STScI | SPACE TELESCOPE SCIENCE INSTITUTE

Instrument Science Report WFC3 2024-03

WFC3/UVIS Guide Star Failure Classification with Machine Learning

M. Jones, F. Dauphin

April 30, 2024

Abstract

The Wide Field Camera 3 (WFC3) onboard the Hubble Space Telescope (HST) has captured over 310,000 images in its near 15-year lifetime. Some of these images are affected by guide star failures, which can cause a smearing of the sources in the image. Although the images are manually flagged by WFC3 team members for such anomalies, machine learning is more practical for observatories that will be far more data rich, and where manual flagging will be inefficient or even impossible. In order to remedy this problem, we trained a convolutional neural network (CNN) to identify WFC3/UVIS images affected by guide star failures. The CNN's training and validation data were taken from May 2009 to May 2022. We developed a data processing pipeline to log-scale, down-sample, and normalize the images. Our best model achieved true negative and true positive rates of 90% *and* 91% *on our validation data. We investigate the model's misclassifications, deployment tests, and rotational dependency. In addition, we present shortcomings from other trained models and ideas for future work. Our code and model parameters can be found on Deepwfc3's GitHub.*

1. Introduction

The Wide Field Camera 3 (WFC3) detector on the Hubble Space Telescope (HST) has provided scientists with important scientific data since its installation during Servicing Mission 4 in 2009. The WFC3 instrument consists of two detectors: UVIS and IR. Here, we focus on the UVIS detector, which is made up of two [2051x4096] pixel CCD chips (Marinelli and Dressel, 2024). Since WFC3/UVIS is a CCD, its observations can suffer from anomalies, such as reflected light and satellite trails, that affect the quality of the captured data (Gosmeyer, The Quicklook Team, 2017). In this report, we focus on the guide star failure anomaly, specifically for WFC3/UVIS (Sahu et al. 2021).

Guide star failures (GS fails) occur when the telescope is unable to lock onto a guide star either before or during an observation. This results in a rolling motion in the telescope during imaging, which creates parallel streaks across an image from any light sources being observed. As of December 2022, 3619 WFC3 observations have been flagged as affected by a GS fail, 1335 of which were UVIS observations for general observers.

