5], Richard J. Wenstrup[2], and Mengjia Xu[6]

[1]Qognitive, Inc., Miami Beach, FL, USA
[2]Epic Sciences, San Diego, CA, USA
[3]Stony Brook University, Department of Physics and Astronomy, Stony Brook, NY, USA
[4]Wayne State University, Department of Mathematics, Detroit, MI, USA
[5]Cornell University, Department of Statistics and Data Science, Ithaca, NY, USA
[6]New Jersey Institute of Technology, Department of Data Science, Newark, NJ, USA
[*]Corresponding authors (giuseppe.dicaro@qognitive.io, vahagn.kirakosyan@qognitive.io)

ABSTRACT

The accurate prediction of chromosomal instability from the morphology of circulating tumor cells (CTCs) enables real-time detection of CTCs with high metastatic potential in the context of liquid biopsy diagnostics. However, it presents a significant challenge due to the high dimensionality and complexity of single-cell digital pathology data. Here, we introduce the application of Quantum Cognition Machine Learning (QCML), a quantum-inspired computational framework, to estimate morphology-predicted chromosomal instability in CTCs from patients with metastatic breast cancer. QCML leverages quantum mechanical principles to represent data as state vectors in a Hilbert space, enabling context-aware feature modeling, dimensionality reduction, and enhanced generalization without requiring curated feature selection. QCML outperforms conventional machine learning methods when tested on out of sample verification CTCs, achieving higher accuracy in identifying predicted large-scale state transitions (pLST) status from CTC-derived morphology features. These preliminary findings support the application of QCML as a novel machine learning tool with superior performance in high-dimensional, low-sample-size biomedical contexts. QCML enables the simulation of cognition-like learning for the identification of biologically meaningful prediction of chromosomal instability from CTC morphology, offering a novel tool for CTC classification in liquid biopsy.

1 Background

Unlike traditional tissue tests[1,2], cell-based liquid biopsy assays enable selection of individual CTCs for the analysis of chromosomal instability using next-generation sequencing by quantification of large-scale state transitions (LST)[3–9]. Chromosomal instability is a genomic characteristic of cancer cells that drives tumor evolution and metastatic potential[10–19]. However, whole genome sequencing assays are laborious, requiring a complex workflow that invariably results in a considerable turnaround time that sometimes is not compatible with clinical practice[20]. A previous study has shown that we can partially predict chromosomal instability in individual cells by developing algorithms that analyze a range of features, including cell shape, size, morphology, and protein levels, from images of CTCs using an automated digital pathology pipeline[3]. Predicting chromosomal instability through morphology offers significant advantages; it can significantly reduce turnaround times compared to whole-genome assays, providing crucial information about the genomic characteristics of CTCs in a patient in a shorter timeframe[3]. Timely information on the presence of CTCs with the highest metastatic potential may be critical for making optimal clinical decisions.

A key challenge in predicting chromosomal instability through morphology is the utilization of a machine-learning method that accurately classifies morphology patterns from all CTC features and provides a generalization and reproducibility, compatible with potential validation for clinical use[21–24]. Key limitations of commonly used machine learning techniques in biology applications, such as support vector machines (SVMs) with Gaussian kernels, include the following[21–24]: 1) The increase in dimensionality that arises from combinations of multiple features exponentially complicates the prediction task, as often seen with cell morphologies. 2) SVMs struggle when classes significantly overlap or when there is label noise, resulting in support vectors that distort the generalization, leading to high misclassification rates and overfitting on independent datasets. 3) The decision boundary learned by a nonlinear SVM is often not biologically interpretable. The biological explainability of the underlying models is crucial to enhancing reproducibility[21–24].

Quantum Cognition Machine Learning (QCML)[25–28] is an emerging field that introduces a novel approach to machine

learning, grounded in the mathematical principles of quantum theory. In QCML, data points are represented as quantum states in a complex Hilbert space, while features and target variables are modeled as Hermitian operators or "observables". The "observables" are learned by optimizing a particular objective function over the observables' parameters. Although QCML models use similar objective functions and evaluation metrics as classical machine learning models, they differ fundamentally in how data is represented and how functional dependencies among features are parameterized. QCML models are versatile, capable of handling numerical and categorical data, as well as missing and/or noisy data. They create a global quantum manifold model (in the sense of quantum geometry[29]) of the original data manifold, that is robust to noise and able to generalize well beyond training samples[25,27,28]. Part of this adeptness at controlling variance stems from the fact that the number of parameters of a QCML model scales linearly with the number of features, thus achieving logarithmic economy of representation. For the first time we introduce the QCML Positive Operator-Valued Measure (POVM). QCML POVM is an extension of QCML that allows one to forecast the probability density function of a target, as opposed to point estimate. It naturally lends itself to our problem setup with pLST forecasting where we can produce both real-valued forecasts of pLST (based on expected mean/median) and probability forecasts of pLST being above/below a certain threshold.

We hypothesize that QCML's ability to generalize gives it an advantage in computing morphology-predicted pLST compared to typical machine learning algorithms like SVMs. Here, we test whether the advantage of QCML may help prevent overfitting and improve prediction performance and reproducibility in metastatic breast cancer cells. QCML outperforms SVM and other machine learning methods when predicting pLST CTC in out of sample verification which achieves the highest balanced accuracy and specificity compared to SVM with Gaussian Kernel, the best performing among classical models. Additionally, among the classification models, QCML POVM achieves the highest AUC-ROC score, a threshold-independent performance metric.

2 Methods

2.1 Data Acquisition and Preparation

As previously published[4], the CTC assay for metastatic breast cancer follows a non-enrichment strategy where all nucleated cells from patient blood are deposited onto slides and stained using immunofluorescence. We used previously published high-resolution digital image analysis technology of CTCs for metastatic breast cancer[4,30]. The pipeline processes high-resolution fluorescence images acquired via the ZEISS^{TM} Axio automated scanning system. High-resolution imaging is performed and an automated algorithm scans the data to identify rare candidates for CTCs among millions of white blood cells[4,30]. The BRIA machine learning framework filters out non-CTCs and artifacts, reducing the number of candidates for pathologist review[4,30]. A multiscale feature enhancement algorithm helps identify nuclei while cell segmentation occurs across various fluorescence channels (CK, DAPI, CD45/CD31)[4,30]. The extracted morphological and molecular characteristics serve as input for a machine learning algorithm trained to identify presumptive CTCs for the validation by experts[4,30].

The genomic profiling of CTC was performed as previously published[4]. A maximum of 5 CTCs per patient are prioritized by a board-certified pathologist for genomic profiling[4]. The selected cells undergo lysis and DNA extraction, followed by amplification of the whole genome and library preparation using the SEQPLEX-I kit[4]. Low-pass genome sequencing is used to evaluate chromosomal instability by quantification of LST. A computational pipeline was used to evaluate copy number variations from CTC sequencing data, following principles similar to standard whole-genome sequencing workflows[6]. Sequencing reads generated on the Illumina platform were mapped to the hg38 human reference genome[31], and read counts were aggregated in 1-Mb intervals across the genome. Quality control metrics were calculated to exclude samples with low sequencing depth, poor alignment quality, or excessive coverage variability[4]. Only high-quality samples were retained for analysis. To normalize genomic coverage, bin-level read depth was scaled relative to the mean autosomal signal, allowing for correction of chromosome-wide copy number variation[4].

A total of 112 morphology features were extracted as previously published[30]. Eight morphological characteristics were extracted from each of the nuclear and cell masks: 1) size, 2) roundness, 3) elongation, and 4) the first Hu moment[32], to measure a more subtle shape variability. We next computed 44 intensity features from nuclear and cell masks across the three DAPI, CK and CD45/CD31 channels: 1) MFI, 2) lower, 3) median, and 4) upper quartiles, 5) interquartile range, as well as co-Localizations between channels. 70 texture features are extracted to characterize image patterns. Gabor filters were extracted for localized frequency and orientation information in images and to detect irregularities in repeating textures with a fractal feature approximation. Gabor filters were applied with 16 distinct parameter features combinations, comprising four orientations of the filter ($\theta = 0°$, $45°$, $90°$, and $135°$), two wavelengths (spatial frequency) ($\lambda = 0.1$ and 0.4), and two standard deviations as Gaussian width ($\sigma = 1$ and 3), selected based on cell size. For each filtered image, the mean and standard deviation are computed, capturing orientation- and scale-specific frequency content. Another set of features is computed using Laws' texture energy measures[33]. This involves generating ordered multiplications of one-dimensional filters—L5 (Level), E5 (Edge), S5 (Spot), and R5 (Ripple)—to detect various spatial patterns, with corresponding statistical descriptors calculated from the filtered outputs. The remaining six features are derived using the Local Binary Pattern (LBP) method, which encodes local

texture variations such as edges, corners, and uniform regions. For each image channel, an LBP-transformed image is generated, and inter-channel relationships are quantified using correlation and normalized mutual information, resulting in six final texture features[30].

