JGR Space Physics

RESEARCH ARTICLE
10.1029/2020JA027808

Auroral Image Classification With Deep Neural Networks

Andreas Kvammen[1], Kristoffer Wickstrøm[1], Derek McKay[2,3], and Noora Partamies[4,5]

[1]Department of Physics and Technology, UiT-The Arctic University of Norway, Tromsø, Norway, [2]NORCE Norwegian Research Centre AS, Tromsø, Norway, [3]Finnish Centre for Astronomy with ESO, FINCA, University of Turku, Turku, Finland, [4]Department of Arctic Geophysics, The University Centre in Svalbard, Longyearbyen, Norway, [5]Birkeland Centre for Space Science, Bergen, Norway

Key Points:
- Auroral images were labeled into seven classes and used to train and test machine learning classifiers
- The deep neural networks outperformed the k-nearest neighbor and support vector machine classifiers
- The ResNet-50 deep neural network had the highest performance and achieved 92% precision

Correspondence to:
A. Kvammen,
andreas.kvammen@uit.no

Citation:
Kvammen, A., Wickstrøm, K., McKay, D., & Partamies, N. (2020). Auroral image classification with deep neural networks. *Journal of Geophysical Research: Space Physics*, *125*, e2020JA027808. https://doi.org/10.1029/2020JA027808

Received 16 JAN 2020
Accepted 23 SEP 2020
Accepted article online 5 OCT 2020

Abstract Results from a study of automatic aurora classification using machine learning techniques are presented. The aurora is the manifestation of physical phenomena in the ionosphere-magnetosphere environment. Automatic classification of *millions* of auroral images from the Arctic and Antarctic is therefore an attractive tool for developing auroral statistics and for supporting scientists to study auroral images in an objective, organized, and repeatable manner. Although previous studies have presented tools for detecting aurora, there has been a lack of tools for classifying aurora into subclasses with a high precision (>90%). This work considers seven auroral subclasses: *breakup*, *colored*, *arcs*, *discrete*, *patchy*, *edge*, and *faint*. Six different deep neural network architectures have been tested along with the well-known classification algorithms: k-nearest neighbor (KNN) and a support vector machine (SVM). A set of clean nighttime color auroral images, without clearly ambiguous auroral forms, moonlight, twilight, clouds, and so forth, were used for training and testing the classifiers. The deep neural networks generally outperformed the KNN and SVM methods, and the ResNet-50 architecture achieved the highest performance with an average classification precision of 92%.

1. Introduction

Spectacular auroral displays can be seen on the night sky at high latitudes if the solar wind, magnetospheric, and ionospheric conditions are opportune. The auroral excitation processes are activated by energetic electrons and protons from the magnetosphere. The charged particles follow the magnetic field from the plasma sheet down to their magnetic footprint in the ionosphere where the energy is dissipated by ionization, excitation, and heating of thermospheric particles. The auroral displays, with typical emission intensity peaks at altitudes between 100 and 130 km, are therefore an indicator of dynamical processes that occur much further out into the Earth's magnetotail. Understanding how different ionospheric and magnetospheric conditions are manifested in the shape, color, intensity, and time evolution of the aurora is not well understood, even over 100 years after the first big auroral imaging campaigns in Bossekop, Norway (Störmer, 1913).

Since the first auroral imaging campaigns, millions of auroral images have been taken in the Arctic and Antarctic regions, and auroral scientists now have access to more data than what is possible to search and analyze by visual inspections. During recent decades, it has been demonstrated that machine learning methods are a valuable and highly applicable tool for automatically classifying large image data sets, for instance, by letter, brain tumor, and facial recognition. Machine learning techniques are, however, not widely used within the auroral research community. Automatic classification of millions of images captured every year will make it easier for scientists to study the images that are of interest in an organized, objective, and repeatable manner. This will further make statistical studies easier to conduct. For example, probability distributions of different auroral structures will facilitate studies of the temporal evolution of the aurora under different geomagnetic condition. In addition, statistical studies of the ionosphere-magnetosphere environment can be conducted by investigating the occurrence of auroral classes which are related to physical processes in the magnetosphere, such as the occurrence and evolution of *patchy aurora* that can be statistically studied in relation to the pitch angle scattering due to wave-particle interactions in the ionosphere-magnetosphere system.

Machine classification of aurora is a difficult task due to the transparent nature of the emission and thus the soft boundaries of the observed forms. Manual classification of auroral images is also a challenging task

since no two auroral events are alike and the aurora is a very dynamic phenomena which changes rapidly and manifests differently depending on geomagnetic conditions. An additional complication is that there is no clear consensus about how many nighttime auroral classes exist and what they are. Nevertheless, some automatic classification and feature extraction techniques have been developed and tested over the last two decades, for instance, using k-nearest neighbor (KNN) (Syrjäsuo & Donovan, 2002, 2004), Fourier descriptors (Syrjäsuo et al., 2007), support vector machines (SVMs) (Rao et al., 2014), and polynomial fitting (Partamies et al., 2015). However, none of the developed tools have achieved broader usage, possibly due to the techniques not being general enough, low classification accuracy, or few and/or unwanted auroral labels. It should also be noted that machine learning techniques have been extensively used for dayside aurora classification in the cusp region using hidden Marcov models (Yang et al., 2012) and to find key features detected by the cycle-consistent generative adversarial network (CycleGAN) (Yang et al., 2019). However, the developed dayside tools cannot be used on nightside aurora images without additional training due to the differences in auroral morphology. In a recent study, Clausen and Nickisch (2018) presented automatic nighttime auroral classification results by employing a pretrained deep neural network (DNN) and 5,824 labeled images. Clausen and Nickisch (2018) achieved $81.7 \pm 0.1\%$ classification accuracy, distributed into six labels: *arc*, *discrete*, *diffuse*, *cloudy*, *moon*, and *clear/no aurora*. Furthermore, when clustering together the auroral classes (*arc*, *discrete*, and *diffuse*) and the nonauroral classes (*cloudy*, *moon*, and *clear/no aurora*), an accuracy of $95.60 \pm 0.03\%$ was achieved for detecting *aurora* versus *no aurora* conditions.

Following the previous work, the main purpose of this study was to classify color (RGB) nighttime auroral images (manually preselected to contain clear skies) with higher precision and into more labels than those in the previous studies. Classification error rates of <10% are considered sufficient for operational purposes (Syrjasuo & Partamies, 2011); this study therefore aimed at an average classification precision of >90%. An additional objective was to define auroral labels which represent an exclusive production mechanism or characterize a physical property of the aurora when possible. Finally, several neural networks and machine learning techniques were tested and compared in this study. The pretrained DNN considered in Clausen and Nickisch (2018) is among the evaluated classifiers. In total, 14,030 auroral images were labeled. Out of these, 3,854 auroral images did not contain clearly ambiguous auroral forms and were therefore selected for network training and testing. The methodology is described in section 2. A comparison of the performance of the different algorithms and an analysis of misidentified images is presented in section 3, with a more general discussion in section 4.