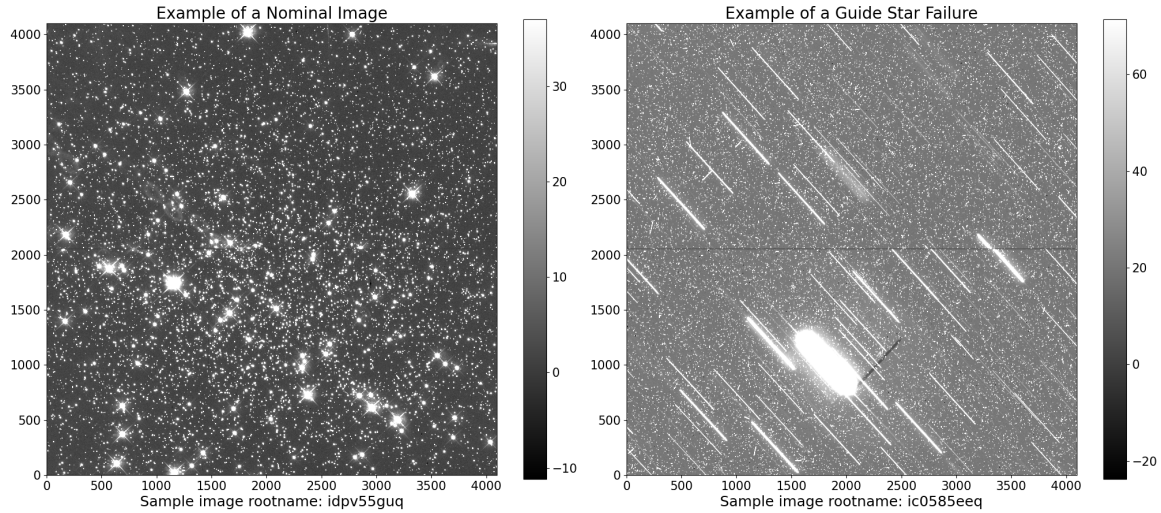

Fig. 1.—*Sample of a nominal image and guide star failure (GS fail) image. The image on the left shows a typical observation, without an anomaly, where the objects in the image are clear and resolved. On the right is an example of an observation affected by guide star failure, which shows the characteristic parallel streaks across the image due to the telescope moving during observation.*

Figure 1 shows how a guide star failure can affect an image. In the image without any anomalies (i.e. nominal), the stars are clear point sources. Conversely, in the GS fail image,

there are several streaks in the image due to the telescope moving during the exposure, causing the integrated light that would have landed in a specific area of the detector to instead spread along the drift direction, compromising scientific data quality.

Currently, GS fails, and other types of anomalies, are manually flagged by members of the WFC3 team. By flagging these defects on the detector, the team can better understand its behavior over time, and monitor it for any changes. This manual method has succeeded in accurate flagging in a timely manner, but as current telescopes continue to age, anomalies such as GS fails will become more and more common. Additionally, newer telescopes will produce much larger quantities of data, making it much harder for telescope team members to continue to keep up with the manual image quality assurance. To help prepare for the future of telescope observations, it is imperative to find novel ways of automating image outlier detection. One way that the automation of this process can be achieved is by training a machine learning model to detect different types of anomalies, such as blobs and figure 8 ghosts in Dauphin et al. 2021 and 2022. In this report, we build on these previous works by focusing on guide star failure identification for UVIS observations.

2. Data

2.1. Data Set Creation

The inital data that was used to build our training and validation data sets included all non-proprietary UVIS general observer (GO) calibrated images, or any observations taken before December 2022. In particular, the training data set contained images taken between May 1, 2009 and May 1, 2021, and the validation data set contained images taken between May 1, 2021 and May 31, 2022. These observations were assessed and labeled by the WFC3 Quicklook team as an image without any anomalies (nominal) or as containing at least one anomaly.

The entire set of general observer (GO) images taken within the training and validation frame had 88,822 calibrated science images, without sky background subtraction (i.e. SCI extensions of FLT images). To simplify our data set, images with an anomaly other than a guide star failure were excluded, which resulted in 60,487 images. Additionally, since moving target observations [1] can look similar to GS fails, those images were excluded from the data set, which left 44,241 images remaining. The first data set we built then had a

[1]Moving target observations deliberately slew the telescope to follow a relatively close source, e.g. a Solar System object.

training set with 39,477 nominal images, and 438 guide star failures as well as a validation set with 4,111 nominal images, and 215 guide star failures. This data set had a significant class imbalance, since there were 20 to 100 times more nominal images in the training and validation sets. Imbalanced data sets can cause the models trained on them to be biased towards the majority class, i.e. nominal images, or have a more difficult time learning the features that define a minority class, i.e. guide star failures.

After creating the first data set, any images that were spatial scans or grism images were removed from the data leaving 42,112 images remaining. Using the remaining observations, we created a second training set consisting of 37,674 nominal and 417 guide star failure images, and a second validation set consisting of 3,841 nominal images and 180 guide star failures. Similarly to the first data set, there was a significant amount of class imbalance in this data set.

2.2. Data Processing

Although our data sets were calibrated for science, we processed them further for our modeling purposes. We needed to resolve nonphysical pixel values, properly scale our images to make features prominent, and resize our images for reasonable compute resources. The data was processed using the following procedure:

1. Set any pixels less than 1 e- to 1 e- to remove negative flux pixels and have real valued pixels after log-scaling.
2. Logarithmically scale the image.
3. Resize the image to (256,256) the image using bi-cubic interpolation.
4. Min/max scale the images' pixels to a range of [0,1] using the following formula:

$$\frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x is the original pixel of an image, x_{min} is the minimum pixel of an image, and x_{max} is the maximum pixel of an image. Examples of the data processing pipeline for nominal and GS fails are shown in Figure 2.