All patient data were analyzed retrospectively and completely anonymized. All procedures conducted in studies involving samples from human participants adhered to the ethical standards set by the institutional research committee of Epic Sciences, which obtained informed consent from all participants.

2.2 Cross-Validation

For the training of both QCML and classical machine learning (ML) methods, we adopted a "case-agnostic approach ", treating each cell as an independent observation. We perform 5-fold cross-validation with 5 repetitions, each using a different random seed for the data split. For hyperparameter tuning, we use the training set of 166 CTCs from 51 patients. We then run the optimized models on the full dataset of 227 CTCs from 73 patients with the same cross-validation process and report the average in-sample and out-of-sample performance.

2.3 Quantum Cognition Machine Learning (QCML)

QCML[25–28] is a recently introduced machine learning approach grounded in the principles of quantum cognition (for an overview of quantum cognition, refer to[34]). QCML models represent data observations as quantum states in complex Hilbert space. Recall that in quantum mechanics, a *state* is a unit-norm vector in a Hilbert space, defined up to an overall phase. We use the bra-ket notation, representing states by kets such as $|\psi\rangle$. The inner product between two states $|\psi_1\rangle$ and $|\psi_2\rangle$ is denoted by the bra-ket $\langle\psi_1|\psi_2\rangle$. A measurement of a quantum observable, represented by a Hermitian operator M, in the state $|\psi\rangle$ yields an eigenvalue m_i of M with probability given by the squared magnitude of the overlap with the corresponding eigenstate $|m_i\rangle$: $|\langle m_i|\psi\rangle|^2$. The expression $\langle\psi|M|\psi\rangle$ gives the expected value of the random variable associated with measuring M in the state $|\psi\rangle$ [35, I.2.2].

In QCML, for each vector $\mathbf{x}_t \in \mathbb{R}^K$ belonging to a data set consisting of $t = 1,\ldots,T$ observations, we define an error Hamiltonian as

$$H(\mathbf{x}_t) = \frac{1}{2}\sum_k (A_k - \mathbf{x}_{t,k}\cdot I)^2. \tag{1}$$

The operators A_k are a fixed set of quantum observables for $k = 1,\ldots,K$, where each observable is represented by a Hermitian operator on an N-dimensional Hilbert space. In Equation (1), I denotes the $N \times N$ identity matrix. Each of these K quantum observables can be viewed as a 'quantization' of a corresponding feature of the original K-dimensional data set. The vector $\mathbf{x}_t$ then can be mapped to a quantum state $|\psi_t\rangle$ by finding the ground state (i.e., the eigenstate associated with the lowest eigenvalue) of the error Hamiltonian (1). This results in a representation of data into quantum states (i.e., normalized vectors in a complex Hilbert space). Conversely, for an arbitrary quantum state $|\psi\rangle$, its 'position' can be defined as the K-dimensional real vector

$$\mathbf{x}(\psi) = \big(\langle\psi|A_k|\psi\rangle\big)_{k=1}^K \in \mathbb{R}^K.$$

In quantum theory, this vector represents the expected outcomes of measuring the observables A_k in the quantum state $|\psi\rangle$. As a result, with a set of quantum observables $\{A_k\}$, we can convert data into quantum states by finding the ground state $|\psi_t\rangle$ for each data point $\mathbf{x}_t$, and we can also extract information from any quantum state $|\psi\rangle$ by calculating its position $\mathbf{x}(\psi)$.

In an unsupervised setting, training a QCML model involves iterative updates to the observables $\{A_k\}$ so that the ground states $|\psi_t\rangle$ 'cohere' to the data, that is, the distance between $\mathbf{x}_t$ and its position $\mathbf{x}(\psi_t)$ is minimized, as well as the variance of the measurement. During optimization, we can use different forms of a loss function: the distance, the total energy of the error Hamiltonian, or a combined loss function, as discussed in detail in[25].

In the supervised setting, which is the main focus of this article, the training process differs from unsupervised case[28]. The target variable $y \in \mathbb{R}$ is assigned a N-dimensional quantum 'forecast' observable B. Given a data point $\mathbf{x}_t$ the corresponding forecast, measured in quantum state ψ_t, is given by

$$\widehat{y}_t = \langle\psi_t|B|\psi_t\rangle\,.$$

During the training process, the quantum observables $\{A_k\}$ and B are updated at each iteration to minimize a loss function $\mathscr{L}(\widehat{y}_t, y_t)$. The loss function can take various forms such as mean absolute error, mean squared error, cross-entropy, etc. Note that in case of mean absolute error, the non-differentiability of the loss function does not add further complexity to the algorithm, since the mapping from $\mathbf{x}_t$ to its ground state ψ_t is already non-differentiable, the singular points corresponding to the locus of degeneracy of the error Hamiltonian (1). This framework can be easily adapted to handle multiple target variables by introducing separate quantum 'forecast' observables for each target. Below is the summary of the training algorithm:

QCML univariate regression model training

- Randomly initialize feature operators $\{A_k\}$ and target operator B.
- Iterate over training data and operators until desired convergence:
  1: Generate error Hamiltonian $H(\mathbf{x}_t)$
  2: Holding A_k constant, find the ground state $|\psi_t\rangle$ of $H(\mathbf{x}_t)$
  3: Generate the forecast $\widehat{y}_t = \langle\psi_t|B|\psi_t\rangle$
  4: Calculate gradients of the loss function $\mathscr{L}(\widehat{y}_t, y_t)$ w.r.t A_k and B
  5: Update A_k and B via gradient descent

The implementation details of these steps vary based on how the operators A_k and B are parameterized, and there are multiple options for loss functions and optimization methods. The Hilbert space dimension N is a hyperparameter that can be tuned through cross-validation. While larger N values generally reduce the loss, they may cause overfitting and poor generalization, whereas smaller dimensions typically result in higher bias but lower variance.[25]. For practical purposes, it's also best to keep N small to maintain computational efficiency.

To this end, the main goal is to have a model which produces binary forecast (classification) for a cell being LST positive (LST+) corresponding to LST parameter LST>12, where the cutoff of 12 is based on previously published analytical validation data of the metastatic breast cancer platform[4]. We also want the model to 1) produce real-valued LST forecasts, and 2) have the ability to control the balance between specificity and sensitivity. To achieve this, we build a QCML-based regression model and designed a mixed-loss function that incorporates both L1 and cross-entropy components, effectively capturing both regression and classification errors. Additionally, the cross-entropy component allows for a varying weight on the positive class to achieve the desired specificity/sensitivity balance. We apply a weight of $w_p = 0.5$ to the positive class to prioritize specificity. Below is the outline on generating the forecasts and probabilities:

1) Generate regression forecast: $\widehat{y}_t = \langle\psi_t|B|\psi_t\rangle$;

2) Form true labels for classification: $y_t^p = \mathbf{1}_{y_t > \theta_{LST}}$, where $\theta_{LST} = 12$ is our LST threshold;

3) Form probability forecast for classification: $\widehat{y_t^p} = \sigma\Big((\widehat{y}_t - \theta_{LST})s_\sigma\Big)$, where $\sigma(x) = \frac{1}{(1+e^{-x})}$ is the sigmoid function and s_σ is a learnable scale parameter.

Then the mixed-loss function takes the following form:

$$\mathscr{L}_{\text{Total}} = \frac{\mathscr{L}_{\text{L1}}}{\mathscr{L}_{\text{L1}}^{(\text{gradient-free})}} + \frac{\mathscr{L}_{\text{CE}}}{\mathscr{L}_{\text{CE}}^{(\text{gradient-free})}}, \tag{2}$$

$$\text{where} \quad \mathscr{L}_{\text{L1}} = \frac{1}{T}\sum_t |\widehat{y}_t - y_t|,$$

$$\mathscr{L}_{\text{CE}} = -\frac{1}{T}\sum_t \left[w_p y_t^p \log\left(\widehat{y_t^p}\right) + (1 - y_t^p)\log\left(1 - \widehat{y_t^p}\right)\right].$$

Here, the "gradient-free" loss $\mathscr{L}_{\text{L1}}^{(\text{gradient-free})}$ and $\mathscr{L}_{\text{CE}}^{(\text{gradient-free})}$ are defined within a gradient-descent-based training framework (PyTorch in our case). This allows us to use the evaluated value of the loss while explicitly excluding its gradients from the optimization process. For $\mathscr{L}_{\text{Total}}$, adjusting each loss component by the corresponding "gradient-free" loss forces each component's loss to be equal to 1. However, the gradient $\nabla\mathscr{L}_{\text{Total}}$ will not necessarily be 0, allowing the model to learn while maintaining consistent weighting between the loss components.

2.4 QCML Positive Operator-Valued Measure

QCML Positive Operator-Valued Measure (POVM) extends QCML to predict probability density functions for targets instead of single point estimates. This extension provides the ability to forecast full probability distributions, which enables the estimation of confidence intervals and offers a more detailed understanding of the predictions. This approach is particularly well-suited for our LST forecasting task, as it allows us to produce both continuous-valued predictions (e.g., expected mean or median LST) and probabilistic forecasts (e.g., the likelihood that LST exceeds a specified threshold).

