MKID Digital Readout Tuning with Deep Learning

R. Dodkins[a,*], S. Mahashabde[a], K. O'Brien[a,b], N. Thatte[a], N. Fruitwala[c], A.B. Walter[c], S.R. Meeker[c], P. Szypryt[c], and B.A. Mazin[c]

[a] *Department of Physics-Astrophysics, University of Oxford, Keble Road, Oxford OX1 3RH, UK*
[b] *Department of Physics, University of Durham, South Rd, Durham, DH1 3LE, UK*
[c] *Department of Physics, University California, Santa Barbara, California 93106, USA*

Abstract

Microwave Kinetic Inductance Detector (MKID) devices offer inherent spectral resolution, simultaneous read out of thousands of pixels, and photon-limited sensitivity at optical wavelengths. Before taking observations the readout power and frequency of each pixel must be individually tuned, and if the equilibrium state of the pixels change, then the readout must be retuned. This process has previously been performed through manual inspection, and typically takes one hour per 500 resonators (20 hours for a ten-kilo-pixel array).

We present an algorithm based on a deep convolution neural network (CNN) architecture to determine the optimal bias power for each resonator. The bias point classifications from this CNN model, and those from alternative automated methods, are compared to those from human decisions, and the accuracy of each method is assessed. On a test feed-line dataset, the CNN achieves an accuracy of 90% within 1 dB of the designated optimal value, which is equivalent accuracy to a randomly selected human operator, and superior to the highest scoring alternative automated method by 10%. On a full ten-kilopixel array, the CNN performs the characterization in a matter of minutes – paving the way for future mega-pixel MKID arrays.

Keywords: instrumentation: detectors; neural networks; supervised learning by classification

1. Introduction

Microwave Kinetic Inductance Detectors, or MKIDs, are a superconducting pair-breaking detector (Day et al., 2003) that operate across the electromagnetic spectrum in different variations such as: Catalano et al. (2016) (microwave) Yates et al. (2011) (far-IR), Mazin et al. (2012) (UV, optical, near-IR), and Ulbricht et al. (2015) (X-ray). UV, optical, near-IR (UVOIR) MKIDs are photon

*Corresponding author
Email address: rupert.dodkins@physics.ox.ac.uk (R. Dodkins)

counting detectors with spectral resolution $R = \lambda/\Delta\lambda \sim 10$ at 1 μm, enabling read-noise free (photon limited), low-resolution spectroimaging without filters or dispersive optics. Compared to other cryogenic detectors, these devices are simple to fabricate and operate, as thousands of these pixels are read out per feed-line using warm electronics. The first MKID camera (ARCONS), when commissioned in 2011, was the largest non-dispersive optical/near-infrared integral field spectrograph fielded by a factor of ten (Mazin et al. 2013; Szypryt et al. 2014; Strader et al. 2016). The current generation of UVOIR MKID devices are up to 20,000 pixels (Meeker et al., 2015), and larger arrays are planned (Marsden et al., 2013). The pixel format of MKIDs in the sub-millimeter and far-infrared are also several kilopixels (Adam et al., 2018) and systems capable of reading out 10^4 pixels have been demonstrated (Baselmans et al., 2017). These large formats make MKIDs a competitive detector technology for high contrast imaging of exoplanets (Meeker et al., 2015), long slit spectroscopy (O'Brien et al., 2014), as well as security and biomedical applications (Jonge et al., 2012).

In order to read out an MKID array, the digital readout system must be tuned for each pixel. Biasing a kilopixel array by manual inspection can take several hours, and is prone to inconstancies between users and human error. Existing approaches for automating this task perform poorly on resonators that are non-ideal, as we will demonstrate in Section 3. We present a machine learning-based package for tuning the bias points for an MKID readout prior to observations. This algorithm has been used to set up the MKID-based high contrast imager–DARKNESS (Meeker et al., 2015) for each of the four observing runs at the Palomar Observatory. This package should be beneficial to any system where biasing many resonators will be required such as: phonon-mediated detectors for neutrinoless double beta-decay or dark matter interactions (Martinez et al., 2017) and frequency multiplexed quantum processors (George et al., 2017).

This manuscript will be structured as follows: In Section 2 we introduce our devices, describe our measurement, and outline the traditional strategies for automated biasing of individual MKIDs. In Section 3 we discuss the various pathologies that can affect resonators and show how the standard ways of automated biasing fail when applied to these scenarios. In Section 4 we present our machine-learning model and in Section 5 we compare the results of the machine-learning model to two alternative automated algorithms and the manual inspection method. In Section 6 we end with the conclusions and future prospects.

2. Resonator Bias Tuning

An MKID device is made from a superconducting film, lithographically patterned into an array of microwave resonators. Modern UVOIR resonators use lumped element designs and thousands are coupled to a single transmission line (see Figure 1) but other geometries (see Zmuidzinas (2012) review) are equivalent for the purposes of this paper. The complex transmission $S_{21}(\omega) = I(\omega) + iQ(\omega)$ (where I and Q are the in-phase and quadrature components) is

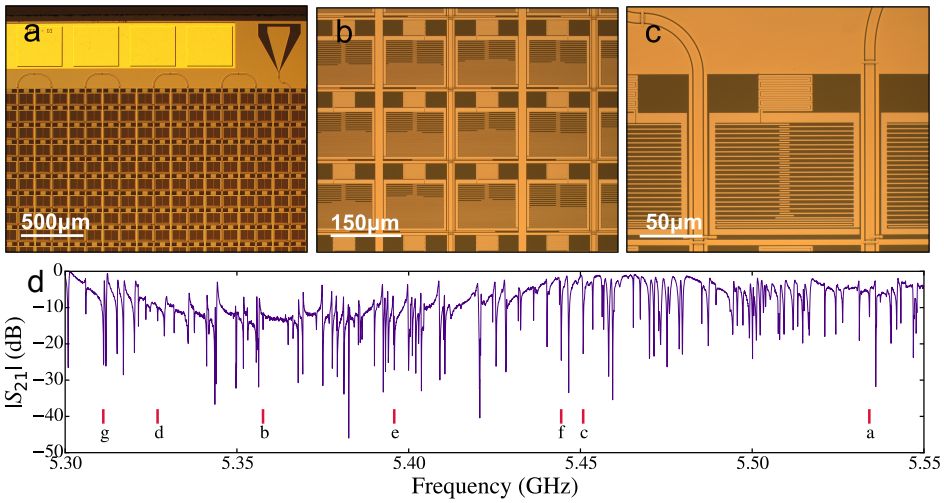

Figure 1: **a-c**: Microscope images, at several levels of magnification, of a single DARKNESS device. The device has several feed-lines each coupled to thousands of resonators. These resonators are lumped element, meaning they have a capacitive and an inductive region, where the inductive region is used as the photosensitive region. **d**: The transmission spectrum of a subsection of a typical feed-line with 2000 resonator pixels in a 4–8 GHz bandwidth sampled at a single readout power using a vector network analyser. The highlighted resonators correspond to those in Figure 3.

a measure of the forward voltage gain across an electrical component. When a probe tone is swept in frequency (ω) across the resonance, the scalar transmission $|S_{21}|$ is minimized at the resonant frequency and is mostly unattenuated off-resonance. Figure 1d shows a subsection of a broadband frequency sweep across a 2000 pixel feed-line, where each dip represents a single resonator with quality factor $Q_r \approx 10^5$.

Figure 2 shows the transmission through the device at frequencies ±450 kHz of the resonant frequency of a single MKID pixel at a range of readout powers. The lowest power sample is approximately -70 dBm. This type of measurement is hereafter termed a resonator powersweep. At low power the I and Q components trace a continuous resonance loop in the complex plane and a continuous dip in the transmission spectrum. At higher powers the resonator exhibits a nonlinear response as a function of driving current, and given sufficient power, the resonator bifurcates into two quasi-stable states, which manifest as a discontinuity close to the resonant frequency (Swenson et al., 2013; Thomas et al., 2015). In this bifurcation regime the resonator is rendered non-functional for photon detection (Strader, 2016).

To read out an MKID pixel, we measure the phase of the transmission at the equilibrium resonant frequency. To optimize the signal-to-noise ratio on this measurement we need to drive the resonators as powerfully as possible to reduce the contribution from the cryogenic amplifier (typically a High-electron-mobility transistor) noise and the effect from Two Level System (TLS) noise

(Gao, 2008). Therefore, to bias a given resonator, the power at which the resonator first bifurcates is identified, and the power 2 dB below this value is chosen. This method is the primary criterion of this manuscript and will hereafter be referred to as Rule #1.