Negative pixel values were set to one electron for each image to make sure that they could be properly log scaled, but also because these values do not provide any scientific information about the images. Then, by logarithmically scaling the pixel values of each image, we extract prominent image features across a wide order-of-magnitude range, making

them easier for the model to identify. In addition, log scaling introduced a more uniform range of pixel values in each image. Next, since the images in our data sets could be up to (4096,4096) pixels in size, and were not all the same size, all images were binned down to a standard size of (256,256). Aside from standardizing the size of all of our samples, images were resized to (256,256) to minimize the computational cost of training our models on the images in our data sets, while still keeping the important features necessary for classification intact. The final measure we took to ensure uniformity across all data sets was to min/max scale them using Equation 1, enforcing every pixel value to a range of [0,1].

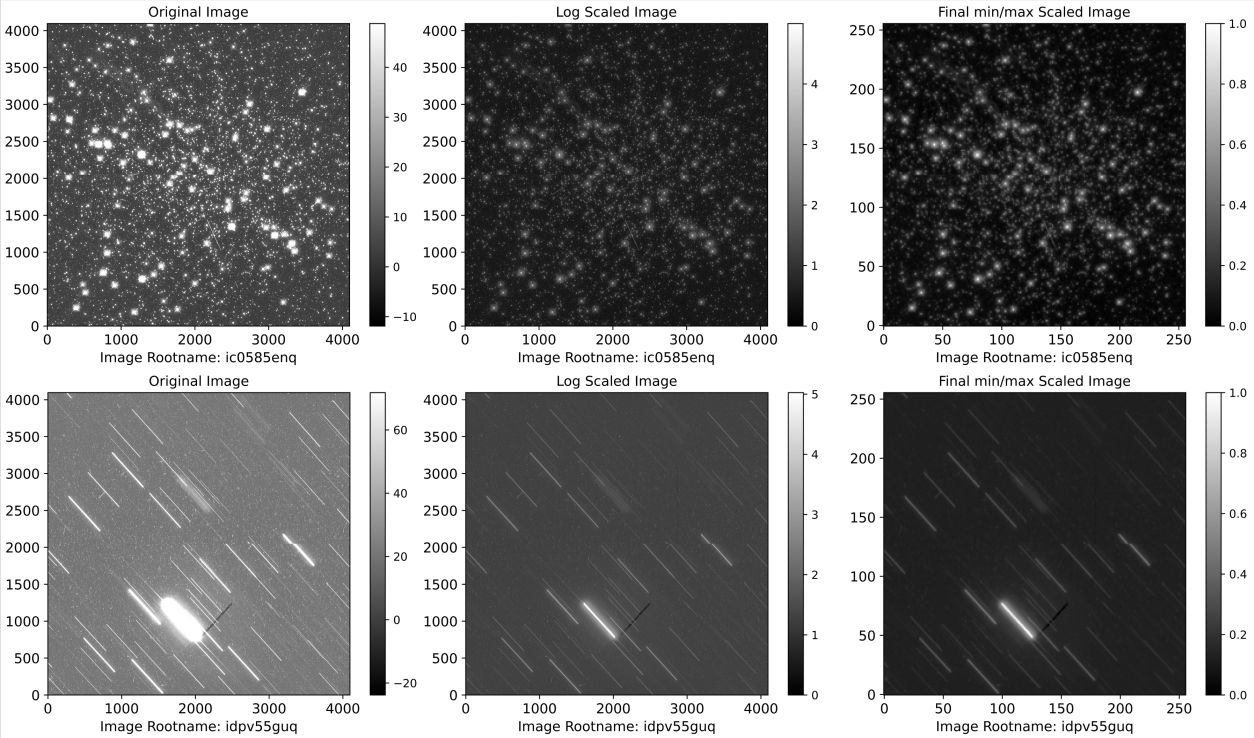

Fig. 2.—*Top row, left to right: Example of the data processing procedure applied to a nominal image. Starting with the original image to the left, values less than 1 are clipped and the image is log scaled. The image is then resized and min-max scaled to standardize pixel values to be between 0 and 1. Bottom row, left to right: The data processing pipeline when applied to an image affected by guide star failure. The same process is applied to make GS fail characteristics in an image stand out.*

2.3. Data Augmentation

As previously mentioned, there was a significant class imbalance between nominal and GS fail images, with the latter only making up about 1.5% of our data sets. In this case, where we did not have as many samples to help the model better learn the features of a GS fail, we create augmented versions of our data, which evened out the number of GS fails to nominal images in our data sets. Data augmentation is a technique to ensure the model generalizes to more data (Maslej-Kresnakova et al., 2021). There were several ways that data could be augmented, such as through random cropping, rotating the images, flipping the images, and shifting the values of the pixels within the image (Wang et al., 2019 and Paillassa et al., 2020). Since we already processed our images, we augmented 10 copies of each image to balance our classes, post-processing, using the following method:

1. Flipping the images vertically with a probability of 50%.
2. Flipping the images horizontally with a probability of 50%.
3. Rotating the images to a random degree between (0,360).
4. Cropping the center of the image to be (180,180) pixels.

Since guide star failures were axially and rotationally invariant, the augmentation process above did not visually affect the differences against the nominal images. Figure 3 illustrates how the data augmentation pipeline affected the same processed images from Figure 2.