Generalized measurements in quantum mechanics are described by a set of operators known as a Positive Operator-Valued Measure (POVM)[35]. A POVM is a collection $\{\hat{F}_k\}$ of positive semi-definite operators acting on the Hilbert space, such that

$\sum_k \hat{F}_k = \hat{I}$. Each operator $\hat{F}_k$ corresponds to a possible measurement outcome. The connection between the quantum state and the measurement outcomes is provided by Born's rule. For a state $|\psi\rangle$, the probability of observing the outcome associated with the POVM element k is given by $p_k = \langle\psi|\hat{F}_k|\psi\rangle$.

QCML POVM allows us to forecast the probability density function $p(y)$ of a continuous target variable y instead of point estimates. Without loss of generality, we will assume that $y \in [-1,1]$. Suppose that we want to generate a probability density function of a continuous variable y conditional on a quantum state $|\psi\rangle$. We introduce a function mapping y into output operators $\hat{Y}(y)$, generally non-Hermitian, such that

$$\int_{-1}^{1} \hat{Y}^\dagger(y)\hat{Y}(y)dy = \hat{I}. \tag{3}$$

The set of operators $\hat{F}(y) = \hat{Y}^\dagger(y)\hat{Y}(y)$, indexed by the continuous parameter y, forms a POVM. By construction, the probability density function $p(y)$ is given by

$$p(y) = \langle\psi|\hat{Y}^\dagger(y)\hat{Y}(y)|\psi\rangle. \tag{4}$$

The POVM elements $\hat{Y}(y)$ can be parametrized in a variety of ways. Here, we suggest parametrization in terms of a finite number of Legendre polynomials[36] $L_k(y)$:

$$\hat{Y}(y) = \sum_{n=0}^{K-1} \hat{A}_n L_n(y)\sqrt{\frac{2n+1}{2}},$$

where, K is the truncation parameter and $\hat{A}_k$ are generally non-Hermitian matrices to be learned. Then Equation (4) becomes:

$$p(y) = \sum_{n,m} \langle\psi|\hat{A}_n^\dagger\hat{A}_m|\psi\rangle L_n(y)L_m(y)\sqrt{\frac{2n+1}{2}}\sqrt{\frac{2m+1}{2}}.$$

Given the orthonormality of Legendre polynomials:

$$\int_{-1}^{1} L_n(z)L_m(z)dz = \frac{2}{2n+1}\delta_{nm}, \tag{5}$$

it follows from Equation (3) that:

$$\sum_{n=0}^{K-1} \hat{A}_n^\dagger\hat{A}_n = \hat{I}.$$

Therefore, the matrices $\hat{F}_n = \hat{A}_n^\dagger\hat{A}_n$ form a POVM.

For a target variable y on arbitrary support $[a,b]$, we consider a PDF $g(y)$ as an initial guess. We use this distribution to do a variable transformation into $z \in [-1,1]$:

$$z = G(y) = 2\int_a^y g(t)dt - 1.$$

Now we construct the PDF to be learned as:

$$p(y) = 2\sum_{nm}\sqrt{\frac{2n+1}{2}}\sqrt{\frac{2m+1}{2}}\langle\psi|\hat{A}_n^\dagger\hat{A}_m|\psi\rangle L_n(G(y))L_m(G(y))g(y).$$

Using (5) and making a variable transformation $z = G(y)$ and $dz = 2g(y)dy$ we can confirm that

$$\int_a^b p(y)dy = \sum_n \langle\psi|\hat{A}_n^\dagger\hat{A}_n|\psi\rangle = 1.$$

Given a special case of $\langle\psi|\hat{A}_n^\dagger\hat{A}_n|\psi\rangle = \delta_{n0}\delta_{m0}$, we get $p(y) = g(y)$.

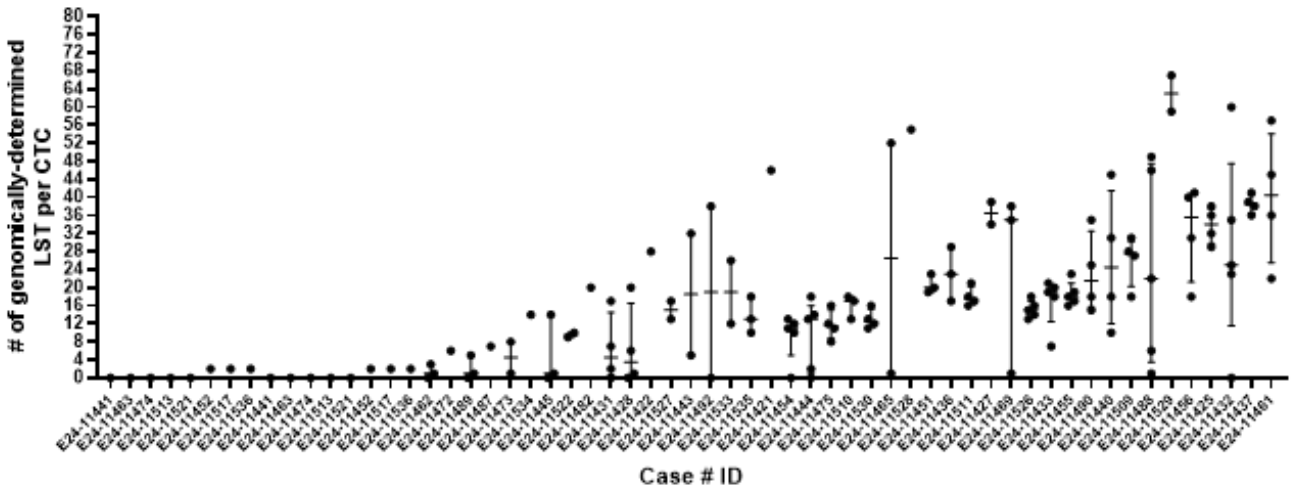


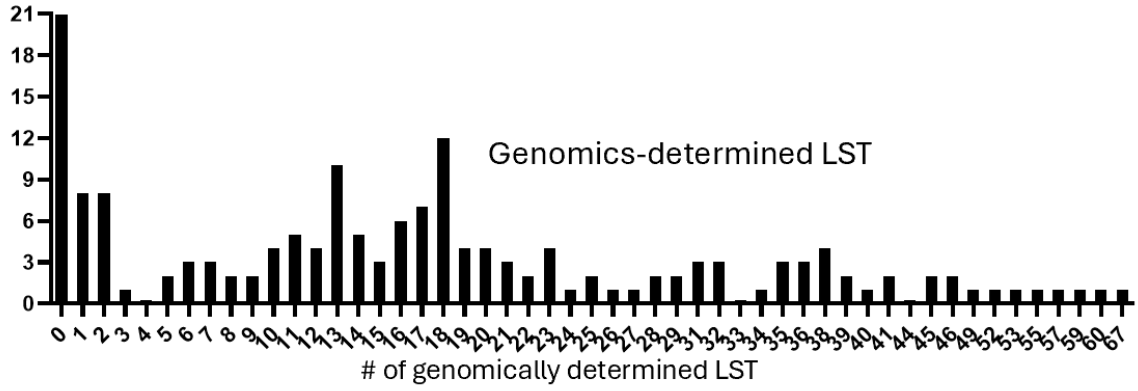
Figure 1. Distribution of genomics-determined LST values across all cases. Scatter plots show the genomically determined LST ground truth values which are shown per CTC, sub-grouped by case ID number and ranked from left to right by increasing LST values. We show the visual representation of the heterogeneity in the distribution of LST values across each case ID.