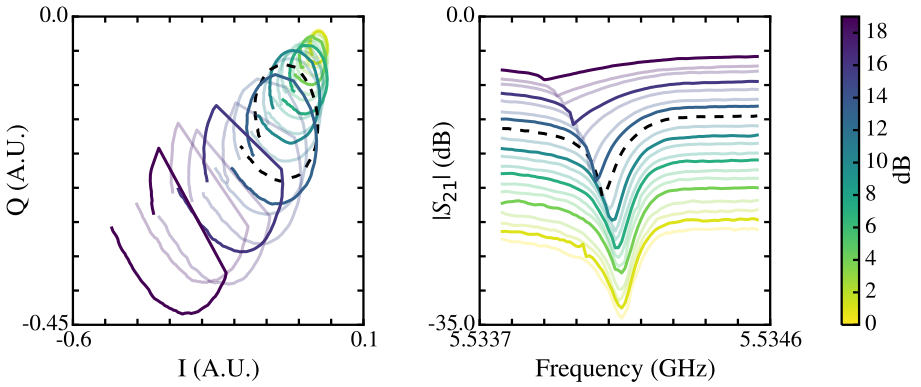

Figure 2: The transmission profile of a resonator, displaying ideal bifurcation behaviour at higher readout powers, plotted in the complex plane (left panel) and the magnitude spectrum (right panel). I and Q are measured in arbitrary units. The colourbar represents the amount of additional power applied to the resonator, and high saturation colours are separated by 3 dB. The lowest bias point is approximately -70 dBm. The black dashed line shows the optimal bias point chosen by an operator, using manual inspection, when applying Rule # 1.

It has been shown by de Visser et al. (2010) that overpowering resonators can be detrimental to device sensitivity because of the increase in generation-recombination noise from readout power heating, while Swenson et al. (2013) have shown that in certain circumstances there could be benefits from operating in the non-linear regime. Ultimately, optimal operating point of a given resonator will depend on several factors such as material, geometry, quality factor and the wavelength of incident light (since UVOIR MKID detectors are not currently limited by generation-recombination noise). This manuscript will assume that sensitivity is maximized by applying Rule #1 (along with the exceptions related to pathologies described later). An investigation into sensitivity optimization of UVOIR MKIDs is saved for later studies.

2.1. Analytical Method

From Khalil et al. (2012), the complex transmission of a resonator, driven in the low power regime, coupled to a transmission line with mismatched input and output impedances, can be described by an asymmetric Lorentzian

$$S_{21}(x_0) = g(x_0)e^{i\phi(x_0)}\left[1 - \frac{\frac{Q_\mathrm{r}}{Q_c}\left(1 + 2jQ_\mathrm{r}\frac{\delta\omega}{\omega_0}\right)}{1 + 2jQ_\mathrm{r}x_0}\right], \tag{1}$$

where the gain $g(x) = g_0 + g_1 x$ and phase $\phi(x) = \phi_0 + \phi_1 x$ factors have been included to account for scaling and orientation of the resonance loop due to, e.g., the length of the transmission line. Q_c is the coupling quality factor, Q_i is the internal quality factor and Q_r is the total quality factor with $Q_\mathrm{r}^{-1} = Q_\mathrm{i}^{-1} + Q_\mathrm{c}^{-1}$. $\delta\omega$ is a frequency offset applied to the resonance to account for any impedance mismatch between the resonator and feed-line (Geerlings, 2013). In the low power regime, the fractional detuning of the sampling frequency ω is

$$x_0 = \frac{\omega - \omega_{\mathrm{r},0}}{\omega_{\mathrm{r},0}}, \tag{2}$$

where $\omega_{\mathrm{r},0}$ is the resonance. If a powersweep measurement contains a pair of colliding resonators, then the fitting function becomes the summation of two asymmetric Lorentzians and the number of free parameters in Equation 1 is doubled.

In the high power regime the nonlinearity of superconducting resonators can be attributed to quasiparticle heating by readout photons (Thomas et al., 2015), or an intrinsic property of the kinetic inductance of superconductors at high current (Swenson et al., 2013; Semenov et al., 2016). The resonators in this manuscript will be analysed in the context of the latter model because there is no major degradation of Q_i at higher powers as expected with the quasiparticle-heating model. To account for the distortion of the S_{21} profile in the nonlinear regime, the fractional detuning becomes

$$x = x_0 + \frac{a}{1 + 4Q_\mathrm{r}^2 x^2}, \tag{3}$$

where a is the nonlinearity parameter

$$a = \frac{2Q_\mathrm{r}^3 P_\mathrm{r}}{Q_\mathrm{c}\omega_\mathrm{r} E_*}, \tag{4}$$

where P_r is the readout power and E_* is the scaling energy from the nonlinearity.

SCRAPS is a superconducting resonator analysis package (Carter et al., 2017). We used this tool to fit Equation 1 to the measured I and Q data of a resonator powersweep. Then, the resulting values for nonlinearity parameter a were used to identify the first bifurcation power, and with Rule #1, an estimate for the optimal bias point was inferred. This analytical method will hereby be referred to as 'AM'.

2.2. Numerical Method

If the primary concern is the degree of bifurcation then a very simple but effective metric is to monitor the separation between the I, Q magnitudes at adjacent frequencies, here termed IQ velocity. At a single readout power, it is defined as

$$v_{IQ}(f) = \sqrt{[Q(f) - Q(f-1)]^2 + [I(f) - I(f-1)]^2}, \tag{5}$$

where f is the frequency index and $(f-1)$ is the previous frequency index. When a discontinuity forms during bifurcation there is a large spike in the v_{IQ} spectrum at the discontinuity frequency, and at the adjacent frequencies the v_{IQ} should remain minimal. High Q_r resonators will show large values of v_{IQ} due to the larger relative frequency sampling compared to the resonator width. However, the v_{IQ} of the adjacent frequencies will also be large. To distinguish spikes in v_{IQ} caused by a high Q_r from that of a discontinuity, the maximum v_{IQ} is compared to the mean of the surrounding values in what is here termed v_{IQ} ratio or VR

$$\mathrm{VR} = \frac{v_{IQ}(M)}{\frac{1}{N}\left(\sum_{f=M-N/2}^{M+N/2}[v_{IQ}(f)] - v_{IQ}(M)\right)}, \tag{6}$$

where M is the frequency index of the maximum v_{IQ} and N is the total number of the adjacent frequency samples. A tradeoff exists whereby a larger N provides more data points to overcome the noise in v_{IQ}, at the cost of increasing the amount of included baseline values around a high Q_r resonator peak and artificially increasing the VR, as well as increasing the likelihood of sampling any glitch features or multiple discontinuities (described in Section 3). Before calculating VR, it is also necessary to smooth v_{IQ} with a low-pass filter because of the measurement noise. After smoothing v_{IQ}, and applying an $N = 6$ in Equation 6, the majority of the false classifications at the lowest bias point are typically removed. This technique, where VR is used to identify the first bifurcation power and apply Rule #1, will hereafter be called the numerical method or NM.

2.3. Manual Inspection

Manual visual inspection 'MI' consists of observing profiles of the resonance loops, the v_{IQ} spectra, and VR, at a range of powers for each resonator, by eye. The value obtained by the NM method is used as a starting estimate and the operator then identifies the first bifurcation power and typically applies Rule #1. Very often, resonators will be afflicted by various pathologies. In these cases, an operator will have to apply Rules #2–4, which are introduced in Section 3.

2.4. Summary

Three methods have been described for classifying the optimal bias point of a resonator powersweep: AM, NM and MI. These methods are summarized in Table 1. The metric thresholds that a resonator sampled at a given power must cross to be considered bifurcated are: $a = 4\sqrt{3}/9 \approx 0.77$ and VR = 3.5, for the AM and NM methods respectively. The a threshold is taken from the bifurcation value in Swenson et al. (2013), and the VR threshold is chosen from experience when applying the MI method, but can be adjusted depending on the amount of I and Q noise.

For both AM and NM methods, the first instance where the metric is above the threshold is used (given that the metric sometimes transitions across the

Table 1: Three methods for evaluating the bias point of a resonator powersweep, the associated metric parameters and the approximate time to implement the technique on a ten-kilopixel MKID array.

Method	Metric	Threshold	Time (min)
Analytical Method (AM)	a	0.77	600
Numerical Method (NM)	VR	3.5	30
Manual Inspection (MI)	I, Q, v_{IQ},VR	–	1200

threshold multiple times), and if this places bias point on or beyond the boundary ($\leq$0 dB), then 1 dB is chosen. If no metric is above the threshold then the location of highest metric is the selected bifurcation power.

Prior to applying the AM and NM methods, the I and Q amplitudes for each of the resonators at each power are filtered to suppress the noise. It was found that the optimal method involved using a convolution-based, running average method across all points excluding the maximum v_{IQ}. This technique meant the amplitude of bifurcation discontinuities remained unaffected. When performing the NM, both the convolution-based smoothing to I and Q, and the low pass filter to v_{IQ}, was applied.

3. Resonator Pathologies

There are several pathologies that can afflict a resonator, and some resonators exhibit multiple pathologies. Figure 3 exemplifies these pathologies. The AM, NM and MI methods have been applied to each example resonator, and the respective estimates of the bias point are shown as vertical lines in the right-most panels.

Each resonator in Figure 3 is misidentified for a specific reason (excluding (a)). One approach to this problem would be to tune a phenomenological model for all eventualities. While time consuming, it is also difficult to create exceptions that don't adversely effect the performance of other pathologies. For this reason, until now, each resonator has been inspected manually, using the NM technique as a first estimate.

In Figure 3a, there is only a 1 dB disagreement between both AM and NM methods and the optimum value using MI. Both AM and NM metrics show a mostly monotonic increase with power past their respective thresholds, and the selection of the bias point is trivial.