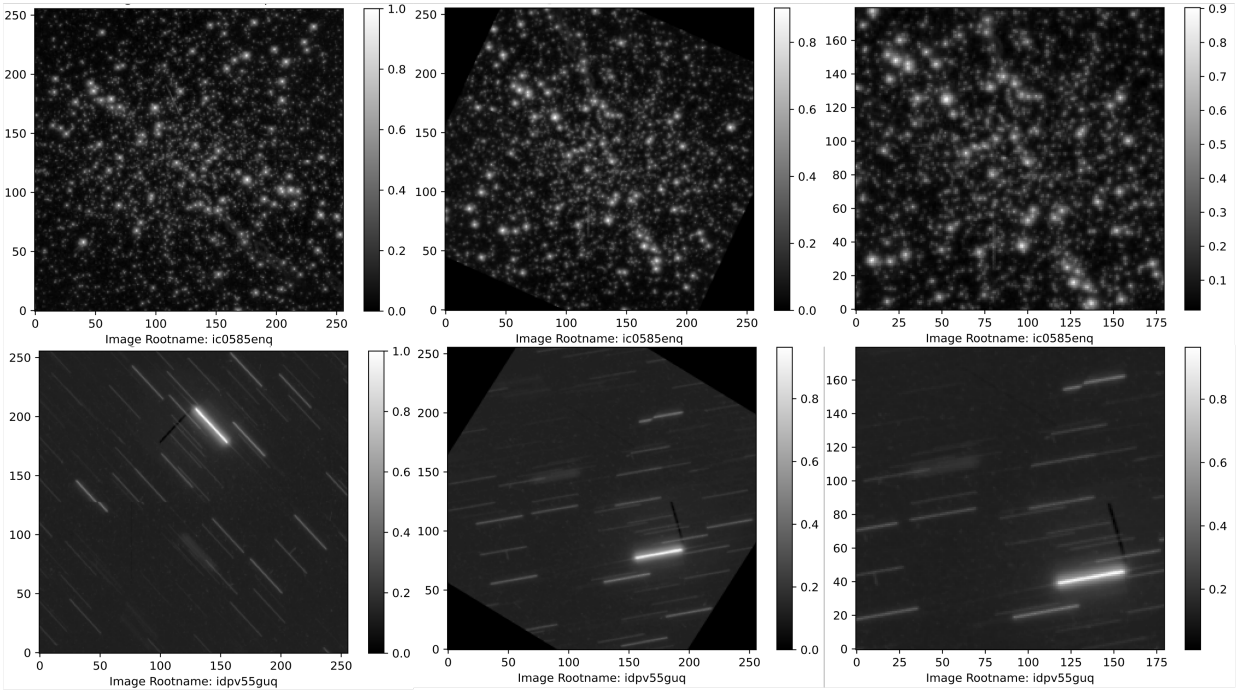

Fig. 3.—*Top row: Example of the data augmentation pipeline when applied to the processed nominal images from Figure 2. Images were first flipped vertically with a probability of 50%, flipped horizontally with a probability of 50%, rotated to a random angle between (0,360) and finally, cropped in the center to be [180,180] pixels to ensure no extra blank space on the edges of the image were present. Bottom row: Example of the data augmentation pipeline when applied to the processed guide star failure from Figure 2. Visual features indicating a GS fail remained prominent.*

3. Methods

3.1. Machine Learning

To automate GS fail classification, we trained a machine learning algorithm to identify images with guide star failures. Machine learning consists of algorithms that do not rely on human input in order to learn and improve its fit of a particular set of data (Lukic et al., 2018). Instead, these algorithms go through cycles of training and validation to learn a mathematical relationship between input data and output data. The cycle begins with training, where the model is shown labeled samples of the data, or examples that have the desired outputs attached. With labeled examples, the model learns important features in the input data. Next, the loss is calculated, which determines how well the

