Data Mining and Knowledge Discovery (2020) 34:1642–1675
https://doi.org/10.1007/s10618-020-00697-6

Comparison of novelty detection methods for multispectral images in rover-based planetary exploration missions

Hannah R. Kerner[1] · Kiri L. Wagstaff[2] · Brian D. Bue[2] · Danika F. Wellington[1] · Samantha Jacob[1] · Paul Horton[1] · James F. Bell III[1] · Chiman Kwan[3] · Heni Ben Amor[1]

Received: 2 September 2019 / Accepted: 3 June 2020 / Published online: 16 June 2020

Abstract

Science teams for rover-based planetary exploration missions like the Mars Science Laboratory Curiosity rover have limited time for analyzing new data before making decisions about follow-up observations. There is a need for systems that can rapidly and intelligently extract information from planetary instrument datasets and focus attention on the most promising or novel observations. Several novelty detection methods have been explored in prior work for three-channel color images and non-image datasets, but few have considered multispectral or hyperspectral image datasets for the purpose of scientific discovery. We compared the performance of four novelty detection methods—Reed Xiaoli (RX) detectors, principal component analysis (PCA), autoencoders, and generative adversarial networks (GANs)—and the ability of each method to provide explanatory visualizations to help scientists understand and trust predictions made by the system. We show that pixel-wise RX and autoencoders trained with structural similarity (SSIM) loss can detect morphological novelties that are not detected by PCA, GANs, and mean squared error autoencoders, but that the latter methods are better suited for detecting spectral novelties—i.e., the best method for a given setting depends on the type of novelties that are sought. Additionally, we find that autoencoders provide the most useful explanatory visualizations for enabling users to understand and trust model detections, and that existing GAN approaches to novelty detection may be limited in this respect.

Responsible editor: Indre Zliobaite.

✉ Hannah R. Kerner
hkerner@asu.edu

[1] Arizona State University, 781 E Terrace Mall, Tempe, AZ 85287, USA

[2] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA

[3] Applied Research, LLC, 9605 Medical Center Drive, Suite 113E, Rockville, MD 20850, USA

Keywords Novelty detection · Unsupervised learning · Space exploration

1 Introduction

The goal of novelty detection approaches is to identify patterns in data that have not been previously observed (Markou and Singh 2003a, b; Chandola et al. 2009; Pimentel et al. 2014). The exact definition of "novelty" varies depending on the application domain and the type of data, but in all cases novel examples differ in some way from "normal" data (Pimentel et al. 2014) and are of particular interest to the user (Chandola et al. 2009). In many real-world applications, novelty detection can provide significant, actionable information, such as a novel feature in a medical image may indicate the presence of a disease or tumor (Schlegl et al. 2017), or novelty in an X-ray scan at airport security may signal the presence of a weapon (Akcay et al. 2018).

One application domain that may greatly benefit from novelty detection techniques is rover-based planetary exploration. Rover missions like the Mars Science Laboratory (MSL) Curiosity rover are operated by mission team members through a system of "tactical planning": the rover is commanded to make observations on the surface of Mars, the rover sends the science data from those observations back to Earth, the science team analyzes the latest data and decides what observations to make on the next sol[1] based on that data, and the process repeats. Since the science team can only communicate with the rover when there is a clear line of sight between the rover or one of three Mars orbiters[2] and the Deep Space Network[3], there are only a few opportunities each day for downlinking new data and uplinking the new plan. Additionally, successive rover drives makes follow-up observation of late-identified science targets increasingly more costly to mission resources to pursue (since the rover would need to reverse course to re-visit the target). These factors require scientists to review the latest science data and identify targets of interest for follow-up analysis in a relatively short amount of time. MSL typically has less than 12 hours for science planning, and the upcoming NASA Mars 2020 rover mission may have as few as five hours (Wilson et al. 2017). There is a need for systems that can rapidly and intelligently extract information of interest from science instrument data to focus on potential discoveries and avoid missed science opportunities. These systems must also provide explanatory visualizations that allow scientists to trust and understand how a system came to its conclusion, a need that has not been explored extensively in prior work. We focused our study on the Mastcam imaging system onboard the MSL rover, which acquires multispectral images in the visible and near-infrared regions of the electromagnetic spectrum (Bell et al. 2017; Malin et al. 2017). A similar camera to Mastcam was onboard the Mars Exploration Rovers Spirit and Opportunity (Pancam, Bell III et al. 2008) and will be onboard the Mars 2020 rover (Mastcam-Z, Bell III et al. 2016). Thus, the ability to detect novel geology in Mastcam multispectral images

[1] The "sol" is the number of Martian days elapsed since MSL began operations on Mars.

[2] Mars Odyssey, the Mars Reconnaissance Orbiter, and the Trace Gas Orbiter (ExoMars).

[3] The Deep Space Network consists of three giant radio antennas in Goldstone (California), Madrid, and Canberra used for deep space communication.

could help increase the scientific return from past, present, and future rover-based Mars exploration missions. This work aims to enable planning and data analysis teams to spend their limited available time on the most promising, or novel, observations.

Since the goal of novelty detection is to identify patterns that have not frequently or ever been observed, in many application domains it is difficult to obtain labeled novel examples. Labeled examples from the typical class may be plentiful but novel labeled examples are more scarce. Thus, a common novelty detection approach is to construct a model based on the typical (non-novel) training examples and identify novel examples as those that are poorly explained by that model compared to typical examples (Pimentel et al. 2014). In this work, we compared the performance of multiple novelty detection methods based on principal component analysis, Reed-Xiaoli (RX) detectors, autoencoder neural networks, and generative adversarial networks for prioritizing images with novel geology in Mastcam multispectral images of the Martian surface. We evaluated the performance of these methods using multiple metrics chosen to represent their performance in operational use, including interpretability of detections (via reconstruction errors/residuals) which is an important factor for ensuring operational uptake of the methods. We present several key findings from our experiments:

- We propose a new autoencoder loss function—the structural similarity index (SSIM), a metric traditionally used for image quality assessment (Wang et al. 2004b)—and show that autoencoders trained with SSIM loss are better suited for detecting morphological novelties while autoencoders trained with conventional mean squared error (MSE) loss are better suited for detecting spectral novelties. In addition, we show that using the SSIM as a regularizing term to the MSE autoencoder loss can provide better novelty detection performance than the convential MSE-only loss.
- We show that pixel spectrum representations for RX can enable better performance on some novelty categories than other methods for which the input representation is the full multispectral image (flattened vector or tensor).
- We show that, of the compared approaches, autoencoders enable the most useful explanatory visualizations for users to understand and trust decisions made by the novelty detection system, but that existing GAN approaches may be limited in their ability to provide useful explanations for this purpose.
- The best novelty detection method for a given application depends on the type of novelties that are sought.

2 Related work

Methods for anomaly detection, including novelty detection and outlier detection, have been surveyed extensively (Markou and Singh 2003a, b; Marsland 2003; Hodge and Austin 2004; Agyemang et al. 2006; Modenesi and Braga 2009; Chandola et al. 2009; Pimentel et al. 2014). While outlier detection approaches aim to identify examples that deviate from the majority of examples in a dataset in an unsupervised manner, novelty detection approaches aim to identify examples that deviate from the examples

seen during training. Thus, novelty detection can be viewed as a one-class classification problem for which the standard approach is to construct a model for the typical (non-novel) training examples and identify novel examples in a test set that are not represented well by that model (Pimentel et al. 2014). Methods designed for outlier and novelty detection can often be used for both problems, thus we cover related work for both outlier and novelty detection in this section.

Kernel methods In one-class classification problems, only *typical* labeled examples are available during training, and novel labeled examples are used only for evaluation. The one-class support vector machine (OC-SVM) is a modification of the popular support vector machine (SVM) method to enable one-class classification (Scholkopf et al. 2000). OC-SVMs learn a decision boundary around only the typical data in the training dataset, though novel examples are often included in validation datasets for tuning hyperparameters since the performance of OC-SVMs depends strongly on the hyperparameter settings (Ma and Perkins 2003; Wang et al. 2004a; Manevitz and Yousef 2001; Munoz-Mari et al. 2010; Pimentel et al. 2014; Erfani et al. 2016; Zenati et al. 2018b). The Support Vector Data Description (SVDD) method is another extension of SVMs and OC-SVMs that automatically optimizes the model hyperparameters using artificially generated unlabeled data in a hypersphere around the typical data, and determines novelty by testing if an example lies within the hypersphere (Pimentel et al. 2014; Campbell and Bennett 2001; Tax and Duin 1999).

Reconstruction-based methods Reconstruction-based methods are another set of one-class classification approaches that characterize the typical class by learning a mapping between typical input examples and a lower-dimensional representation that minimizes the loss between the input and its reconstruction from the lower-dimensional representation. PCA can be used for reconstruction-based novelty detection, in which the reconstruction error between inputs and their inverse transformation from the principal subspace is used as a novelty score (e.g., Kwak 2008; Chandola et al. 2009; Toivola et al. 2010; Wagstaff et al. 2013; Xiao et al. 2013; Jablonski et al. 2015). Diaz and Hollmen (2002) used kernel-based and least-squares based general regression neural networks (GRNNs) for novelty detection and showed that kernel based approaches provided more meaningful and interpretable residuals (reconstruction errors) than least squares approaches. Given the success of deep neural networks at learning complex relationships in high-dimensional data (LeCun et al. 2015), more recent approaches have employed deep learning methods for reconstruction-based novelty detection. Similar to PCA, autoencoder neural networks (Hinton and Salakhutdinov 2006) are trained to minimize the reconstruction error for non-novel (typical) examples, and score the novelty of new inputs using the reconstruction error (e.g., Japkowicz et al. 1995; Thompson et al. 2002; Williams et al. 2002; Manevitz and Yousef 2007; Xiong and Zuo 2016; Richter and Roy 2017; Zhou and Paaenroth 2017; Kerner et al. 2019). Variational autoencoders have also been proposed for novelty detection (An and Cho 2015; Park et al. 2018). Generative adversarial networks (GANs) (Goodfellow et al. 2014), which have been successfully used for learning data-generating distributions for complex datasets (e.g., Antipov et al. 2017; Dong et al. 2018), have been recently proposed for novelty detection (Schlegl et al. 2017; Akcay et al. 2018; Zenati et al. 2018b).

Distribution and density estimation methods The Reed-Xiaoli (RX) method, which computes pixel-wise anomaly scores using the Mahalanobis distance between the pixel and a background distribution (Reed and Yu 1990), and its kernel variants are widely used for unsupervised anomaly detection in multispectral and hyperspectral images (e.g., Kwon and Nasrabadi 2005; Molero et al. 2013; Zhou et al. 2016; Ayhan et al. 2017; Wagstaff et al. 2019). Though RX is usually used for detecting global or local outliers/anomalies, it can be used for novelty detection by computing the background statistics from the typical training dataset as proposed in our study. For data that are Gaussian-distributed, thresholds on the likelihood of data modeled by a Gaussian probability distribution, or a mixture of Gaussians using Gaussian Mixture Models (GMM), can be used to identify novel or outlying examples(Chandola et al. 2009). Similarly, kernel density estimators (KDE) estimate the probability density of a dataset by assigning individual kernels (e.g., Gaussian kernels) to each data point and summing over all the kernels (Silverman 1986). Dense regions where points are close together will contribute more to the density estimate than points in diffuse regions of the feature space, thus outliers can be identified using a threshold on the likelihood under the learned probability distribution (e.g., Desforges et al. 1998; Latecki et al. 2007; Ristic et al. 2008; Laxhammar et al. 2009; Schubert et al. 2014). GMMs and KDEs are examples of probabilistic novelty detection methods, a category that also includes statistical hypothesis tests and box plots (Pimentel et al. 2014). Local outlier factor (LOF) (Breunig et al. 2000) detects outliers in sparse regions of the feature space by computing the local density of each point compared to its nearest neighbors; several modifications of LOF have been proposed (Tang et al. 2002; Chiu and Fu 2003; Papadimitriou et al. 2003; Tang et al. 2007).

Distance-based methods Distance-based methods for novelty detection include nearest-neighbor (e.g., Angiulli and Pizzuti 2002; Ertöz et al. 2003; Bay and Schwabacher 2003; Dongmei Ren et al. 2004; Abe et al. 2006; Yu et al. 2006; Zhang and Wang 2006; Ghoting et al. 2008) and clustering approaches (e.g., Yu et al. 2002; He et al. 2003; Pires and Santos-Pereira 2005; Srivastava and Zane-Ulman 2005; Srivastava 2006; Budalakoti et al. 2006; Clifton et al. 2007; Wang 2009; Filippone et al. 2010; Syed et al. 2010; Kim et al. 2012), but can be problematic for high-dimensional datasets due to their reliance on an appropriate distance metric (Pimentel et al. 2014). Carrera et al. (2015) used convolutional sparse models to learn local structures from typical images and detect novel regions of test images based on the distance between the learned filter and coefficient map as well as the spread of non-zero elements in the coefficient maps.