3.1. No Bifurcation

Figure 3b and 3c display resonators that do not appear to bifurcate. The resonator in Figure 3b has abnormally good power handling ability, which can happen when a resonator has low quality factor Q_{r}, according to equation 4. When an operator is classifying this resonator through MI, they may choose the highest power, based on the degree of bifurcation and a rough extrapolation of the trend with power. However, it is also advantageous not to bias a single

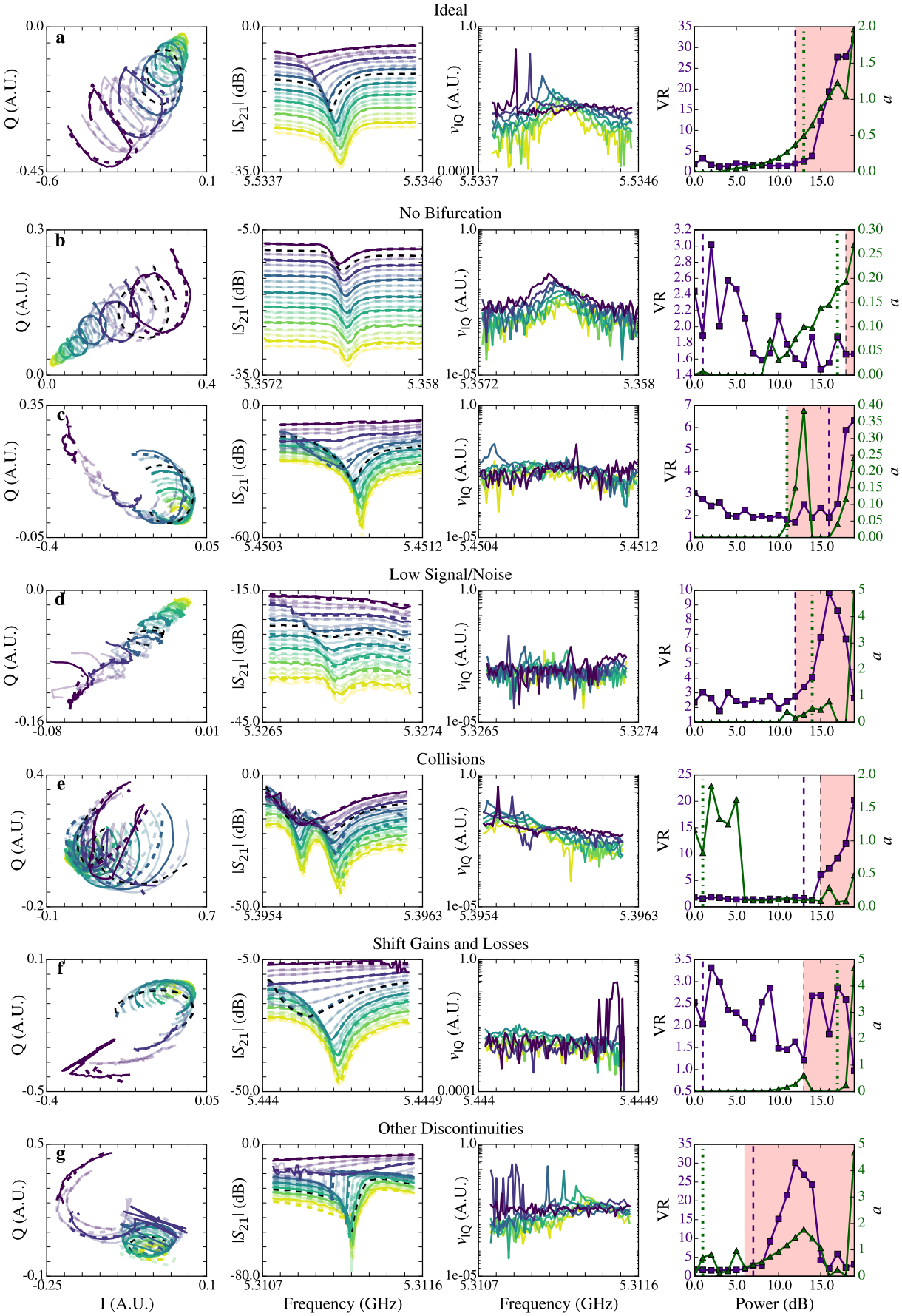

Figure 3: **a-g**: Seven measured resonator powersweeps exemplifying the pathologies described in Section 3. The colour scheme is the same as Figure 2. The first two columns display each resonator's raw transmission (solid lines) superimposed with the AM fit (dashed lines). The black dashed line is the bias point chosen from MI. The v_{IQ} parameter shown in the third column is used in the application of the MI and NM techniques. The observation of a single point above the background signifies a discontinuity and therefore bifurcation. In the right-most column, on either vertical axis of each plot is the metric of bifurcation for the analytical (triangular markers) and numerical (square markers) methods. The dashed and dot-dash vertical lines display the location of the evaluated bias point from the VR and a metrics respectively. The intersection of the coloured and blank zones (also marked with a dashed vertical line) is the bias point chosen by manual inspection (later $\mathrm{MI_A}$). Full agreement between the three methods is only found on the ideal powersweep.

resonator with too much power, otherwise it will impact the dynamic range of the digital readout for lower powered resonators (Strader, 2016). This second criterion is applied at the discretion of the operator and is referred to here as Rule #2.

The AM model has detected a trend towards bifurcation and matches to the bias point from MI well. The I and Q magnitudes, however, do not show any discontinuity from bifurcation, so the NM technique has no bifurcation indicator. In this instance, the bias point is actually vastly underestimated, because the NM metric, VR, has an outlier at low power. This effect can very often manifest at low powers where the signal to noise ratio is low and lead to a misclassification of otherwise ideal resonators. This issue can be mitigated by using a wider window during the initial smoothing step for each resonator, using a larger N when calculating VR, or using a larger VR threshold, however each of these processes result in misclassifications for other types of resonators.

The resonator in Figure 3c shows a degradation in Q_r with increasing readout power until the resonator has lost all appreciable magnitude without any discontinuity forming at any power. This behaviour can sometimes be the result of a second collided resonator sufficiently close, such that it is not discernible from the higher frequency resonator. In order to maintain resonator spectral resolution, one of the lower powers should be chosen during the classification, where the resonance loop shows sufficient curvature and the transmission profile shows sufficient depth, at the discretion of the MI operator. This criterion, hereby known as Rule #3, also helps with the finite readout dynamic range.

3.2. Low Signal/Noise

The resonator in Figure 3d has a low signal-to-noise ratio at all powers, in part because it has a low internal quality factor Q_i compared to its coupling quality factor Q_c, and also because it is in the 5.3 – 5.45 GHz region of the feedline with reduced transmission (see Figure 1d), leading to a shallow resonator transmission dip. This makes the bifurcation power harder to identify, and in severe cases, no bifurcation is visible. With the example resonator shown in Figure 3d, the AM and NM metrics show significant noise and nearly pass their respective thresholds at low power.

For certain measurements, such as speckle-noise suppression of high contrast imaging observations, it is often preferable to retain resonators with these characteristics in order to maximize total pixel count and retain the temporal information. In observations where high spectral resolution is of primary concern (for example measuring galactic redshifts (Marsden et al., 2013)), these pixels can be discarded in post processing. Another reason for retaining these observations is that operators will want to characterize a test array.

3.3. Collisions

Resonators can shift away from the designed frequency by differing amounts (Semenov et al., 2016) because of the non-uniformities in fabrication often causing collisions. Research into different superconductor materials has shown some

progress in reducing this effect (Szypryt et al., 2016). However, the issue persists due to cost and technology limits on readout bandwidth, pushing resonators closer together in frequency space to make larger format arrays.

Figure 3e shows an example of a collision. If a 1 μm photon were incident on the higher frequency resonator pixel, that resonator will shift approximately 50 kHz towards the lower frequency resonator. If resonators are too close originally, then probe tones at both resonant frequencies could measure a phase shift and a false detection may occur in the lower frequency resonator pixel. Typically, the higher frequency resonator of a collided pair within 200 kHz of each other is used in the readout tone generation list. However, the lower frequency resonator can be used if the higher frequency resonator shows exceptionally low Q_i. This criterion, Rule #4 is applied at the discretion of the operator.

Since the neighbor modifies the profile of the chosen resonator, fitting that resonator alone may result in unreliable predictions for a. Fitting both resonators simultaneously is a larger problem requiring more parameters, meaning that the least squares algorithm is more likely to converge on a local minimum. The problem can be alleviated using maximum likelihood estimation from Markov Chain Monte Carlo sampling, at the expense of vastly longer computation times on these powersweeps.

The NM underpredicted the bias point because it does not have a means of associating the discontinuity with the lower frequency resonator, which bifurcates at a higher power. Neither the NM, nor the AM, account for both of the bifurcation metrics simultaneously, the $|S_{21}|$ differences, the quality factors, or the separation between adjacent resonators.

3.4. Sampling Window Gains and Losses

Figure 3f shows some ways in which the choice of sampling window, coupled with resonators that show a large frequency response with power, can make resonator biasing more challenging. The resonance is centred on the sampling window at low powers and undergoes a relatively large translation out of the sampling window at higher powers. This is an example where operator would have to apply Rule #3. Both the AM and NM metrics do not reach their respective thresholds before the resonance translates out of the sampling window and the location of the highest metric is taken. The analytical algorithm has to handle the onset of bifurcation features and then the sudden disappearance of these features. For this reason, the AM technique fits each resonator power sample independently, and no fit is performed to a as a function of power, otherwise the anomalies at high power would skew the bias point estimate. However, this approach can lead to outliers in a at low power providing a false classification as seen in Figure 3e and 3f.