2. Methodology

A description of the data processing, the auroral labels, and the applied machine learning techniques are presented in this section. A detailed description of auroral classification ergonomics and potential biases in the data set is presented in McKay and Kvammen (2020).

2.1. Data Acquisition and Preprocessing

The images used in this study were acquired by the all-sky camera located near Kiruna, Sweden, at 425 m above mean sea level with geographic (latitude and longitude) location: 67.84°N, 20.42°E and CGM location: 64.69°, 102.64°, and operated by the Swedish Institute for Space Physics. The camera is a Nikon D700 equipped with a Nikon Nikkor 8 mm 1:2.8 lens giving almost 180° field of view. The color sensitivity of the detector, from Mauer (2009), is depicted in Figure 1 along with the characteristic 427.8, 557.7, and 630.0 nm auroral lines. The camera was installed in 2009 with the available data set used in this project extending from 2010 to 2019. The exposure time is 6 s, with images taken automatically on each minute. To ease data transfer rates and processing, JPEG images (720×479 pixels) were used, rather than the full-resolution images.

Images containing clouds and moon where excluded, since earlier studies by Rao et al. (2014) and Clausen and Nickisch (2018) showed high accuracy for labeling clouds and moon while being less successful in labeling auroral subclasses. A preprocessing stage was carried out where keograms were manually inspected and areas of potential auroral emission were selected, thus rejecting the bulk of overcast sky conditions. Celestial positions were calculated using the Skyfield software (Rhodes, 2019) and images where the Sun was at an apparent elevation greater than −15°, or where the moon was above the horizon, were automatically rejected.

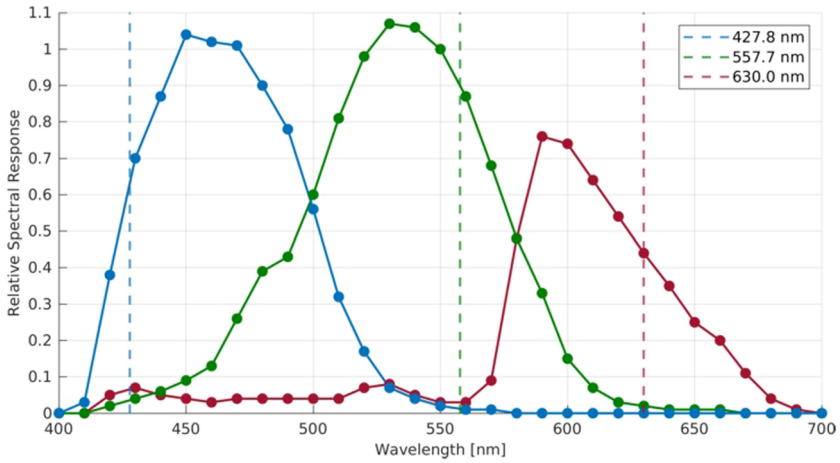

Figure 1. The figure shows the color sensitivity of the blue, green, and red channels of the detector. It is clear that the camera sensitivity is not uniform with respect to the wavelength. The relative response is approximately 0.65 for the blue auroral line at 427.8 nm, 0.90 for the green line at 557.7 nm, and 0.45 for the red line at 630.0 nm.

To prepare the images for labeling and network training, a four-step image processing procedure (see Figure 2) was performed on each image:

1. Rotate the images clockwise 90° to direct the geomagnetic pole toward the top of the image and flip the image along the east-west axis.
2. Filter each image with a 3 × 3 median filter to avoid bias effects from the location of stars, remove bad pixels, and reduce noise.
3. Bin the pixels by using a 2 × 2 averaging window to reduce the size of the images and thus speed up the training process.
4. Crop the images to the central 128 × 128 pixels of the binned image, corresponding to the size of the red frame in the left panel of Figure 2. Apart from further speeding-up the training process, there are three reasons for cropping the images. First, pixels that contain considerably distorted features from the fish-eye lens projection and atmospheric conditions are removed. Second, aurorae look similar toward the horizon and classification of aurorae at small elevation angles is not useful. Lastly, selecting a smaller field of view reduced the number of frames which included several aurora classes; this increased the accuracy of the network and made it easier to both label the images and to choose representative aurora labels.

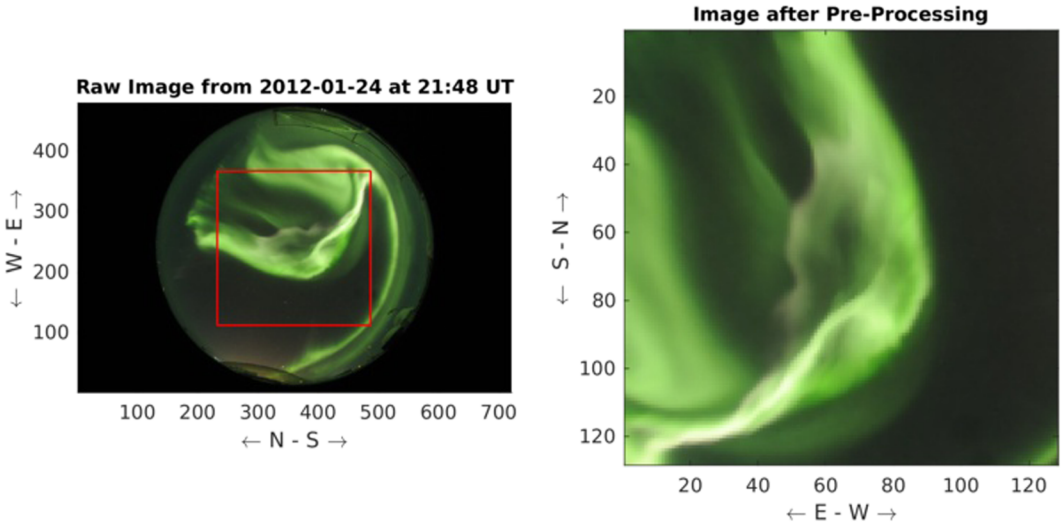

Figure 2. The image to the left depicts a raw (720 × 479 pixels) auroral image. The image to the right shows the same image after the four-step processing procedure. The processed image is rotated, filtered, binned, and cropped. The resolution of the processed image is 128 × 128 pixels. The red frame (left panel) indicates the size of the cropped area.

2.2. Aurora Classification

The processed images were labeled manually. The set of labels used in this study is described below and illustrated in Figure 3. The visual class definition, used for labeling the auroral images, is written in italic, followed by a short physical description and/or explanation. The proposed set of auroral labels was easy to identify under different geomagnetic conditions and applicable to most of the auroral images. In addition, most labels represent an exclusive production mechanisms or characterize a physical property of the aurora. Finally, the labels represent auroral forms which are recognizable at different viewing/elevation angles. Thus, the distribution of the classes is not dependent on the location of the auroral form in the image. This makes statistical interpretations of the results easier. The *edge aurora* class is an exception to the location independence by its definition.

Auroral breakup. Auroral breakup is characterized by bright and large auroral forms which cover most of the frame. Sample images labeled as *auroral breakup* are shown in Row 1 in Figure 3. *Auroral breakup* is an expansion of bright aurora that includes a variety of different large-scale features (e.g., Nishimura et al., 2010). It is characteristic to the substorm expansion phase, which is caused by dynamic processes in the magnetotail (e.g., Xing et al., 2010), leading to enhanced particle precipitation to the ionosphere.