model is performing. Within machine learning, loss functions quantify the difference between the model's prediction and the correct prediction for the example. Common loss functions include cross-entropy for classification, or mean squared error for regression. Once loss is determined, the model undergoes back-propagation to update its parameters for the next round of training. Finally, the model moves on to validation, where it will see labeled samples that it has not seen during training to evaluate how well the model is generalizing.

The type of machine learning algorithm that we used was a neural network algorithm. In this type of algorithm, we set up a series of layers, which were made up of nodes, that each have their own function associated with them. In this set up, the nodes that make up each successive layer bases their inputs off of the outputs from the previous layer. This process allows neural network models to learn more complex relationships and features within a data set, and solve more complex problems.

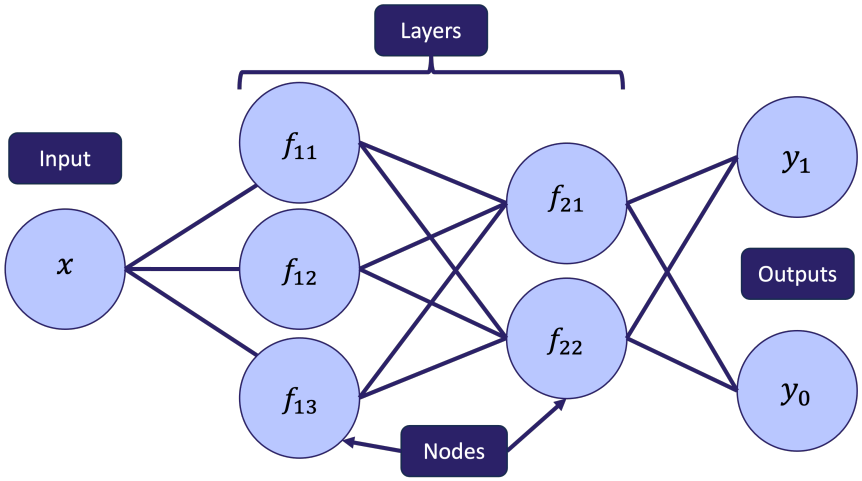

Fig. 4.—*A generalized visual representation of a neural network. Neural networks are built up using nodes, or neurons, that are organized into layers. The x represents the input data (e.g. pixels in an image), and the ys represent the output data (e.g. classification probability). The nodes (i.e. the fs for linear functions with an activation applied) in each successive layer build off of the nodes in the preceding layer, allowing them to create more complex models as the number of neurons and layers is increased.*