At higher powers the resonator in Figure 3e is afflicted by a second pathology. After the original resonator has shifted out of the sampling window, a higher frequency resonator, which is already bifurcated, shifts into the sampling window. AM is triggered by this new resonator, and a bias that is too high is chosen. In other instances when the initial resonator remains in the sampling window, if the second resonator comes within the collision threshold separation

of the first resonator, then the classifier (human or an automated algorithm) could choose a lower power or apply Rule #4. If the two resonators are sufficiently separated then both resonators should be fit simultaneously and apply Rule #4.

In each of these instances the choice of sampling bandwidth and centre is important. A trade off exists whereby a wider sampling window is more likely to capture the entire resonator profile as it evolves, but this increases the risk of contamination from adjacent resonators. In practice, a bandwidth of between 0.5 and 1.5 MHz is used depending on the average Q_r of the resonators. Similarly, centering the sampling window on the initial resonance is useful for observing when higher frequency resonators pass the collision separation threshold. The quality-factor and the amount a resonator will translate at high powers (if at all, see Figure 3b) are not known a-priori, so a sliding sampling window has its own difficulties.

3.5. Other Discontinuities

Sometimes resonators can show behaviour that is wildly different from those on the same feed-line. The resonator in Figure 3g displays two seemingly independent characteristics. The first is a discontinuity forming at frequencies above the resonance, here termed backwards resonators. This type of behaviour appears to correlate with the loss of transmission between 5.3 and 5.45 GHz. This could be because of an impedance mismatch creating a severe asymmetry in the transmission profile (parameterized with $\delta\omega$ in Equation 1). It should be noted that, in the IQ plane, this asymmetry does not affect the bifurcation signature.

The second example of an alternative discontinuity is where a resonator shows multiple discontinuities at higher power. This could be explained as the resonator switching between the available states after bifurcation, or sometimes glitches from digital readout induced errors.

3.6. Summary

Table 2 summarizes the statistics of the chosen types of resonator power-handling behaviour mentioned in this section. In this instance, the main fabrication artifact driving the variation between resonators, and causing the low amount of ideal and adequate resonators, was the film non-uniformity of stoichiometric titanium nitride (TiN) (Szypryt et al., 2016). The statistics of certain types will vary between different arrays and even feed-lines of the same device. Backwards bifurcation may be more specific to this feed-line but collisions are a common occurrence in all feed-lines. There will also be other types of resonator non-ideal behaviour that are not apparent in this example feed-line.

This is why it is important to develop an algorithm that can progressively learn to account for different types of behaviour as it is exposed to them, instead of continually attempting to manually optimize a phenomenological algorithm. This type of problem, where input data has a large variety of cases, but share some embedded commonalities, is well suited to machine learning. With enough

Table 2: Occurrence statistics of the different powersweep pathologies for feed-line two of device Ukko, "Ukko2", identified manually. There are 372 resonators in total and the properties are not mutually exclusive. 'Ideal' resonators show no collisions, sampling window effects, multiple or backwards discontinuities. 'Adequate' resonators may include well separated collisions, backwards discontinuities, shift gain or multiple discontinuities after the bifurcation power. 'Shift Loss' is the where the resonance translates out of the sampling window to lower frequencies. 'Shift Gain' is when an adjacent resonance shifts into the sampling window. 'Backwards' refers to resonators with the discontinuity at frequencies higher than the curve minimum. 'Low Q_r' are those which are difficult to classify because of shallow profile of the resonator.

Type	Absolute	Percentage (%)
Adequate	185	49.9
Shift Loss	133	35.8
Multi. Disc.	108	29.1
Noise	102	27.5
Backwards	87	23.5
Shift Gain	83	22.0
Collision	71	19.1
Ideal	44	11.9
Underpowered	34	9.2
No Bifurcation	16	4.3
Low Q_r	10	2.7
Unusable	2	0.5

training data a deep neural network architecture should be able to develop a sufficiently sophisticated model to bias resonators to a level of human accuracy.

4. Resonator Biasing with Machine Learning

Neural Networks are a class of machine learning algorithms that with emergent behaviour are able to discover and model high level abstractions in multi-dimensional data (Bengio et al., 2007). As these models are exposed to increasing amounts of input data, their accuracy can progressively improve. Deep Convolution Neural Networks (CNNs) have had great success in many areas of computer vision (LeCun et al., 2015), most notably image recognition (Krizhevsky et al., 2012). These networks take a multi-dimensional vector as their input, process the input vector with a series of filters to extract the relevant features, and produce an estimate of the label that identifies the input vector (LeCun et al., 1995). In the case of image recognition tasks, the input vector is the red, green and blue channels of an image, and the label corresponds to the class of object in the image. The filters (or weight vectors) must be "learned" by training the model on many input vectors of known labels, known as a training dataset.

A CNN was utilized as a model for predicting the bias point of resonators, with an architecture similar to how a conventional image-recognition network would be structured. The power and frequency axes can be thought of as the

spatial dimensions of an input image, and the I, Q, v_{IQ} channels play the same role as the red, green and blue channels. The label corresponds to the optimal bias point for the powersweep. For the training and evaluation datasets, previous manual inspection (MI) data were used.

4.1. Architecture

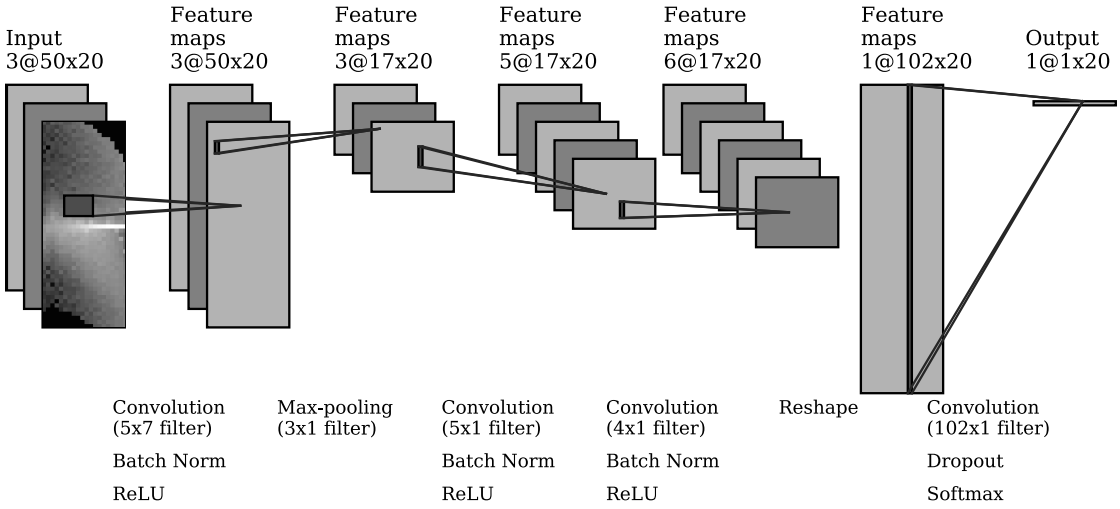

Figure 4: An illustration of the convolution neural network architecture used for classifying powersweeps based on the optimal bias point. The input vector is I, Q and v_{IQ} with 50 frequency samples and 20 power samples. The v_{IQ} vector is shown with example values logarithmically scaled. The output vector contains a probability for each of those power samples. The dark boxes highlighting a small area in each layer display the receptive field of each weight vector. The initial convolution layers are superseded by a batch normalisation layer and a rectified linear activation layer; the final convolution layer instead is superseded by a dropout layer and a softmax activation layer. The penultimate layer is not to scale. This figure was generated by adapting the code from draw_convnet[1]

The chosen design of the CNN was primarily motivated by the risk of overfitting the model to the limited amount of available training data, which would cause the model to not generalize well to unseen data. The architecture of the CNN is shown in Figure 4.

First, each of the layers of the input vector is convolved with a filter that has a spatial extent known as the receptive field and a depth that controls the number of output layers. These filters extract a common feature seen throughout the spatial axis. During convolution, the input vector is zero padded along the boarder so the output vector retains the same size as the input vector. The first convolution step is the only place in the network where the filters have a

[1] https://github.com/gwding/draw_convnet

receptive field that extends across multiple powers. By not using a filter with a receptive field that extends across the power axis (a fully-connected layer in one dimension), the learned filters are spatially invariant, and the CNN is made insensitive to non-uniform bias distributions present in the training data.

The product of the convolution layer then undergoes a technique called batch normalisation that increases the rate of convergence during training and acts as a regularization technique to reduce overfitting (Ioffe and Szegedy, 2015). After the normalisation layer, the vector is then processed by a nonlinear function that acts as a decision boundary depending on the input value, in what is called an activation layer. Rectified Linear Unit (ReLu) activation was chosen for its superior training times on CNNs (Krizhevsky et al., 2012).