Colored aurora. An image is classified as colored aurora if the aurora, of any shape and form, is clearly not monochromatic green but has a prominent red, blue or purple emission. Images classified as *colored aurora* are presented in Row 2 in Figure 3. The typical precipitation energy and the atmospheric composition in the altitude range of 90–130 km result in green (557.7 nm) being the dominant color in most aurorae. Distinct colored aurora occurs when the electron energy distribution has a pronounced low- or high-energy tail, changing the electron penetration depth into the Earth's ionosphere. Blue and purple auroral displays are usually seen when electrons penetrate deeper into the Earth's ionosphere, as compared to electrons causing the green aurora. Red aurora, however, is produced at higher altitudes and characterizes lower energy electron precipitation.

Auroral arcs. Aurorae with the emission distributed in a single or multiple east-west aligned structure/structures spanning across the image are labeled as auroral arcs. Row 3 in Figure 3 illustrates samples of *arcs. Auroral arcs* run parallel to the auroral oval and the magnetic latitudes (Karlsson et al., 2020). They result from quasi-static particle acceleration in a region close to the ionosphere (Lysak et al., 2020), and they magnetically map to the plasma sheet. *Arcs* are typically considered as quiet time auroral forms but exist at all magnetic activity levels as a basic element of the auroral displays.

Discrete-irregular. Auroral emission appears in broken arcs, north-south aligned arcs, vortical structures, or a combination of several discrete shapes. Discrete-irregular auroral forms are not as bright and not as large as auroral breakup forms. Sample images labeled as *Discrete-irregular* are presented in Row 4 in Figure 3. The *Discrete-irregular* class contains a mixture of different physical generation processes which are not easy to untangle.

Patchy aurora. Patchy aurora is characterized by diffuse aurora consisting of irregular shapes which cover large portions of the image. The intensity of the auroral emission in this class is weak. Patchy aurora images are shown in Row 5 in Figure 3. *Patchy aurora* mainly consists of different pulsating aurora structures (Nishimura et al., 2020). Diffuse patches are caused by pitch angle scattering of energetic electrons to the ionosphere. Different plasma waves play a key role in the scattering processes.

Edge aurora. Images with auroral emission occurring only at the edge of the image are labeled as edge aurora. Sample images are seen in Row 6 in Figure 3. *Edge aurora* can be any of the auroral classes above but information about the class is limited by an insufficient number of bright pixels and uncertainty of the form of the aurora outside the image frame. Thus, not attempting to classify these images as, for instance, *breakup*, *arcs*, or *discrete*, makes the classifier more robust. The *edge aurora* label was included as an additional subclass since the location of the aurora in a set of images is often valuable information for determining if aurora is drifting northward, southward, eastward, or westward.

Faint clear. Images which are dark without clearly visible aurora are labeled as faint clear. Images labeled as *faint clear* are presented in the bottom row of Figure 3. *Faint-clear* images indicate very weak electron precipitation and a quiet ionosphere-magnetosphere environment along the field lines overhead.

Images where a mixture of classes existed were labeled by the most dominant feature with priority given from top (highest priority) to bottom (lowest priority), in the auroral class description. Furthermore, two additional labels were used for classifying the entire data set: *unknown-complicated* and *rejected*. Images classified as *unknown-complicated* or *rejected* were not used for training the networks and are therefore not

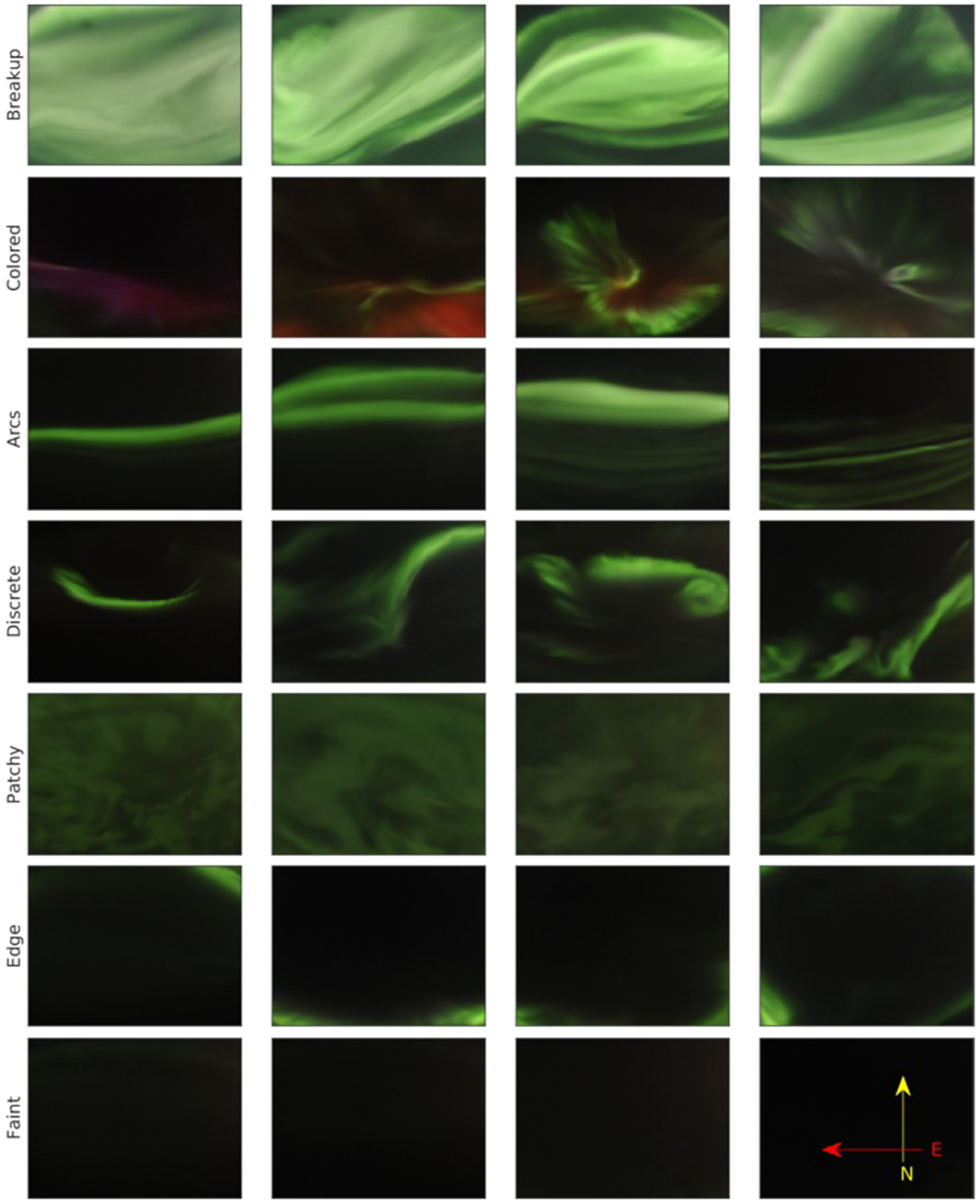

Figure 3. The figure depicts four sample images from each aurora class used for training and testing of the classifiers. The images are processed according to the four-step processing procedure described in section 2.1. The direction with respect to the magnetic pole is indicated by the arrows in the bottom right.

presented in Figure 3. The *unknown-complicated* class was used to exclude images with ambiguous auroral forms which fitted none or several of the auroral labels without a clearly dominant feature. The *rejected* class was used to exclude images with unwanted features, such as clouds and light pollution, which were not detected in the initial preprocessing stage; see section 2.1.