3.2. Convolutional Neural Networks

In order to better tailor our neural network to the problem of identifying features in WFC3/UVIS images, we built and trained a convolutional neural network (CNN) following a

similar method used in Dauphin et al., 2022. We chose CNNs because they use convolutional layers, which are particularly well suited to understanding image data. Convolutional layers help to extract important features in the image by creating feature maps, which are down-sampled versions of the original input data, before using this as an input for the neurons. This allows us to add more layers of complexity to the types of problems that can be solved using a neural network model.

The models presented in this report were all trained using the same architecture which consisted of four convolutional layers, and two fully-connected layers. The convolutional layers had 32, 64, 128, and 256 filters, respectively, and the fully-connected layers had 64, and two neurons. In the context of a CNN, filters refer to the kernels used to convolve the image to create feature maps within each layer of the convolutional neural network. For more machine learning related vocabulary, see appendix of Dauphin et al. 2021.

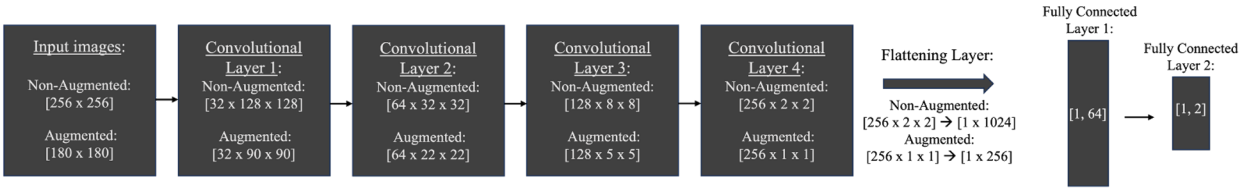

Fig. 5.—*The convolutional neural network architecture used in the models for this project. There are four convolutional layers with 32, 64, 128, and 256 filters, respectively, and two fully-connected layers with 64 and 2 neurons. This architecture expands on the neural network in Figure 4 by adding convolutional layers to the structure, which will create feature maps (down-sampled versions of the images) that will learn and extract important features of an image. These feature maps can then be flattened and used as an input for the traditional neural network structure, which corresponds with the "fully-connected" layers in this figure.*

3.3. Model Training

All of the models trained for 100 epochs, using a batch size of 128, cross-entropy loss, and the Adam optimization function (Kingma, Ba, 2014). For the purposes of this report, we focus on our best trained model, which we call Model 1, that was trained on a non-augmented data set that excluded spatial scans and grism images. By training on non-augmented data, and excluding spatial scans and grism images, we ensured that the model specifically learned what a GS fail looked like, rather than identifying other types of images that may have similar features. During training of non-augmented data sets, a random sample of nominal images was taken to match the number of GS fail images in order to balance our classes.

For comparison, two other models were trained using the same architecture as Model 1, but used different training and validation sets. Model 2 trained on augmented data that included spatial scans and grism images. Since this model was trained on augmented data, the number of nominal images chosen for training was 10 times the size of the GS fail set to match with the 10 augmented versions of each GS fail image. This model was trained to assess how accurate our model will be when trained using augmented data. The inclusion of spatial scans and grism images in our data sets for this model may affect the ability of the model to learn to identify GS fails specifically, since those images were the result of an intentional effect similar to an (unintentional) guide star failure. Model 3 trained on non-augmented data that included spatial scans and grism images. The results of this model help determine how including spatial scans and grism images in our data set will affect the model's ability to accurately predict whether or not an image is nominal or a guide star failure.