3 Results

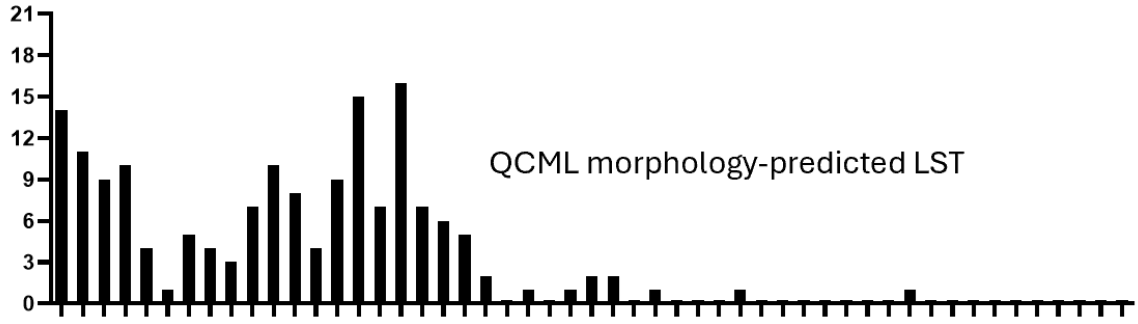
3.1 CTC Dataset Description

A total of 227 available cells were identified across all patients. These cells were 1) selected as candidated CTC by a trained pathologist, 2) sequenced, and 3) met the genomic quality control (QC) metrics established by our pipeline[4]. On average, each patient had 3.25 sequenced CTCs (with a standard deviation of 1.87; the maximum was eight, and the minimum was 1). This data established the ground truth for measuring chromosomal instability based on the number of LST. We divided the overall cohort into a training set, which included 51 patients and 166 CTCs. Figure 1 shows the heterogeneity of LST values across each case in the training set. Figure 2a shows the genomically determined LST values distribution for all cases, organized by case ID number and ranked in ascending order based on LST values. We identify that 73% (37 out of 51) of the cases in the training set had at least one LST+ CTC. Following that, digital pathology features, including cell morphology and fluorescence intensity levels were extracted from each CTC image. As detailed in Methods Section, and previously published[30] 112 morphology features were extracted: 8 from nuclear and cell masks (size, roundness, elongation, and the first Hu moment) and 44 intensity features from DAPI, CK, and CD45/CD31 channels (MFI, lower, median, upper quartiles, interquartile range, and co-localizations). Additionally, 70 texture features were obtained using Gabor filters, applied with 16 parameters to capture frequency content, orientation, and irregularities with fractal approximations. Laws' texture energy measures were used for detecting spatial patterns, and six features were derived from the Local Binary Pattern method to analyze local texture variations across image channels.

To understand the linear relationships between metastatic breast cancer CTCs digital pathology features and their ground truth genomically determined LST, we calculated Pearson's correlations between them. We found that CTCs with a higher degree of chromosomal instability, represented by higher LST values, were significantly correlated with a larger nuclear $(r = 0.22, P = 0.032)$ and cellular $(r = 0.15, P = 0.04)$ morphology size, which is a measure of the overall pixel area of a segmented nucleus and cell (Figure 3a and 3b). It was also substantially correlated with nuclear fractal features which measure shape complexity and heterogeneity $(r = 0.24, P = 0.001)$ (Figure 3c). These results are consistent with nuclear enlargements, spatial disorganization and pleomorphism which may be due to polyploidy, and multinucleation which is expected to occur in genomically unstable cells[3,37,38]. Several CD45/CD31 intensity values were correlated with lower LST with the most significant being cellular low quartile range LQI $(r = -0.21, P = 0.005)$ (Figure 3d). CD45/CD31 is a negative CTC marker that is typically down-regulated in cancer cells of epithelial origin from solid tumors. LST values trended with a lower expression of several DAPI intensities which were most correlated as expressed by mean fluorescent intensity (MFI) in the nuclear mask $(r = -0.21, P = 0.0072)$ (Figure 3e) and the inter quartile range IQI in the cellular mask $(r = -0.23, P = 0.002)$ (Figure 3f). DAPI binds strongly to A-T rich regions of double stranded DNA[39]. However, the inverse correlation of DAPI nuclear and cellular intensities result may be explained by the fact that genomically unstable cells are expected to show unpredictable intensity patterns due to chromatin remodeling, micronuclei, or fragmentation[39–41]. Also, DAPI intensity can be affected by cell cycle phases as G2/M cells are expected to have more DNA content than G1 phase



(a) Genomically determined LST.



(b) Morphology-predicted LST.

Figure 2. Count of CTCs with specific LST values. The bar plot comparison illustrates the number of CTCs ranked from left to right by increasing LST values. The bar plot at the top displays the LST values ground truth, which is the genomically determined LST per CTC, while the bottom plot shows the morphology-predicted LST values computed by QCML. The count of CTCs for the ground truth genomically determined LST follows the same bimodal trend as that of the morphology-predicted LST.

which are typically altered during chromosomal instability[39–41]. Cross-channel Local Binary Pattern (LBP) for CK-DAPI and for CK-CD45/CD31 channel pairs, which is a measure of cross channel correlation and similarity across local binary pattern for channel pairs within a segmented cell[30,42], were found to be significantly inversely correlated with the extent of LST numbers ($r = -0.19, P = 0.01; r = -0.21, P = 0.008$) (Figure 3g and 3h). The cross-channel LBP is a measure of colocalization between proteins of a CTC calculated by comparing pixels intensity in the same position for each of the two channels. In doing so, one can capture subtle spatial relationships and structural changes such as nuclear deformities and cytoskeletal remodeling[30,42]. Therefore, a lower texture cross-channel LBP suggests spatial discordance between nuclear and cytoplasmic structure, which is expected during instability-driven morphological shifts.

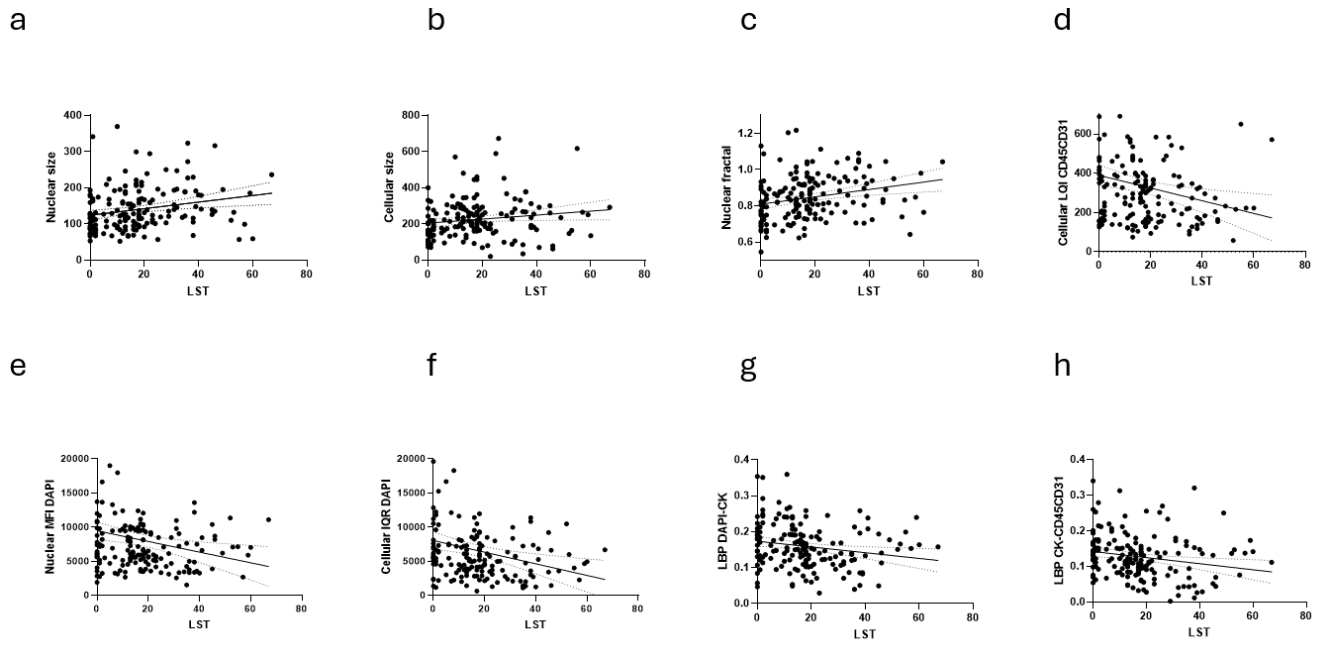


Figure 3. Linear correlation of cellular and nuclear morphology and protein intensity features with the extent of LST as determined by genome sequencing. In the scatter plots each dot is an individual CTC and graphs show the best-fit line with the 95% confidence bands of linear correlation between the extent of LST as a measure of chromosomal instability and a) nuclear size, b) cellular size, c) nuclear fractal, d) cellular LQI CD45/CD31, e) nuclear MFI DAPI, f) cellular IQR DAPI, g) LBP DAPI-CK, h) LBP CK-CD45/CD31.

3.2 Regression Models

We train a QCML model along with classical machine learning models (see Table 1) following the cross-validation procedure described in Section 2.2. The problem is set up as a regression task in which we produce a real-valued LST forecast and then apply a cutoff of LST > 12 to produce a binary forecast of CTCs being LST+. As described in Equation (2) in Section 2.3, our QCML model employs a mixed loss function and allows for varying weights on the positive class, enabling a balance between specificity and sensitivity. The incidence of false positive cases is particularly detrimental in clinical settings and needs to be minimized as much as possible. It may erroneously signal a lack of response from current treatment and lead to an unnecessary change in an otherwise effective line of therapy. For this reason, a weight of 0.5 was applied to the positive class to prioritize specificity and reduce the occurrence of false positive cases. To test the hypothesis that QCML can improve prediction performance and reproducibility across independent datasets, we compared it to 10 different classical machine learning models, including Linear Support Vector Machine (SVM), Neural Networks, Random Forests, XGBoost, Nearest Neighbors, RBF SVM, AdaBoost, Logistic Regression, and Naive Bayes[21] (Table 1).