Two of the authors classified the entire data set, one of the authors labeled the images in consecutive order and the other in random order. Only the images with agreeing labels were used for training the network. This was done in order to reduce labeling bias and noise. Thus, a clean training and testing data set was produced by only using the images with agreeing auroral labels and excluding the images with ambiguous auroral forms (labeled as *unknown-complicated*), unwanted features (labeled as *rejected*), and disagreeing

Table 1
The Table Summarizes the Image Data Sets Used in This Study

Year	Number of images	Label	Training	Test	Number of images
2010	479	*Breakup*	113	33	146 (1.04%)
2011	1,506	*Colored*	97	36	133 (0.95%)
2012	2,498	*Arcs*	636	195	831 (5.92%)
2013	0	*Discrete*	150	44	194 (1.38%)
2014	1,684	*Patchy*	1,691	459	2,150 (15.32%)
2015	2,835	*Edge*	167	50	217 (1.55%)
2016	2,037	*Faint*	146	29	175 (1.25%)
2017	1,733	*Unknown-complicated*	—	—	2,630 (18.75%)
2018	996	*Rejected*	—	—	1,078 (7.68%)
2019	262	Disagreement	—	—	6,476 (46.16%)
Total	14,030	Total	3,000	846	14,030 (100.00%)

Note. To the left, the number of auroral images each year that satisfies the initial conditions described in section 2.1 (no clouds, moon below the horizon, and the Sun below −15°). To the right, the number of training, testing, and total images per class. Note that images labeled as *Unknown-complicated*, *Rejected*, or with disagreeing labels were not used for training and testing the classifiers.

labels. The clean data set contained 3,846 (27%) of the 14,030 images in the initial data set, spanning over the Years 2010 to 2019. Finally, the clean data were split into a training and test data set. The training set contains 3,000 images and was used for training the classifiers, while the testing set contains 846 images and was used as an independent test set to evaluate the performance of the classifier. The number of images each year and in each label is presented in Table 1 along with number of images in the training and test sets. Note that although 46% of the images had disagreeing labels, the most disagreement was whether or not an image was suitable for training/testing. The experts agreed on 95% of the labels on images that both experts considered suitable for training/testing (i.e., both experts labeled as *breakup*, *colored*, *arcs*, *discrete*, *patchy*, *edge*, or *faint*), as illustrated in Figure 3.

The motivation for constructing a clean data set is to avoid "confusing" the network by using ambiguous images during training and testing. Note that labels defined in auroral observers classification guides, such as the labels proposed in the International Auroral Atlas (IAA, C.D.W., 1964), were not applicable for training and testing of our data set. The studied data set is not large enough to reliably train and test networks with numerous (>10) and very specific auroral subclasses while still maintaining an acceptable number (∼100) of training images. In addition, labels defined by the temporal characteristics of the aurora cannot be implemented in our study since all considered classifiers are time invariant.

The labels containing less than ∼100 training images is at the lower limit for network training. Note that an insufficient number of training images might cause nonconvergence during training of the classifier. In the early stages of the study, some readjustment of the classes was necessary to ensure there were enough representative samples for the anticipated classes. There were also some false starts, where preselection (i.e., removal of images based on other criteria, such as the presence of the moon) caused some classes to be too few in number.

2.3. DNNs

In machine learning, an object is commonly described by a set of values referred to as features. For images, these features are the raw pixel values of the image. Machine learning algorithms seek to find patterns in the given set of features such that some task can be solved in an optimal way. Determining what features to present the algorithm is crucial in designing a good algorithm. Historically, the main approach has been to design methods that extract features from raw data that are assumed to provide good discriminative power (Guyon & Elisseeff, 2003). However, the design of such features can be challenging and might require a significant amount of domain knowledge. In contrast, DNNs automatically learn which features in the training data are important to solve the desired task (LeCun et al., 2015). The DNNs achieve this automatic extraction by transforming the data through a cascade of nonlinear transformations, which results in a representation that is suited for the problem at hand. Each transformation is commonly referred to as a hidden layer, which

contains parameters that must be optimized. These parameters are optimized using gradient descent, where the gradients are obtained through the back propagation algorithm (Theodoridis & Koutroumbas, 2009b).

Currently, there exists a vast amount of different DNN architectures. This study focuses on some widely used and well-known architectures, as the goal was to demonstrate that DNNs can be effective in auroral classification. The following DNNs were used for classifying the aurora images:

- VGG: A widely used family of convolutional neural networks (CNNs) that have demonstrated a high performance on a number of tasks (Simonyan & Zisserman, 2014). Dropout (Srivastava et al., 2014) was included to regularize the model. Later versions also included batch normalization (Ioffe & Szegedy, 2015). Different versions of the VGG can be created by adding more layers to the network. In this work, we evaluated two versions of the VGG. First, the 16-layer version titled the VGG-16, then the 19-layer version named VGG-19.
- AlexNet: Often considered the breakthrough of deep learning, AlexNet is a CNN consisting of five convolutional layers and three fully connected layers (Krizhevsky et al., 2012). Each layers is followed by a Rectified Linear Unit (ReLU) activation function. Similarly with VGG, Dropout was included to regularize the model.
- ResNet: Residual networks (ResNets) (He et al., 2016) include skip connections between the hidden layers of the networks. Such skip connections ease the flow of gradients in the network (Balduzzi et al., 2017) such that more hidden layers can be included in the network, a process that has shown to increase performance. As with the VGG, adding more layers results in different version of the ResNet. In this study, we evaluated two versions of the ResNet. First, the 18-layer version titled ResNet-18, then the 50-layer version named ResNet-50.
- Clausen and Nickisch (2018) approach: Additionally, we evaluated the performance of the approach used by Clausen and Nickisch (2018). Clausen and Nickisch (2018) used a pretrained deep learning model, an inception model (Szegedy et al., 2017), as a feature extractor that was combined with a linear classifier.

All models were trained using a cross-entropy loss and the Adam optimizer (Kingma & Ba, 2015). The models were developed using the deep learning framework Pytorch (Paszke et al., 2017) on a Tesla K80 GPU.

3. Results

This section presents the classification scores for the KNN method, the SVM method, and six different deep learning-based models on the aurora data set. Further, the confusion matrix and the presoftmax representation obtained by the highest performing model is visualized. Examples of images that were incorrectly classified by the highest performing model are shown and discussed. Lastly, a comparison between the class wise accuracy of a deep learning-based approach to a traditional machine learning approach is presented.