Other methods The Isolation Forest algorithm Liu et al. (2008) is a method based on random forests that "isolates" individual examples by recursively and randomly partitioning their features. The number of partitions required to isolate the example tends to be smaller (shallower trees) for novel examples than for typical examples. Novelty detection is closely related to the problem of zero-shot learning, which aims to classify examples from classes not seen during training and can involve detection of out-of-distribution examples that deviate from the training samples. Bhattacharjee et al. (2019) investigated the use of an autoencoder for novelty detection to detect examples from classes not seen in the training dataset. Lee et al. (2018) identified test samples far from training samples (i.e., novel or adversarial samples) using the Maha-

lanobis distance between the test sample and a class-conditional Gaussian distribution computed from the hidden representation at the end of a deep neural network.

An important limitation of prior studies is that they have largely evaluated novelty detection methods using datasets that contain non-image/relatively low-dimensional data, grayscale images, or color (RGB) images. Furthermore, many of these datasets are benchmark datasets and do not emulate real-world applications that would greatly benefit from operational uptake of novelty detection methods. In this work, we evaluated four novelty detection methods involving PCA, RX, autoencoders, and GANs for a real-world scientific multispectral image dataset acquired by the Mastcam instrument onboard the Mars Science Laboratory (Curiosity) rover for the real-world task of prioritizing images for review during planning of science operations. We focused our study on reconstruction-based methods because the residuals (reconstruction errors) provide a clear way of identifying what the novel features in the input are and directly relates to the novelty score for the image. We evaluated the performance of these methods using multiple metrics chosen to represent their performance in operational use, including interpretability of detections (via reconstruction errors/residuals). By evaluating existing novelty detection methods using a challenging real-world dataset of multispectral images, we demonstrated the tradeoffs in performance and interpretability between the methods and identified limitations not previously explored in prior studies.

3 Dataset

One instrument the MSL rover uses to make geologic observations on Mars is the mast camera, or "Mastcam," a pair of CCD imagers mounted on the rover's mast ∼2 meters above the surface (Grotzinger et al. 2012; Bell et al. 2017; Malin et al. 2017). A similar camera to Mastcam called "Mastcam-Z" will be onboard the Mars 2020 rover (Bell III et al. 2016). Each of the Mastcam cameras, or "eyes," has an eight-position filter wheel enabling images to be acquired in "true color" (Bayer pattern broadband red, green, and blue) and with six narrow-band spectral filters spanning ∼400-1100 nm (visible to short-wave near-infrared) (Bell et al. 2017). The imagers have different focal lengths: 34 mm for the left eye and 100 mm for the right eye, thus they are referred to as "M-34" and "M-100" respectively. Some of the band wavelengths also differ between each eye (Bell et al. 2017). For this reason, we considered images from the M-34 and M-100 as two separate datasets for which two separate novelty detection systems should be developed. In this study, we elected to use the M-100 (right eye) dataset because there were more multispectral images acquired using the right eye than the left eye in the period of the traverse we studied. Examples of novel geology in Mastcam images include iron meteorites (Wellington et al. 2017a; Johnson et al. 2014) and broken rocks that expose mineralogy under the dusty surface.

Our reconstruction-based novelty detection experiments require two datasets: one that represents the typical geology of Mars—which will be used for training, validation, and testing the models–and one that contains expert-identified novel examples, which will be used for testing only. To construct these datasets, we considered all Mastcam multispectral images that were acquired between sols (Martian days since landing) 1 to 1666 using all six narrow-band spectral filters (sols 1-1666 correspond to Earth dates

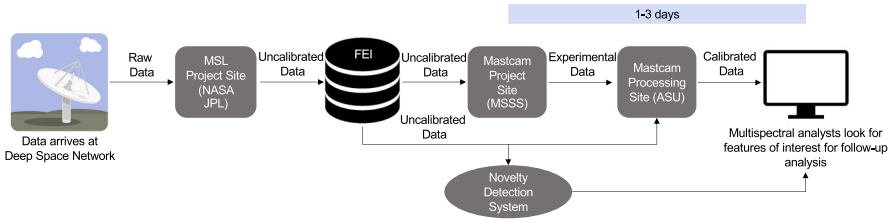

Fig. 1 Illustration of the Mastcam image processing pipeline

August 6, 2012-April 14, 2017). The earliest Mastcam image product that is available during tactical planning is the uncalibrated thumbnail version of the full-resolution image. When Mastcam images are acquired by the MSL rover, the full-resolution images are stored on the rover's onboard computer. The rover first downlinks low-resolution thumbnail versions of the Mastcam images, and later (as bandwidth allows) sends the full-resolution versions of the multispectral images. Figure 1 shows the flow of Mastcam image data once it arrives at the Deep Space Network. When data is sent from the rover to the Deep Space Network, the raw image data is sent to the MSL project team at the Jet Propulsion Laboratory, who processes the raw data and publishes the uncalibrated images to File Exchange Interface (FEI) servers that members of the Mastcam instrument team subscribe to. After the uncalibrated data is received by the primary Mastcam project site (Malin Space Science Systems, MSSS), it is processed to create "experimental data records" which contain additional non-image data (e.g., header information) that are not included in the uncalibrated image products. At the start of the next planning day after data is downlinked, these experimental data are sent to the Mastcam calibration team at Arizona State University (ASU), who creates radiometrically-calibrated versions of the data that are used by members of the MSL science team for tactical planning. By using the uncalibrated rather than radiometrically-calibrated versions of Mastcam images, a novelty detection system could deliver insights about potential observations of interest to the tactical planning team 1-3 days earlier than in the current, fully-manual process. While it is possible that novelty detection methods could achieve improved performance using full-resolution and/or calibrated observations, our goal is to provide insights about the novelty of new observations as early as possible *before* tactical planning begins. Thus, we chose to use the uncalibrated thumbnail versions of full-resolution multispectral images since these products are the earliest available products for tactical science planning.

These images constitute a dataset of 477 multispectral (6-band) thumbnails. Two co-authors (D. Wellington and S. Jacob) with experience reviewing Mastcam multispectral products reviewed the images in this source dataset and labeled regions they considered geologically interesting or "novel" with respect to the typical geology of Mars. These considerations were based on published results in the literature (e.g., Rice et al. 2013; Léveillé et al. 2014; Johnson et al. 2014, 2015; Wellington et al. 2017a, b; Wellington 2018) and science team discussions during the period of tactical planning in which the images in this dataset were acquired. These labels were in the

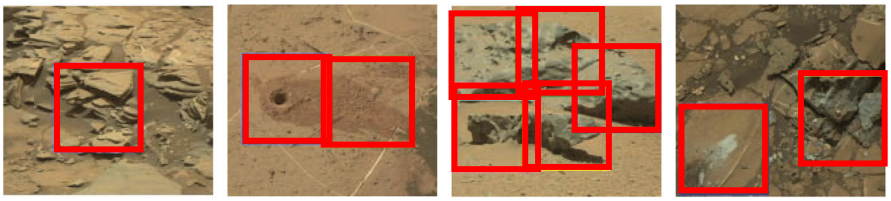

Fig. 2 Example bounding box labels around novel geology in Mastcam images

Table 1 Number of 64 × 64 × 6 image tiles in training, validation, and test datasets

Dataset	Num. typical tiles	Num. novel tiles
Train	9302	0
Validation	1386	0
Test	426	430

form of 64 × 64 × 6-pixel bounding boxes (e.g., Fig. 2[4]). They identified 237 novel bounding boxes within 156 of the 477 source images. We classified the remaining 321 source images as containing only typical geology, and partitioned them into training, validation, and test sets using a 80%/10%/10% split respectively (after randomization). To increase the number of typical images available for training and validation, we sub-sampled 64 × 64 × 6-pixel tiles from images in the training and validation datasets using a sliding window with a 16-pixel stride size. This resulted in 9,302 typical training tiles and 1,386 typical validation tiles. We used a larger stride size of 32 pixels to sub-sample tiles from the test images to reduce overlap between images in the test dataset, resulting in 345 typical test tiles.

The 156 images containing novel bounding box labels were used for testing only. We again sub-sampled 64 × 64 × 6-pixel tiles from these images using a 32-pixel stride. Any tiles that overlapped with the center of a novel bounding box label were classified as novel test tiles (430 tiles), while tiles that had no overlap with novel bounding box labels were classified as typical test tiles and were added to the typical test dataset (81 tiles). Table 1 summarizes the number of novel and typical images in each dataset.

To further assess model performance on different types of novel geology in Mastcam images, we divided the novel test dataset into 8 sub-classes based on input from expert Mastcam multispectral analysts: meteorite, float rock, bedrock, vein, broken rock, dump pile, drill hole, and dust removal tool (DRT) spot (Fig. 3). Meteorites are fragments of rocks from meteors originating elsewhere in the solar system that entered the Mars atmosphere and landed on the surface; meteorites are discovered serendipitously on Mars during rover exploration missions and depending on the type of meteorite can be obvious or subtle in Mastcam images (Wellington 2018). Float rocks are loose rocks transported to their current location by other geologic processes,

[4] Since the images in the Mastcam dataset are multispectral (6-band), in all figures displaying examples from the dataset we display channels 2 (805 nm), 0 (527 nm), and 1 (447 nm) of the input image as red, green, and blue (respectively).

Fig. 3 Eight categories of novel geology in Mastcam multispectral image dataset. Images shown are subframes from thumbnail images

Table 2 Number of image tiles in each sub-class of the novel test dataset

Sub-class	Num. tiles
DRT spot	111
Dump pile	93
Broken rock	76
Drill hole	62
Meteorite	34
Vein	30
Float rock	18
Bedrock	11

thus their composition resembles that of the source material rather than surrounding material (Wellington 2018). In contrast, bedrock describes material that formed in place. Veins are materials (usually light-toned) that fill fractures in rocks. The broken rock sub-class describes rocks that have been broken or crushed (by the rover wheel or other instrument) to expose the fresh interior of the rock. The dump pile, drill hole, and DRT spot sub-classes all describe features that are created by the rover performing contact science on the surface. Dump piles are the drill sample material analyzed by the CheMin and/or SAM (Blake et al. 2012; Grotzinger et al. 2012; Mahaffy et al. 2013) instruments dumped back onto the surface to make those instruments available for a new sample. Drill holes are holes created when the rover drills into a rock, and the tailings are the removed material surrounding the hole. DRT spots are elliptical spots where dust has been removed using the rotating dust removal tool on the rover; the contrast between the dust-removed area and the background varies depending on the color of the rock that underlies the dust and the thickness of the dust layer covering the rock. Table 2 gives the number of novel test tiles included in each category. The complete dataset can be accessed at https://doi.org/10.5281/zenodo.1486195.

4 Methods

In this novelty detection application, many labeled examples of typical geology and relatively few examples novel geology on the Martian surface are available. We chose four methods to evaluate for detecting novel geology in rover-based multispectral images using only typical examples during training: PCA, RX detectors, autoencoders, and GANs. All of these methods except RX are reconstruction-based methods, which we chose for their ability to enable visualization of novel detections within the image to

aid interpretation. We chose to evaluate RX and PCA because they are well established novelty detection methods and serve as informative baselines for the deep learning methods. We chose to evaluate autoencoders and GANs because, compared to traditional methods, deep learning methods have been shown to exhibit better performance for many high-dimensional image datasets. Code for each method can be accessed at https://github.com/JPLMLIA/mastcam-noveltydet. We discuss the details of our implementation for each method below.

4.1 PCA

PCA defines a linear projection of the data onto a principal subspace that retains maximal variance in the data. The principal components are the eigenvectors of the data covariance matrix, which can be computed using singular value decomposition (SVD) (Tipping and Bishop 1999):

$$\boldsymbol{U} = \text{SVD}(\boldsymbol{\Sigma}) \tag{1}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix for a dataset $\boldsymbol{X}$. PCA can be used to reduce the dimensionality of data by retaining the top k principal components (first k columns of $\boldsymbol{U}$) and projecting the data onto the k-dimensional principal subspace:

$$z = \boldsymbol{U}_r^T \boldsymbol{x} \tag{2}$$

where the columns of U_r contain the k principal components, $\boldsymbol{x}$ is a vector of the pixel intensities in the image, and $\boldsymbol{z}$ is the reduced-dimension representation of x in the principal subspace. The inverse transformation reconstructs the original data points from their representation in the principal subspace:

$$\hat{\boldsymbol{x}} = \boldsymbol{U}_r \boldsymbol{z} = \boldsymbol{U}_r \boldsymbol{U}_r^T \boldsymbol{x} \tag{3}$$

where $\hat{\boldsymbol{x}}$ is the reconstruction of image $\boldsymbol{x}$. When PCA is used to reduce the dimensionality of data, the reconstruction error between X and $\hat{X}$ can be interpreted as a novelty score (Wagstaff et al. 2013):

$$a_{\text{PCA}}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = ||\boldsymbol{x} - \hat{\boldsymbol{x}}||_2 \tag{4}$$

We used the Scikit-learn Python package for applying PCA (Pedregosa et al. 2011).

