IOP Publishing Mach. Learn.: Sci. Technol. 6 (2025) 015006 https://doi.org/10.1088/2632-2153/ada087

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
13 July 2024

REVISED
24 October 2024

ACCEPTED FOR PUBLICATION
16 December 2024

PUBLISHED
13 January 2025

PAPER

Automation of quantum dot measurement analysis via explainable machine learning

Daniel Schug[1,2], Tyler J Kovach[3], M A Wolfe[3], Jared Benson[3], Sanghyeok Park[3], J P Dodson[3], J Corrigan[3], M A Eriksson[3] and Justyna P Zwolak[2,4,5,*]

[1] Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742, United States of America
[2] National Institute of Standards and Technology, Gaithersburg, MD 20899, United States of America
[3] Department of Physics, University of Wisconsin-Madison, Madison, WI 53706, United States of America
[4] Joint Center for Quantum Information and Computer Science, University of Maryland, College Park, MD 20742, United States of America
[5] Department of Physics, University of Maryland, College Park, MD 20742, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: jpzwolak@nist.gov

Keywords: explainable machine learning, explainable boosting machines, semiconductor quantum dots

Abstract

The rapid development of quantum dot (QD) devices for quantum computing has necessitated more efficient and automated methods for device characterization and tuning. Many of the measurements acquired during the tuning process come in the form of images that need to be properly analyzed to guide the subsequent tuning steps. By design, features present in such images capture certain behaviors or states of the measured QD devices. When considered carefully, such features can aid the control and calibration of QD devices. An important example of such images are so-called *triangle plots*, which visually represent current flow and reveal characteristics important for QD device calibration. While image-based classification tools, such as convolutional neural networks (CNNs), can be used to verify whether a given measurement is *good* and thus warrants the initiation of the next phase of tuning, they do not provide any insights into how the device should be adjusted in the case of *bad* images. This is because CNNs sacrifice prediction and model intelligibility for high accuracy. To ameliorate this trade-off, a recent study introduced an image vectorization approach that relies on the Gabor wavelet transform (Schug *et al* 2024 *Proc. XAI4Sci: Explainable Machine Learning for Sciences Workshop (AAAI 2024) (Vancouver, Canada)* pp 1–6). Here we propose an alternative vectorization method that involves mathematical modeling of synthetic triangles to mimic the experimental data. Using explainable boosting machines, we show that this new method offers superior explainability of model prediction without sacrificing accuracy. This work demonstrates the feasibility and advantages of applying explainable machine learning techniques to the analysis of QD measurements, paving the way for further advances in automated and transparent QD device tuning.

1. Introduction

In many machine learning (ML) applications, there has been a longstanding trade-off between the accuracy and interoperability of candidate models [2–4]. This is evident in the extreme example of deep neural networks (DNNs), which can offer excellent accuracy for many problems, often surpassing existing methods [5–7]. Yet, the best-performing ML models are limited in their interpretability due to the number of inaccessible layers. Alternatively, simple techniques like linear models or decision trees allow the user to fully comprehend the internal weights. However, these techniques often cannot model the complex relationships seen in modern datasets. For image data, there has been considerable progress toward finding a middle ground, typically through explaining complex models with surrogates such as the local interpretable model-agnostic explanations (LIME) [8] and Shapley [9].

In contrast with many black-box ML models, explainable boosting machines (EBMs) are a glass-box method that enables the model to be directly interpretable rather than relying on surrogate explanations [10]. Specifically, EBMs extend generalized additive models to include pairwise interactions, allowing one to observe the relationship between features. EBMs often provide accuracy on par with many black-box models with the additional advantage of enhanced intelligibility, which makes them an appealing replacement for other models, especially in applications of consequence, such as medicine [11–13] or finance [14]. However, to date, EBMs have not been adapted to any other data type than tabular data.

For spatial data, such as images, interpretability is more challenging. This is partly due to their composition by structures highly correlated at multiple scales in the two-dimensional (2D) space. This feature is one of the main reasons why convolutional neural networks (CNNs) [5] and, more recently, vision transformers (ViTs) [15] have quickly become the dominant ML approach for many computer vision tasks. However, the black-box nature of CNNs and ViTs makes their use prohibitive in applications where a solid understanding of model predictions is necessary. While making CNNs and ViTs more interpretable has been a very active area of research, techniques proposed to date vary in utility in many applications, especially as the depth of the neural network increases [16–18].

Our previous work developed a methodology that addresses some of these interpretability concerns for measurement data by combining image vectorization with EBMs. Using EBMs as models for image data poses numerous challenges, the principal of which is the mapping from images to a vector representation that could be used directly with EBMs. To achieve this goal, we used the Gabor Wavelet transform and a constrained optimization procedure to extract key image features from the data. We also applied custom feature engineering to tailor this process to the particular dataset [19]. To ensure that the resulting model produces human-agreeable interpretations, we relied on domain knowledge and understanding of the physical systems under investigation to inform the feature extraction process. Here, we demonstrate that the same approach can be successfully applied to assist in the tune-up of accumulation mode Si/Si_xGe_{1-x} quantum dot (QD) devices. We also propose an alternative image vectorization method involving the generation of synthetic data to approximate the experimentally acquired scans [1]. We then show that both methods result in comparable performance, but the latter produces more intuitive and easier to interpret features.

The paper is organized as follows: in section 2, we provide a brief overview of the scientific context of the problem. Section 2.1 describes the problem of tuning QD devices and introduces the concept of triangle plots. Data used for benchmarking is discussed in section 2.2. The two vectorization methods, the Gabor filterbank approach and the synthetic data modeling approach, are described in section 2.3 and section 2.5, respectively. Finally, the EBMs performance on experimental data vectorized using both methods, as well as using a hybrid approach where the dominant Gabor filter is combined with synthetic data, is presented in section 3. We conclude with a discussion of the future direction in section 4.

2. Background and methods

In this section, we introduce the scientific context of our problem and the data utilized in our experiments. We also present an overview of the signal processing methods used in the *Gabor filterbank approach* and the *synthetic data modeling approach*.

2.1. QD tuning problem and triangle plots

Arrays of QDs–interconnected islands of electrons confined in a semiconductor heterostructure with unique properties that allow them to act as artificial atoms–are a leading candidate for use as qubits, the fundamental information carriers in quantum computers [20]. We use a type of gate-defined QD device, as shown in figure 1(a), in which three layers of overlapping gates fabricated on top of a Si/Si_xGe_{1-x} heterostructure are used to form and control QDs [21, 22]. Scaling these systems to large arrays suitable for quantum computations is a challenging task, as with the number of QDs, the number of gates needed to control them grows, making the manual tuning process unfeasible. An autotuning framework incorporating ML tools was originally proposed and validated off-line in [23] using premeasured experimental scans capturing a large range of gate voltages and then deployed online (i.e. *in situ*) to tune a double QD in real-time in [24]. A detailed description of the tuning process is available in [25].