Models	Training (in-sample)			Verification (out of sample)		
	Sensitivity	Specificity	Balanced Accuracy	Sensitivity	Specificity	Balanced Accuracy
QCML	92% ± 2%	64% ± 5%	78% ± 2%	84% ± 9%	57% ± 10%	70% ± 8%
SVM Gaussian Kernel	95% ± 1%	53% ± 5%	74% ± 2%	90% ± 7%	45% ± 11%	68% ± 7%
Elastic Net	95% ± 1%	51% ± 7%	73% ± 3%	88% ± 6%	40% ± 11%	64% ± 6%
Linear SVM	93% ± 2%	58% ± 4%	75% ± 2%	84% ± 7%	46% ± 13%	65% ± 8%
XGBoost	100% ± 0%	99% ± 1%	100% ± 1%	83% ± 7%	43% ± 11%	63% ± 7%
MLP	98% ± 2%	92% ± 5%	95% ± 2%	74% ± 11%	50% ± 12%	62% ± 8%
AdaBoost	97% ± 1%	69% ± 4%	83% ± 2%	91% ± 5%	39% ± 10%	65% ± 5%
Nearest Neighbors 5	95% ± 1%	44% ± 6%	70% ± 3%	89% ± 8%	36% ± 12%	63% ± 8%
Random Forest	98% ± 1%	69% ± 4%	84% ± 2%	93% ± 6%	34% ± 8%	63% ± 6%
Nearest Neighbors 32	98% ± 1%	18% ± 11%	58% ± 5%	98% ± 4%	12% ± 9%	55% ± 4%
Linear Regression	94% ± 2%	75% ± 4%	85% ± 2%	67% ± 11%	52% ± 11%	60% ± 6%

Table 1. In sample and out of sample performance of QCML and classical ML models forecasting LST+. Showing the average and standard deviation of sensitivity, specificity and balanced accuracy across 25 folds (5-fold with 5 repeats).

QCML shows the highest out-of-sample specificity (57%) concordance to the ground truth while achieving a high sensitivity of 84% and outperforms the rest of the models in terms of balanced accuracy (70%). The results also confirm QCML's capacity to generalize compared to classical models; it shows a smaller disconnect between in-sample and out-of-sample performance, while most of the classical models overfit in-sample and experience substantial reduction in performance out-of-sample. In Figure 2b we also show the distribution of QCML morphology-predicted LST values across all CTCs which follows a similar trend as the count of CTCs of the genomically-determined LST (Figure 2a).

3.3 Classification Models

Here, as opposed to a regression model, we set up the problem as a classification task where we forecast a binary target of CTCs being LST+. Although this approach does not generate real-valued forecasts, it allows evaluating models with various probability thresholds to target a specific balance between specificity and sensitivity. Additionally, it allows measurement of threshold-independent metrics like AUC-ROC to summarize the performance across all possible classification thresholds. For QCML we use the QCML Positive Operator-Valued Measure model to produce probability forecasts, where we model the LST target based on an exponential transformation and use a Legendre polynomial parametrization. Table 2 shows the performance of QCML POVM in conjunction with classical machine learning models based on a probability threshold of 0.6[43,44].

Model	Training (in-sample)				Verification (out of sample)			
	Sensitivity	Specificity	Balanced Accuracy	ROC AUC	Sensitivity	Specificity	Balanced Accuracy	ROC AUC
QCML POVM	95% ± 1%	78% ± 5%	86% ± 3%	0.951	77% ± 10%	57% ± 11%	67% ± 8%	0.763
XGBoost	100% ± 0%	100% ± 0%	100% ± 0%	1.000	78% ± 7%	55% ± 12%	66% ± 8%	0.747
Random Forest	97% ± 1%	100% ± 0%	98% ± 1%	0.999	70% ± 9%	63% ± 10%	67% ± 7%	0.744
Nearest Neighbors 32	83% ± 4%	58% ± 6%	70% ± 2%	0.784	82% ± 10%	53% ± 11%	68% ± 8%	0.737
Nearest Neighbors 5	70% ± 3%	89% ± 3%	80% ± 2%	0.884	63% ± 9%	71% ± 10%	67% ± 7%	0.734
RBF SVM	81% ± 3%	64% ± 4%	72% ± 2%	0.816	74% ± 8%	58% ± 12%	66% ± 7%	0.715
Neural Net	100% ± 0%	100% ± 0%	100% ± 0%	1.000	77% ± 10%	56% ± 10%	67% ± 6%	0.715
Linear SVM	84% ± 3%	70% ± 5%	77% ± 2%	0.860	75% ± 9%	59% ± 12%	67% ± 7%	0.713
AdaBoost	11% ± 9%	100% ± 0%	56% ± 4%	1.000	7% ± 9%	94% ± 7%	51% ± 3%	0.697
Logistic Regression	87% ± 2%	86% ± 3%	86% ± 2%	0.942	73% ± 10%	56% ± 8%	65% ± 7%	0.677
Naive Bayes	65% ± 14%	69% ± 13%	67% ± 3%	0.739	60% ± 15%	60% ± 17%	60% ± 8%	0.643

Table 2. In sample and out of sample performance of QCML POVM and classical ML classification models forecasting LST+. Showing the average and standard deviation of sensitivity, specificity and balanced accuracy across 25 folds (5-fold with 5 repeats). Also showing the average ROC AUC score per model.

QCML POVM achieves the highest ROC AUC score of (0.763) as shown in Table 2 and Figure 4 with ROC AUC curves for top performing models. Additionally, QCML POVM is able to generate full probability densities of LST values as shown in Figure 5.

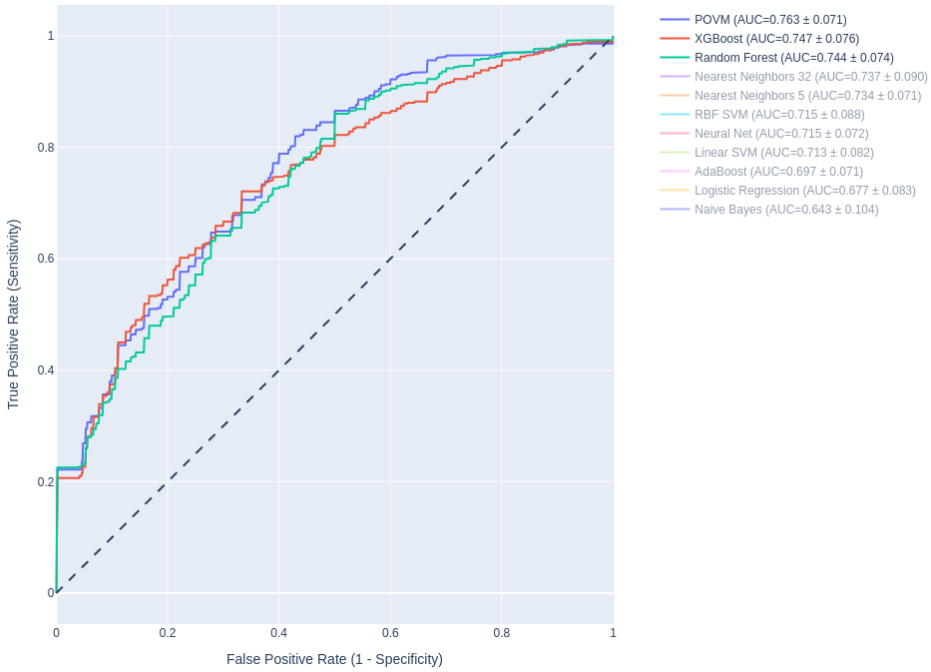


Figure 4. ROC AUC curves for the top three performing models: QCML POVM (blue), XGBoost (red), and Random Forest (green). At ∼70% specificity all three models achieve similar sensitivity. At lower specificity, QCML POVM is outperforming both XGBoost and Random Forest (i.e., it can achieve higher sensitivity for a fixed specificity). At higher specificity, QCML POVM is on par with XGBoost and outperforms Random Forest.