4.2 RX detector

The RX detector is commonly used to detect anomalous pixels in multispectral or hyperspectral images. RX assigns an anomaly score to each pixel that is the Mahalanobis distance between the pixel and a background distribution estimated from the data (Reed and Yu 1990; Chang and Chiang 2002). The background is usually defined to be all pixels in the image except the pixel under test, or a window of pixels around

the pixel under test. Since our goal is to identify novel features with respect to the dataset rather than within individual images, we instead computed the RX anomaly score for each pixel with respect to the background of the entire training dataset, i.e.:

$$a_{\mathrm{RX_p}}(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \boldsymbol{\mu}_t)^T \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_t) \tag{5}$$

where $\boldsymbol{x}_i \in \mathbb{R}^{1\times m}$ is the spectrum of pixel i across m multispectral bands, $\boldsymbol{\mu}_t \in \mathbb{R}^{1\times m}$ is the mean spectrum computed from all pixel spectra in the training dataset, and $\boldsymbol{\Sigma}_t \in \mathbb{R}^{m\times m}$ is the covariance matrix computed from all pixel spectra in the training dataset. Since Eq. 5 computes the RX anomaly score for each pixel in the image, not the entire image, we computed a representative RX anomaly score for an image as the mean RX score across all pixels in the image.

The RX anomaly score can also be computed for an image using the flattened multispectral image vector as the input representation rather than the pixel spectrum. This representation is equivalent to the input representation for PCA (Sect. 4.1). In this formulation, we compute the RX anomaly score for an image with respect to the background of the entire training dataset using the following equation (the subscript "f" indicates a flattened image representation):

$$a_{\mathrm{RX_f}}(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \boldsymbol{\mu}_t)^T \boldsymbol{\Sigma}_t^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_t) \tag{6}$$

where $\boldsymbol{x}_i \in \mathbb{R}^{1\times n}$ is the flattened multispectral image vector, n is the number of pixels in the multispectral image, $\boldsymbol{\mu}_t \in \mathbb{R}^{1\times n}$ is the mean image vector computed from all images in the training dataset, and $\boldsymbol{\Sigma}_t \in \mathbb{R}^{n\times n}$ is the covariance matrix computed from all images in the training dataset.

The inverse covariance matrix in Eqs. 5 and 6 projects the input data along the principal components with the least variance (lowest eigenvectors) and novelty is assessed as the distance from the mean in this space. Similarly, the reconstruction error in PCA assesses novelty as the distance of the input data from the mean in the space of the low-variance principal components (Chang and Chiang 2002). The primary difference between PCA and RX is that the reconstruction error in PCA measures the residual information along the components $k+1$ to n where k is number of high-variance components retained in the projection matrix and n is the number of input data features (n is the number of pixels for the flattened image representation), whereas the inverse covariance matrix in RX includes all components (Chang and Chiang 2002; Wagstaff et al. 2019).

4.3 Autoencoder

An autoencoder is a type of neural network that learns the salient features in a dataset. A convolutional autoencoder (CAE) consists of an encoder network to map (compress) inputs to a low-dimensional encoding and a decoder network to reconstruct inputs from the encoding (also called the "bottleneck" representation) using convolutional layers (Masci et al. 2011). We used a CAE with three convolutional layers in the encoder and three transposed convolutional layers in the decoder. We used 5×5 convolution

Table 3 CAE architecture. Layers beginning with "E" and "D" are part of the encoder and decoder, respectively

Layer	Dimension
Input	$64 \times 64 \times 6$
E1	$64 \times 64 \times 12$
E2	$32 \times 32 \times 8$
E3 (Bottleneck)	$16 \times 16 \times 3$
D1	$32 \times 32 \times 8$
D2	$64 \times 64 \times 12$
D3 (Output)	$64 \times 64 \times 6$

kernels in all layers with a stride size of 1 pixel in the first and last layers and 2 pixels in all other layers. Table 3 gives the size of representations at each layer of the CAE. The dimension of the bottleneck representation is $16 \times 16 \times 3$, thus inputs are compressed by a factor of 32 before being reconstructed by the decoder. The CAE is trained using a dataset of typical images.

CAE loss functions The most common loss function used to minimize error between CAE inputs and reconstructions is mean squared error (MSE) (Masci et al. 2011; Xiong and Zuo 2016; Richter and Roy 2017; Kerner et al. 2019), defined as:

$$\mathrm{E}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = \frac{1}{NMK} \sum_{k=1}^{K} \sum_{j=1}^{M} \sum_{i=1}^{N} (x_{ij}^{k} - \hat{x}_{ij}^{k})^2 \tag{7}$$

where x_{ij}^{k} and $\hat{x}_{ij}^{k}$ are the pixel intensities at row i, column j, and band k of the input and reconstructed images $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ respectively; N and M are the spatial dimensions of each image; and K is the number of multispectral bands. Other loss functions have been proposed in prior work, including binary cross-entropy (in which outputs for each pixel are interpreted as a probability) (Alain and Bengio 2014; Creswell et al. 2017) and mutual information (Hjelm et al. 2019).

One limitation of MSE loss is that two images with the same MSE can have very different spatial distributions of pixel errors, e.g., the error can be dispersed as noise throughout the image or might distort the structure of subjects in the image (Wang and Bovik 2009). Diaz and Hollmen (2002) also showed that least-squares based methods can lead to trivial solutions and ineffective residuals for explaining abnormal deviations. The structural similarity index (SSIM) was originally proposed for image quality analysis to overcome these limitations of MSE (Wang et al. 2004b). SSIM measures the degradation of structural information between an image and its compressed version (e.g., using JPEG compression) based on the assumption that the human visual system focuses on structural information in a scene (Wang et al. 2004b; Wang and Bovik 2009). Considering the CAE as a compression function and the reconstructed image as the lossy form of the input image, we propose to use SSIM to optimize the weights of the CAE such that SSIM is maximized. SSIM between a multispectral input image $\boldsymbol{X}$ and its reconstruction $\hat{\boldsymbol{X}}$ is defined as:

$$\mathrm{S}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = \frac{1}{K} \sum_{k=1}^{K} \frac{(2\mu_{\boldsymbol{X}_k}\mu_{\hat{\boldsymbol{X}}_k} + C_1)(2\sigma_{\boldsymbol{X}_k\hat{\boldsymbol{X}}_k} + C_2)}{(\mu^2_{\boldsymbol{X}_k} + \mu^2_{\hat{\boldsymbol{X}}_k} + C_1)(\sigma^2_{\boldsymbol{X}_k} + \sigma^2_{\hat{\boldsymbol{X}}_k} + C_2)} \tag{8}$$

where K is the number of bands in $\boldsymbol{X}$, $\mu_{\boldsymbol{X}_k}$ and $\mu_{\hat{\boldsymbol{X}}_k}$ are the mean pixel intensities in band k of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$; $\sigma^2_{\boldsymbol{X}_k}$ and $\sigma^2_{\hat{\boldsymbol{X}}_k}$ are the variances in pixel intensities in band k of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$; $\sigma_{\boldsymbol{X}_k\hat{\boldsymbol{X}}_k}$ is the covariance between pixel intensities in band k of $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$; and $C_1 = 0.01R$ and $C_2 = 0.03R$ are small constants to ensure positive values, where $R = 255$ is the dynamic range of the pixel intensities (Wang et al. 2004b). Because we want to maximize SSIM when training the autoencoder, the SSIM loss function minimizes the negative SSIM.

Finally, we propose to use a third loss function that combines MSE and SSIM loss:

$$\mathrm{H}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = -\mathrm{S}(\boldsymbol{X}, \hat{\boldsymbol{X}}) + \lambda \mathrm{E}(\boldsymbol{X}, \hat{\boldsymbol{X}}) \tag{9}$$

where λ is a constant weighting factor used to balance the magnitude of MSE loss with that of SSIM, since SSIM values range from (-1, 1].[5]. In Sect. 5.1, we compare the performance a CAE trained with MSE loss, SSIM loss, and a hybrid loss combining MSE and SSIM for separating novel and typical examples. We did not consider binary cross-entropy loss because we wish to interpret the reconstructions in each pixel as the "expected" signal, compared to the input ("observed") signal. We used the TensorFlow library in Python for implementation (Abadi et al. 2015).

CAE novelty scores When CAEs are used for novelty detection, the novelty score for a test image is typically chosen to be the MSE (Eq. 7) or l_2-norm between the input and reconstructed image (Richter and Roy 2017; Zhou and Paaenroth 2017; Kerner et al. 2019):

$$a_{\mathrm{CAE}}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = ||\boldsymbol{X} - \hat{\boldsymbol{X}}||_2 \tag{10}$$

In prior work, we observed that much of the errors between pixels in the input and reconstructed image are due to noise in the reconstruction rather than a spatial or spectral feature that was poorly reconstructed (Kerner et al. 2019). To combat this, we propose a new novelty score that captures the number of large errors between input and reconstructed images, which we refer to as "outlier count." The outlier count is computed as the number of errors in each pixel between the input and reconstructed image that are above the mean error, i.e., the number of pixels for which the following inequality is true:

$$(x^k_{ij} - \hat{x}^k_{ij})^2 > \frac{1}{NMK} \sum_{k=1}^{K} \sum_{j=1}^{M} \sum_{i=1}^{N} (x^k_{ij} - \hat{x}^k_{ij})^2 \tag{11}$$

[5] https://www.tensorflow.org/api_docs/python/tf/image/ssim.

where x_{ij}^k, $\hat{x}_{ij}^k$, N, M, and K are defined as in Eq. 7. In Sect. 5.1, we compare the effectiveness of two novelty scores—l_2-norm (Eq. 10) and outlier count (Eq. 11)—for separating novel and typical test examples.

4.4 GAN

Generative adversarial networks (GANs) are a type of neural network that learns the data-generating distribution for a dataset via minimax optimization of two networks (Goodfellow et al. 2014). The generator network $G(\boldsymbol{z})$ samples from a d-dimensional normal distribution (where d is the size of the latent vector $\boldsymbol{z}$) and tries to reconstruct an image $\hat{\boldsymbol{X}}$ that resembles images in the training dataset. The discriminator network $D(\boldsymbol{X})$ tries to distinguish training images from generated images by classifying inputs as *real* or *fake*. The discriminator *minimizes* the binary cross-entropy loss during training, while the generator simultaneously *maximizes* the discriminator loss:

$$C_{\text{dis}}(y, \hat{y}) = y\log(\sigma(\hat{y})) - (1 - y)\log(1 - \sigma(\hat{y})) \tag{12}$$

where y is the binary label (*real* or *fake*) for the input image, $\hat{y}$ is the logit output from $D(\boldsymbol{X})$, and $\sigma(\hat{y})$ is the sigmoid function used to map logits to the interval [0, 1]:

$$\sigma(\hat{y}) = \frac{1}{1 + e^{-\hat{y}}} \tag{13}$$

In a bi-directional GAN (BiGAN), a third network called the encoder network $E(\boldsymbol{X})$ is simultaneously trained to map images to a latent vector $\boldsymbol{z} \in \mathbb{R}^d$ (similar to the encoder network of the CAE) (Donahue et al. 2017). Thus in a BiGAN, the generator learns to reconstruct images from an encoded vector as opposed to randomly-sampled vector $\boldsymbol{z}$, i.e., $G(E(\boldsymbol{X}))$. We used the BiGAN approach to novelty detection proposed in (Zenati et al. 2018b, a), in which the BiGAN is trained using a dataset of typical images and the dimension of $\boldsymbol{z}$ is 1×100. Table 4 describes the BiGAN architecture. The encoder network uses 5×5 convolution kernels while the generator and discriminator networks use 4×4 convolution kernels. All convolutional layers use 2×2-pixel strides except the first layer of the encoder network. The novelty score of an image is defined as:

$$a_{\text{GAN}}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = (1 - \alpha)\mathcal{L}_G + \alpha\mathcal{L}_D \tag{14}$$

where $\mathcal{L}_G$ is the generator loss $||\boldsymbol{X} - G(E(\boldsymbol{X}))||$, $\mathcal{L}_D$ is the discriminator feature loss $||f(\boldsymbol{X}) - f(\hat{\boldsymbol{X}})||$ where $f(\cdot)$ represents the feature activations at the last layer of the discriminator network (preceding the logit output layer), and α is a constant weighting factor between the two terms (Zenati et al. 2018a). We used $\alpha = 0.1$ as in Zenati et al. (2018a).