This study evaluates all classifiers by calculating the precision, recall, and F1 score. Precision is defined as follows:

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}, \tag{1}$$

and measures the classifiers ability for not labeling positive samples as negative. A true positive is when the model correctly predicts the positive class. A false positive is when the model incorrectly predicts the positive class. Recall is defined as follows:

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}, \tag{2}$$

and measures the classifiers ability to find positive samples. A false negative is when the model incorrectly predicts the negative class. F1 score is defined as follows:

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

and acts as a weighted average of precision and recall.

Table 2
The Table Summarizes the Precision, Recall, and F1 Score for Different Classifiers on the Aurora Data Set

Algorithm	Precision	Recall	F1 score	# of parameters (M)
3NN	0.84 ± 0.0	0.56 ± 0.0	0.58 ± 0.0	0
5NN	0.66 ± 0.0	0.54 ± 0.0	0.53 ± 0.0	0
Linear SVM	0.78 ± 0.0	0.70 ± 0.0	0.72 ± 0.0	≈0
Clausen and Nickisch (2018)	0.88 ± 0.01	0.87 ± 0.01	0.88 ± 0.01	43
VGG-16	0.84 ± 0.02	0.80 ± 0.03	0.81 ± 0.02	138
VGG-19	0.82 ± 0.04	0.78 ± 0.03	0.79 ± 0.03	143
AlexNet	0.88 ± 0.03	0.88 ± 0.03	0.87 ± 0.03	60
ResNet-18	**0.92 ± 0.02**	0.87 ± 0.05	0.89 ± 0.04	11
ResNet-50	**0.92 ± 0.03**	**0.89 ± 0.04**	**0.90 ± 0.03**	25

Note. The reported scores are the average over 10 runs with different random initialization. The last column lists the number of parameters, in millions, for each classifier, rounded to the nearest whole million. The bold entries mark the highest performance.

3.1. Comparison of the Classification Performances

The performance results of the six deep learning-based models described in section 2.3 are displayed in Table 2. Additionally, a baseline using two well-known machine learning classifiers is provided, namely, a KNN classifier (Theodoridis & Koutroumbas, 2009a) and a SVM classifier (Cortes & Vapnik, 1995). For both classifiers, the histogram of oriented gradients method (Dalal & Triggs, 2005) is used for extracting features from the images. For the KNN classifier, the results when considering the three and five nearest neighbors are reported. For the SVM, the results are for where a linear kernel is utilized. All deep learning-based models outperform the KNN and SVM baseline. The ResNet-50 achieved the highest score across all metrics.

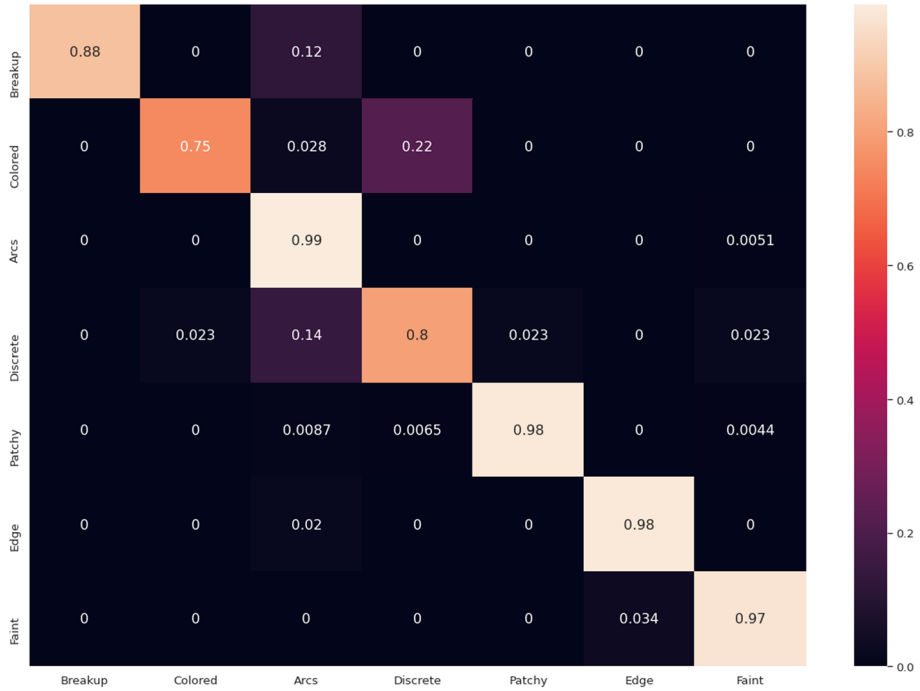

Figure 4. The figure shows the confusion matrix for the ResNet-50 network on the test data. The diagonal displays the percentage of correctly classified images for each class, that is, images where the network automatically classified an auroral image similarly to the manually labeled test data. The off-diagonal elements show the percentage of images from a given class erroneously classified as another class. The results show that the model achieves a high accuracy on most classes, but has some difficulties with separating *colored aurora* and *discrete-irregular* aurora.

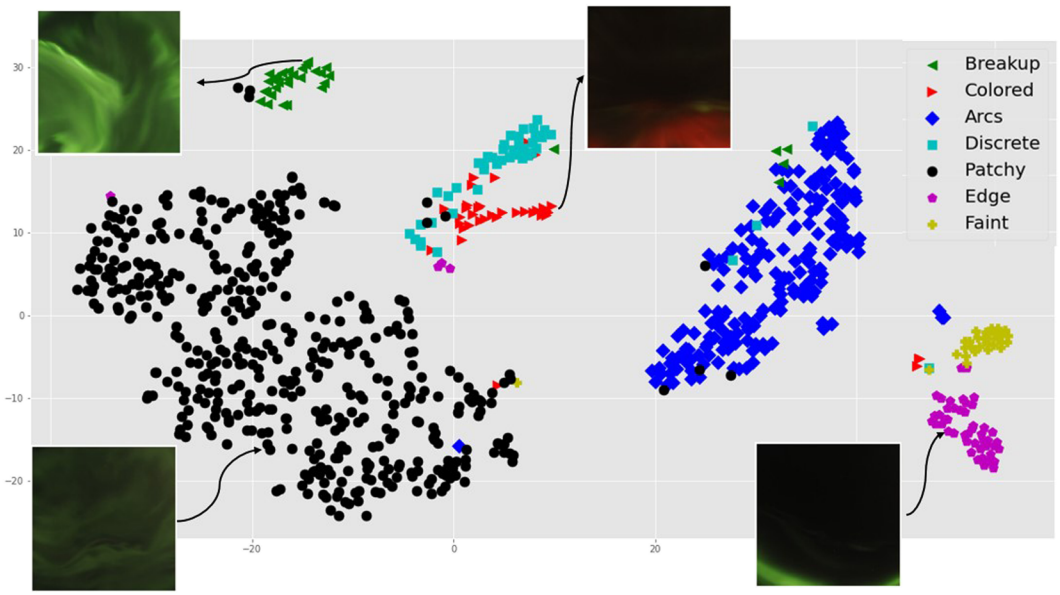

Figure 5. The figure presents the presoftmax representation of the aurora test data produced by the ResNet-50 network and projected down to two dimensions using T-SNE. Images from some of the classes are also displayed in the plot. The figure shows that the model has found a suitable representation where the classes are clustered together and are easily separable.