One sub-problem within the tuning procedure is determining the voltage placement of gates to allow the formation of isolated current channels inside the 2D electron gas formed at the intersection of the Si and Si_xGe_{1-x} layers in the heterostructure, see figure 1(a). One way to achieve proper gate voltages is by sweeping certain gates, which we call screening gates, [S_1, S_2, S_3, and S_4 in figure 1(a)] until the currents begin flowing in the correct channels. The bright bars at the edges of the images in figures 1(b) and (c), called *walls*, are due to the current flowing strictly under those gates. The region with the ridged pattern figure 1(b) is the area of

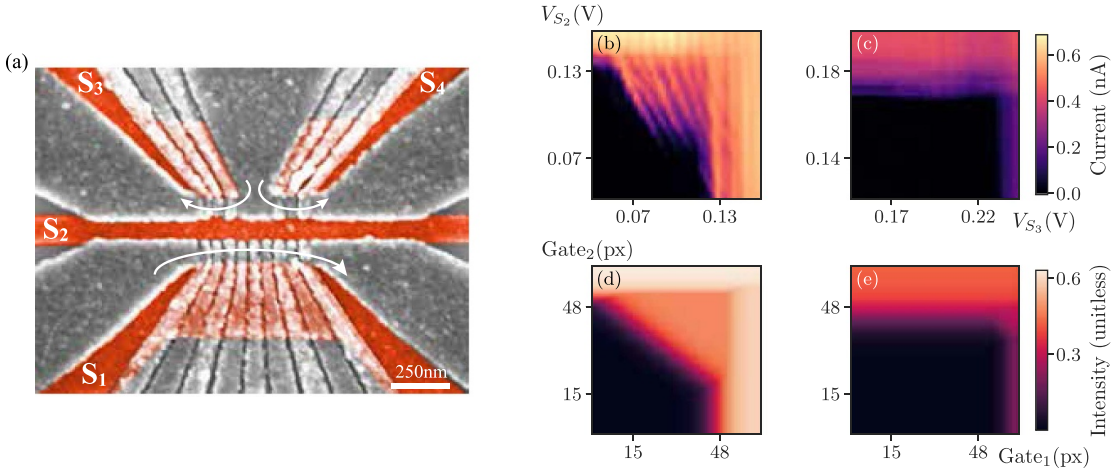

Figure 1. (a) False-colored SEM micrograph of a typical Si/Si$_x$Ge$_{1-x}$ heterostructure device. The red highlighted gates are the screening gates that are swept for the triangle plot. This quad-QD device has three distinct current channels between the screening gates marked with white arrows. One of the channels (in the lower half of the device) contains four QDs in series while the other two channels (in the upper half of the device) are single QD channels for charge-based single-electron-transistor readout. (b) An example of a good triangle plot with high visibility in the triangle region above the background and (c) a corresponding synthetic triangle plot. (c) An example of a bad triangle plot with little to no current in the triangle region and (d) a corresponding synthetic triangle plot.

interest—it indicates that the current is not flowing under any screening gates, but instead, it is flowing between them. If there is no current in this region, as shown in figure 1(c), it is necessary to increase the voltage of gates over the current channel to accumulate more electrons in the 2D electron gas and try again.

In gate-defined QD devices, the area of interest usually takes the form of a triangle—thus the name *triangle plots*—and indicates the formation of an isolated current channel. Due to the physics principles at play, the only possible orientation of the triangle plot is the one with the right angle in the top-right corner of the image. The patterned texture inside this region is due to charge defects and variability in gate uniformity near the current channel [26]. Combining this with the fact that there is no complete analytical form for the triangle plot at this time makes physical modeling of the measurement impossible. While researchers can easily visually identify these regions, finding them automatically using simple threshold analysis is challenging; largely due to experimental variations for different voltage configurations and when measuring different pairs of gates. In moving towards automating the tuning process, it is desirable to have an algorithm that can (1) predict whether the triangle plot indicates a well-behaving current channel and (2) explain this prediction.

2.2. Triangle plots dataset

The images in the triangle plots dataset represent various scales and orientations of image data as well as variations in the triangle region size. About 90% of the images were taken at the high temperature of 1.3 K, instead of roughly 100 mK, which means that the Coulomb blockade and other gate turn-ons are much less steep than in typical measurements. Since the high-intensity features (e.g. the Coulomb blockade fringes), while prominent to the eye, are less predictive of correct device performance, we focus on the larger-scale features, such as the presence or absence of the triangle region.

For our study, images that contain the triangle region, as shown in figure 1(a), are considered *good*. Images that do not contain the triangle region are considered *bad*. We use 902 experimentally acquired triangle plots, of which a total of 210 are labeled as having good triangles and 692 are labeled as having bad triangles.

The images range in size from 31×31 to 171×151 pixels. Prior to analysis, all images are resized to 64×64 pixels, utilizing a bicubic interpolation provided by the Pillow image library [27]. This is done to enable using fixed-parameter filterbanks and synthetic triangles with comparable parameter ranges. While bicubic rescaling can distort the angle of the diagonal region, the majority of the samples that needed preprocessing were in the bad class and thus had minimal to no diagonal activity. In the good class, the maximum observed aspect distortion is within about 13%, which corresponds to about 3.5° difference and does not present significant errors in the context of the Gabor filters. Each image is then vectorized following the procedure described in the Gaber filterbank approach (section 2.3) and the synthetic data modeling approach (section 2.5). The QFlow Triangles dataset containing all experimental images as well as the corresponding synthetic data and Gaber filterbanks is available at Zenodo [28].

2.3. Gabor filterbank approach

To vectorize the image data, we utilize the 2D Gabor wavelet transform, an oriented multi-scale representation that is shown in [29] to be a model for complex neurons in mammalian vision. Gabor wavelet transform has seen frequent use in numerous computer vision tasks and serves as an economical and effective feature transform. As such, this representation seems ideal for extracting oriented features, as well as textures, from image data.

Definition. The 2D Gabor Kernel for parameters $p = \{\sigma_x, \sigma_y, \lambda, \theta\}$ for $(x, y) \in \mathbb{R}^2$ is defined as

$$G_p(x,y) = \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} e^{i\lambda(x\sin(\theta)+y\cos(\theta))}, \tag{1}$$

where σ_x and σ_y represent scale in x and y, and θ and λ are the wave direction and wavelength. We further denote the convolutional application of kernel G_p to an image $u(x,y)$ to be $G_p(x,y) * u(x,y)$.

Specifically, we consider small filterbanks of Gabor wavelets to capture scales and orientations directly relevant to the triangle plots image data. Relying on wavelets leads to a significantly more compact representation.

Definition. The 2D Gabor filterbank for discrete set of N parameters $P = \{p_0, \ldots, p_N\}$

$$G_P(x,y) = \left\{G_{p_i}(x,y)\right\}_{i=0}^{N}. \tag{2}$$

We denote the application of the filterbank G_P to an image u to be $G_P * u = \{G_{p_i}(x,y) * u(x,y)\}_{i=0}^{N}$

In practice, we attempt to construct the set of parameters P such that the Fourier transform of the filterbank G_P supports prominent frequencies observed in the Fourier transform of the image u. This can be accomplished with optimization or other filterbank construction techniques [30–32].