Table 4 BiGAN architecture (Zenati et al. 2018a)

Layer	Dimension
Encoder	
Input	$64 \times 64 \times 6$
E1	$64 \times 64 \times 64$
E2	$32 \times 32 \times 128$
E3	$16 \times 16 \times 256$
Output	1×100
Generator	
Input	1×100
G1	1×256
G2	1×32768
G3	$32 \times 32 \times 64$
Output	$64 \times 64 \times 6$
Discriminator	
Input	$64 \times 64 \times 6$
D1	$32 \times 32 \times 64$
D2	$16 \times 16 \times 128$
D3	$8 \times 8 \times 256$
D4	1×512
Output	1×1

5 Experiments

5.1 Autoencoder loss function

In Sect. 4.3, we described three loss functions for tuning the encoder and decoder weights of the CAE for novelty detection: mean squared error (MSE) (Eq. 7), structural similarity (SSIM) (Eq. 8), and a hybrid loss combining MSE and SSIM (Eq. 9). We compared the hybrid loss function performance using $\lambda = 0.1, 0.01, 0.001$. We trained the CAE until validation loss converged for each of the three compared loss functions (batch size = 100) and tested each model on the test dataset containing typical and novel image examples. Figure 4 shows the receiver operating characteristics (ROC) curves for the CAE with MSE, SSIM, and hybrid loss combined with the l_2-norm and outlier count novelty scores. ROC curves illustrate the tradeoff between the true positive rate and the false positive rate for a range of threshold settings (in this case, a threshold on the novelty score to separate novel from typical examples) (Krzanowski and Hand 2009). Points along the ROC curve that have higher true positive rates and lower false positive for each threshold value will be closer the upper left quadrant of the plot and span a larger area under the curve. Thus, the area under the curve (AUC) computed from each ROC curve (Table 5) is often used to summarize a model's discrimination performance and compare multiple models before selecting a threshold (Rosset 2004).

The CAE-SSIM and CAE-Hybrid ($\lambda = 0.01$) methods combined with the outlier count novelty score tied for the highest AUC score (0.65) for the test dataset. However,

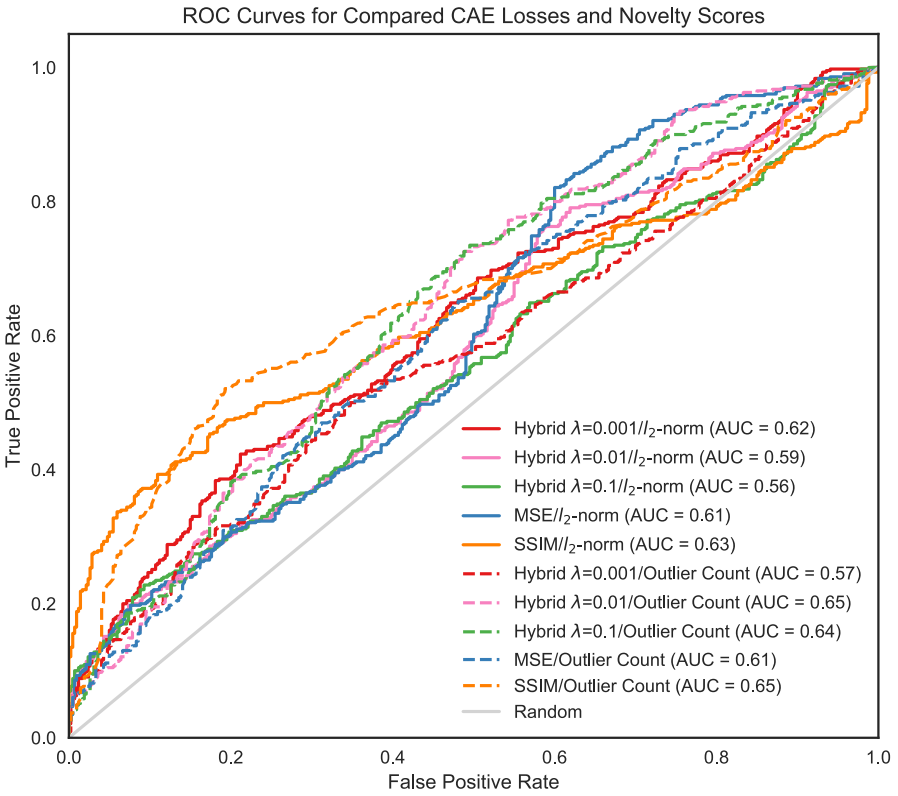

Fig. 4 ROC curves for CAE trained with each of three loss functions and two novelty scores (Color figure online)

Table 5 ROC AUC score for CAE with MSE, SSIM, and hybrid loss functions on test dataset (highest score in bold)

Loss function	Novelty score	AUC
MSE	l_2-norm	0.61
Hybrid ($\lambda = 0.1$)	l_2-norm	0.56
Hybrid ($\lambda = 0.01$)	l_2-norm	0.59
Hybrid ($\lambda = 0.001$)	l_2-norm	0.62
SSIM	l_2-norm	0.63
MSE	Outlier Count	0.59
Hybrid ($\lambda = 0.1$)	Outlier Count	0.64
Hybrid ($\lambda = 0.01$)	**Outlier Count**	**0.65**
Hybrid ($\lambda = 0.001$)	Outlier Count	0.57
SSIM	**Outlier Count**	**0.65**

comparing the ROC curves for these two methods (orange vs. pink dashed lines) shows that the CAE-SSIM method reached its maximum true positive rate at a much lower false positive rate than the CAE-Hybrid method. In other words, a higher true positive rate for the CAE-Hybrid method would come at the expense of more false positives than for the CAE-SSIM method (with respect to this test dataset). In practice, for a novelty detection system being used operationally for tactical planning, it is important for the ratio of correctly prioritized novel observations to incorrectly prioritized ones to be high in order to accelerate image review while also maintaining the user's trust in the system—thus, between two models with equivalent AUC scores, the model with the lower false positive rate at its maximal true positive rate is preferred (CAE-SSIM in this experiment).

5.2 Novelty detection performance

We evaluated each novelty detection method described in Sect. 4 on the combined novel and typical test dataset. We evaluated the CAE using the SSIM, MSE, and hybrid ($\lambda = 0.01$) loss functions. We used the outlier count novelty score for CAE-SSIM and CAE-Hybrid, and the l_2-norm novelty score for CAE-MSE (i.e., the scores that gave the best performance for these loss functions; see Fig. 4 and Table 5). We computed the ROC AUC score for each method using the entire novel test dataset as well as each of the 8 novel sub-classes (Fig. 3, Table 2), combined with the typical test dataset. We report these scores in Fig. 5 and Table 6. We labeled the results from the PCA method as "PCA (flat)" to emphasize that the input representation for this method is the flattened multispectral image vector (as in the RX (flat) method), in contrast with the CAE and GAN methods for which inputs are tensors and the RX (pixel) method for which inputs are pixel spectra. The gray dashed line in Fig. 5 indicates the AUC score for a random choice of novel or typical for each example.

For the entire novel test dataset combined with the typical test dataset, the RX (pixel) method had the best performance and RX (flat) had the worst performance (no

Table 6 ROC AUC scores for combined novel and typical test dataset overall and for each novel sub-class

Novelty class	CAE-Hybrid	CAE-MSE	CAE-SSIM	GAN	PCA^{f}	RX^{f}	RX^{p}
All	0.65	0.61	0.66	0.59	0.50	0.49	**0.72**
Meteorite	0.50	0.77	0.56	0.74	0.78	0.45	**0.97**
Float	0.44	0.82	0.37	0.78	**0.87**	0.56	0.57
Bedrock	0.66	0.87	0.16	**0.88**	0.87	0.42	0.53
Vein	0.58	0.94	0.13	**0.96**	0.95	0.36	0.22
Broken rock	0.64	0.88	0.38	0.84	**0.90**	0.49	0.42
Dump pile	0.66	0.48	**0.88**	0.49	0.47	0.50	0.83
Drill hole	0.81	0.49	0.79	0.52	0.51	0.51	**0.90**
DRT spot	0.64	0.39	0.85	0.34	0.41	0.52	**0.86**

The superscript "f" indicates the input representation is a flattened image vector and "p" indicates the input representation is a pixel spectrum. Bold text indicates the highest AUC score in each category

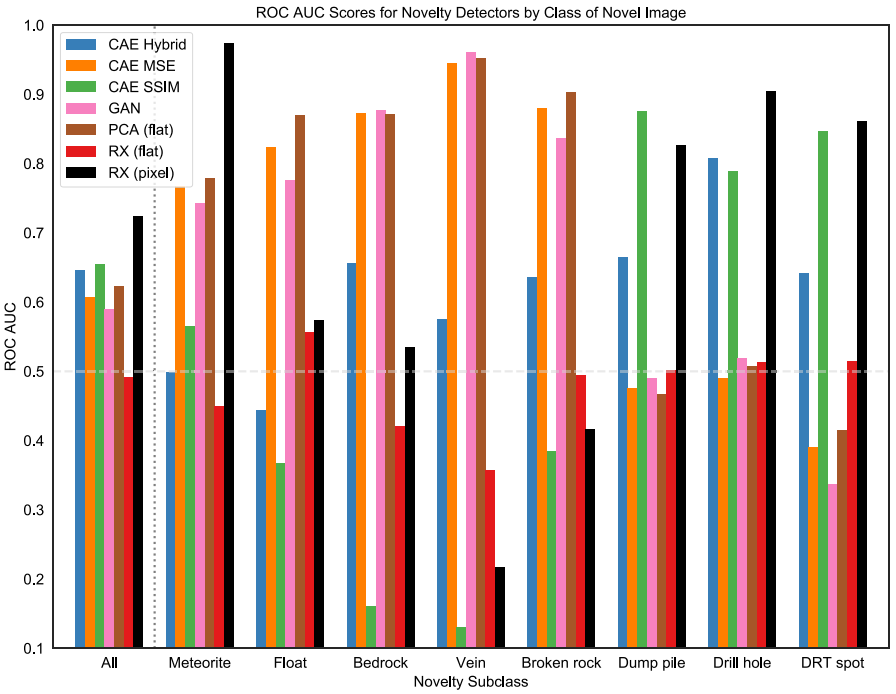

Fig. 5 ROC AUC scores for combined novel and typical test dataset overall and for each novel sub-class (Color figure online)

better than random guessing). In the float, bedrock, vein, and broken rock categories, the CAE-MSE, GAN, and PCA methods performed comparably well, while the CAE-SSIM, CAE-Hybrid, RX (flat), and RX (pixel) methods performed significantly worse (AUC scores near or worse than random). In the drill hole, DRT spot, and dump pile categories, all methods except CAE-SSIM, CAE-Hybrid, and RX (pixel) performed poorly. In the meteorite category, RX (pixel) had the best performance followed by comparable performance by the CAE-MSE, PCA, and GAN methods; CAE-SSIM, CAE-Hybrid, and RX (flat) had the lowest performance in the meteorite category. Because the DRT spot, drill hole, and dump pile categories have the highest frequency in the novel test dataset than the other categories (Table 2), high performance scores in these categories for RX (pixel), CAE-SSIM, and CAE-Hybrid result in higher AUC scores when using the entire novel dataset, despite poor performance in several other categories. These results reveal three groups of novel image categories based on model performance: one that contains the drill hole, DRT spot, and dump pile categories; one that contains the float, bedrock, vein, and broken rock categories; and one that contains the meteorite category. We will explore explanations for the differences in model performance for these three categories further in Sect. 7.

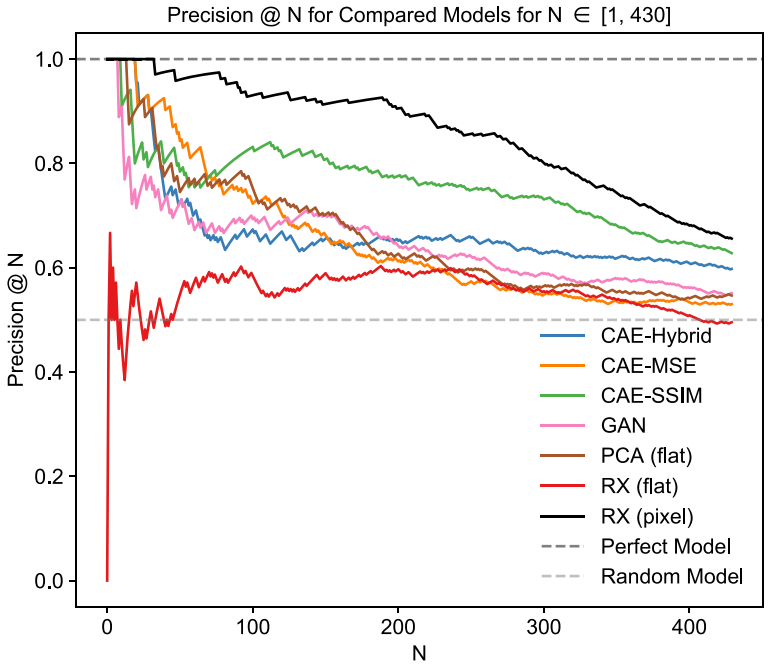

Fig. 6 Precision at N for each method up to $N = 430$ (number of novel images in test dataset) (Color figure online)

Table 7 Precision at $N = 20$ and fraction of novel examples in the bottom 20 images by novelty score

Method	Precision at $N = 20$	False Negatives at $N = 20$
RX (pixel)	**1.0**	**0.45**
CAE-MSE	**1.0**	**0.80**
PCA (flat)	0.90	0.05
CAE-Hybrid	0.90	0.30
CAE-SSIM	0.80	0.50
GAN	0.75	0.10
RX (flat)	0.55	0.70