Figure 4 shows the confusion matrix for the ResNet-50 on the test set of the aurora data. From this figure, it is evident that some classes are well understood by the network. Specifically, almost all samples from the *auroral arcs*, *patchy aurora*, *edge aurora*, and *faint-clear* classes in the test set are classified correctly. On the other hand, the *colored aurora* is more challenging and is partly classified as *discrete-irregular* aurora. Also, both *auroral breakup* and *discrete-irregular* are partly classified as *auroral arcs*.

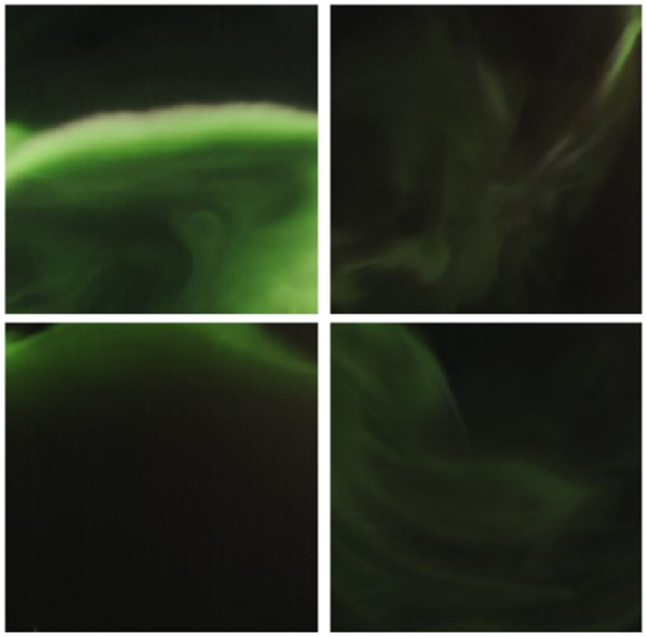

Figure 6. This figure depicts test images that were incorrectly classified by the ResNet-50 network. (top left) *Auroral breakup* classified as *arcs*. (top right) *Colored aurora* classified as *discrete-irregular*. (bottom left) *Edge aurora* classified as *arcs*. (bottom right) *Patchy aurora* classified as *discrete-irregular*. All images are processed according to the four-step processing procedure described in section 2.1.

3.2. Details of the Highest Performing Model (ResNet-50)

Figure 5 displays a two-dimensional presoftmax representation of the test data of the aurora data set, produced by the ResNet-50. The representation was projected down to two dimensions using the t-distributed stochastic neighbor embedding (T-SNE) dimensionality reduction technique (van der Maaten & Hinton, 2008). The figure also contains some examples of the actual images that are represented by the two-dimensional points. It is from this representation that the network determines what class to assign a sample image to. From this figure, it is clear that some classes are well separated from the other classes. For instance, the cluster of *edge aurora* and the cluster of *faint-clear* samples toward the rightmost part of the figure are compactly represented and well separated from the other classes. However, some classes are more mixed together, which corroborates the findings in Figure 4.

Figure 6 shows four examples of incorrectly classified images from the auroral test data. The incorrect classifications are generally sensible, and it can be seen how the algorithm may have opted for the incorrect classification. Possible interpretations are as follows: Figure 6 (top left) shows *auroral breakup* classified as *arcs*; although the breakup is imminent, the image is still dominated by the bright arc. Figure 6 (top right) shows *colored aurora* classified as *discrete-irregular* aurora; the distinction of color is subtle compared to other colored aurorae in the training set. Additionally, the strong non-east-west feature in the top right is found in *discrete-irregular* aurora cases. Figure 6 (bottom left) shows *edge aurora*

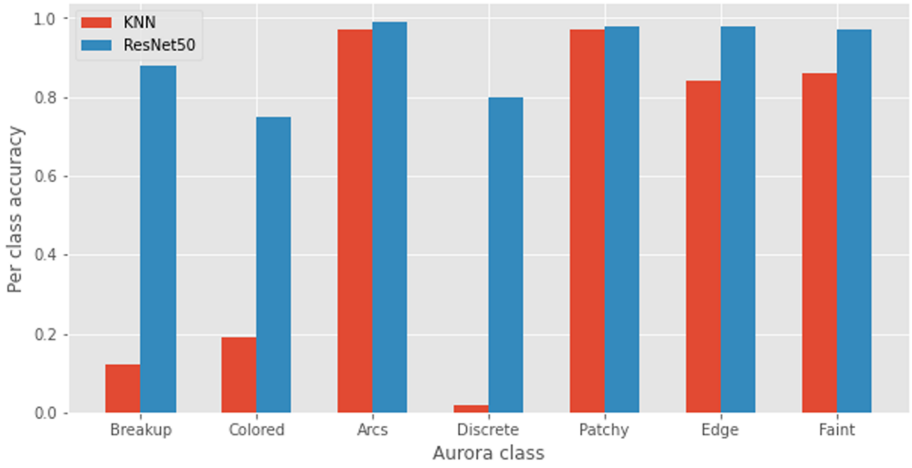

Figure 7. The figure shows the class wise accuracy comparison between a 3NN-classifier and the ResNet-50 network. The figure illustrates how the deep learning-based model outperform the traditional machine learning method in locating all classes. The ResNet-50 classifier handles particularly the breakup, colored, and discrete aurora classes much better than the KNN-classifiers.

classified as *arcs*; in fact, this is an arc, but it is on the edge of the field of view. The intensity of the arc is not high, so the brightness in the corners is not sufficient for the algorithm to opt for that class. Figure 6 (bottom right) shows *patchy aurora* classified as *discrete-irregular*; there is a sharp edge feature in the center-top of the image, which is generally not present in *patchy aurora* cases.

To examine what type of images the deep learning-based models are capable of recognizing compared to the traditional machine learning approach, we compare the class wise accuracy of the 3NN-classifier and the ResNet-50. Figure 7 displays the class-wise accuracy on the test images of the aurora data set. The accuracy of some classes is comparable between the two methods, but particularly for the *auroral breakup*, *colored aurora*, and *discrete-irregular* auroral classes it is clear that the deep learning approach is superior. These classes are typically more challenging to recognize, which seems to suggest that the deep learning-based approach is capable of identifying more complicated structures in auroral images.