To produce the vector representation of the image, we take advantage of heuristics. In particular, the main factor allowing the discrimination of good and bad triangles is the presence of the pattern of line-like features at approximately 45° orientation corresponding to the interaction between the two gates, see figure 1(b). The image is vectorized by taking the L^2-norm of each filter in the Gabor filterbank with this orientation and at different scales. This achieves a single numerical measure of the filter's response to the given image, which is ultimately descriptive due to the images' simple structural content. The specific scales used were obtained through an iterative process. At each iteration, an EBM model is trained on an available filterbank, starting with one that consists of a large number of possible scales and is refined by eliminating scales that are not discriminative. The process is repeated until no further reduction of the filterbank size is observed. The resulting filterbank contains scales determined to be highly discriminative.

In our case, the final filterbank consists of six filters at scales $\sigma_x = \sigma_y = 4, 8, 16$, and orientation 45° and −45° degrees, where λ is fixed to 1. We also perform additional feature engineering in the form of extracting the estimated location of narrow, edge-like Gabor filters to localize the boundaries of the walls and triangle regions. This technique is used to absorb important spatial features for later use in classification. While other techniques for vectorizing the data based on orientation or scale are possible and, in some instances, advantageous, we chose Gabor because it enables encapsulating the specific diagonal behavior in very few filters. Such vectorization is desirable to ensure that the prediction interpretation is given in terms of relevant features.

Figure 2(b) shows a weighted sum of the filters of the final filterbank for a scan shown in figure 2(a), with weights defined by the contribution of each term in the EBM. While the resulting filterbank is an incomplete representation of the data, it covers primary discriminant support in the frequency domain between classes. The features are designed to measure the extent of response of the diagonal component of the triangle plot at different scales, where we operate under the intuition that good triangles will have a greater associated response at the ±45° orientations than bad triangles, within a region resembling a fringed isosceles right triangle. The location of the edge-like filters acts as a further sieve for the presence and quality of a triangle region. While other vectorization techniques might provide more complete representations of the data and, in some cases, offer superior classification performance, this effect can be diminishing while harming the overall interpretability.

2.4. Synthetic triangle plots

To enable the creation of interpretable features, we generate a set of crude synthetic triangle plots. The synthetic data share certain salient features visible in experimental triangle plots, namely the presence of walls of varying width and a possible diagonal region. It is important to note that while the resulting images

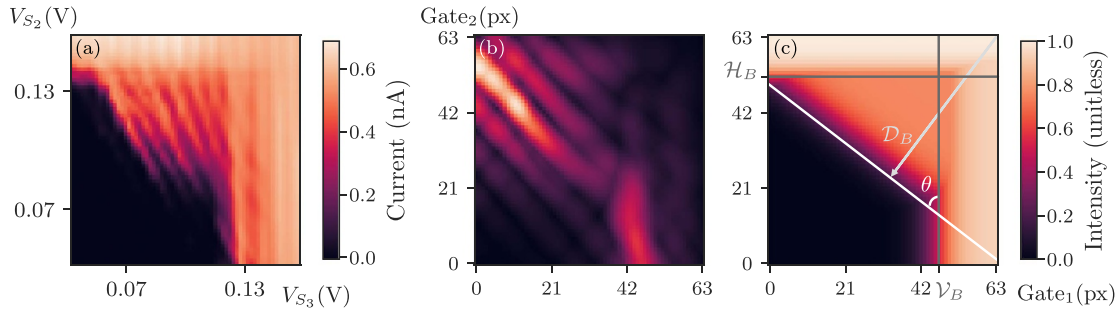

Figure 2. (a) A sample experimentally acquired scan, the same as in figure 1(b). (b) The Gabor filters for the scan shown in (a), weighted by their importance as indicated by the EBM model. (c) Synthetic fit to scan shown in (a). The $\mathcal{H}_B$, $\mathcal{V}_B$, and $\mathcal{D}_B$ are the horizontal, vertical, and diagonal optimal boundaries, respectively, and the θ indicates the orientation of the diagonal boundary.

are visually similar to experimental data, as depicted in figures 1(c) and (e), this approach does not produce physically realizable data.

For generating the synthetic triangles, we utilize the 2D sigmoid function,

$$s_{\mathrm{m,r,b}}(x) = \frac{\mathrm{m}}{1+\mathrm{e}^{\mathrm{r}(x-\mathrm{b})}}, \tag{3}$$

where m, r, and b control the magnitude, rate, and shift of the sigmoid, respectively. The motivation for this choice is the resemblance of the idealized walls and triangle region to the step function. In practice, the transition from no signal to wall and triangle should be smooth. Triangle plots are assumed to have a horizontal and/or vertical wall corresponding to individual gates, as well as a possible third diagonal region linking the walls. To construct the 2D triangle plots, we compose three sigmoids:

$$\Delta_{(x,y)} = \max\left(s_{\mathrm{h}}(x), s_{\mathrm{v}}(y), s_{\mathrm{d}}(x\sin\theta + y\cos\theta)\right) \tag{4}$$

where s_{h}, s_{v}, and s_{d} are the parameters for the horizontal, vertical, and diagonal sigmoid, respectively, with θ indicating the orientation of the diagonal sigmoid. In practice, we evaluate this function on an evenly spaced grid of (x, y) points.

It is important to restrict the parameters used to define the synthetic triangles to physically reasonable ranges. The simple constraints we chose include a requirement that the sigmoids cannot be infinitely dilated and that the walls must be in the image and have positive heights. The criteria used to determine the presence of a triangle region are derived from the definition of the triangle plot: we require that $s_{\mathrm{d}}(x\sin(\theta) + y\cos(\theta))$ is greater than s_{h} and s_{v}. We also impose a constraint that the orientation of the diagonal sigmoid is sufficiently diagonal, with the cutoff defined as $0 \leqslant \theta \leqslant \frac{\pi}{2}$.

2.5. Synthetic data modeling approach

The synthetic triangles are used to model the experimental triangle plots by deriving from experimental data parameters defining a most similar synthetic triangle. The resulting vector provides a powerful and compact representation for ML. Moreover, this vectorization approach has a key advantage in that the vector features correspond directly to the original image and that fewer assumptions about the frequency content of the data are made.

We start by defining an appropriate measure of image similarity. The experimentally acquired triangle plots contain considerable noise, texture, and other high-frequency features that are not essential in characterizing the overall structure. The synthetic triangles can serve as a representation of the boundaries of the vertical, horizontal, and diagonal walls defined by a shifted, scaled, and dilated sigmoid function. Since the synthetic and experimentally acquired plots are assumed to be structurally similar by design, both can be transformed into a region of the frequency domain that they are expected to share.

In practice, this involves applying an ideal low-pass filter, such as a Gaussian filter. To focus on the details of the structure, we use the magnitude gradient of the low-pass filtered image. While this can be thought of as a simple edge detection, we do not have *a priori* knowledge about the appropriate scale of the Gaussian for any particular image. The presence of the low-pass filter makes this approach notably more insensitive to certain classes of noise, which is desirable for our use case. To quantify the similarity between an experimentally acquired and synthetic triangle plot, we use the L^2-norm applied to both transformed images, assuming that the two images are visually similar if their distance in this transformed space is small.