Method with highest precision at $N = 20$ in bold

5.3 Novelty ranking

During a science planning cycle for Mastcam or other instruments onboard the MSL Curiosity rover, science team members process and analyze the latest images downlinked from the rover's onboard computer to determine targets of interest for follow-up analysis. Thus, in practice, science team members could benefit from novelty detection algorithms that rapidly prioritize the most interesting observations, e.g., by ranking new images by novelty score. To evaluate the performance of each novelty detection method in this prioritization context, we sorted the images in the combined novel and

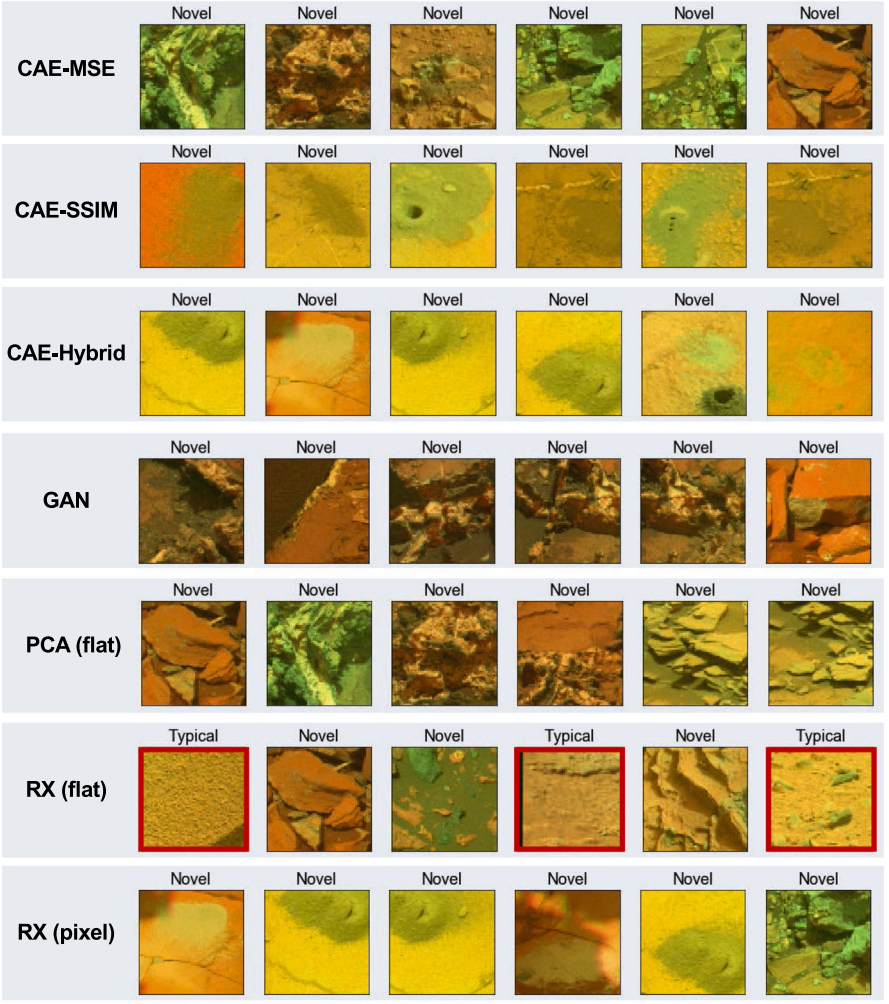

Fig. 7 Test images with 6 **highest** novelty scores for each method. Red frames indicate false positives (Color figure online)

typical test dataset by novelty score in descending order and calculated the "precision at N" (P@N) for $N \in [1, 430]$ (Fig. 6), where $N = 430$ is the number of novel examples in the test dataset. Precision at N is the proportion of correct results in the top N ranks (Campos et al. 2016):

$$P@N = \frac{\#\text{true positives}}{N} \tag{15}$$

Figure 6 shows that the RX (pixel) method has the highest P@N for all values of N. The CAE, GAN, and PCA methods show similar trends in which P@N scores start high then gradually decrease, with the exception of CAE-SSIM which increases

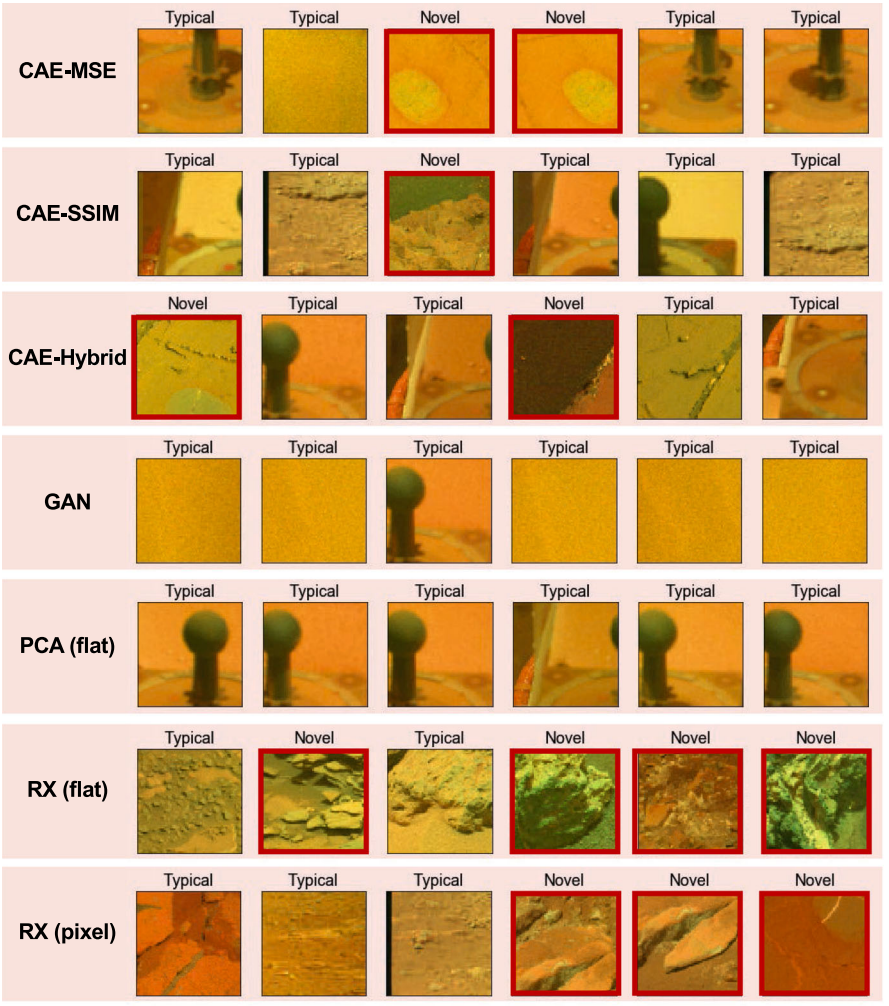

Fig. 8 Test images with 6 **lowest** novelty scores for each method. Red frames indicate false negatives (Color figure online)

steeply around $N = 60$. To help science planning teams focus their limited available time on the most promising observations from a newly-downlinked set of multispectral image observations, it is more important to have high P@N for low values of N with as few false negatives (representing missed novelties) as possible. Table 7 shows the precision at $N = 20$ as well as the fraction of novel examples in the *bottom* 20 images by novelty score (false negatives at N). The RX (pixel) and CAE-MSE methods have the highest precision at $N = 20$, but the RX (pixel) method has a lower false negative rate. PCA has the lowest fraction of false negatives at $N = 20$ and a $P@N = 20$ score close to 1.0 (the overall best P@N score). Thus, we concluded that the RX (pixel), CAE-MSE, and PCA methods had the best performance as measured by P@N.

In Fig. 7, we show the 6 images with the highest novelty scores using each method, and the lowest novelty scores in Fig. 8. Figure 7 shows that all methods except RX(flat) correctly identified only novel examples in the top 6 ranks. Figure 8 shows that the PCA and GAN methods all correctly identified only typical examples in the bottom 6, while all other methods had false negatives in the bottom 6 images. The images ranked least novel by the PCA and GAN methods primarily contain features that occur frequently in the training dataset—the calibration target (black cylindrical object with sphere on top) and sand.

6 Explanations

For the proposed novelty detection methods to be useful in practice, they must also provide explanatory visualizations that allow scientists to trust and understand *why* an image was identified as novel and *what* features within the image are novel. Since the GAN, CAE, and PCA methods are reconstruction-based methods in which the novelty score of the overall image is some measure of the similarity between the input and reconstructed images, the residual between the input and reconstructed images can be used as a visualization of the features in the input image (both in the spatial and spectral dimensions) that were poorly reconstructed by the model (as in Kerner et al. 2019; Diaz and Hollmen 2002, e.g.). We defined the residual, or error map, $\delta(\mathbf{X}, \hat{\mathbf{X}})$ as a $64 \times 64 \times 6$ tensor containing elements $(x_{ij}^k - \hat{x}_{ij}^k)^2$ for $i = 1, ..., N$, $j = 1, ..., M$, and $k = 1, ..., K$ where $N = 64$ is the number of rows, $M = 64$ is the number of columns, and $K = 6$ is the number of channels in each multispectral image (Kerner et al. 2019). RX is not a reconstruction-based method, but does RX compute an anomaly score for each pixel in the image which can be visualized as single-channel image. In Fig. 9, we show the reconstructions and explanatory visualizations for each model (error map for the GAN, CAE, and PCA methods; pixel-wise anomaly scores for RX) for an example from the novel test dataset that contains a nickel-iron meteorite. In all error maps except RX, we show the error in bands 2, 0, and 1 (same bands as shown for the input and reconstructed image).

Figure 9 shows that similar explanations are produced using all three CAE methods for the example shown, where most high-error pixels in the error map correspond to the novel meteorite in the input image (though different bands have higher errors between the three methods). The PCA error map also shows high-error pixels that coincide with

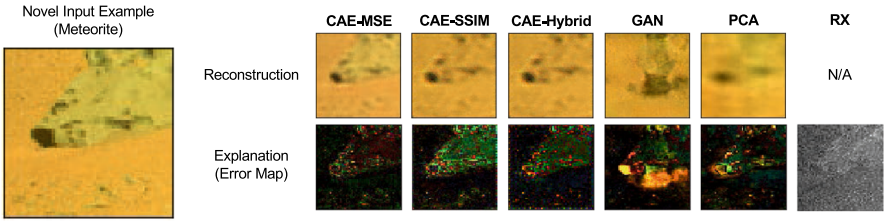

Fig. 9 Novel input example containing a partial meteorite with reconstructions and explanatory visualizations (error maps) for each novelty detection method

the meteorite pixels, but they appear less uniformly distributed over the meteorite area than in the CAE error maps, and the most error appears focused around the shadowed point of the meteorite. In the RX visualization, it is possible to make out the structure of the meteorite, but the error map is noisy and the meteorite does not clearly stand out from the background as in the other methods.

The GAN error map shows some overlap between high-error pixels and the meteorite, however there are large clusters of high-error pixels that do not coincide with the meteorite. The GAN reconstruction of the input image appears to show features similar to the Mastcam calibration target and a rock, instead of a lower-resolution version of the meteorite as in the other methods. During training, the encoder network of the BiGAN learns to map typical images from the training dataset to latent vectors $z \in \mathbb{R}^{100}$ based on activations in feature maps learned during training that enable the generator network to produce a realistic-looking image similar to those in the training dataset from z. This has the result that similar images in the training dataset will be nearby in the latent z-space, and clusters in the latent space should contain images with similar features (e.g., the calibration target or sand) (Donahue et al. 2017). Given a typical test image that shares characteristics with other typical images in the training dataset, we would expect the encoder to map the input image to a representation that is nearby similar images in the latent space. However, given a novel test image, we should not expect the encoder to map the input to a meaningful encoding z, since the feature extraction (convolutional) layers of the encoder were tuned to extract features common in the typical training images. Consequently, the generated image may not appear similar to the novel input image because it was conditioned on spurious activations in the encoder. We will explore this further in Sect. 7.

When analyzing multispectral images, scientists typically use a spectral analysis tool to inspect the spectrum in a pixel of group of pixels within the image. The spectrum is a plot of the wavelength on the x-axis and reflectance on the y-axis. Scientists compare the observed spectrum to known spectral patterns and characteristics for different materials to come up with interpretations for the observed data (Wellington et al. 2017a). While Fig. 9 shows the residual error between the input and its reconstruction as an image, we can also visualize the residual error between the input and its reconstruction for individual pixel values in each multispectral channel. Thus the residual for a single pixel across all channels represents the magnitude and direction of the novelty in each wavelength of the reflectance spectrum.