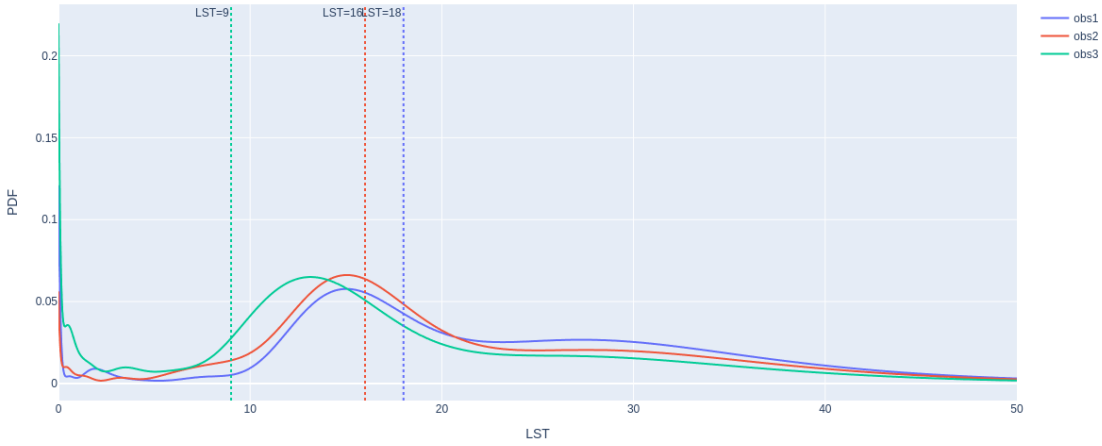


Figure 5. QCML POVM produces full probability distribution of LST for all observations. Showing predicted LST distribution for a sample of 3 CTCs. Dotted vertical lines show the actual LST for each CTC. Having probability distribution allows one to forecast both LST quantiles as well as probabilities for LST being above/below a threshold.

3.4 Gradient-based Feature Importance

QCML has several ways of identifying important features, one of them being the gradient-based feature importance which is based on the impact of changes in input features on the final output of the model. We use this approach to rank CTCs morphology features and protein expressions which were considered to be of high importance for the prediction. As shown in Figure 6, the top 10 CTCs morphology and protein content features that the QCML classifier used as predictor of LST are measures of:

(i) Protein correlation and colocalization between channels such as Cross-channel (LBP) for CK vs DAPI and CK vs CD45/CD31 and cellular colocalization between CK-DAPI;

(ii) Intensity features of the DAPI cellular mask;

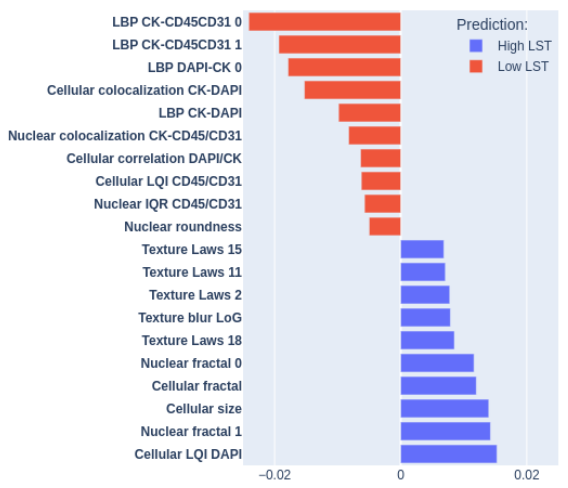
(iii) Fractal nuclear and cellular features.

1) Cross-channel LBP and Cell colocalization features in the QCML model were predictors of low chromosomal instability which is in line with those features' ability to measure cellular remodeling and spatial discordance between nuclear and cytoplasmic structures[42].

2) Interestingly, QCML weighed the signal of the lower DAPI quartile intensity from the cellular mask which indicates the cytoplasmic signal as one of the most important features in predicting higher levels of chromosomal instability (LST+). DAPI stains chromatin and double-stranded DNA which in normal cells typically resides in the nucleus and, for this reason, does not normally stain the cytoplasm[39]. However, QCML data make sense with the underlying biology of cancer cells, as abnormal DAPI signals can appear in the cytoplasmic region due to the presence of micronuclei or nuclear envelope rupture, which typically occurs concomitantly with chromosomal instability[38,41,45,46], thus supporting the results of QCML.

3) QCML identified cellular size, which is biologically relevant, as the presence of cytoplasmic micronuclei and multinucleation is permitted by larger-sized cells and is expected to occur in genomically unstable cells[47,48]. Furthermore, cell size as a predictor of increased chromosomal instability is consistent with previous reports showing that CTC with larger metastatic breast cancer size were associated with the worst patient outcomes[49–52] and for this reason larger cells are expected to be more likely to have increased chromosomal instability.

4) Fractal features are approximations of irregularity and complexity in cellular and nuclear shape, suggesting abnormal nuclear contours that are expected to occur during chromosomal instability[38,53].



Features	Importance	Prediction
LBP CK-CD45CD31 0	0.024	Low LST
LBP CK-CD45CD31 1	0.019	Low LST
LBP DAPI-CK 0	0.018	Low LST
Cellular colocalization CK-DAPI	0.015	Low LST
Cellular LQI DAPI	0.015	High LST
Nuclear fractal 1	0.014	High LST
Cellular size	0.014	High LST
Cellular fractal	0.012	High LST
Nuclear fractal 0	0.012	High LST
LBP DAPI-CK 1	0.010	Low LST

Figure 6. QCML gradient-based feature importance. Left: Top morphological and protein intensity features associated with high LST (blue) and low LST (red) predictions. **Right:** Top feature names based on absolute importance, absolute importance score values, and model predictions (i.e., High LST or Low LST).

3.5 Visualizing CTCs with QCML distance

As mentioned in Section 2.3, QCML represents each observation as a quantum state. This allows one to have a natural notion of proximity between observations, since proximity between quantum states can be defined as *quantum fidelity* [35, III.9]

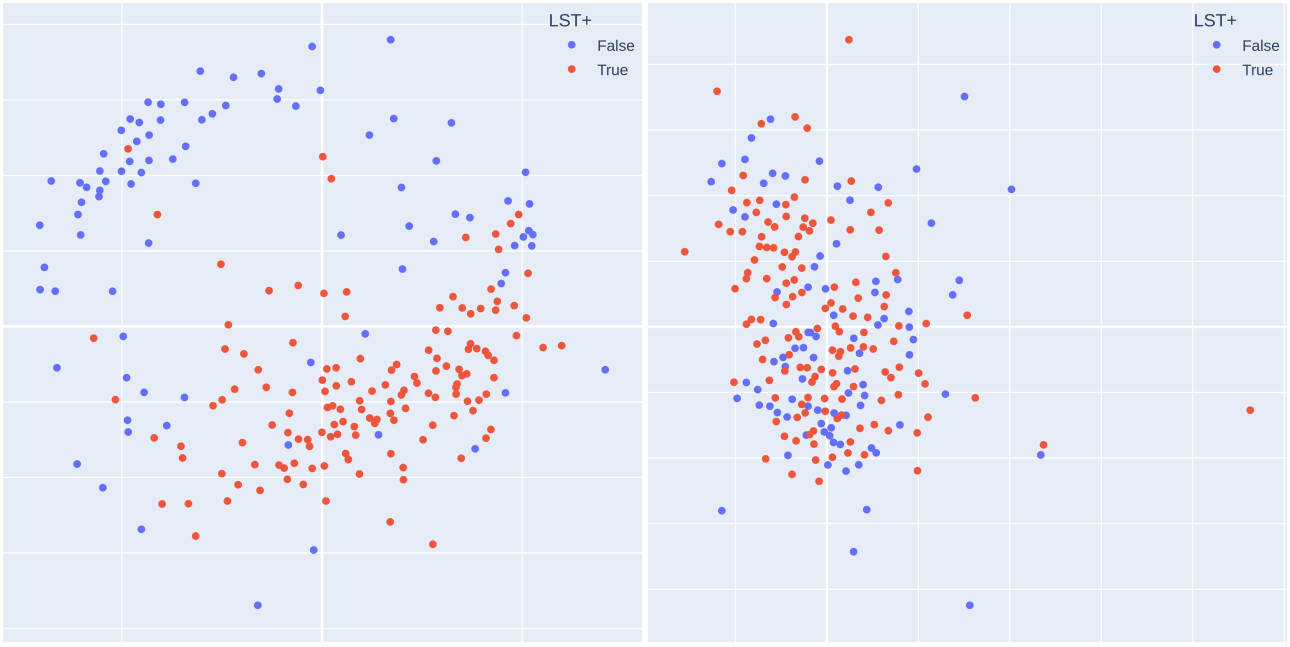
$$f(\psi_1, \psi_2) = |\langle \psi_1 | \psi_2 \rangle|^2,$$

which can be interpreted as the probability of identifying the state ψ_1 with the state ψ_2, when performing a quantum measurement designed to test whether a given quantum state is equal to ψ_2 (or vice versa). In the context of QCML, this type of proximity can be used to define a similarity measure on the data. Given the mapping from an observation to quantum state $\mathbf{x}_t \to |\psi_t\rangle$, we can define the *QCML distance* between two data points $\mathbf{x}_t, \mathbf{x}_{t'}$ as

$$d_Q(\mathbf{x}_t, \mathbf{x}_{t'}) = 1 - f(\psi_t, \psi_{t'}) = 1 - |\langle \psi_t | \psi_{t'} \rangle|^2. \tag{6}$$

Note that in contrast to the standard Euclidean distance between two data points, the QCML distance is a type of supervised similarity measure, since the representation of the data in quantum states ψ_t has been optimized using the training targets.