4. Results

4.1. Model 1 Performance

We assessed Model 1's performance using loss and accuracy metrics during the training and validation. The loss and accuracy metrics calculated for the first 30 epochs of training and validation of Model 1 are shown in Figure 6 below.

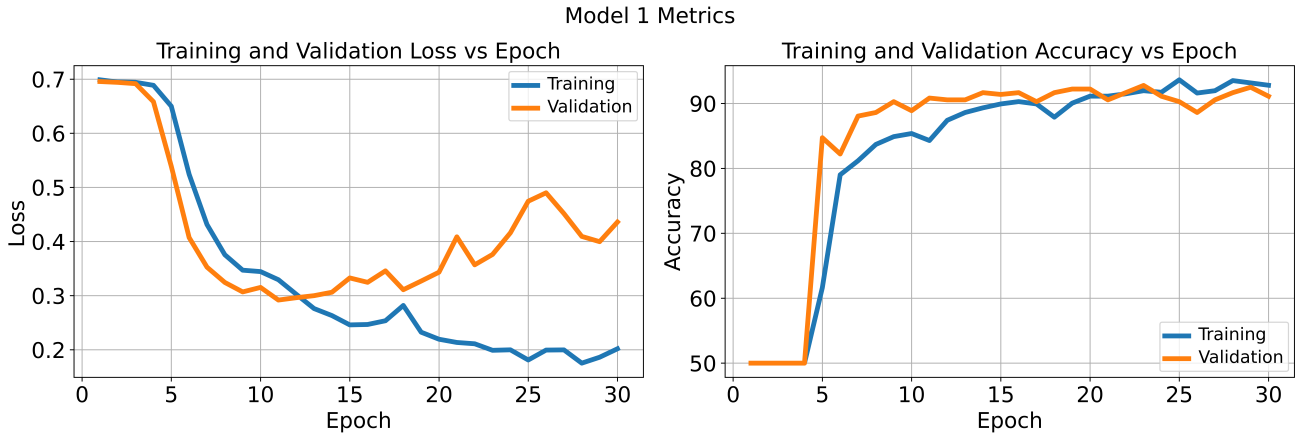

Fig. 6.—*Plots of epoch vs. loss (left) and accuracy (right) metrics for Model 1 training and validation. Epoch 20 was chosen as the stopping point for Model 1 because loss was relatively minimal in both training and validation near that epoch. The divergence of training and validation loss after epoch 20 indicates that Model 1 "memorized" the training set, and stopped generalizing to the validation data.*

After epoch 20 of training, the loss metrics for training and validation began to significantly diverge, which indicated over-fitting in the model, or that it did not generalize to other sets of unseen data. Based on these metrics, we chose the model parameters at epoch 20 to be the final version of this model, since the training and validation loss had not yet significantly diverged, and the accuracy was up to 90%.

To help us better understand the behavior of our model, we analyzed the model's confusion matrix. In a confusion matrix, we determine the rates of true positive, true negative, false positive, and false negative for our model. Model 1 was relatively accurate in identifying both nominal and GS fail images with true negative and positive rates of 0.9 and 0.91, respectively. The classification threshold for all of our models was chosen to be 0.5 to optimize the performance of our models. Different thresholds may be chosen to decrease the number of false negatives that occur to prevent any images from being missed, but we leave that for future work during deployment.