Figure 10 (left) shows the novel image and CAE-MSE error map containing an iron meteorite from Fig. 9 with the pixels having the highest error (most novelty) and lowest error (least novelty) indicated in purple and green respectively. On the right in Fig. 10, we plotted the spectrum of values across all bands in these two pixel locations from the input image compared to the reconstructed image. Using this visualization, scientists can quickly identify the direction and magnitude of novelty in each band, and combine this information with their domain expertise to make a geological interpretation about the novelty. In this example, comparing the novel pixel input and reconstructed spectrum shows lower reflectance in filter 3 (805 nm) and higher reflectance in filter 6 (1013 nm) than was expected (reconstructed) by the CAE-MSE model. The lower reflectance in the 805 nm band indicates that the rock is less red, or less dusty, which is consistent with the dark-toned appearance of iron meteorites. The

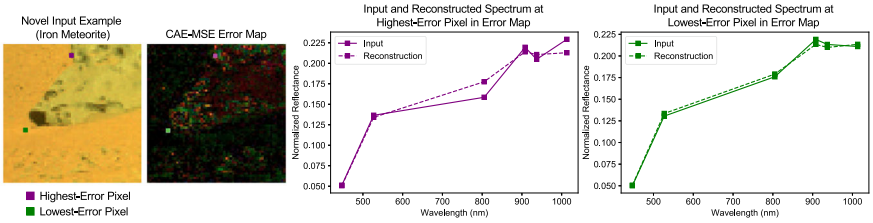

Fig. 10 Novel input image and error map for example containing a partial iron meteorite, with pixel locations with the highest (most novel) and lowest (least novel) reconstruction error indicated in purple and green (respectively). Plots compare the normalized input and reconstructed spectrum for the novel and typical pixel locations. The reflectance value in each pixel was normalized by dividing by the total reflectance across all values in the spectrum (Color figure online)

higher reflectance in the 1013 nm band results in a positive slope from filter 5 (937 nm) to filter 6, which is consistent with an increase in near-infrared reflectance values that is typical of iron meteorites relative to native Martian materials (Gaffey 1976; Wellington et al. 2017a). In contrast, comparing the input and reconstructed spectrum for the least novel (typical) pixel shows that the observed (input) and expected (reconstructed) spectrum show minimal differences. This type of explanation is enabled by a reconstruction-based approach since it requires a reconstructed signal to compare with the input signal, thus this explanation is not available for the RX detector (which is not a reconstruction-based approach).

7 Discussion

7.1 CAE loss functions

In Sect. 5.2, we found that the CAE-SSIM had better performance than all methods except RX (pixel) for detecting novel examples in the DRT spot, drill hole, and dump pile sub-classes, but worse performance in other sub-classes compared to those other methods. The only difference between the three CAE methods is the loss function used for training the model. In Fig. 11, we show an example novel image from the vein sub-class (left) and from the DRT spot sub-class (right). For each image, we show the reconstruction (output) from the CAE-MSE, SSIM loss, and hybrid ($\lambda = 0.01$) loss as well as the error map between the input and reconstructed images. The error maps show the features in the input image that were not reconstructed by each model, and thus were not optimized by the loss function. In the vein example (Fig. 11, left), the mineralogy of the light-toned vein is the novel feature, thus if a model detects the vein as novel we would expect to see colored pixels that align spatially with the vein in the input image. Both the CAE-MSE and CAE-Hybrid loss detect that the vein is novel. While the CAE-SSIM error map includes the vein, it also includes the entire rock, making it difficult to distinguish if the rock or the vein is being detected as novel. In the morphologically novel example (Fig. 11, right), the bright ellipse where dust was removed by the DRT is the novel feature, a spatial pattern that did not occur in typical images in the training dataset. When the CAE is trained using MSE loss, the

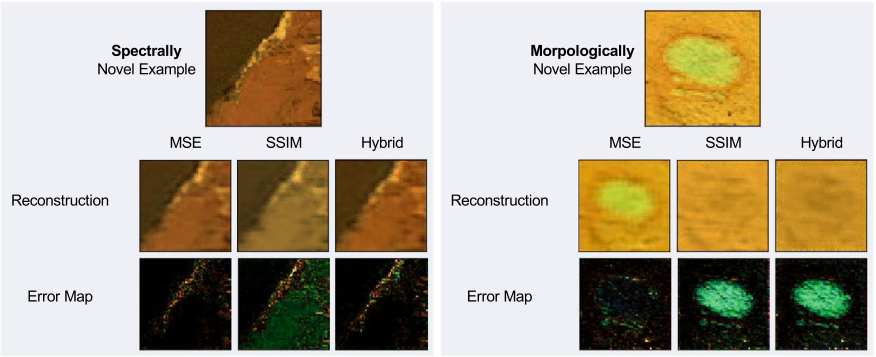

Fig. 11 Spectrally (left) and morphologically (right) novel examples from the novel test dataset with CAE reconstructions and error maps

model reconstructs the shape of the image well with some blurring at the borders of the ellipse; however, when the CAE is trained to optimize SSIM or a combination of SSIM and MSE, the CAE does not reconstruct the DRT spot at all.

This suggests that optimizing SSIM when training the CAE enables the preservation of spatial structure between the input and reconstruction during training at the expense of spectral information, causing the latent feature maps encode primarily spatial information. This causes morphological novelties like the DRT spot in Fig. 11 to be poorly (or not at all) reconstructed at test time. In contrast, MSE is a measure of the mean difference in pixel intensity between the input and reconstruction and does not measure any spatial relationships, thus encouraging the latent features encode primarily spectral information. This causes spectral novelties to be poorly reconstructed at test time. The hybrid loss aims to leverage the strengths of MSE and SSIM to detect both spectrally and morphologically novel features, but choosing the optimal setting for λ may depend on the specific application or dataset.

7.2 RX pixel representations

The results in Fig. 4 showed that, of the compared methods, only RX (pixel) and the CAEs trained with structural similarity (SSIM and Hybrid) had good performance for the dump pile, drill hole, and DRT spot categories. In the previous section, we discussed why this might be the case for CAEs with SSIM vs. MSE loss by differentiating between spectral and morphological novelties. Since the input representation for RX (pixel) is a single pixel spectrum without any spatial context, this does not explain the superior performance of RX in these categories. We observed that for all novel categories except meteorite, when RX (pixel) scores were high, scores for the PCA, GAN, and CAE-MSE methods were low, and vice versa. The input representation is the flattened multispectral image vector for PCA and the multispectral image tensor for the GAN and CAE-MSE methods. Both representations include spatial context. The good performance of RX (pixel) and poor performance of other methods in the dump pile, drill hole, and DRT spot categories suggests that spatial context does

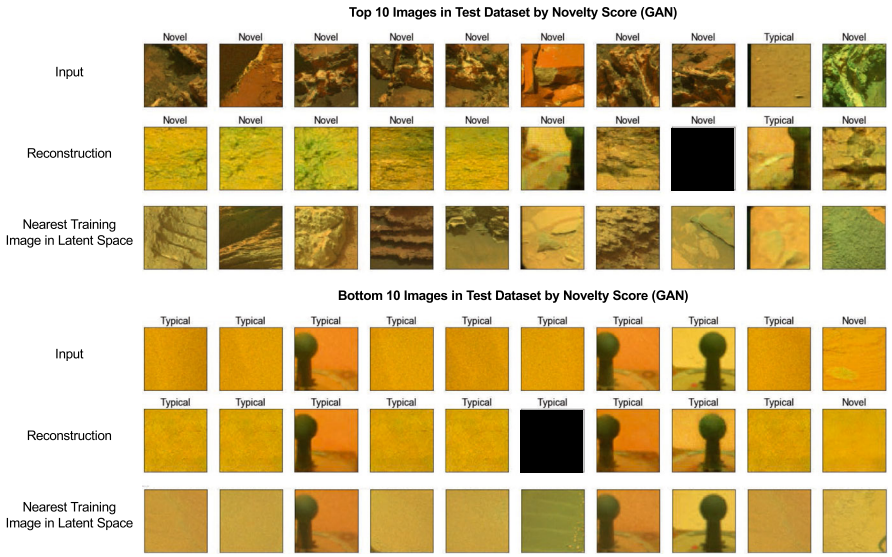

Fig. 12 Images from the test dataset with the highest (top) and lowest (bottom) novelty scores using the GAN method as well as the GAN reconstruction (generator output) for each input and the nearest image from the training dataset in the latent space. Labels above input images indicate the image label (novel or typical)

not help identify these novelties (except when the model is trained to encode spatial information as with SSIM). Conversely, the good performance of the PCA, GAN, and CAE-MSE methods and poor performance of RX (pixel) in the float, bedrock, vein, and broken rock categories suggests that spatial context *does* help identify novelties in these categories. This difference in performance could also be a result of way we computed the novelty score for the whole image based on the pixel RX scores for this method, i.e., by computing the average RX score across all pixels in the image. This could result in high novelty scores for images that either contain spectra that are strong outliers with respect to most other pixels in the image (and hence bias the average), or in which the novel feature spans a large fraction of the total pixels in the image (thus the novel pixel scores are not diluted by a large number of low non-novel pixel scores). This could also explain the good performance of RX (pixel) for the meteorite category, since meteorites exhibit distinct spectral signatures in the near-infrared (Gaffey 1976; Wellington et al. 2017a) and the meteorite covers a large portion of the frame in most images in our dataset (see Figs. 3 and 9, e.g.).

7.3 GAN reconstructions

We discussed in Sect. 6 that we should not expect the BiGAN encoder network to map novel images to latent vectors z that enable the generator network to reconstruct an image that appears similar to the novel input image, since it is likely conditioned on spurious activations in the encoder network since the encoder was trained with typical training images. To test this hypothesis, in Fig. 12 we show the 10 images with the

highest (top) and lowest (bottom) novelty scores in the test dataset using the GAN method—representing the "most novel" and "most typical" images in the test dataset according to the GAN model—as well as the reconstruction (generator output) and nearest image from the training dataset in the latent space (using Euclidean distance). Most of the typical images contain the calibration target or sand, which are frequently observed in the training dataset, and the reconstructed images are very similar to the input images. There is one novel image in the 10 lowest-scoring images containing a DRT spot. While the reconstructed image shows that the novel feature (the DRT spot) was not preserved (as would be expected since the training dataset did not contain any DRT spots), the difference in pixel intensities corresponding to the DRT spot may not have been large enough to result in a high novelty score. Most of the novel images in Fig. 12 (top) contain veins, except the sixth image which contains a broken rock, and the ninth image which contains a typical image likely misclassified as novel because of the black stripe (an image artifact) on the left side of the image. Images with this black stripe were filtered out of the training dataset, so this feature when seen in test images might (for good reason) be mistaken as novel. The reconstructions for the vein images are visually similar to each other, but bear little similarity to the input images. The nearest (typical) training images in the latent space to these novel inputs are also dissimilar, but appear to either have similar overall coloring or to contain linear features, which suggests these features might be extracted by the convolutional layers in the encoder network. In examples such as these, the explanatory visualizations we discussed in Sect. 9 may not be useful to a scientist who wishes to understand which features in the input image were considered novel by the detector. This is an important limitation of the GAN method for novelty detection, and GAN approaches that enable explanatory visualizations of detections could be a valuable topic for future work.

There is one image in in both the top 10 and bottom 10 images in Fig. 12 where the reconstructed images contain all zeros. Though these images appear nearly identical to the other images in the row (of veins for the novel examples and sand for the typical examples), the reconstruction is anomalous and the nearest training images are not similar to those of similar input images in each row. This suggests that there may also be some instability to the GAN approach for novelty detection.

8 Conclusions

There has been limited prior work exploring novelty detection methods for multispectral images and scientific data. In this work, we compared the performance of autoencoder, GAN, PCA, and RX approaches for prioritizing images with novel geologic features in multispectral images of the Martian surface acquired by the Mastcam imaging system in order to accelerate tactical planning for the Mars Science Laboratory (MSL) Curiosity rover. We found that the RX (pixel) method had the best overall performance as measured by ROC AUC score and Precision @ N, but may not provide the most effective explanatory visualizations for allowing users to understand the features in an image that were detected as novel. For the CAE methods, we showed that maximizing structural similarity (SSIM) during training enables accurate detection of morphologically novel features that are not detected by most other methods

(including the CAE trained with mean squared error). For images with spectral novelties, we found that the CAE with MSE loss and PCA had the best performance. The CAE methods were also shown to enable more effective explanatory visualizations for novel detections at the image and pixel spectrum level. Finally, we demonstrated that existing GAN approaches to novelty detection may be limited in their ability to enable explanatory visualizations of detections, which are critical for their practical use in novelty detection applications. In a future study, we plan to investigate the impact of the choice of autoencoder loss function on novelty detection performance for different types of novelties (e.g., spectral vs. morphological) as well as the interpretability of the residual image [e.g., as shown in Diaz and Hollmen (2002) for general regression neural networks].

Following this study of the comparison of novelty detection methods, we are actively developing an interface for these methods into the tactical planning process for MSL and Mars 2020 to assess their benefit in a practical setting. In addition, we are investigating the use of these and other novelty detection methods for identifying targets for follow-up analysis onboard the rovers, as a novelty-based variant of the Autonomous Exploration for Gathering Increased Science (AEGIS) autonomous targeting system (Francis et al. 2017). While training time is longer for the deep learning methods presented in this study (autoencoder and GAN) than for PCA and RX, all methods have similar inference times for test examples using a GPU (a Tesla M60 GPU was used for this study), thus are equally suitable for ground-based analysis. However, the size and complexity of the deep learning models would pose a problem for their implementation on the rovers' onboard computers, thus we will evaluate PCA, RX, and other methods that are more computationally efficient. Finally, we plan to extend this work to explore novelty detection systems for orbital remote sensing images of Mars, Earth, and other planets to prioritize images that contain rare surface features or atmospheric phenomena.