The maximum pooling technique down-samples the spatial axes, prioritizing higher values, to allow larger scales to be probed by an equivalently sized filter during the next convolution. Different architectures were explored containing pooling layers at different stages. Ultimately, a single pooling layer was applied early in the network that only reduces the frequency axis. The subsequent convolution layers then detect the higher level, more abstract features, and add more parameters that increase the model sophistication. Each of these convolution layers is superseded by batch normalisation and ReLU layers.

The feature map vector with a depth of six is flattened along the power axis. The data along the frequency axis is then combined to produce a scalar value for each bias. This is achieved by applying a filter with receptive field that extends across the extent of frequency dimension. During training, in order to lessen overfitting, the regularisation technique known as dropout is applied with a 50% probability of removing a given filter unit during each pass (Srivastava et al., 2014). The final layer is the softmax function that converts the scalar values for each bias into a probability. The class with the largest probability is the selected bias point.

4.2. Input data

The lists of bias points for feed-lines found through MI, and originally created for biasing arrays for observations, make up the training data. Table 3 summarizes all the input feed-line datasets used. In total, 4549 powersweeps from ten feed-line datasets were available for training (and evaluating) the CNN. These feed-line datasets came from several arrays, with data from two feed-lines, at most, for each array. Some feed-line datasets are repeat readings of powersweeps on different cool-downs. These devices are made up of different materials, three were TiN on silicon and one was platinum silicide (PtSi) on sapphire.

Each input vector was preprocessed to help the CNN converge on the relevant properties. The v_{IQ} magnitudes of each powersweep were normalized to their maxima across all powers. The I and Q magnitudes were both scaled to the maximum of $|S_{21}|$ across all powers. The power axis of all feed-line datasets were trimmed to 20 by removing the lowest power samples – and accordingly any powersweep with bias point in that range. The spectral axis was trimmed to a window (50 samples) centred on the resonance at each power. It was found that this step increased the accuracy of the CNN by several per cent despite

Table 3: A summary of all the available feed-line datasets used in training and evaluating the full CNN. Each dataset is comprised of measured powersweeps and corresponding bias point estimates from MI that function as the labels. 'F.L.' is the assigned feed-line number and 'Repeat' is index of the powersweep measurements. Feed-line 2 of Ukko has been classified by several operators to evaluate the different biasing methods in Section 5.2.

Device	F.L.	Repeat	Material	Resonators	MI labels
Morpheus	2	i	TiN	198	1
Morpheus	5	i	TiN	192	1
Varuna	2	i	TiN	674	1
Faceless	3	i	PtSi	339	1
Faceless	3	ii	PtSi	390	1
Faceless	3	iii	PtSi	390	1
Faceless	2	i	PtSi	632	1
Faceless	2	ii	PtSi	632	1
Ukko	1	i	TiN	730	1
Ukko	2	i	TiN	372	4

the loss of information about the magnitude of the resonance translation with readout power. The location of the resonant frequency was taken as the location of maximum v_{IQ} (at any power sample, fewer than 1% of resonators have a maximum v_{IQ} that is not within 50kHz of the resonance, as chosen by MI, due to excessive noise).

The label vector for each powersweep is a probability density function (PDF) with a lognormal distribution and a maximum at the MI bias point. The lognormal profile accounts for the fact that it is more detrimental to overpower a resonator and lower powers are sometimes preferable. These labels only classify a powersweep based on readout power. A CNN could be conceived that intelligently tunes both the bias point and frequency by means of a multilabel classifier architecture. This concept is saved for later work.

Data augmentation is the process of creating additional training data by performing label preserving transformations on the original training data, or synthesizing new data based on a model. Data augmentation was explored to both increase the total amount of training data and correct for the non-uniform distribution of training data for each class. These label preserving transformations included resonator phase transformations (changing the ratio of Q and I) and shifting the input data in the spectral domain by several units. No appreciable increase in accuracy was observed, and the extra training data increased the training time of the CNN. The inability of augmented data to increase the CNN accuracy is indicative of either, inconsistencies in the original training data, or the fact the CNN has learned to be insensitive to these transformations. Synthesizing powersweeps based on the AM model, and assuming some relation for each of the parameters with power, may have increased the accuracy of the CNN.

4.3. Training and Evaluation

The CNN was implemented using the TENSORFLOW machine learning library (Abadi et al., 2016) on a computer with 16 Intel Xeon cores running at 2.2 GHz. Initially, the weights of the CNN were allocated normally distributed random values. For each training step, batches of up to 50 input powersweeps were fed into the CNN, and the difference between the true and predicted classifications (known as loss) was measured. The weights are then adjusted in the direction that minimizes the loss, which is found using the ADAM gradient descent method (Kingma and Ba, 2014). As this process is repeated and CNN begins to converge on the optimal set of values, the magnitude of the correction applied to the weights was exponentially decreased to prevent overshoot.

For each investigation, separate *evaluation*, *training* and *testing* datasets were created. The evaluation dataset contains all the measured powersweeps from a single feed-line, selected for the purpose of evaluating the CNN against the other bias selection methods. The remaining nine feed-line datasets are pooled together and split (typically 95:5) to produce the training and testing data respectively. The test data helps evaluate the performance CNN on unseen data from the same feed-lines.

Ensemble techniques are where multiple models are trained separately, usually on the same training data, and combined to achieve a greater accuracy. For example, bagging involves training different models in parallel, and since the trained models are non-deterministic, the combination of the predictions should act to cancel some of the biases each model has from overfitting. This technique was used on the full CNN in Section 5.1.2.

5. Results and Analysis

A misclassication by 1 dB is often permissible because of the subjectivity that arises for some powersweeps. For example, there is some subjectivity inherent for some pathologies such as 'No Discontinuity' or 'Shift Loss' powersweeps. Another type of subjectivity is due to uncertainty in the bifurcation power because of the limited power sampling. In the event of the bifurcation power existing between two samples, a more aggressive operator may choose a higher bias point.

Therefore, a Boolean classification accuracy within 1 dB is used to assess the performance of CNN (or any of the described evaluation methods). First, the Boolean accuracy is measured for each powersweep p according to

$$a_p(m_\mathrm{e}|m_\mathrm{t}) = \begin{cases} 100, & \text{if } |b(m_\mathrm{t}) - b(m_\mathrm{e})| \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

where b is the bias class, m_e is the evaluation method (here CNN), method m_t is assumed to produce the true values. Then, the mean a_p for all powersweeps

in a feed-line dataset, yields the accuracy parameter of m_e given m_t

$$A(m_\mathrm{e}|m_\mathrm{t}) = \frac{1}{P}\sum_{p}^{p=P} a_p. \tag{7}$$

5.1. CNN Training

5.1.1. Input Data Investigation

To investigate the consistency of the feed-line datasets, the CNN was trained and evaluated on different combinations of them. The evaluation dataset was sequentially set to each feed-line and a CNN model was trained and tested on a combination of the powersweeps from the remaining feed-lines. Each time 3565 and 180 powersweeps were randomly selected for training and testing respectively. Each CNN was trained for 800 steps of 50 batches and then evaluated. Figure 5 displays each CNN's performance.

The average train data accuracy parameter is higher than the average test or evaluation data accuracy parameter , suggesting that the CNN could benefit from additional training data or more aggressive regression techniques. The evaluation accuracies for Face3ii, Face3iii and Face2i are appreciably low. The Face3ii and Face2i powersweeps had a large amount of I and Q noise because of an artifact in the digital readout – Face3iii and Face2ii are the repeat measurements after the resolution of the problem. Face3ii and iii contain higher frequency resonators which also results in more I and Q noise. Interestingly, the evaluation accuracy parameter achieved with Varu2 is among the highest despite the CNN having no experience with powersweeps from that array, indicating the utility of the CNN for future arrays.

The evaluation accuracy parameter of Ukko2 is not far the mean achieved across all feed-lines, indicating that the difficulty of that feed-line is fairly representative of all feed-line datasets measured. Furthermore, Ukko2 is preferable for the comparison between the bias selection methods because it has a uniform distribution of bias points. For a typical feed-line, the distribution of bias points peaks around the fifth bias, because the resonators are designed to be as identical as possible. If the CNN overfits to these classes it will be most apparent in Ukko2 measurements. Therefore, Ukko2 was chosen as the evaluation dataset for all remaining investigations.

5.1.2. Full CNN

To investigate the training data requirements, subsets of non-Ukko2 powersweeps were randomly selected for training at increasing amounts, and a new CNN algorithm was trained each time. The testing data were also non-Ukko2 powersweeps, but the full 212 powersweeps were used for each CNN. This process continued until the CNN was trained on all 3965 powersweeps for 1200 training steps. The batch size for each step matched the amount of training data until this passed 50 powersweeps, then the number of batches remained at 50 for each step. The accuracies achieved at each training step, on each subset, are displayed in Figure 6.