Definition. Let U and V be two images subject to some transform F with parameters q. The similarity measure between U and V is defined as

$$\mathcal{S}_q^F(U,V) = ||F_q(U) - F_q(V)||_2. \tag{5}$$

The desired vector representation of an experimentally acquired image $\mathcal{I}_{\text{exp}}$ is obtained by optimizing the parameters defining the corresponding synthetic image $\mathcal{I}_{\text{sim}}$ such that $\mathcal{S}_q^F(\mathcal{I}_{\text{exp}},\mathcal{I}_{\text{sim}})$ is minimized subject to boundary constraints. We set up the optimization problem as:

$$q^*,\sigma^* = \underset{q,\sigma}{\text{argmin}} \left\{ \lambda \mathcal{S}_{q,\sigma}^{\nabla_\Gamma}(\mathcal{I}_{\text{exp}},\mathcal{I}_{\text{sim}}) + (1-\lambda)\mathcal{S}_q^I(\mathcal{I}_{\text{exp}},\mathcal{I}_{\text{sim}}) + \epsilon(q) \right\}, \tag{6}$$

where $\mathcal{S}_\sigma^{\nabla_\Gamma}$ is the similarity measure using the gradient of the Gaussian transform, $\nabla_\Gamma(U) = \nabla(\Gamma_\sigma(U))$; $\mathcal{S}^I$ is the similarity measure using the identity transform, $I(U) = U$; λ is a hyperparameter giving us control over the balance between trusting $\mathcal{S}_\sigma^{\nabla_\Gamma}$ and $\mathcal{S}_I$ similarity measures; and

$$\epsilon(q) = \begin{cases} C, & \text{if q violates boundary constraints,} \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

for some arbitrarily large penalty constant C. In practice, it is desirable to make this penalty larger than the other terms of the objective function so that the constraints are satisfied.

Since the initial guess for optimization is determined solely based on the raw image, we use the differential evolution global optimizer implemented in SciPy [33] to remedy this limitation. This makes our method less sensitive to initial values than if we had used a local optimizer but incurs a greater computational cost. A local optimizer might still be viable in cases where the phenomenological model is simpler, e.g. with fewer parameters or with guaranteed convexity. In an experimental context, we have access to additional data tied to each image to aid in establishing initial guesses for optimization.

The optimal parameters, as well as the value of the objective function using those parameters, define the final vector representation of image $\mathcal{I}_{\text{exp}}$:

$$\boldsymbol{\nu}(\mathcal{I}_{\text{exp}}) = \left[\sigma^*, \mathcal{H}_{(B,M,R)}(q^*), \mathcal{V}_{(B,M,R)}(q^*), \mathcal{D}_{(B,M,R,\theta)}(q^*), \mathcal{F}(q^*)\right] \tag{8}$$

where B, M, and R denote the boundary, magnitude, and rate for the optimal horizontal $\mathcal{H}$, vertical $\mathcal{V}$, and diagonal $\mathcal{D}$ component, respectively; θ indicates the orientation of the diagonal boundary; σ^* is the optimal scale, $q^* = [\mathcal{H},\mathcal{V},\mathcal{D}]$[see. Equation (6)]; and $\mathcal{F}$ is the value of the objective function for the optimized parameters (the fit fitness). For a visualization of these features, see figure 2(c).

3. Results

We present the results of experiments with the Gabor filterbank approach, the synthetic data modeling approach, and the hybrid approach combining the synthetic features with the single most informative Gabor feature. The transformation procedures described in sections 2.3 and 2.5 are applied to each experimentally observed triangle to produce the final vector of features $\boldsymbol{\nu}(\mathcal{I}_{\text{exp}})$. In each case, the vectorized features $\boldsymbol{\nu}(\mathcal{I}_{\text{exp}})$ are used to train and test an EBM model. Since the dataset is quite imbalanced in terms of good and bad data, the training is performed using five six-fold stratified cross-validations, with 10% of the data withheld for testing purposes. To determine the optimal k-fold configuration, we performed the stratified cross-validation for k ranging between 1 and 10, as we found that $k = 6$ achieved the most performant set of models.

To improve the accuracy and interpretability at a modest training time cost, we utilize the smoothing and greedy rounds parameters [34]. For a more thorough discussion of this, see [35]. We carry out five six-fold stratified cross-validations and report averaged results as well as type I and type II errors for each method. For completeness, we also include the averaged confusion matrices.

The results from all experiments, presented in table 1, show a relatively close performance for all three methods, with the Gabor fitting approach only slightly outperforming the more intuitive synthetic data modeling approach, at $92.5(1.2)\%$ vs. $90.5(1.0)\%$, respectively. Interestingly, the inclusion of a single, most important Gabor filter in the vector of features obtained using synthetic data modeling brings the EMBs performance back up to $91.7(8)\%$. Importantly, the model resulting from the hybrid vectorization retains the majority of the performance while the features in the hybridized vector remain well aligned with the physical intuition of the problem. This validates that the optimal fit captures the location and average intensity of the walls and diagonal components. Modifying the diagonal component function to capture the ridged pattern observed in the experimental data could further improve the performance.

Table 1. The results of five six-fold stratified cross-validation for the experimentally acquired data. The accuracy, type I error, and type II error are reported for each method. Corresponding confusion matrices are included for completeness. The value(uncertainty) notation is used to express uncertainties. All uncertainties herein reflect the uncorrelated combination of single-standard deviation statistical and systematic uncertainties.

Model	Accuracy [%]	Type I [%]	Type II [%]	Confusion matrix [counts]		
Gabor filterbank	92.5(1.2)	4.9 (0.9)	2.6 (0.7)	*real* ↓ / *pred* →	Good	Bad
				Good	406 (4)	14 (4)
				Bad	27 (5)	99 (5)
Synthetic triangles	90.5(1.0)	5.8 (0.5)	3.6 (0.6)	*real* ↓ / *pred* →	Good	Bad
				Good	400 (3)	20 (3)
				Bad	32 (3)	94 (3)
Hybrid approach	91.7(8)	5.0 (0.8)	3.3 (0.2)	*real* ↓ / *pred* →	Good	Bad
				Good	402 (1)	18 (1)
				Bad	27 (4)	4.3 (4)