4. Discussion

In this study, the DNNs generally outperformed the KNN and SVM techniques. Overall, from the presented results in Table 2 where several architectures, namely, the Clausen and Nickisch (2018) method, AlexNet, ResNet-18, and ResNet-50, achieved close to 90% average precision, infer that aurora classification is a suitable task for DNNs. It has been demonstrated that it is possible to classify specific auroral forms such as *auroral arcs*, *edge aurora*, *patchy aurora*, and *faint clear* with a high precision (>90%), as seen in Figure 4. The networks were in general less successful in classifying *auroral breakup*, *colored aurora*, and *discrete-irregular* aurora, likely reasons are too few and/or too ambiguous training images. In addition, from Figure 5 it is clear that colored aurora has an overlap with *discrete-irregular* aurora. This overlap occurs when the aurora does not have a clearly pronounced colored emission. Note that it might be possible to achieve a higher label separability using unsupervised clustering methods. However, an additional goal of this study was to define auroral labels that represents exclusive production mechanisms or characterize physical properties of the aurora when possible. The resulting clusters attained by an unsupervised clustering method will most likely not satisfy this goal. An unsupervised classifier might therefore not be applicable for most scientific purposes where the aim is to study the physics of the aurora and its production mechanisms.

The DNN architectures VGG-16 and VGG-19 had the worst performance. These models have a significantly higher amount of parameters, as presented in Table 2, and it might be that the data set is not large enough for training models of such a size. Also, the ResNet models are generally known to outperform AlexNet and the VGG-based models (He et al., 2016), often attributed to their ability to propagate gradients effectively even for very deep networks (Balduzzi et al., 2017).

Clausen and Nickisch (2018) used a pretrained DNN to extract features and then trained a linear classifier using these features. This means that the parameters of the pretrained DNN are not optimized for handling

auroral images but for other types of images. In contrast, the models used in this study, VGG, AlexNet, and ResNet, are only trained using auroral images. From Table 2 it is evident that the pretrained approach gives better performance than the traditional machine learning algorithms but does slightly worse than the best DNNs trained on only auroral images. This suggests that training the models specifically for classifying images of aurora might improve their capability to detect different types of aurora. However, a larger data set is needed to validate this proposition.

An analysis of the network performance on data from other cameras still remains to be done. Note that the color sensitivity of the detector, presented in Figure 1, is a camera-specific feature. Thus, the RGB images acquired by other cameras should be adjusted to account for the differences in the wavelength-dependent light response before being used by the classifier. In addition, the training and testing was done using images without clearly ambiguous images, as described in section 2. Hence, a proper analysis of how ambiguous images is classified also needs to be investigated. The network will ideally classify the most dominant auroral feature. However, an evaluation of the performance on ambiguous images will itself be subject to biases and subjective interpretations, as described by McKay and Kvammen (2020). It is irrational to expect a DNN to classify ambiguous images correctly if not even auroral experts can agree on what the correct label is. Thus, a common consensus about the auroral morphology and the criteria for each class needs to be introduced before progress can be made on classifying ambiguous images. Alternatively, one could interpret ambiguous auroral images as a mixture of more common classes (e.g., *auroral breakup*, *arcs*, and *patchy aurora*) and label ambiguous images with multiple classes.

Future endeavors in aurora classification with DNNs should investigate the dimension space using different cameras and auroral events, for instance, by T-SNE maps. Further improvements to the classifier can likely be achieved by including the time dimension. The time dimension information can be included by combining the CNNs with recurrent neural networks (RNNs) or by using the CNNs directly on a data set consisting of labeled stacks of consecutive images with, for example, ∼10 images in each stack, the appropriate number of images in each stack depends on the imaging frequency, the exposure time, and the field of view. The inclusion of the time dimension also allows for classifying labels with a distinctive temporal behavior (e.g., *pulsating aurora*). From the data set used in this study, it can be concluded that images labeled as *colored aurora* and *discrete-irregular* are not well separated by the classifiers. Other labeling sets should therefore be considered. It should be noted that classifying colored aurora is of particular interest for detecting exciting phenomena such as sunlit aurora (Shiokawa et al., 2019; Størmer, 1929), stable auroral red (SAR) arcs (Mendillo et al., 2016; Rees & Roble, 1975), and Strong Thermal Emission Velocity Enhancement (STEVE) (Gallardo-Lacourt et al., 2018; MacDonald et al., 2018), which are characterized by pronounced colored emissions. In addition, both *auroral arcs* and *patchy aurora* achieved a high precision and are common auroral forms. Subdividing these classes, for instance, *arcs* into *single arc* and *multiple arcs*, might therefore be advantageous. Furthermore, auroral omega bands are quite common and distinctive auroral forms (Partamies et al., 2017) which might be possible to classify with a high precision.

5. Conclusion

This paper presents the results of an extensive study of automatic aurora classification using different machine learning techniques. Seven auroral classes were considered: *auroral breakup*, *colored aurora*, *auroral arcs*, *discrete-irregular*, *patchy aurora*, *edge aurora*, and *faint-clear*. The classifiers were both trained and tested on clean colored (RGB) auroral images, without clearly ambiguous auroral forms and unwanted features, such as clouds and light pollution. Six DNN architectures were tested along with the well-known machine learning classifiers KNN and SVM. The ResNet-50 DNN architecture achieved the highest performance with an average classification precision of 92%.

Overall, the conclusion is that automatic auroral image classification is a suitable task for DNNs. The DNNs generally outperformed the KNN and SVM techniques. However, progress in this field of study is constrained by biases and subjective interpretations (McKay & Kvammen, 2020). It is irrational to expect the DNNs to classify an auroral image correctly if auroral researchers cannot agree on what the correct aurora label is. High precision (>90%) classification of clearly ambiguous auroral images can therefore not be readily achieved before a common consensus about the auroral morphology and the criteria for each class is established. The use of data without clearly ambiguous auroral forms for automatic aurora classification,

mainly labeled into physically meaningful definitions, might be the first step in sorting auroral structures in a morphological space.

Data Availability Statement

The image data archive are freely accessible at https://www2.irf.se/allsky/data.html, and the processed image data set and code used in this paper are available online (at https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/SSA38J).

Acknowledgments
The authors would like to thank Urban Brändström and the Swedish Institute of Space Physics for providing the original auroral image data. However, the users are obliged to contact the Kiruna Atmospheric and Geophysical Observatory before using the images. Andreas Kvammen is supported by the Tromsø Research Foundation. The work by Derek McKay is partly supported by the Academy of Finland project number 322535. The work by Noora Partamies is partly supported by the Research Council of Norway under contract 287427 and a CoE contract 223252. The authors would also like to thank Björn Gustavsson for valuable suggestions and comments.