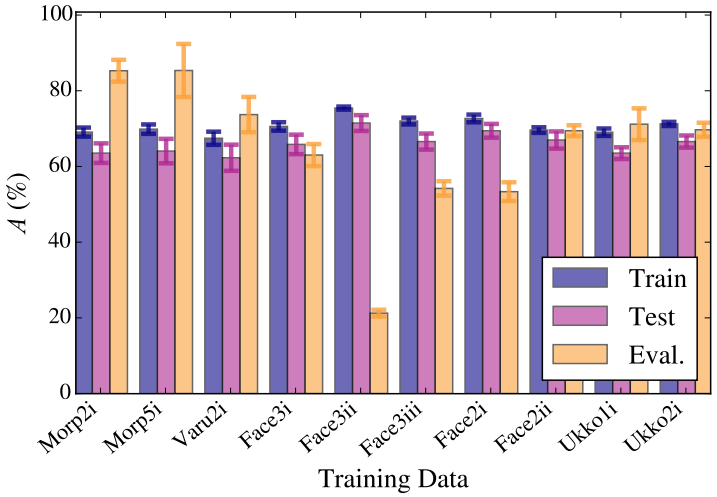

Figure 5: The CNN was trained on 3565 powersweeps from nine of the feed-line datasets, and using those powersweeps as m_t (Train), 180 unseen powersweeps from the same feed-lines as m_t (Test), and all of the powersweeps from the remaining feed-line dataset as m_t (Evaluation), the accuracy parameter A was evaluated from the CNN's predictions (m_e). This process was repeated for each of the ten feed-line datasets in turn. When Ukko2 was used as the evaluation dataset, the classifications from first of the four MI operators were used, $m_\mathrm{t} = \mathrm{MI_A}$. Each time the CNN was trained for 800 steps of 50 batches. These measurements were repeated ten times and the errorbars are one standard deviation from the mean of the accuracies achieved. The evaluation accuracy parameter achieved with Ukko2 indicates that that dataset is fairly representative of all feed-line datasets.

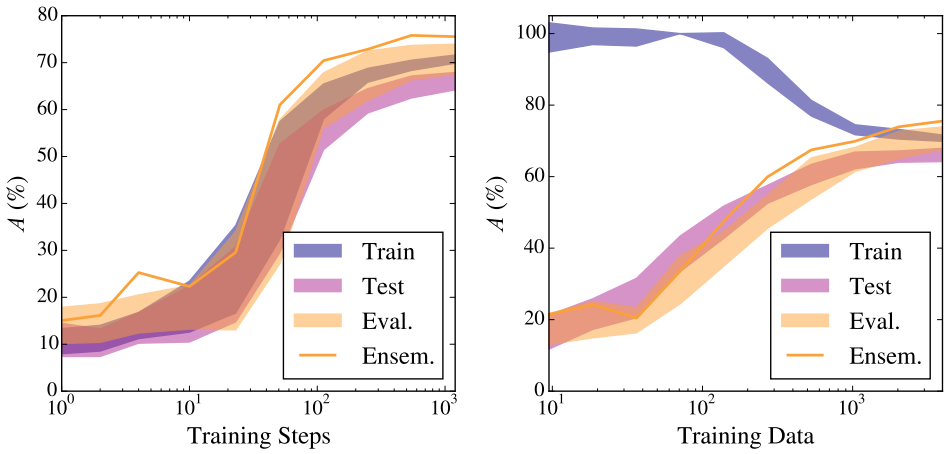

Figure 6: The CNN was trained on increasing amounts of randomly selected powersweeps from the non-Ukko2 datasets, and using those powersweeps as m_t (Train), 212 unseen powersweeps from the non-Ukko2 datasets as m_t (Test), and all of the powersweeps from Ukko2 as m_t (Evaluation), the accuracy parameter A was evaluated from the CNN's predictions (m_e). For the Ukko2 labels, the first of the four MI operators were used, $m_\mathrm{t} = \mathrm{MI}_\mathrm{A}$. For each training step 50 powersweeps from the training dataset were randomly selected. The Ensemble curve takes a median of all the predictions made on the evaluation dataset. **left**: The accuracy parameter of the CNN on the full amount of powersweeps from the training, testing and evaluation datasets, measured at different amounts of training steps. **right**: The achieved accuracy parameter after full amount of training steps on different amounts of training data. The extent of the filled regions are one standard deviation from the mean, of the range of accuracies achieved when repeating each set of measurements ten times, hence why A can go above 100%. The ensemble accuracies on Ukko2 demonstrate that sufficient accuracies can be achieved on minimal amounts of training data.

There is a large range of final accuracies on the evaluation dataset achieved by different instances of the CNN. The ensemble technique is therefore very advantageous in guaranteeing that the optimum accuracy is achieved. It also has an appreciable impact of accuracy parameter $A = 5\%$ above the final mean.

When investigating the effect of the amount of training data, initially the training accuracy is 100%. At these amounts of training data, the number of parameters in the CNN is sufficient to essentially store the powersweeps, rather than develop a model to make predictions on them. Similarly, the test and evaluation accuracies, when using low amounts of training data, have comparatively large variety. The plateaus at large amount of training data (and training steps) indicates higher accuracies require more standardized training data or more aggressive regularization. Interestingly, using the ensemble technique on just 200 training powersweeps is sufficient to reach over accuracy parameter $A = 60\%$.

5.2. Accuracy Comparison

The MI feed-line datasets will suffer from human error because of telescope deadlines, lack of experience from some operators as well as the subjectivity inherent in some types of powersweeps. In order to create a more accurate dataset for comparison, the evaluation feed-line was classified four times, each time by a different operator. To account for the subjective systematic offsets between the four MI datasets, A was measured for each of the six combinations. It was found that increasing half of labels of $\mathrm{MI_D}$ by 1 dB (increasing the average by 0.5 dB) maximized the total A of the six combinations. Three of these datasets were then combined by taking a median of the classifications for each powersweep, to produce the $\mathrm{MI_{av}}$ dataset. The remaining dataset, $\mathrm{MI_B}$, remained independent from $\mathrm{MI_{av}}$, to evaluate the MI performance against the automated methods. If any of the operators decided a powersweep was insufficient for classification, the powersweep and label were omitted from the evaluation dataset. This took the amount of powersweeps down from 377 to 340.

Confusion matrices are two-dimensional histograms used to compare the labels evaluated by two different classifier algorithms. If m_e agrees with m_t for a given powersweep, that powersweep (a true positive) will lie on the $b(m_\mathrm{e}) = b(m_\mathrm{t})$ diagonal. The false positives are all the values in the corresponding row, and the false negatives lie along the corresponding column (both excluding those located on the true positives diagonal).

Figure 7 shows the confusion matrices for two MI datasets, the results from the four methods when compared to the 'true' dataset $\mathrm{MI_{av}}$, and a new method that uses predictions from both NM and AM. In the top left of each panel is the accuracy parameter A. The uncertainty on A was taken to be the standard error from the a_p distribution, and the uncertainty from the quantization error in classification, added in quadrature.

$\mathrm{MI_B}$ tended to classify powersweeps more conservatively than $\mathrm{MI_A}$ by approximately 0.2 dB. When comparing all six pairs of MI datasets the A ranged from 79% to 85% for the confusion matrix shown. The primary reason for the disagreement between operators is the subjectivity of the classifications. This

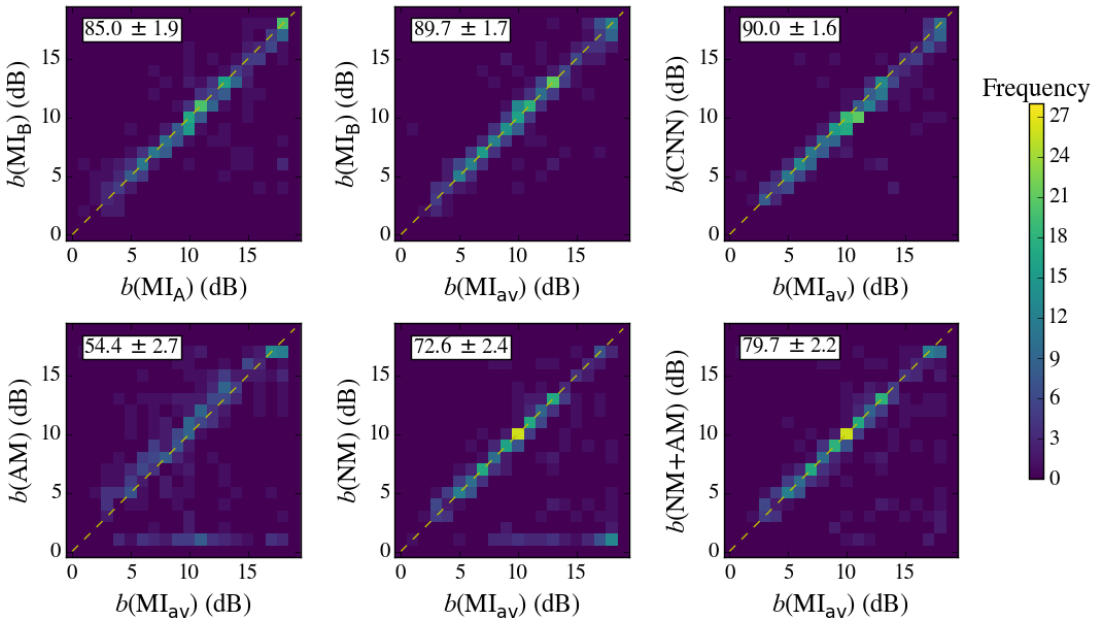

Figure 7: Confusion matrices comparing the evaluated bias points from different methods. $\mathrm{MI_{av}}$ is created using three of the four operators and is assumed to have the true bias values. $\mathrm{MI_A}$ and $\mathrm{MI_B}$ were the first two operators to use MI to classify the evaluation dataset. b(NM+AM) refers to a method that combines some of the predictions from both methods. The colour of the datapoints corresponds to the number of powersweeps in that bin (frequency). The dashed yellow diagonal line are true positive classifications. The accuracy parameter A, shown in the top left of each plot is the percentage of powersweeps that lie within 1dB of the true positives diagonal.

inconsistency in the training data places a limit on the attainable accuracy that the CNN can achieve.