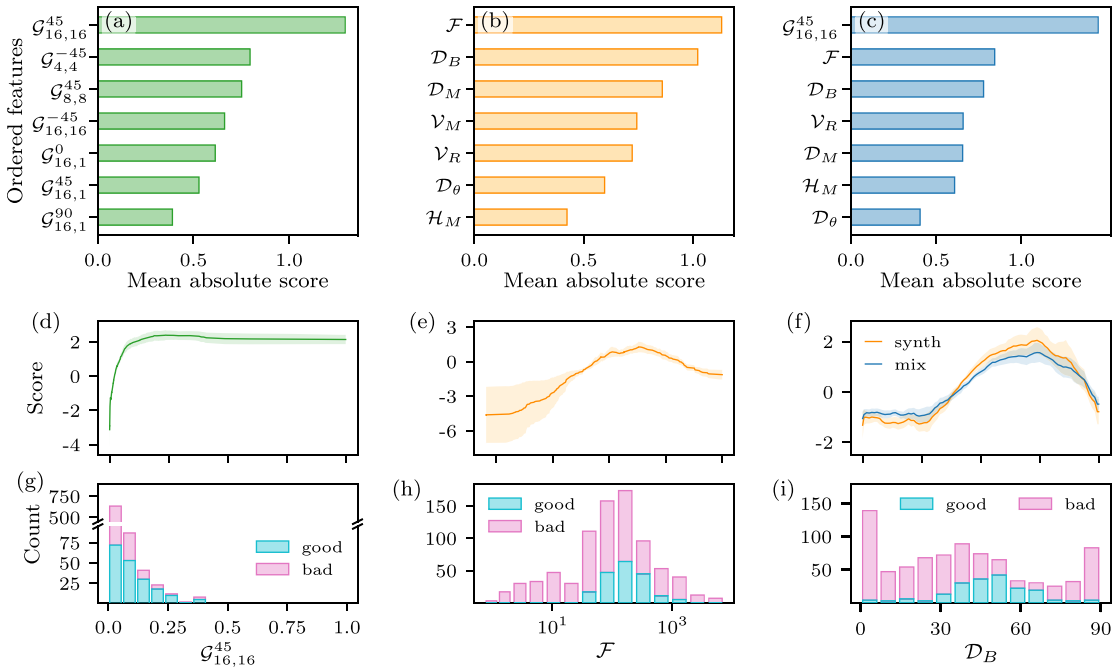

Figure 3. The feature importance plot for (a) the Gabor filterbank vectorization approach, (b) the synthetic data modeling vectorization approach, and (c) the hybrid vectorization approach. The mean absolute score refers to the extent to which a feature contributes to all predictions. Feature curve for (d) the oriented Gabor feature $\mathcal{G}^{45}_{16,16}$, (e) the fit fitness feature $\mathcal{F}$ of the synthetic triangle fit, and (f) the diagonal boundary component $\mathcal{D}_B$ of the synthetic triangle fit for the synthetic data modeling approach shown in orange and for the hybrid approach shown in blue. The x-axis indicates the value attained by the feature [normalized to $(0, 1)$ in (d) and (e)], and the score indicates the log odds towards good (positive) and bad (negative) classes for each respective feature. The histograms in (g)–(i) depict the relative class distribution as a function of the normalized feature value, with cyan representing good triangles and pink representing bad triangles.

3.1. Prediction interpretability

To understand the EBM model decisions, we rely on the EBM's feature importance, as well as the plots of individual features. The feature importance for all three cases we consider is shown in figures 3(a)–(c). The mean absolute score metric indicates the extent of the overall feature contribution to the model across all predictions for each feature in the model. We also consider the feature curves, depicted in figures 3(d)–(f), which show the log odds of being in a good vs. bad class, respectively, as a function of the feature values [normalized to $(0, 1)$ for plots (d) and (e)]. The accompanying histograms, shown in figures 3(g)–(i), provide the relative class distribution as a function of feature values.

Examining the overall feature importances for the Gabor method, shown in figure 3(a), we see that the four dominant roles are occupied by oriented features $\mathcal{G}^{\theta}_{\sigma_x,\sigma_y}$ with $\theta = 45$. The most important is a broad scale filter $\mathcal{G}^{45}_{16,16}$, one of the filters characterizing the triangle region of the plots.

Examining the feature curve for the dominant filter, shown in figure 3(d), we see that the model associates a strong 45° response with good triangles. The curve also reveals a threshold in the filter response

below which a triangle is expected to be bad. The histogram in figure 3(d), showing the distribution of the good and bad features, further supports the discriminative nature of this filter, with 90.4% of bad data falling in the first bin. It is notable that despite the higher performance of the Gabor-based technique, each $\mathcal{G}^{\theta}_{\sigma_x,\sigma_y}$ filter reveals only simple information about the extent of the presence of 45° information. Thus, despite being highly discriminative, the information obtained from this family of filters does not directly provide information on how to adjust the experimental setup.

The feature importance plot for the synthetic data modeling approach, shown in figure 3(b), reveals that this model also prioritizes features characterizing the triangle region, with $\mathcal{D}_B$ (the diagonal boundary component) and $\mathcal{D}_M$ (the diagonal magnitude component) being among the top three filters. The most important feature here is the fit fitness $\mathcal{F}$, quantifying how close the vectorized representation matches the experimental data. Looking at the feature curve for $\mathcal{F}$ depicted in figure 3(e), we see that for a modest cost associated with fitting, that is when $10^2 \lessapprox \mathcal{F} \lessapprox 10^3$, the log odds of belonging to the good class are positive. For bad triangles, the associated cost is either disproportionately low or high. On the low end of the spectrum, this is likely due to the problem of fitting a bad triangle with only two rectangular regions being a significantly simpler optimization problem. On the high end, the exceptionally high cost is generally associated with poor quality of the data.

The second most important feature in the synthetic data modeling approach is $\mathcal{D}_B$. The $\mathcal{D}_B$ feature curve, shown in orange in figure 4(f), indicates that the log odds that a triangle region belongs to a good plot increase with $\mathcal{D}_B$ once it extends past 28. The log odds of a triangle region belonging to a good plot are positive when $35 < \mathcal{D}_B < 85$. The $\mathcal{D}_B$ feature quantifies the extent to which the triangle region stretches in the screening gate-screening gate plot, as depicted in figure 2(c). When $\mathcal{D}_B < 35$, it is likely that the fitted triangle region overlaps with the horizontal and/or vertical sigmoid, and thus $\mathcal{D}_B$ would not be a reliable discriminant between good and bad class. While theoretically vector $\mathcal{D}_B$ can extend to the bottom left corner in figure 2(c), for large values of $\mathcal{D}_B$ the hypotenuse becomes relatively short which makes finding a reliable fit challenging. This dependency is confirmed by the feature plot, indicating that for $\mathcal{D}_B > 85$, the log odds of a triangle region belonging to a good class fall below 0. The importance of the $\mathcal{D}_B$ filter is further evidenced by its dominant role in the hybrid approach, where it occupies the third most important position, see figure 4(c).

It is important to emphasize that, due to the way EBMs are trained, the feature curves are not the same as the marginal distribution of feature values. Rather, each curve is affected by the complete vector of features $\boldsymbol{\nu}(\mathcal{I}_{\text{exp}})$ used to train the model. This means that individual features cannot be treated as entirely independent from one another. This is confirmed by the slight difference between the shape of the $\mathcal{D}_B$ curve for the synthetic data modeling approach vs. the hybrid approach, shown in figure 3(f) in orange and blue, respectively.

The hybrid model achieves comparable performance to the Gabor-based model while including only a single Gabor feature in addition to the synthetic features. It thus retains the improved intelligibility of the synthetic features without a significant reduction in accuracy.

EBMs also give us the ability to examine predictions for individual images. Figure 4(a), (b), and (c) show an example of a bad image, its synthetic fit, and Gabor filter representation, respectively. The synthetic triangle fit effectively captures the structural content of the experimental image, exhibiting no triangle region. The relative feature importance plot, shown in figure 4(d), also agrees with this observation, indicating $\mathcal{V}_B$ (the vertical boundary component) and $\mathcal{H}_B$ (the horizontal boundary component) as the most dominant features contributing to the negative (i.e. bad class) prediction.

Similarly, for good triangle regions, shown in figures 4(e) and (i), the synthetic triangle fit also correctly captures the structural content, as depicted in figure 4(f) and (j), respectively. In the relative feature importance plots, shown in figures 4(h) and (i), respectively, the top three dominant features include the Gabor filter $\mathcal{G}^{45}_{16,16}$, $\mathcal{D}_B$, and $\mathcal{F}$, which is consistent with the global prediction for the good images class, albeit the order of the latter two features is swapped compared to figure 3(c). As expected, all three features strongly contribute to the positive class prediction.