Given a distance matrix, we can visualize the observations in a two-dimensional space using multidimensional scaling (MDS). This is a common dimensionality reduction technique that can be used to visualize high-dimensional data in two dimensions. In a nutshell, MDS finds a mapping of the high-dimensional data into two dimensions that minimizes the matrix norm of the difference between the distance matrix of the original data and that of the two-dimensional transformation. Using MDS we plot the high-dimensional CTCs data in two dimensions, as shown in Figure 7.



(a) QCML distance

(b) Euclidean distance

Figure 7. Multi-dimensional scaling visualization based on a distance matrix of CTCs. Red color represents LST+ CTCs, while blue color represents stable CTCs. Using QCML distance (a) one can achieve a better separation between LST+ and stable CTCs.

Although the plots in two-dimensional MDS target space do not offer a completely faithful representation of the distances between CTCs, they help qualitatively illustrate some of the differences between QCML and Euclidean distance. Specifically, the QCML distance is much better at finding a separation between genomically unstable and stable CTCs.

4 Discussion

Here, we applied Quantum Cognition Machine Learning (QCML) to digital pathology-derived morphological features and protein expression levels from CTCs, to enable prediction of chromosomal instability. The QCML-based pLST model outperformed conventional machine learning approaches in predictive accuracy. The ultimate objective of deploying a morphology prediction of LST algorithm in liquid biopsy CTC assays is to enable real-time detection of CTCs with elevated level of chromosomal instability and high metastatic potential. As a result, pLST detection can reduce diagnostic latency and circumvent the long turnaround times associated with whole genome sequencing in clinical workflows.

Chromosomal instability is a key molecular driver of tumor heterogeneity, which in turn supports the high plasticity and evolutionary adaptability of cancer cells in response to environmental pressures, ultimately enabling metastasis[10,11,13–19]. Therefore, the detection of cancer cells through preconceived expert knowledge of their expected biological phenotype may be a limitation, as it may not account for tumor evolution and the emergence of new CTC phenotypes[54–56]. To address this challenge, the present study employed a previously published[4,30], systematic and quantifiable approach to nuclear and cellular segmentation using digital pathology to extract the broadest set of morphological, texture and intensity-based characteristics[4,30]. This approach provides an optimal foundation for the application of advanced machine learning models, including QCML. Looking ahead, future research should explore quantum-assisted feature mapping to identify complex and subtle patterns directly from raw image pixel data to see whether such an approach may potentially surpass the performance of biologically

informed feature extraction. This direction may further enhance the ability to detect chromosomal instability in CTCs without relying on guided feature extraction.

One of the key challenges in applying machine learning to single cell diagnostics is the inherent biological complexity, which becomes especially problematic when the number of features (e.g., genes, or digital pathology features) far exceeds the number of samples[57,58]. This imbalance, common in genomics, proteomics, and digital pathology, can lead to overfitting and poor generalization[57]. Classical statistical methods, such as Bayesian inference, often require data volumes that grow exponentially with the number of features, making them impractical in such high-dimensional biological settings[21–23]. To address this, modern approaches in genomics and digital pathology typically reduce dimensionality by selecting curated "signatures" based on preconceived biological knowledge. Rather than relying on statistical associations on all available features, these curated features reflect a mechanistic understanding of biological systems. The interpretability of the features typically reduces the overfitting to the training data and improves the robustness and reproducibility of the model[21–23]. In other words, we apply our human cognition to design predictive models that make sense and draw conclusions ignoring irrelevant information. However, such feature engineering is a way to offset machine learning limitations by leveraging human intervention and its understanding of the biological problem[57,59].

It has been proposed that machine learning algorithms should learn representations of the data by disentangling explanatory factors, mimicking human cognition in understanding disease mechanisms[57,59]. In a similar attitude, more advanced and recent approaches leverage principles from quantum theory to address high-dimensional data representation by simulating cognition[25,28]. By adopting the formalism of quantum probabilities, particularly the uncertainty principle, data can be encoded as vectors within a Hilbert space, where no state corresponds to an exact position of the features configuration. Therefore, through a simulation of quantum principles using classical computers, we enable an intrinsic reduction in feature representation, offering a novel way to manage complexity and dimensionality in biomedical data analysis.

Following these principles, in the present study, QCML was applied to abstract out the features that are the most fundamental to estimate the intrinsic dimensions of the CTC morphology data. By doing this, and without human curation, QCML learned a gradient of feature importance that was found *posthoc* to be biologically and mechanistically involved with chromosomal instability in cancer cells. QCML's prediction of chromosomal instability classification abstracted a model for instability-driven morphological shifts where CTCs are larger in cellular size with higher spatial discordance between nuclear and cytoplasmic structures. The texture cross-channel measures of colocalization identified by QCML suggest that CTCs with chromosomal instability may be more likely to have poorly aligned nuclear, subcellular and cytoskeletal textures. Those indicate structural rearrangements and nuclear pleomorphism which have been previously linked to genomically unstable tumor cells[60]. In addition, the morphological manifestation of chromosomal instability, which can be perceived as lower cellular integrity, has been shown to provide functions that could ultimately be evolutionary advantageous for cancer[41]. QCML findings were also corroborated by the evidence that the cellular localization of DAPI intensity and size is important for the prediction of chromosomal instability. In cancer cells, the distinction between nuclear and cellular (cytoplasmic) DAPI expression becomes especially important, as abnormalities in DAPI localization and intensity (e.g. small, round DAPI-positive bodies in cytoplasm) can reveal hallmarks of chromosomal instability, and architecture defects[38,39,41,45,46].

Future studies will be required to validate these findings and establish whether the presence of CTC with predicted chromosomal instability classified by QCML can predict patient survival with better performance compared to conventional methods.

References

1. Gradishar, W. J. *et al.* Breast cancer, version 3.2022, nccn clinical practice guidelines in oncology. *J Natl Compr Canc Netw* **20**, 691–722, DOI: 10.6004/jnccn.2022.0030 (2022).

2. Wolff, A. C., American Society of Clinical Oncology & College of American Pathologists. Recommendations for her2 testing in breast cancer. *Arch Pathol Lab Med* **138**, 241–256, DOI: 10.5858/arpa.2013-0953-SA (2014).

3. Schonhoft, J. D. *et al.* Morphology-predicted large-scale transition number in circulating tumor cells identifies a chromosomal instability biomarker associated with poor outcome in castration-resistant prostate cancer. *Cancer Res* **80**, 4892–4903, DOI: 10.1158/0008-5472.CAN-20-1216 (2020).

4. Di Caro, G. *et al.* A novel liquid biopsy assay for detection of erbb2 (her2) amplification in circulating tumor cells (ctcs). *J Circ Biomark* **13**, 27–35, DOI: 10.33393/jcb.2024.3046 (2024).

5. Di Cosimo, S. *et al.* Low-pass whole genome sequencing of circulating tumor cells to evaluate chromosomal instability in triple-negative breast cancer. *Sci Rep* **14**, 20479, DOI: 10.1038/s41598-024-71378-3 (2024).

6. Greene, S. B. & Dago, A. E. Chromosomal instability estimation based on sequencing. *PLoS One* **11**, DOI: 10.1371/journal.pone.0165089 (2016).

7. Tellez-Gabriel, M., Ory, B., Lamoureux, F., Heymann, M. F. & Heymann, D. Tumour heterogeneity: The key advantages of single-cell analysis. *Int J Mol Sci* **17**, DOI: 10.3390/ijms17122142 (2016).

8. Malihi, P. D. *et al.* Single-cell circulating tumor cell analysis reveals genomic instability as a distinctive feature of aggressive prostate cancer. *Clin Cancer Res* **26**, 4143–4153, DOI: 10.1158/1078-0432.CCR-19-4100 (2020).

9. Brown, L. C. *et al.* Circulating tumor cell chromosomal instability and neuroendocrine phenotype by immunomorphology and poor outcomes in men with mcrpc treated with abiraterone or enzalutamide. *Clin Cancer Res* **27**, 4077–4088, DOI: 10.1158/1078-0432.CCR-20-3471 (2021).

10. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability–an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* **11**, 220–8, DOI: 10.1038/nrm2858 (2010).

11. Sansregret, L., Vanhaesebroeck, B. & Swanton, C. Determinants and clinical implications of chromosomal instability in cancer. *Nat Rev Clin Oncol* **15**, 139–150, DOI: 10.1038/nrclinonc.2017.198 (2018).

12. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell* **168**, 613–628, DOI: 10.1016/j.cell.2017.01.018 (2017).

13. Duijf, P. H. G. *et al.* Mechanisms of genomic instability in breast cancer. *Trends Mol Med* **25**, 595–611, DOI: 10.1016/j.molmed.2019.04.004 (2019).

14. Hanahan, D. Hallmarks of cancer: New dimensions. *Cancer Discov* **12**, 31–46, DOI: 10.1158/2159-8290.CD-21-1059 (2022).

15. Eccleston, A. Targeting cancers with chromosome instability. *Nat Rev Drug Discov* **21**, 556, DOI: 10.1038/d41573-022-00111-4 (2022).