Despite the inconsistency of the training data, the accuracy parameter achieved by the CNN was 90%, which is similar to the randomly selected operator, $\mathrm{MI_B}$. (If operator A, C or D is used as the independent operator and $\mathrm{MI_{av}}$ is recreated accordingly, the maximum difference between the accuracy parameter achieved by the CNN and MI is 2%). Only a handful of powersweeps are classified with more than 1 dB above $\mathrm{MI_{av}}$ ensuring that the vast majority of resonators are not bifurcated and will correctly operate as photon detectors. The accuracy of the CNN method is also substantially higher than AM and NM, and showing no bias towards certain optimal operating points. The reason for the apparent performance improvement of the CNN between Figures 6 and 7 is because of the elimination of 37 of the questionable powersweeps from, and the improved accuracy of, the $\mathrm{MI_{av}}$ labels compared to the $\mathrm{MI_A}$ labels, which are used as m_t. This makes all the methods appear more accurate. (If $m_\mathrm{t} = \mathrm{MI_A}$, as it was in Figure 6, the CNN's $A = 76\%$ is still 25% and 12% above that achieved by AM and NM, respectively.)

AM classified 20% of powersweeps with >1 dB power above the $\mathrm{MI_{av}}$, which would have likely rendered them unusable. Both AM and NM tend to falsely classify powersweeps in the second lowest bias. The additional noise in I and Q at these powers can trigger a false result on different pathologies (as shown in Figure 3). If this occurs in any of the lowest four power indices, because of the choice of implementation of Rule #1 (described in Section 2.4), the bias point prediction will be this value.

NM achieves an impressive 47% of powersweeps along the true positives diagonal. This could be attributed to the fact that the MI operators used the NM predictions as a first guess, and so for 'adequate' powersweeps (49.9% from Table 2) the operator would tend not modify those predictions. In an effort to extract the best predictions from both AM and NM, a new model was created that substituted the predictions from AM when $b(\mathrm{NM}) = 1$ dB. This method achieves an accuracy parameter $A = 80\%$, at the cost of the extra computation time of AM compared to NM.

Figure 8 shows how the different type of powersweep effects the accuracy parameter of each method. The accuracies of the AM and the NM are highest for ideal powersweeps as expected. The NM is vastly superior on powersweeps showing multiple discontinuities, since these are not accounted for in the AM model. Conversely, the AM does better on underpowered powersweeps and those with no discontinuity possibly because the AM technique can identify other features for the onset of bifurcation other than the formation of a discontinuity. It is apparent that the average fit with the collision model for AM was insufficient. The accuracy of this method might therefore benefit from Markov Chain Monte Carlo sampling of the larger parameter space and maximum likelihood estimation for the parameters, at the expense of increasing the time to perform the method even further.

The CNN has superior accuracy parameter, and smaller uncertainty, across all but two powersweep types. This is partly because of the higher sophistica-

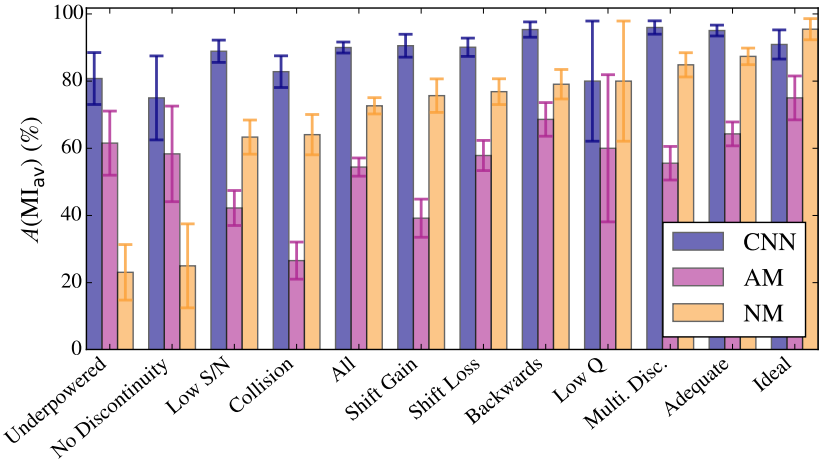

Figure 8: The accuracies of each of the automated methods when $m_\mathrm{t} = \mathrm{MI_{av}}$. The categories are the same powersweep types shown in Table 1. The associated error is the standard error from the a_p distribution and the reading error added in quadrature.

tion of the model, but also because, on powersweeps such as 'underpowered', the model has the advantage of having been exposed to these often-featureless powersweeps and learned the convention of classifying them as a high bias point.

5.3. Timing Comparison

Table 4 shows predicted times to tune a ten-kilopixel array using each of the described methods by extrapolating the times values determined on Ukko2. 'Rate' is a measure of the number of correctly classified powersweeps per minute if tuning an array for the first time. For a typical observing run, some feed-lines will require multiple retunings while others require only one round of tuning. In

Table 4: The operator and computation times are for biassing 1000 pixels (half a feedline). The observation run total ('Obs. Run') assumes that the full 10,000 pixel array would need to be completely tuned two times, and that retuning is 75% faster for AM and MI. The timing was performed on a 16 core Intel Xeon processor. It was assumed that the AM method could be optimized to achieve the same accuracy parameter in 20% of the time. Rate is calculated from the time it takes to tune 1000 pixels and the accuracy parameter A. The timing each method are rounded to the nearest significant value. The time to take the power-sweep measurements are not included.

Method	Operator	Computation	Obs. Run	A (%)	Rate (m^{-1})
MI	2h	30s	25h	90	7
AM	15m	1h	16h	54	7
NM+AM	1m	11m	3h	80	66
NM	–	3m	1h	73	242
CNN	–	1m	20m	90	900

other runs the device may need to be swapped, or there needs to be alterations to the digital readout hardware or software settings. Depending on the severity of the changes to the resonator properties, the retuning can be made faster by using values from the first tuning round as starting points. It was found that retuning, using either AM or MI, can be performed as quick as 25% of the time of the initial tuning process. This speed enhancement has been applied to those methods for the retunings.

For the AM, some operator time is required to identify which powersweeps need to be fit with the resonator collision model (18 parameters instead of 9). A first guess can be acquired by fitting all powersweeps with the standard AM model, and then manually checking those with the goodness-of-fit parameter beyond some threshold.

The timing and the accuracy parameter displayed for the NM include the low pass filtering of v_{IQ}. The method can be made three times faster at the expense of roughly accuracy parameter $A = 20\%$. When using MI, each powersweep is first classified using the NM for a first estimate to guide the operator. The majority of the time spent classifying powersweeps is then by the operator.

For the CNN method, the time shown is the time to classify the powersweeps. The creation of the training set and subsequent training would be done prior to an observation-run, and both processes do not have to be repeated for each new device, so those times are not included in the total observing run time. The CNN can be made ten times faster by not applying the ensemble averaging, at the expense of approximately accuracy parameter $A = 5\%$. The time to train the CNN with 1000 steps on the full training dataset took just three minutes. Furthermore, trained algorithms can be made publicly available to be used by other research laboratories[2].

For a typical telescope observing run, the total time spent tuning powersweeps with the CNN algorithm is a factor of 75 faster than using the current technique of MI, and correctly classifies more powersweeps per minute than any other of the investigated methods.

6. Conclusions

Analytical models make accurate predictions within their assumptions. Outside of this, either a more sophisticated model or human input is required. A neural network operates in the middle ground between these two techniques: it creates a very sophisticated (3×10^8 parameters) model using the human intuition (and accordingly any human error) encompassed in the training data. It finds the most optimal parameters to solve any type of powersweep as efficiently as possible, for a given number of training steps.

It has been shown that a relatively simple CNN (total training time of three minutes) can characterize a kilo-pixel MKID array in under 1 minute, with approximately equivalent 1 dB accuracy to the method of MI, and fewer over powered classifications, on a feed-line where the majority of powersweeps show non-ideal power handling behaviour. Roughly equivalent accuracies can be achieved

on completely unseen devices, and competitive accuracies can be achieved on just 200 powersweeps when using ensemble training. The CNN markedly outperforms the alternative automated methods across all powersweep types and classes. This accuracy is achieved in spite of the inconsistencies present in the training data. This further justifies the implementation of a rigorous automated algorithm to optimize the quality of MKID observational data.

The CNN algorithm has been used to set up the device and readout for each of the four previous DARKNESS observation runs at Palomar, saving potentially 100 human hours. The more and reliable training data that is accumulated, the more accurate this MKID tuning algorithm should become, which is implemented as part of the open source DARKNESS Digital Readout pipeline[2]. The ability to reliably automate the biasing of kilo-pixel arrays is vitally important as the community works towards the eventual goal of realizing mega-pixel MKID cameras (Marsden et al., 2013). Additionally, experiments can now be conducted on current devices which were unfeasible before, such as: measuring the change of the bias point of resonators across multiple feed-lines with the CNN, after recycling a adiabatic demagnetisation refrigerator; or using the CNN to tune a full array, and using photon data, measure the sensitivity at a range of powers around those bias points, to investigate the validity of Rule #1 on a statistically meaningful sample of resonators, free from human error.