In all three examples, there is a direct observable relationship between the fit parameters and the visual features of the synthetic triangle. In practical applications, these visual features can lead to thresholding decisions and derived quantities for choosing operating points in the voltage space. For example, in the case of figure 4(i), a good operating point might be right inside the triangle region above the diagonal boundary, bisecting the area of the triangle at around $(V_{S_3}, V_{S_2}) = (0.09, 0.09\,\text{V})$ [at point (25,25) in figure 4(j)]. Using the $\mathcal{D}_B$ and $\mathcal{D}_\theta$ with the knowledge of the boundaries of the image in voltage space, along with the $\mathcal{V}_B$ and $\mathcal{H}_B$, we can create a normal vector describing the direction in which the triangle is expanding. This normal vector describes the relative strengths of the associated screening gates. Ideally, the normal vector of the triangle would point symmetrically towards the bottom left corner, meaning that both screening gates have equal action on the forming current channel. But, more often than not, this vector tends to vary. By re-scaling the normal vector and defining the magnitude to begin at $(\mathcal{V}_B,\mathcal{H}_B)$ with the end pointing to the

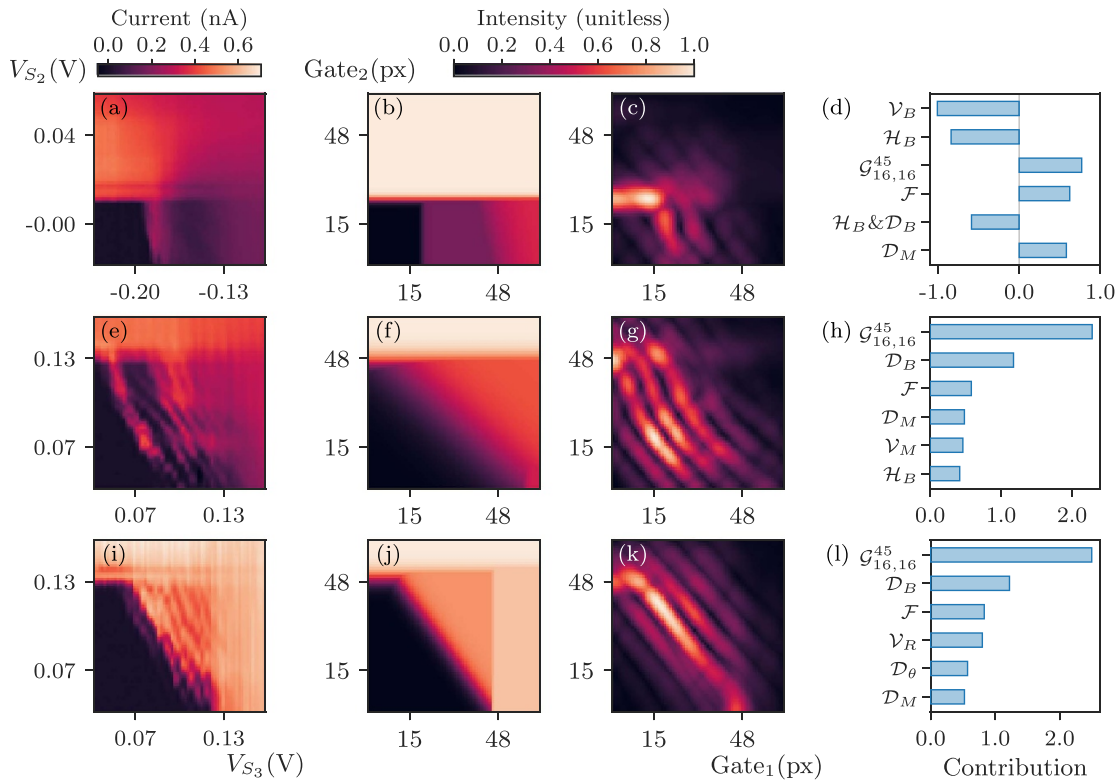

Figure 4. (a) A sample experimentally acquired bad scan. (e), (i) Two sample experimentally acquired good scans. (b), (f), (j) The fitted synthetic triangle plots for scans (a), (e), and (i), respectively. (c), (g), (k) Gabor-filter-based local interpretation of scans is shown in (a), (e), and (i), respectively. (d), (h), (l) The feature importance for the hybrid vectorization method for scans (a), (e), and (i), respectively. The intercept coefficient $\alpha = 4$ is removed from plots (d), (h), and (l) for clarity.

diagonal boundary of the triangle, the vector can now encode size information. If we then reduce the magnitude by 10%, we have an initial guess, inside the bounds of the triangle, for the operating point. We can also use the vector to calculate the area of the triangle in voltage space, which can be thresholded to ensure it is large enough to proceed.

If, on the other hand, the triangle plot looks like in figure 4(a), and no triangle is predicted, we can enhance the chance of detecting one by increasing the voltage of gates over the current channel and taking the triangle plot once more. Hence, there are many derivable quantities from this model that could be used iterably to function as feedback in a larger tune-up procedure beyond just determining triangle existence.

4. Conclusion and outlook

In this work, we demonstrate an alternative approach to vectorizing image data for use with EBMs that relies on generating synthetic data that best fits the experimentally acquired scans. We show that the new method is better adapted to the complexity of the triangle plots dataset. While the original vectorization method, invoking the Gabor filterbank features, produces a model that performs well with the classification task, its interpretability and usefulness are limited as the resulting model does not comprehend the underlying structure of the data. The alternative method, relying on fitting synthetic triangle plots to the experimental data, retains comparable accuracy, failing in about 2% more cases than the Gabor filterbank approach. However, unlike the Gabor approach, it produces features that are strongly aligned with heuristics-based intuition about the data, resulting in qualitatively superior explanations. As such, the features can be tied directly back to the scientific problem and used to adjust the experimental system to produce triangle plots of desirable quality. Future work might include integrating this analysis directly into an automated real-time QD tuning system.

It is reasonable to assume that the Gabor technique excels due to its ability to represent the ridged pattern in the triangle region which is considered an important characteristic of a good scan. We see that including a single Gabor filter–capturing the presence and prevalence of a diagonal activity–allows us to recover nearly all of the performance without compromising the model's interpretability. Alternatively, to avoid calculating the Gabor filterbank, this behavior could also be encapsulated within the synthetic approach by representing this pattern with a sinusoidal term in the synthetic triangle model.

The one disadvantage of the synthetic data modeling approach, when weighed against the Gabor filterbank, approach is the greater computational cost of the former. This is primarily due to the reliance on a global optimizer, whereas the Gabor-based approach only requires a small number of fixed filters. To remedy this, alternative similarity functions that would improve convergence and be less reliant on a global optimizer could be explored. However, care must be taken to ensure that the qualitative fitness is not reduced.