16. Drews, R. M. *et al.* A pan-cancer compendium of chromosomal instability. *Nature* **606**, 976–983, DOI: 10.1038/s41586-022-04789-9 (2022).

17. Sanchez-Vega, F. *et al.* Oncogenic signaling pathways in the cancer genome atlas. *Cell* **173**, 321–337 e10, DOI: 10.1016/j.cell.2018.03.035 (2018).

18. Lee, J. K., Choi, Y. L., Kwon, M. & Park, P. J. Mechanisms and consequences of cancer genome instability: Lessons from genome sequencing studies. *Annu. Rev Pathol* **11**, 283–312, DOI: 10.1146/annurev-pathol-012615-044446 (2016).

19. Sansregret, L. & Swanton, C. The role of aneuploidy in cancer evolution. *Cold Spring Harb Perspect Med* **7**, DOI: 10.1101/cshperspect.a028373 (2017).

20. Liu, Z., Zhu, L., Roberts, R. & Tong, W. Toward clinical implementation of next-generation sequencing-based genetic testing in rare diseases: Where are we? *Trends Genet.* **35**, 852–867, DOI: 10.1016/j.tig.2019.08.006 (2019).

21. Sommer, C. & Gerlich, D. W. Machine learning in cell biology - teaching computers to recognize phenotypes. *J Cell Sci* **126**, 5529–39, DOI: 10.1242/jcs.123604 (2013).

22. Weiskittel, T. M. *et al.* The trifecta of single-cell, systems-biology, and machine-learning approaches. *Genes (Basel)* **12**, DOI: 10.3390/genes12071098 (2021).

23. Xu, J. *et al.* Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. *Hum Genet.* **138**, 109–124, DOI: 10.1007/s00439-019-01970-5 (2019).

24. Li, Y., Wu, X., Fang, D. & Luo, Y. Informing immunotherapy with multi-omics driven machine learning. *NPJ Digit. Med* **7**, 67, DOI: 10.1038/s41746-024-01043-6 (2024).

25. Candelori, L. *et al.* Robust estimation of the intrinsic dimension of data sets with quantum cognition machine learning (2024). https://arxiv.org/abs/2409.12805.

26. Musaelian, K. *et al.* Quantum cognition machine learning: AI Needs Quantum (2024). https://www.qognitive.io/papers/QCML-Qognitive,Inc.pdf.

27. Samson, R. *et al.* Quantum cognition machine learning: financial forecasting (2024). Risk.net.

28. Rosaler, J. *et al.* Supervised similarity for high-yield corporate bonds with quantum cognition machine learning (2025). 2502.01495.

29. Steinacker, H. C. *Quantum Geometry, Matrix Theory, and Gravity* (Cambridge University Press, 2024).

30. Schwab, E. *et al.* Fully automated ctc detection, segmentation and classification for multi-channel if imaging. In *Medical Optical Imaging and Virtual Microscopy Image Analysis*, 55–65 (2025).

31. Human genome assembly grch38 (2013).

32. Ming-Kuei, H. Visual pattern recognition by moment invariants. *IRE Transactions on Inf. Theory* **8**, 179–187, DOI: 10.1109/TIT.1962.1057692 (1962).

33. Laws, K. *Textured Image Segmentation* (University of Southern California, 1980).

34. Pothos, E. M. & Busemeyer, J. R. Quantum cognition. *Annu. Rev. Psychol.* **73**, 749–778, DOI: 10.1146/annurev-psych-033020-123501 (2022). 10.1146/annurev-psych-033020-123501.

35. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge University Press, 2000).

36. Weber, H.-J. & Arfken, G. B. *Mathematical methods for physicists*, vol. 148 (Elsevier Academic Cambridge, MA, USA, 2005).

37. Abel, J. *et al.* Ai powered quantification of nuclear morphology in cancers enables prediction of genome instability and prognosis. *NPJ Precis. Oncol.* **8**, 134, DOI: 10.1038/s41698-024-00623-9 (2024).

38. Chow, K. H., Factor, R. E. & Ullman, K. S. The nuclear envelope environment and its cancer connections. *Nat Rev Cancer* **12**, 196–209, DOI: 10.1038/nrc3219 (2012).

39. Ferro, A. *et al.* Blue intensity matters for cell cycle profiling in fluorescence dapi-stained images. *Lab Invest* **97**, 615–625, DOI: 10.1038/labinvest.2017.13 (2017).

40. Pentzold, C., Kokal, M., Pentzold, S. & Weise, A. Sites of chromosomal instability in the context of nuclear architecture and function. *Cell Mol Life Sci* **78**, 2095–2103, DOI: 10.1007/s00018-020-03698-2 (2021).

41. Lim, S., Quinton, R. J. & Ganem, N. J. Nuclear envelope rupture drives genome instability in cancer. *Mol Biol Cell* **27**, 3210–3213, DOI: 10.1091/mbc.E16-02-0098 (2016).

42. Ojala, T., Pietikäinen, M. & Harwood, D. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognit.* **29**, 51–59 (1996).

43. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35, DOI: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3 (1950).

44. Perkins, N. J. & Schisterman, E. F. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am. J. Epidemiol.* **163**, 670–675, DOI: 10.1093/aje/kwj063 (2006).

45. Krupina, K., Goginashvili, A. & Cleveland, D. W. Causes and consequences of micronuclei. *Curr Opin Cell Biol* **70**, 91–99, DOI: 10.1016/j.ceb.2021.01.004 (2021).

46. Kalsbeek, D. & Golsteyn, R. M. G2/m-phase checkpoint adaptation and micronuclei formation as mechanisms that contribute to genomic instability in human cells. *Int J Mol Sci* **18**, DOI: 10.3390/ijms18112344 (2017).

47. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer* **1**, 800–810, DOI: 10.1038/s43018-020-0085-8 (2020).

48. Cancer Genome Atlas Research Network. Electronic address, e. d. s. c. & Cancer Genome Atlas Research, N. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell* **171**, 950–965 e28, DOI: 10.1016/j.cell.2017.10.014 (2017).

49. Mu, Z. *et al.* Prognostic values of cancer associated macrophage-like cells (caml) enumeration in metastatic breast cancer. *Breast Cancer Res Treat* **165**, 733–741, DOI: 10.1007/s10549-017-4372-8 (2017).

50. Tang, C. M. *et al.* Blood-based biopsies-clinical utility beyond circulating tumor cells. *Cytom. A* **93**, 1246–1250, DOI: 10.1002/cyto.a.23573 (2018).

51. Baak, J. P., Van Dop, H., Kurver, P. H. & Hermans, J. The value of morphometry to classic prognosticators in breast cancer. *Cancer* **56**, 374–82, DOI: 10.1002/1097-0142(19850715)56:2<374::aid-cncr2820560229>3.0.co;2-9 (1985).

52. Pienta, K. J. & Coffey, D. S. Correlation of nuclear morphometry with progression of breast cancer. *Cancer* **68**, 2012–6, DOI: 10.1002/1097-0142(19911101)68:9<2012::aid-cncr2820680928>3.0.co;2-c (1991).

53. Zimmermann, A. *Nucleus, Nuclear Structure, and Nuclear Functions: Pathogenesis of Nuclear Abnormalities in Cancer*, 3071–3087 (Springer International Publishing, 2017).

54. Alix-Panabieres, C. & Pantel, K. Challenges in circulating tumour cell research. *Nat Rev Cancer* **14**, 623–31, DOI: 10.1038/nrc3820 (2014).

55. Micalizzi, D. S., Maheswaran, S. & Haber, D. A. A conduit to metastasis: circulating tumor cell biology. *Genes Dev* **31**, 1827–1840, DOI: 10.1101/gad.305805.117 (2017).

56. Joosse, S. A., Gorges, T. M. & Pantel, K. Biology, detection, and clinical implications of circulating tumor cells. *EMBO Mol Med* **7**, 1–11, DOI: 10.15252/emmm.201303698 (2015).

57. Uhler, C. Building a two-way street between cell biology and machine learning. *Nat Cell Biol* **26**, 13–14, DOI: 10.1038/s41556-023-01279-6 (2024).

58. Ji, Y., Lotfollahi, M., Wolf, F. A. & Theis, F. J. Machine learning for perturbational single-cell omics. *Cell Syst* **12**, 522–537, DOI: 10.1016/j.cels.2021.05.016 (2021).

59. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* **35**, 1798–828, DOI: 10.1109/TPAMI.2013.50 (2013).

60. Fu, Y. *et al.* Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer* **1**, 800–810, DOI: 10.1038/s43018-020-0085-8 (2020).

Author contributions

G.D.C. designed the study, and drafted the manuscript. V.K. designed and performed the QCML analysis. E.T.L extracted the image features, with all authors contributing to the research and providing feedback and advice.

Data Availability

The underlying data is under restricted access in compliance with ethical principles.

Competing Interests

The authors declare that they have no competing interests. Research support for this study was funded by Epic Sciences and Qognitive.