Acknowledgements

We acknowledge the support of National Science Foundation grant AST-1308556 for the construction of the DARKNESS instrument. The MKID arrays described were fabricated with support from NASA grant NNX16AE98G. RD acknowledges the Science Technology Facilities Council support through the Doctoral Training and Long Term Attachment grants ST/M50371X/1. KO acknowledges PRD grant number ST/M003868/1. NT acknowledges the support of HARMONI grant ST/N002717/1.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al., 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. preprint arXiv:1603.04467 .

Adam, R., Adane, A., Ade, P., André, P., Andrianasolo, A., Aussel, H., Beelen, A., Benoit, A., Bideaud, A., Billot, N., et al., 2018. The nika2 large-field-of-view millimetre continuum camera for the 30 m iram telescope. Astronomy & Astrophysics 609, A115.

[2]https://github.com/abwalter/MkidDigitalReadout

Baselmans, J., Bueno, J., Yates, S., Yurduseven, O., Llombart, N., Karatsu, K., Baryshev, A., Ferrari, L., Endo, A., Thoen, D., et al., 2017. A kilopixel imaging system for future space based far-infrared observatories using microwave kinetic inductance detectors. Astronomy & Astrophysics 601, A89.

Bengio, Y., LeCun, Y., et al., 2007. Scaling learning algorithms towards ai. Large-scale kernel machines 34, 1–41.

Carter, F.W., Khaire, T.S., Novosad, V., Chang, C.L., 2017. scraps: An Open-Source Python-Based Analysis Package for Analyzing and Plotting Superconducting Resonator Data. IEEE Transactions on Applied Superconductivity 27, 1–5. doi:10.1109/TASC.2016.2625767.

Catalano, A., Adam, R., Ade, P., André, P., Aussel, H., Beelen, A., Benoît, A., Bideaud, A., Billot, N., Bourrion, O., et al., 2016. The nika2 commissioning campaign: performance and first results. preprint arXiv:1605.08628 .

Day, P.K., LeDuc, H.G., Mazin, B.A., Vayonakis, A., Zmuidzinas, J., 2003. A broadband superconducting detector suitable for use in large arrays. Nature 425, 817–821. URL: http://www.nature.com/doifinder/10.1038/nature02037, doi:10.1038/nature02037.

Gao, J., 2008. The physics of superconducting microwave resonators. Ph.D. thesis. California Institute of Technology.

Geerlings, K.L., 2013. Improving coherence of superconducting qubits and resonators. Yale University.

George, R.E., Senior, J., Saira, O.P., Pekola, J., de Graaf, S., Lindström, T., Pashkin, Y.A., 2017. Multiplexing superconducting qubit circuit for single microwave photon generation. Journal of Low Temperature Physics 189, 60–75.

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:1502.03167 [cs] URL: http://arxiv.org/abs/1502.03167. 01099 arXiv: 1502.03167.

Jonge, C.d., Baryshev, A.M., Ferrari, L., Yates, S.J.C., Baselmans, J.J.A., Endo, A., 2012. Development of a passive stand-off imager using MKID technology for security and biomedical applications, in: 2012 37th International Conference on Infrared, Millimeter, and Terahertz Waves, pp. 1–2. doi:10.1109/IRMMW-THz.2012.6380107.

Khalil, M., Stoutimore, M., Wellstood, F., Osborn, K., 2012. An analysis method for asymmetric resonator transmission applied to superconducting devices. Journal of Applied Physics 111, 054510.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. preprint arXiv:1412.6980 .

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

LeCun, Y., Bengio, Y., et al., 1995. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361, 1995.

Marsden, D.W., Mazin, B.A., O'Brien, K., Hirata, C., 2013. Giga-z: A 100,000 OBJECT SUPERCONDUCTING SPECTROPHOTOMETER FOR LSST FOLLOW-UP. The Astrophysical Journal Supplement Series 208, 8. URL: http://stacks.iop.org/0067-0049/208/i=1/a=8?key=crossref.67caf0d5ab9df225b42df73326f3802a, doi:10.1088/0067-0049/208/1/8.

Martinez, M., Bellini, F., Cardani, L., Casali, N., Castellano, M.G., Colantoni, I., Cosmelli, C., Cruciani, A., DAddabbo, A., Di Domizio, S., et al., 2017. Phonon-mediated kids as light detectors for rare event search: The calder project. IEEE Transactions on Applied Superconductivity 27, 1–5.

Mazin, B.A., Bumble, B., Meeker, S.R., O'Brien, K., McHugh, S., Langman, E., 2012. A superconducting focal plane array for ultraviolet, optical, and near-infrared astrophysics. Optics Express 20, 1503–1511.

Mazin, B.A., Meeker, S.R., Strader, M.J., Szypryt, P., Marsden, D., van Eyken, J.C., Duggan, G.E., Walter, A.B., Ulbricht, G., Johnson, M., Bumble, B., O'Brien, K., Stoughton, C., 2013. ARCONS: A 2024 Pixel Optical through Near-IR Cryogenic Imaging Spectrophotometer. Publications of the Astronomical Society of the Pacific 125, 1348–1361.

Meeker, S., Mazin, B., Jensen-Clem, R., Walter, A., Szypryt, P., Strader, M., Bockstiegel, C., 2015. Design and Development Status of MKID Integral Field Spectrographs for High Contrast Imaging. Adaptive Optics for Extremely Large Telescopes 4 - Conference Proceedings 1. URL: http://escholarship.org/uc/item/217686nz, doi:10.20353/K3T4CP1131701.

O'Brien, K., Thatte, N., Mazin, B., 2014. Kidspec: an mkid based medium resolution integral field spectrograph, in: SPIE Astronomical Telescopes+ Instrumentation, International Society for Optics and Photonics. pp. 91470G–91470G.

Semenov, A., Devyatov, I., de Visser, P., Klapwijk, T., 2016. Coherent excited states in superconductors due to a microwave field. Physical review letters 117, 047002.

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. Journal of machine learning research 15, 1929–1958.

Strader, M., Archibald, A., Meeker, S., Szypryt, P., Walter, A., van Eyken, J., Ulbricht, G., Stoughton, C., Bumble, B., Kaplan, D., et al., 2016. Search for optical pulsations in psr j0337+ 1715. Monthly Notices of the Royal Astronomical Society 459, 427–430.

Strader, M.J., 2016. Digital Readout for Microwave Kinetic Inductance Detectors and Applications in High Time Resolution Astronomy. Ph.D. thesis. California Institute of Technology.

Swenson, L.J., Day, P.K., Eom, B.H., Leduc, H.G., Llombart, N., McKenney, C.M., Noroozian, O., Zmuidzinas, J., 2013. Operation of a titanium nitride superconducting microresonator detector in the nonlinear regime. Journal of Applied Physics 113, 104501. doi:10.1063/1.4794808.

Szypryt, P., Duggan, G.E., Mazin, B.A., Meeker, S.R., Strader, M.J., Eyken, J.C.v., Marsden, D., O'Brien, K., Walter, A.B., Ulbricht, G., Prince, T.A., Stoughton, C., Bumble, B., 2014. Direct detection of SDSS J0926+3624 orbital expansion with ARCONS. Monthly Notices of the Royal Astronomical Society 439, 2765–2770. URL: http://mnras.oxfordjournals.org/content/439/3/2765, doi:10.1093/mnras/stu137.

Szypryt, P., Mazin, B.A., Ulbricht, G., Bumble, B., Meeker, S.R., Bockstiegel, C., Walter, A.B., 2016. High quality factor platinum silicide microwave kinetic inductance detectors. Applied Physics Letters 109, 151102. URL: http://aip.scitation.org/doi/abs/10.1063/1.4964665.

Thomas, C.N., Withington, S., Goldie, D.J., 2015. Electrothermal model of kinetic inductance detectors. Superconductor Science and Technology 28, 045012. URL: http://stacks.iop.org/0953-2048/28/i=4/a=045012, doi:10.1088/0953-2048/28/4/045012.

Ulbricht, G., Mazin, B.A., Szypryt, P., Walter, A.B., Bockstiegel, C., Bumble, B., 2015. Highly multiplexible thermal kinetic inductance detectors for x-ray imaging spectroscopy. Applied Physics Letters 106, 251103. URL: http://aip.scitation.org/doi/full/10.1063/1.4923096, doi:10.1063/1.4923096.

de Visser, P., Withington, S., Goldie, D., 2010. Readout-power heating and hysteretic switching between thermal quasiparticle states in kinetic inductance detectors. Journal of Applied Physics 108, 114504.

Yates, S., Baselmans, J., Endo, A., Janssen, R., Ferrari, L., Diener, P., Baryshev, A., 2011. Photon noise limited radiation detection with lens-antenna coupled microwave kinetic inductance detectors. Applied Physics Letters 99, 073505.

Zmuidzinas, J., 2012. Superconducting Microresonators: Physics and Applications. Annual Review of Condensed Matter Physics 3, 169–214. doi:10.1146/annurev-conmatphys-020911-125022.