Finally, while all data used in this work comes from a quad-QD device with an overlapping gate architecture, we expect these results to be generalizable to an entire subclass of devices in Si/Si_xGe_{1-x} with different gate designs [21, 36, 37]. Likewise, other electronic materials that can host gate-defined QDs, such as bi-layer graphene [38] and silicon-MOS structures [39], can benefit from ML-enabled characterization and tuning as proposed in this work. The only requirement is that the tuning procedure must involve confining electrons to a 1D channel with a pair (or pairs) of gates so that other additional gates can then divide that channel into a chain of QDs.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.14549897. Figure source files are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.14589568.

Acknowledgments

J P Dodson present address: HRL Laboratories, LLC, 3011 Malibu Canyon Road, Malibu, CA 90265, USA. J Corrigan present address: Intel Corp., Hillsboro, OR 97124, USA. We acknowledge Patrick Walsh and Emily Joseph for experimental assistance. We acknowledge HRL Laboratories, LLC for support and L F Edge for providing one of the Si/Si_xGe_{1-x} heterostructures used in this work. J C acknowledges support from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1747503 and the Graduate School and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation. This research was sponsored in part by the Army Research Office (ARO) under Grant Nos. W911NF-24-2-0043, W911NF-23-1-0110 and W911NF-17-1-0274. We acknowledge the use of facilities supported by NSF through the UW-Madison MRSEC (DMR-2309000). The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Government or the ARO. The US Government is authorized to reproduce and distribute reprints for Government purposes, notwithstanding any copyright noted herein. Any mention of equipment, instruments, software, or materials; it does not imply recommendation or endorsement by the National Institute of Standards and Technology.

Appendix

A.1. Misclassified data: a post-hoc analysis

While the synthetic triangles provide a good representation of the crude features captured in experimental data for the majority of the data, there are cases where the vectorized representation is misclassified by the EBM. There are two types of misclassifications: classifying a triangle plot with a well-pronounced triangle region as bad [false-negative, see figures A1 panels (a)–(c)] and classifying a plot without well-defined triangle region as good [false-positive, see figure A1 panels (d)–(e)].

There are several reasons for such misclassifications. In figures A1(a) and (b), the plots are excessively accumulated, which would usually indicate a good sample if the user were to zoom out, and there is not sufficient diagonal activity for the sample to be valid. In figure A1(a) the scan window seems too small for the size of the triangle plot. In figure A1(b), on the other hand, we see the shadow of a second channel influencing the readout. This is a rare breakdown of the model's assumptions where the channels of the QD device were assumed to be independent. The ohmic potential is typically proportional to the maximum current in the image. With high accumulation, the ohmic potential can drift dynamically depending on multiple channels simultaneously since the measured current takes the path of the lowest resistance. This issue only plagues the upper channel of this device with charge sensors, see figure 1(a), since they are not separated by a gate-like isolation of the quad-QD channel on the bottom of the device. To resolve this with the same model, the experimental setup would need to know exactly when the ohmic potential started drifting and compensate for both halves of the charge sensor to keep exactly half the current flowing out through both channels equally. In figure A1(c), the slope indicating the formation of coulomb blockade is

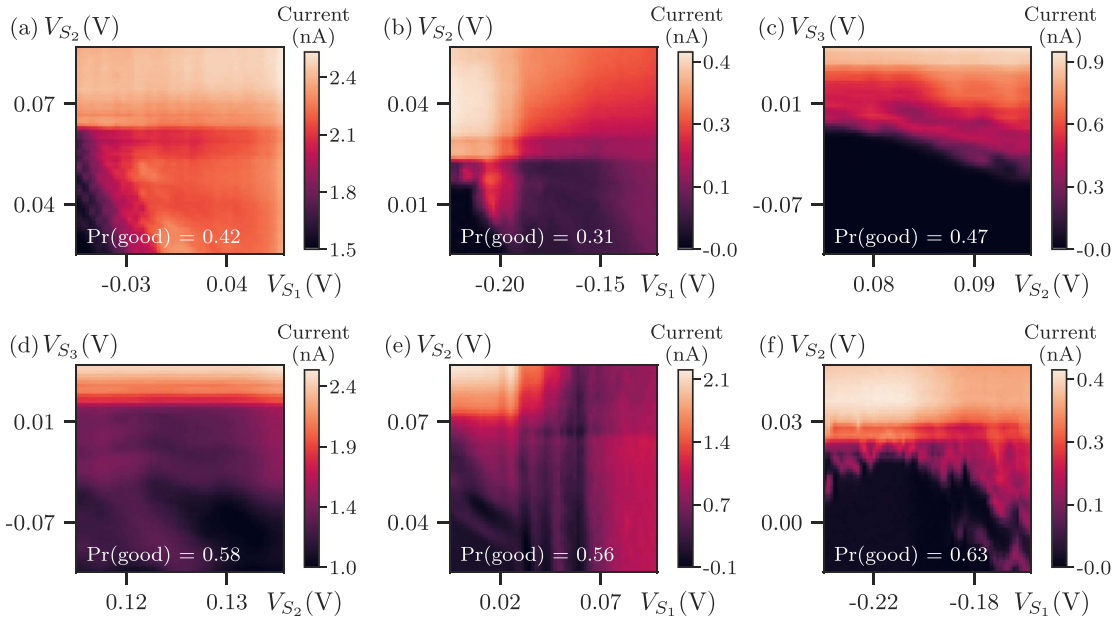

Figure A1. Sample ambiguous triangle plots. (a)–(c) Good-class triangle plots classified by the model as bad. (d)–(e) Bad-class triangle plots classified by the model as good.

relatively weak but technically visible, which, to an experimentalist, indicates that the triangle plot is good. However, in all three plots, one of the channels seems to be missing, which contributed to them being classified as bad even though the triangle region is present in all of them.

In figure A1(d), there is a substantial region with no current inside the triangle region, which seems to suggest that the channel is not accumulated correctly. In figure A1(e), the device is tuned to the wrong regime for the triangle plot where the blockade is controlled by one screening gate more than the other. This can be seen by the fringes being vertical rather than diagonal. Because plots in figures A1(d) and (e) have erroneous dropout regions, they should be considered bad. However, because the presence of the significant diagonal behavior has been vectorized through the synthetic triangle modeling, the EBM classified these images as good. In the plot depicted in figure A1(f), the strong diagonal and ridged pattern indicates a possibly shocked device. Yet, the diagonal rigid region is incorrectly vectorized as a proper triangle, which again leads to incorrect classification by EBM.