Acknowledgements This work was funded in part by NASA STTR #80NSSC17C0035 and NASA/JPL funding from the Mars Science Laboratory Mastcam instrument investigation. It was carried out (in part) at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration and funded through the Internal Strategic University Research Partnerships (SURP) program.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado G, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Kaiser L, Kudlur M, Levenberg J, Man D, Monga R, Moore S, Murray D, Shlens J, Steiner B, Sutskever I, Tucker P, Vanhoucke V, Vasudevan

V, Vinyals O, Warden P, Wicke M, Yu Y, Zheng X (2015) TensorFlow: Large-scale machine learning on heterogeneous distributed systems. https://doi.org/10.1038/nn.3331

Abe N, Zadrozny B, Langford J (2006) Outlier detection by sampling with accuracy guarantees. In: ACM SIGKDD international conference on knowledge discovery and data mining (KDD). pp 767–772. https://doi.org/10.1145/1150402.1150501

Agyemang M, Barker K, Alhajj R (2006) A comprehensive survey of numeric and symbolic outlier mining techniques. Intell Data Anal 10(6):521–538. https://doi.org/10.3233/IDA-2006-10604

Akcay S, Atapour-Abarghouei A, Breckon T (2018) GANomaly: Semi-supervised anomaly detection via adversarial training. In: Asian conference on computer vision (ACCV). pp 622–637

Alain G, Bengio Y (2014) What regularized auto-encoders learn from the data generating distribution. J Mach Learn Res 15:3743–3773

An J, Cho S (2015) Variational autoencoder based anomaly detection using reconstruction probability. Tech. rep., SNU Data Mining Center

Angiulli F, Pizzuti C (2002) Fast outlier detection in high dimensional spaces. In: European conference on principles of data mining and knowledge discovery (PKDD). pp 15–27. https://doi.org/10.1007/3-540-45681-3_2

Antipov G, Baccouche M, Dugelay JL (2017) Face aging with conditional generative adversarial networks. In: IEEE international conference on image processing. pp 2089–2093. https://doi.org/10.1109/ICIP.2017.8296650

Ayhan B, Dao M, Kwan C, Chen HM, Bell JF, Kidd R (2017) A novel utilization of image registration techniques to process Mastcam images in Mars rover with applications to image fusion, pixel clustering, and anomaly detection. IEEE J Sel Top Appl Earth Observ Remote Sens 10(10):4553–4564. https://doi.org/10.1109/JSTARS.2017.2716923

Bay SD, Schwabacher M (2003) Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: ACM SIGKDD international conference on knowledge discovery and data mining (KDD). pp 29–38. https://doi.org/10.1145/956750.956758

Bell JF, Godber A, McNair S, Caplinger MA, Maki JN, Lemmon MT, Van Beek J, Malin MC, Wellington D, Kinch KM, Madsen MB, Hardgrove C, Ravine MA, Jensen E, Harker D, Anderson RB, Herkenhoff KE, Morris RV, Cisneros E, Deen RG (2017) The Mars Science Laboratory Curiosity rover Mastcam instruments: preflight and in-flight calibration, validation, and data archiving. Earth Space Sci 4(7):396–452. https://doi.org/10.1002/2016EA000219

Bell III J, Calvin W, Farrand W, Greeley R, Johnson J, Joliff B, Morris R, Sullivan R, Thompson S, Wang A, Weitz C, Squyres S (2008) Mars Exploration Rover Pancam multispectral imaging of rocks, soils, and dust in Gusev Crater and Meridiani Planum. In: Bell III J (ed) The martian surface: composition, mineralogy, and physical properties, chap 13. pp 281–314

Bell III JF, Maki JN, Mehall GL, Ravine MA, Caplinger MA (2016) Mastcam-Z: Designing a geologic, stereoscopic, and multispectral pair of zoom cameras for the NASA Mars 2020 rover. In: 3rd International workshop on instrumentation for planetary missions, vol 1980

Bhattacharjee S, Mandal D, Biswas S (2019) Autoencoder based novelty detection for generalized zero shot learning. In: 2019 IEEE international conference on image processing (ICIP). pp 3646–3650. https://doi.org/10.1109/ICIP.2019.8803562

Blake D, Vaniman D, Achilles C, Anderson R, Bish D, Bristow T, Chen C, Chipera S, Crisp J, Des Marais D, Downs RT, Farmer J, Feldman S, Fonda M, Gailhanou M, Ma H, Ming DW, Morris RV, Sarrazin P, Stolper E, Treiman A, Yen A (2012) Characterization and calibration of the CheMin mineralogical instrument on Mars Science Laboratory. Space Sci Rev 170(1–4):341–399. https://doi.org/10.1007/s11214-012-9905-1

Breunig MM, Kriegel HP, Ng RT, Sander J (2000) LOF: Identifying density-based local outliers. In: ACM SIGMOD international conference on management of data. pp 93–104. https://doi.org/10.1145/342009.335388

Budalakoti S, Srivastava AN, Akella R, Turkov E (2006) Anomaly detection in large sets of high-dimensional symbol sequences. Tech. Rep. TM-2006-214553, NASA Ames Research Center

Campbell C, Bennett KP (2001) A linear programming approach to novelty detection. In: Advances in neural information processing systems (NIPS). pp 395–401

Campos GO, Zimek A, Sander J, Campello RJGB, Micenková B, Schubert E, Assent I, Houle ME (2016) On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Min Knowl Discov 30(4):891–927. https://doi.org/10.1007/s10618-015-0444-8

Carrera D, Boracchi G, Foi A, Wohlberg B (2015) Detecting anomalous structures by convolutional sparse models. In: International joint conference on neural networks (IJCNN). pp 1–8. https://doi.org/10.1109/IJCNN.2015.7280790

Chandola V, Banerjee A, Kumar V (2009) Anomaly detection. ACM Comput Surv 41(3):1–58. https://doi.org/10.1145/1541880.1541882

Chang CI, Chiang SS (2002) Anomaly detection and classification for hyperspectral imagery. IEEE Trans Geosci Remote Sens 40(6):1314–1325. https://doi.org/10.1109/TGRS.2002.800280

Chiu AL, Fu AWc (2003) Enhancements on local outlier detection. In: International database engineering and applications symposium. pp 298–307. https://doi.org/10.1109/IDEAS.2003.1214939

Clifton DA, Bannister PR, Tarassenko L (2007) A framework for novelty detection in jet engine vibration data. Key Eng Mater 347:305–310. https://doi.org/10.4028/www.scientific.net/KEM.347.305

Creswell A, Arulkumaran K, Bharath AA (2017) On denoising autoencoders trained to minimise binary cross-entropy. arXiv preprint arXiv:1708.08487

Desforges MJ, Jacob PJ, Cooper JE (1998) Applications of probability density estimation to the detection of abnormal conditions in engineering. J Mech Eng Sci 212(8):687–703. https://doi.org/10.1243/0954406981521448

Diaz I, Hollmen J (2002) Residual generation and visualization for understanding novel process conditions. In: International joint conference on neural networks. IJCNN'02 (Cat. No.02CH37290). pp 2070–2075. https://doi.org/10.1109/IJCNN.2002.1007460

Donahue J, Krahenbuhl P, Darrell T (2017) Adversarial feature learning. In: International conference on learning representations (ICLR). pp 1–18

Dong HW, Hsiao WY, Yang LC, Yang YH (2018) MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In: AAAI conference on artificial intelligence. pp 34–41

Erfani SM, Rajasegarar S, Karunasekera S, Leckie C (2016) High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. Pattern Recogn 58:121–134. https://doi.org/10.1016/J.PATCOG.2016.03.028

Ertöz L, Steinbach M, Kumar V (2003) Finding topics in collections of documents: a shared nearest neighbor approach. Clust Inf Retr 11:83–103. https://doi.org/10.1007/978-1-4613-0227-8_3

Filippone M, Masulli F, Rovetta S (2010) Applying the possibilistic c-means algorithm in kernel-induced spaces. IEEE Trans Fuzzy Syst 18(3):572–584. https://doi.org/10.1109/TFUZZ.2010.2043440

Francis R, Estlin T, Doran G, Johnstone S, Gaines D, Verma V, Burl M, Frydenvang J, Montaño S, Wiens R et al (2017) Aegis autonomous targeting for chemcam on mars science laboratory: deployment and results of initial science team use. Sci Robot 2(7):eaan4582. https://doi.org/10.1126/scirobotics.aan4582

Gaffey MJ (1976) Spectral reflectance characteristics of the meteorite classes. J Geophys Res 81(5):905–920. https://doi.org/10.1029/JB081i005p00905

Ghoting A, Parthasarathy S, Otey ME (2008) Fast mining of distance-based outliers in high-dimensional datasets. Data Min Knowl Discov 16(3):349–364. https://doi.org/10.1007/s10618-008-0093-2

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems (NIPS). pp 2672–2680

Grotzinger JP, Crisp J, Vasavada AR, Anderson RC, Baker CJ, Barry R, Blake DF, Conrad P, Edgett KS, Ferdowski B, Gellert R, Gilbert JB, Golombek M, Gómez-Elvira J, Hassler DM, Jandura L, Litvak M, Mahaffy P, Maki J, Meyer M, Malin MC, Mitrofanov I, Simmonds JJ, Vaniman D, Welch RV, Wiens RC (2012) Mars Science Laboratory mission and science investigation. Space Sci Rev 170(1–4):5–56. https://doi.org/10.1007/s11214-012-9892-2

He Z, Xu X, Deng S (2003) Discovering cluster-based local outliers. Pattern Recogn Lett 24(9–10):1641–1650. https://doi.org/10.1016/S0167-8655(03)00003-5

Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507. https://doi.org/10.1126/science.1127647

Hjelm RD, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A, Bengio Y (2019) Learning deep representations by mutual information estimation and maximization. In: International conference for learning representations (ICLR)

Hodge VJ, Austin J (2004) A survey of outlier detection methodologies. Artif Intell Rev 22(2):85–126. https://doi.org/10.1007/s10462-004-4304-y

Jablonski JA, Bihl TJ, Bauer KW (2015) Principal component reconstruction error for hyperspectral anomaly detection. IEEE Geosci Remote Sens Lett 12(8):1725–1729. https://doi.org/10.1109/LGRS.2015.2421813

Japkowicz N, Myers C, Gluck M (1995) A novelty detection approach to classification. In: International joint conference on artificial intelligence (IJCAI), vol 1. pp 518–523

Johnson JR, Bell JFI, Gasnault O, Le Mouelic S, Rapin W, Bridges J, Wellington DF (2014) First iron meteorites observed by the Mars Science Laboratory (MSL) rover Curiosity. In: American geophysical union (AGU) fall meeting

Johnson JR, Bell J III, Bender S, Blaney D, Cloutis E, DeFlores L, Ehlmann B, Gasnault O, Gondet B, Kinch K, Lemmon M, Le Mouélic S, Maurice S, Rice M, Wiens R (2015) ChemCam passive reflectance spectroscopy of surface materials at the Curiosity landing site, Mars. Icarus 249:74–92. https://doi.org/10.1016/J.ICARUS.2014.02.028

Kerner HR, Wellington DF, Wagstaff KL, Bell JF, Ben Amor H (2019) Novelty detection for multispectral images with application to planetary exploration. In: AAAI conference on artificial intelligence. pp 9484–9491. https://doi.org/10.1609/aaai.v33i01.33019484

Kim D, Kang P, Cho S, Hj Lee, Doh S (2012) Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. Expert Syst Appl 39(4):4075–4083. https://doi.org/10.1016/J.ESWA.2011.09.088

Krzanowski WJ, Hand DJ (2009) ROC curves for continuous data. CRC Press, Boca Raton

Kwak N (2008) Principal component analysis based on l1-norm maximization. IEEE Trans Pattern Anal Mach Intell 9:1672–1680. https://doi.org/10.1109/TPAMI.2008.114

Kwon H, Nasrabadi NM (2005) Kernel RX-algorithm: a nonlinear anomaly detector for hyperspectral imagery. IEEE Trans Geosci Remote Sens 43(2):388–397. https://doi.org/10.1109/TGRS.2004.841487

Latecki LJ, Lazarevic A, Pokrajac D (2007) Outlier detection with kernel density functions. Mach Learn Data Min Pattern Recogn. https://doi.org/10.1007/978-3-540-73499-4_6

Laxhammar R, Falkman G, Sviestins E (2009) Anomaly detection in sea traffic—a comparison of the Gaussian Mixture Model and the Kernel Density Estimator. In: International conference on information fusion

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539

Lee K, Lee K, Lee H, Shin J (2018) A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Advances in neural information processing systems. pp 7167–7177