References

Balduzzi, D., Frean, M., Leary, L., Lewis, J. P., Ma, K. W.-D., & McWilliams, B. (2017). The shattered gradients problem: If resnets are the answer, then what is the question? In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 342–350), Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR. Retrieved from http://proceedings.mlr.press/v70/balduzzi17b.html

C.D.W. (1964). International Auroral Atlas. Published for the International Union of Geodesy & Geophysics, Edinburgh (University Press), 1963. pp. 17; 2 Figures; 52 black and white, 4 colour plates. 45s. *Quarterly Journal of the Royal Meteorological Society*, *90*(386), 502–502. https://doi.org/10.1002/qj.49709038624

Clausen, L. B. N., & Nickisch, H. (2018). Automatic classification of auroral images from the Oslo Auroral THEMIS (OATH) data set using machine learning. *Journal of Geophysical Research: Space Physics*, *123*, 5640–5647. https://doi.org/10.1029/2018JA025274

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886–893). San Diego, CA, USA, USA: IEEE. https://doi.org/10.1109/CVPR.2005.177

Gallardo-Lacourt, B., Nishimura, Y., Donovan, E., Gillies, D. M., Perry, G. W., Archer, W. E., et al. (2018). A statistical analysis of STEVE. *Journal of Geophysical Research: Space Physics*, *123*, 9893–9905. https://doi.org/10.1029/2018JA025368

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182. Retrieved from http://dl.acm.org/citation.cfm?id=944919.944968

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). Las Vegas, NV, USA: IEEE. https://doi.org/10.1109/CVPR.2016.90

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37* (pp. 448–456), ICML'15: JMLR.org. JMLR.org. Retrieved from http://dl.acm.org/citation.cfm?id=3045118.3045167

Karlsson, T., Andersson, L., Gillies, D. M., Lunch, K., Marghitu, O., Partamies, N., et al. (2020). Quiet, discrete auroral arcs—Observations. *Space Science Reviews*, *216*, 16.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Retrieved from http://arxiv.org/abs/1412.6980

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems* (Vol. 1, pp. 1097–1105). NIPS'12. USA: Curran Associates Inc. http://dl.acm.org/citation.cfm?id=2999134.2999257

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

Lysak, R., Echim, M., Karlsson, T., Marghitu, O., Rankin, R., Song, Y., & Watanabe, T.-H. (2020). Quiet, discrete auroral arcs: Accelerations mechanisms. *Space Science Reviews*, *216*, 92.

MacDonald, E. A., Donovan, E., Nishimura, Y., Case, N. A., Gillies, D. M., Gallardo-Lacourt, B., et al. (2018). New science in plain sight: Citizen scientists lead to the discovery of optical structure in the upper atmosphere. *Science advances*, *4*(3), eaaq0030.

Mauer, C. (2009). Measurement of the spectral response of digital cameras with a set of interference filters. University of Applied Sciences Cologne. Department of Media- and Phototechnology.

McKay, D., & Kvammen, A. (2020). Auroral classification ergonomics and the implications for machine learning. *Geoscientific Instrumentation, Methods and Data Systems*, *9*(2), 267–273.

Mendillo, M., Baumgardner, J., & Wroten, J. (2016). SAR arcs we have seen: Evidence for variability in stable auroral red arcs. *Journal of Geophysical Research: Space Physics*, *121*, 245–262. https://doi.org/10.1002/2015JA021722

Nishimura, Y., Lessard, M., Katoh, Y., Miyoshi, Y., Grono, E., Partamies, N., et al. (2020). Diffuse and pulsating aurora. *Space Science Reviews*, *216*, 4.

Nishimura, Y., Lyons, L., Zou, S., Angelopoulos, V., & Mende, S. (2010). Substorm triggering by new plasma intrusion: THEMIS all-sky imager observations. *Journal of Geophysical Research*, *115*, A07222. https://doi.org/10.1029/2009JA015166

Partamies, N., Juusola, L., Whiter, D., & Kauristie, K. (2015). Substorm evolution of auroral structures. *Journal of Geophysical Research: Space Physics*, *120*(7), 5958–5972. https://doi.org/10.1002/2015JA021217

Partamies, N., Weygand, J. M., & Juusola, L. (2017). Statistical study of auroral omega bands. *Annales Geophysicae*, *35*(5), 1069–1083. https://doi.org/10.5194/angeo-35-1069-2017

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in PyTorch.

Rao, J., Partamies, N., Amariutei, O., SyrjÃđsuo, M., & van de Sande, K. E. A. (2014). Automatic auroral detection in color all-sky camera images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(12), 4717–4725. https://doi.org/10.1109/JSTARS.2014.2321433

Rees, M. H., & Roble, R. G. (1975). Observations and theory of the formation of stable auroral red arcs. *Reviews of Geophysics*, *13*(1), 201–242.

Rhodes, B. (2019). Skyfield v1.10. Retrieved from http://rhodesmill.org/skyfield/

Shiokawa, K., Otsuka, Y., & Connors, M. (2019). Statistical study of auroral/Resonant-Scattering 427.8-nm emission observed at subauroral latitudes over 14 years. *Journal of Geophysical Research: Space Physics*, *124*, 9293–9301. https://doi.org/10.1029/2019JA026704

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958. Retrieved from http://jmlr.org/papers/v15/srivastava14a.html

Störmer, C. (1913). On an auroral expedition to bossekop in the spring of 1913. *The Astrophysical Journal*, *38*, 311.
Størmer, C. (1929). The distribution in space of the sunlit aurora rays. *Nature*, *123*(3090), 82–83.
Syrjäsuo, M., & Donovan, E. (2002). Analysis of auroral images: Detection and tracking. *Geophysica*, *38*(1–2), 3–14.
Syrjäsuo, M. T., Donovan, E. F., Qin, X., & Yang, Y. H. (2007). Automatic classification of auroral images in substorm studies. In *8th International Conference on Substorms (ICS8)* (pp. 309–313).
Syrjäsuo, M. T., & Donovan, E. F. (2004). Diurnal auroral occurrence statistics obtained via machine vision. *Annales Geophysicae*, *22*(4), 1103–1113. https://doi.org/10.5194/angeo-22-1103-2004
Syrjasuo, M., & Partamies, N. (2011). Numeric image features for detection of aurora. *IEEE Geoscience and Remote Sensing Letters*, *9*(2), 176–179.
Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 4278–4284). Palo Alto, CA: AAAI Press.
Theodoridis, S., & Koutroumbas, K. (2009a). Chapter 2—Classifiers based on Bayes decision theory. In S. Theodoridis & K. Koutroumbas (Eds.), *Pattern recognition (fourth edition)* (4th ed., pp. 13–89). Boston: Academic Press. https://doi.org/10.1016/B978-1-59749-272-0.50004-9
Theodoridis, S., & Koutroumbas, K. (2009b). Chapter 4—Nonlinear classifiers. In S. Theodoridis & K. Koutroumbas (Eds.), *Pattern recognition (fourth edition)* (4th ed., pp. 151–260). Boston: Academic Press. https://doi.org/10.1016/B978-1-59749-272-0.50006-2
van der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.
Xing, X., Lyons, L., Nishimura, Y., Angelopoulos, V., Larson, D., Carlson, C., et al. (2010). Substorm onset by new plasma intrusion: THEMIS spacecraft observations. *Journal of Geophysical Research*, *115*, A10246. https://doi.org/10.1029/2010JA015528
Yang, Q., Liang, J., Hu, Z., & Zhao, H. (2012). Auroral sequence representation and classification using hidden Markov models. *IEEE Transactions on Geoscience and Remote Sensing*, *50*(12), 5049–5060. https://doi.org/10.1109/TGRS.2012.2195667
Yang, Q., Tao, D., Han, D., & Liang, J. (2019). Extracting auroral key local structures from all-sky auroral images by artificial intelligence technique. *Journal of Geophysical Research: Space Physics*, *124*, 3512–3521. https://doi.org/10.1029/2018JA026119