ORCID iDs

Daniel Schug https://orcid.org/0009-0001-3758-501X
Tyler J Kovach https://orcid.org/0009-0007-0807-7300
Jared Benson https://orcid.org/0009-0009-1673-5259
J P Dodson https://orcid.org/0000-0003-4265-5024
M A Eriksson https://orcid.org/0000-0002-3130-9735
Justyna P Zwolak https://orcid.org/0000-0002-2286-3208

References

[1] Schug D, Yerramreddy S, Caruana R, Greenberg C and Zwolak J P 2024 Explainable classification techniques for quantum dot device measurements *Proc. XAI4Sci: Explainable Machine Learning for Sciences Workshop (AAAI 2024)* (*Vancouver, Canada*) pp 1–6
[2] Luo Y, Tseng H-H, Cui S, Wei L, Ten Haken R K and El Naqa I 2019 Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling *BJR Open* **1** 20190021
[3] Zhang H, Yu Y, Jiao J, Xing E, Ghaoui L E and Jordan M 2019 Theoretically principled trade-off between robustness and accuracy *Proc. of the 36th Int. Conf. on Machine Learning, Volume 97 of Proc. Machine Learning Research*, ed Chaudhuri K and Salakhutdinov R (PMLR) pp 7472–82
[4] Baryannis G, Dani S and Antoniou G 2019 Predicting supply chain risks using machine learning: the trade-off between performance and interpretability *Future Gener. Comput. Syst.* **101** 993–1004
[5] Krizhevsky A, Sutskever I and Hinton G E 2012 Imagenet classification with deep convolutional neural networks *Advances in Neural Information Processing Systems* vol 25, eds Pereira F, Burges C J, Bottou L and Weinberger K Q (Curran Associates, Inc)
[6] He K, Zhang X, Ren S and Sun J 2015 Delving deep into rectifiers: surpassing human-level performance on imagenet classification *Proc. IEEE Int. Conf. on Computer Vision* pp 1026–34

[7] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (IEEE) pp 770–8
[8] Ribeiro M T, Singh S and Guestrin C 2016 why should I trust you?: explaining the predictions of any classifier *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 1135–44
[9] Lundberg S M and Lee S-I 2017 A unified approach to interpreting model predictions *Advances in Neural Information Processing Systems* vol 30
[10] Lou Y, Caruana R, Gehrke J and Hooker G 2013 Accurate intelligible models with pairwise interactions *Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 623–31
[11] Caruana R, Lou Y, Gehrke J, Koch P, Sturm M and Elhadad N 2015 Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission *Proc. 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 1721–30
[12] Letham B, Rudin C, McCormick T H and Madigan D 2015 Interpretable classifiers using rules and bayesian analysis: building a better stroke prediction model *Ann. Appl. Stat.* **9** 1350–71
[13] Ustun B and Rudin C 2015 Supersparse linear integer models for optimized medical scoring systems *Mach. Learn.* **102** 349–91
[14] Israel R, Kelly B and Moskowitz T 2020 Can machines 'learn' finance? *J. Invest. Manag.* **18** 23–36
[15] Dosovitskiy A *et al* 2021 An image is worth 16x16 words: transformers for image recognition at scale *Int. Conf. on Learning Representations*
[16] Zhang Q-S and Zhu S-C 2018 Visual interpretability for deep learning: a survey *Front. Inform. Tech. El.* **19** 27–39
[17] Kim S, Nam J and Ko B C 2022 ViT-NeT: interpretable vision transformers with neural tree decoder *Proc. 39th Int. Conf. on Machine Learning, Volume 162 of Proc. of Machine Learning Research*, ed Chaudhuri K, Jegelka S, Song L, Szepesvari C, Niu G and Sabato S (PMLR) pp 11162–72
[18] Pan B, Panda R, Jiang Y, Wang Z, Feris R and Oliva A 2021 Ia-red^2: interpretability-aware redundancy reduction for vision transformers *Advances in Neural Information Processing Systems* vol 34, eds M Ranzato, A Beygelzimer, Dauphin Y, Liang P S and Vaughan J W (Curran Associates, Inc) pp 24898–911
[19] Schug D, Kovach T J, Wolfe M A, Benson J, Park S, Dodson J P, Corrigan J, Eriksson M A and Zwolak J P 2023 Extending explainable boosting machines to scientific image data *Proc. Machine Learning and the Physical Sciences Workshop (NeurIPS 2023)* (*New Orleans, LA, USA*) pp 1–6
[20] Burkard G, Ladd T D, Pan A, Nichol J M and Petta J R 2023 Semiconductor spin qubits *Rev. Mod. Phys.* **95** 025003
[21] Zajac D M, Hazard T M, Mi X, Nielsen E and Petta J R 2016 Scalable gate architecture for a one-dimensional array of semiconductor spin qubits *Phys. Rev. Appl.* **6** 054013
[22] Dodson J P, Holman N, Thorgrimsson B, Neyens S F, MacQuarrie E F, McJunkin T, Foote R H, Edge L F, Coppersmith S N and Eriksson M A 2020 Fabrication process and failure analysis for robust quantum dots in silicon *Nanotechnology* **31** 505001
[23] Kalantre S S, Zwolak J P, Ragole S, Wu X, Zimmerman N M, Stewart M D and Taylor J M 2019 Machine learning techniques for state recognition and auto-tuning in quantum dots *npj Quantum Inf.* **5** 1–10
[24] Zwolak J P, McJunkin T, Kalantre S S, Dodson J P, MacQuarrie E R, Savage D E, Lagally M G, Coppersmith S N, Eriksson M A and Taylor J M 2020 Autotuning of double-dot devices in situ with machine learning *Phys. Rev. Appl.* **13** 034075
[25] Zwolak J P and Taylor J M 2023 Colloquium: advances in automation of quantum dot devices control *Rev. Mod. Phys.* **95** 011006
[26] Ye F, Ellaboudy A, Albrecht D, Vudatha R, Jacobson N T, and Nichol J M 2021 Characterization of individual charge fluctuators in Si/SiGe quantum dots (arXiv:2401.14541)
[27] Clark A 2015 Pillow (PIL Fork) Documentation
[28] Schug D, Kovach T, Benson J, Eriksson M and Zwolak J P 2024 QFlow Triangles: quantum dot triangle plots data for machine learning *Zenodo* (https://doi.org/10.5281/ZENODO.14549896)
[29] Daugman J G 1988 Complete discrete 2-D gabor transforms by neural networks for image analysis and compression *IEEE Trans. Acoust. Speech Signal Process.* **36** 1169–79
[30] Daugman J 2004 How Iris Recognition Works *IEEE Trans. Circuits Syst. Video Technol.* **14** 21–30
[31] Krüger V and Sommer G 2002 Gabor wavelet networks for efficient head pose estimation *Image Vis. Comput.* **20** 665–72
[32] Kwolek B Face Detection Using Convolutional Neural Networks and Gabor Filters 2005 *et al Artificial Neural Networks: Biological Inspirations– Icann 2005* (*Series Title: Lecture Notes in Computer Science*) vol 3696 (Springer) pp 551–6
[33] Jones E, Oliphant T and Peterson P *et al* 2001 SciPy: open source scientific tools for Python (accessed 7 February 2024)
[34] Nori H, Jenkins S, Koch P, and Caruana R 2019 Interpretml: a unified framework for machine learning interpretability (arXiv:1909.09223)
[35] Nori H, Caruana R, Bu Z, Hanwen Shen J, and Kulkarni J 2021 Accuracy, interpretability, and differential privacy via explainable boosting (arXiv:2106.09680)
[36] Neyens S *et al* 2024 Probing single electrons across 300-mm spin qubit wafers *Nature* **629** 80–85
[37] Philips S G J *et al* 2022 universal control of a six-qubit quantum processor in silicon *Nature* **609** 919–24
[38] Eich M *et al* 2018 Spin and valley states in gate-defined bilayer graphene quantum dots *Phys. Rev.* X **8** 031023
[39] Dumoulin Stuyck N *et al* 2024 Silicon spin qubit noise characterization using real-time feedback protocols and wavelet analysis *Appl. Phys. Lett.* **124** 114003