Léveillé RJ, Bridges J, Wiens RC, Mangold N, Cousin A, Lanza N, Forni O, Ollila A, Grotzinger J, Clegg S, Siebach K, Berger G, Clark B, Fabre C, Anderson R, Gasnault O, Blaney D, Deflores L, Leshin L, Maurice S, Newsom H (2014) Chemistry of fracture-filling raised ridges in Yellowknife Bay, Gale Crater: window into past aqueous activity and habitability on Mars. J Geophys Res Planets 119(11):2398–2415. https://doi.org/10.1002/2014JE004620

Liu FT, Ting KM, Zhou ZH (2008) Isolation forest. In: IEEE international conference on data mining (ICDM). pp 413–422. https://doi.org/10.1109/ICDM.2008.17

Ma J, Perkins S (2003) Time-series novelty detection using one-class support vector machines. In: International joint conference on neural networks, vol 3. pp 1741–1745. https://doi.org/10.1109/IJCNN.2003.1223670

Mahaffy P, Webster CR, Atreya SK, Franz H, Wong M, Conrad PG, Harpold D, Jones JJ, Leshin LA, Manning H, Owen T, Pepin RO, Squyres S, Trainer M, Science Team T MSL (2013) Abundance and isotopic composition of gases in the Martian atmosphere from the Curiosity rover. Science 341(6143):263–266. https://doi.org/10.1126/science.194.4271.1298

Malin MC, Ravine MA, Caplinger MA, Tony Ghaemi F, Schaffner JA, Maki JN, Bell JF, Cameron JF, Dietrich WE, Edgett KS, Edwards LJ, Garvin JB, Hallet B, Herkenhoff KE, Heydari E, Kah LC, Lemmon MT, Minitti ME, Olson TS, Parker TJ, Rowland SK, Schieber J, Sletten R, Sullivan RJ, Sumner DY, Yingst AR, Duston BM, McNair S, Jensen EH (2017) The Mars Science Laboratory (MSL) Mast cameras and Descent imager: investigation and instrument descriptions. Earth Space Sci 4(8):506–539. https://doi.org/10.1002/2016EA000252

Manevitz L, Yousef M (2007) One-class document classification via neural networks. Neurocomputing 70(7–9):1466–1481. https://doi.org/10.1016/J.NEUCOM.2006.05.013

Manevitz LM, Yousef M (2001) One-class SVMs for document classification. J Mach Learn Res 2:139–154

Markou M, Singh S (2003a) Novelty detection: a review—part 1: statistical approaches. Sig Process 83(12):2481–2497. https://doi.org/10.1016/J.SIGPRO.2003.07.018
Markou M, Singh S (2003b) Novelty detection: a review—part 2: neural network based approaches. Sig Process 83(12):2499–2521. https://doi.org/10.1016/J.SIGPRO.2003.07.019
Marsland S (2003) Novelty detection in learning systems. Neural Comput Surv 3(2):157–195
Masci J, Meier U, Cireşan D, Schmidhuber J (2011) Stacked convolutional auto-encoders for hierarchical feature extraction. In: International conference on artificial neural networks (ICANN): artificial networks and machine learning. pp 52–59. https://doi.org/10.1007/978-3-642-21735-7_7
Modenesi AP, Braga AP (2009) Analysis of time series novelty detection strategies for synthetic and real data. Neural Process Lett 30(1):1–17. https://doi.org/10.1007/s11063-009-9106-4
Molero JM, Garzon EM, Garcia I, Plaza A (2013) Analysis and optimizations of global and local versions of the RX algorithm for anomaly detection in hyperspectral data. IEEE J Sel Top Appl Earth Observ Remote Sens 6(2):801–814. https://doi.org/10.1109/JSTARS.2013.2238609
Munoz-Mari J, Bovolo F, Gomez-Chova L, Bruzzone L, Camp-Valls G (2010) Semisupervised one-class support vector machines for classification of remote sensing data. IEEE Trans Geosci Remote Sens 48(8):3188–3197. https://doi.org/10.1109/TGRS.2010.2045764
Papadimitriou S, Kitagawa H, Gibbons P, Faloutsos C (2003) LOCI: fast outlier detection using the local correlation integral. In: International conference on data engineering. pp 315–326. https://doi.org/10.1109/ICDE.2003.1260802
Park D, Hoshi Y, Kemp CC (2018) A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder. IEEE Robot Autom Lett 3(3):1544–1551. https://doi.org/10.1109/LRA.2018.2801475
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
Pimentel MA, Clifton DA, Clifton L, Tarassenko L (2014) A review of novelty detection. Sig Process 99:215–249. https://doi.org/10.1016/J.SIGPRO.2013.12.026
Pires A, Santos-Pereira C (2005) Using clustering and robust estimators to detect outliers in multivariate data. In: International conference on robust statistics. https://doi.org/10.1007/978-3-642-57489-4_41
Reed I, Yu X (1990) Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. IEEE Trans Acoust Speech Signal Process 38(10):1760–1770. https://doi.org/10.1109/29.60107
Ren D, Wang B, Perrizo W (2004) RDF: A density-based outlier detection method using vertical data representation. In: IEEE international conference on data mining (ICDM). pp 503–506. https://doi.org/10.1109/ICDM.2004.10010
Rice MS, Bell III JF, Godber A, Wellington DF, Fraeman AA, Johnson JR, Kinch KM, Malin MC, Grotzinger JP, the MSL Science Team (2013) Mastcam multispectral imaging results from the Mars Science Laboratory investigation in Yellowknife Bay. In: European planetary science congress (EPSC), vol 8
Richter C, Roy N (2017) Safe visual navigation via deep learning and novelty detection. In: Robotics: science and systems (RSS). https://doi.org/10.15607/RSS.2017.XIII.064
Ristic B, La Scala B, Morelande M, Gordon N (2008) Statistical analysis of motion patterns in AIS data: anomaly detection and motion prediction. In: International conference on information fusion. pp 1–7. https://doi.org/10.1109/ICIF.2008.4632190
Rosset S (2004) Model selection via the AUC. In: International conference on machine learning (ICML). https://doi.org/10.1145/1015330.1015400
Schlegl T, Seebock P, Waldstein SM, Schmidt-Erfurth U, Langs G (2017) Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International conference on information processing in medical imaging (IPMI). pp 146–157. https://doi.org/10.1007/978-3-319-59050-9_12
Scholkopf B, Williamson R, Smola A, Shawe-Taylor J, Platt J, Holloway R (2000) Support vector method for novelty detection. In: Advances in neural information processing systems (NIPS). pp 582–588
Schubert E, Zimek A, Kriegel HP (2014) Generalized outlier detection with flexible kernel density estimates. In: SIAM international conference on data mining (SDM). pp 542–550. https://doi.org/10.1137/1.9781611973440.63
Silverman BW (1986) Density estimation for statistics and data analysis. Monogr Stat Appl Probab. https://doi.org/10.1002/bimj.4710300745

Srivastava A, Zane-Ulman B (2005) Discovering recurring anomalies in text reports regarding complex space systems. In: IEEE aerospace conference. pp 3853–3862. https://doi.org/10.1109/AERO.2005.1559692

Srivastava AN (2006) Enabling the discovery of recurring anomalies in aerospace problem reports using high-dimensional clustering techniques. In: IEEE aerospace conference. https://doi.org/10.1109/AERO.2006.1656136

Syed Z, Saeed M, Rubinfeld I (2010) Identifying high-risk patients without labeled training data: anomaly detection methodologies to predict adverse outcomes. In: AMIA annual symposium. pp 772–776

Tang J, Chen Z, Fu AWc, Cheung DW (2002) Enhancing effectiveness of outlier detections for low density patterns. In: Pacific-Asia conference on knowledge discovery and data mining (PAKDD). pp 535–548. https://doi.org/10.1007/3-540-47887-6_53

Tang J, Chen Z, Fu AW, Cheung DW (2007) Capabilities of outlier detection schemes in large datasets, framework and methodologies. Knowl Inf Syst 11(1):45–84. https://doi.org/10.1007/s10115-005-0233-6

Tax DM, Duin RP (1999) Support vector domain description. Pattern Recogn Lett 20(11–13):1191–1199. https://doi.org/10.1016/S0167-8655(99)00087-2

Thompson B, Marks R, Choi J, El-Sharkawi M, Huang MY, Bunje C (2002) Implicit learning in autoencoder novelty assessment. In: International joint conference on neural networks, vol 3. pp 2878–2883. https://doi.org/10.1109/IJCNN.2002.1007605

Tipping ME, Bishop C (1999) Probabilistic principal component analysis. J R Stat Soc Ser B (Stat Methodol) 61(3):611–622. https://doi.org/10.1111/1467-9868.00196

Toivola J, Prada MA, Hollmén J (2010) Novelty detection in projected spaces for structural health monitoring. In: International symposium on intelligent data analysis (IDA). pp 208–219. https://doi.org/10.1007/978-3-642-13062-5_20

Wagstaff KL, Lanza NL, Thompson DR, Dietterich TG, Gilmore MS (2013) Guiding scientific discovery with explanations using DEMUD. In: AAAI conference on artificial intelligence. pp 905–911

Wagstaff KL, Doran G, Davies A, Anwar S, Chakraborty S, Cameron M, Daubar IJ, Phillips C (2019) Enabling onboard detection of events of scientific interest for the Europa Clipper spacecraft. In: ACM SIGKDD international conference on knowledge discovery and data mining (KDD). pp 2191–2201. https://doi.org/10.1145/3292500.3330656

Wang CH (2009) Outlier identification and market segmentation using kernel-based clustering techniques. Expert Syst Appl 36(2):3744–3750. https://doi.org/10.1016/J.ESWA.2008.02.037

Wang Y, Wong J, Miner A (2004a) Anomaly intrusion detection using one class SVM. In: IEEE SMC information assurance workshop. pp 358–364. https://doi.org/10.1109/IAW.2004.1437839

Wang Z, Bovik A (2009) Mean squared error: Love it or leave it? A new look at signal fidelity measures. IEEE Signal Process Mag 26(1):98–117. https://doi.org/10.1109/MSP.2008.930649

Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004b) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612. https://doi.org/10.1109/TIP.2003.819861

Wellington DF (2018) Mars in the visible to near infrared: two views of the red planet. Ph.D. thesis, Arizona State University

Wellington DF, Bell JF III, Johnson JR, Kinch KM, Rice MS, Godber A, Ehlmann BL, Fraeman AA, Hardgrove C (2017a) Visible to near-infrared MSL/Mastcam multispectral imaging: Initial results from select high-interest science targets within Gale Crater, Mars. Am Mineral 102(6):1202–1217. https://doi.org/10.2138/am-2017-5760CCBY

Wellington DF, Bell III JF, Johnson JR, Rice MS, Fraeman AA, Horgan B (2017b) VIS/NIR spectral differences of materials within Gale Crater, Mars: parameterization of MSL/Mastcam multispectral observations. In: 48th Lunar and planetary science conference (LPSC)

Williams G, Baxter R, He H, Hawkins S, Lifang Gu (2002) A comparative study of RNN for outlier detection in data mining. In: IEEE international conference on data mining (ICDM). pp 709–712. https://doi.org/10.1109/ICDM.2002.1184035

Wilson M, Trosper J, Abilleira F (2017) NASA Mars 2020 Landed Mission Development. Tech. rep., Jet Propulsion Laboratory, California Institute of Technology

Xiao Y, Wang H, Xu W, Zhou J (2013) L1 norm based KPCA for novelty detection. Pattern Recogn 46(1):389–396. https://doi.org/10.1016/J.PATCOG.2012.06.017

Xiong Y, Zuo R (2016) Recognition of geochemical anomalies using a deep autoencoder network. Comput Geosci 86:75–82. https://doi.org/10.1016/J.CAGEO.2015.10.006

Yu D, Sheikholeslami G, Zhang A (2002) FindOut: Finding outliers in very large datasets. Knowl Inf Syst 4(4):387–412. https://doi.org/10.1007/s101150200013

Yu JX, Qian W, Lu H, Zhou A (2006) Finding centric local outliers in categorical/numerical spaces. Knowl Inf Syst 9(3):309–338. https://doi.org/10.1007/s10115-005-0197-6

Zenati H, Foo CS, Lecouat B, Manek G, Ramaseshan Chandrasekhar V (2018a) Efficient GAN-based anomaly detection. In: International conference on learning representations (ICLR)

Zenati H, Romain M, Foo CS, Lecouat B, Chandrasekhar VR (2018b) Adversarially learned anomaly detection. In: IEEE international conference on data mining (ICDM). pp 727–736. https://doi.org/10.1109/ICDM.2018.00088

Zhang J, Wang H (2006) Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance. Knowl Inf Syst 10(3):333–355. https://doi.org/10.1007/s10115-006-0020-z

Zhou C, Paaenroth RC (2017) Anomaly detection with robust deep autoencoders. In: ACM SIGKDD international conference on knowledge discovery and data mining (KDD). https://doi.org/10.1145/3097983.3098052

Zhou J, Kwan C, Ayhan B, Eismann MT (2016) A novel cluster Kernel RX algorithm for anomaly and change detection using hyperspectral images. IEEE Trans Geosci Remote Sens 54(11):6497–6504. https://doi.org/10.1109/TGRS.2016.2585495