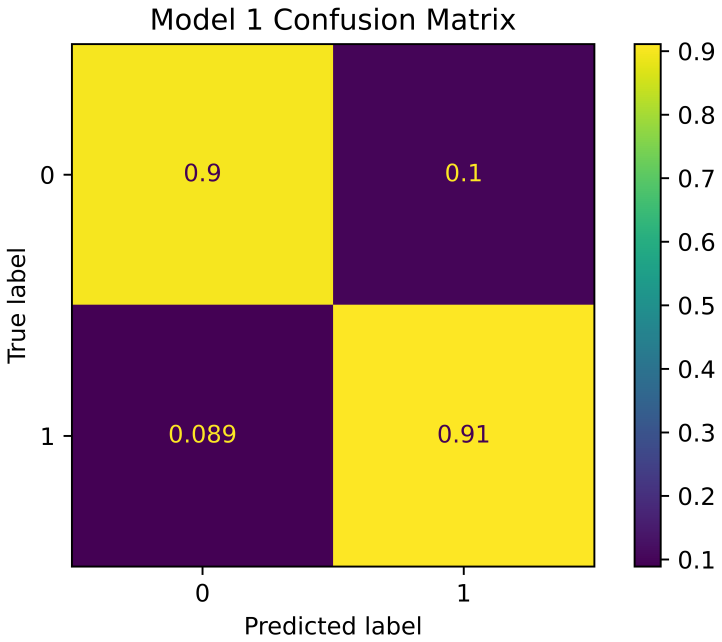

Fig. 7.—*The confusion matrix for Model 1 calculated over the validation data set at epoch 20 of training. The top left square to the bottom right square indicate the true negative (0.9), the false positive (0.1), false negative (0.089), and true positive rates (0.91). Model 1's strong performance illustrates its high level of understanding the difference between a nominal image and a GS fail image.*

4.2. Model 1 Sample Assessment and Prediction Confidence Distribution

We also assessed Model 1 by analyzing the distribution of the model's prediction probabilities and specific examples of images that the model classified correctly and incorrectly. Based on the results of the confusion matrix in Figure 7, we expect that the histogram of prediction confidence will show that Model 1 is classifying the images in the data set with a high level of confidence. Figure 8 illustrates the difference in distribution width; the prediction probability distribution for nominal images was wide while the probabilities for GS fails were concentrated near unity. This distribution supports our earlier results, and also shows that the model is very confident identifying the features that make an image a guide star failure.

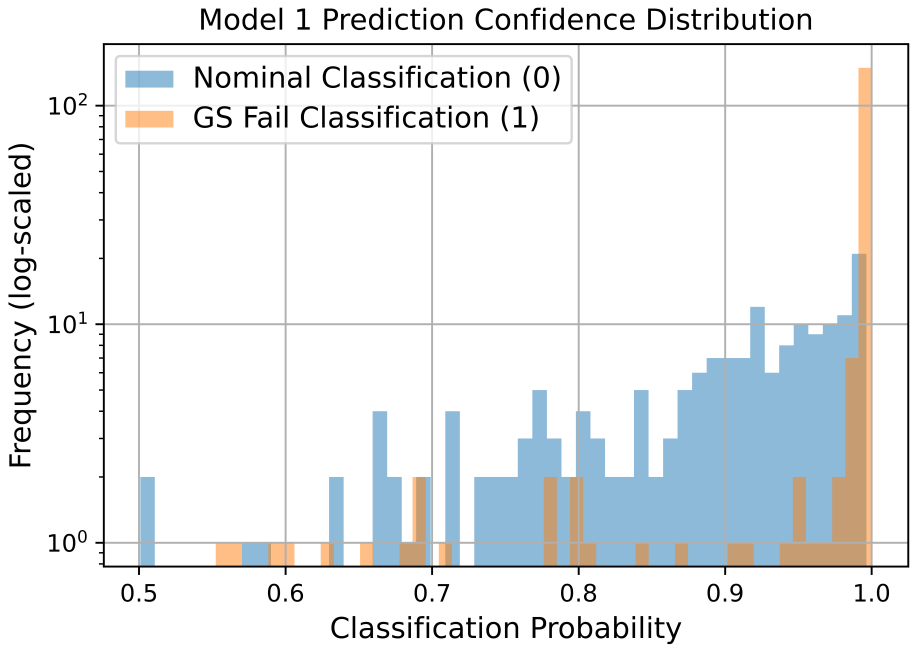

Fig. 8.—*The distribution of prediction confidence for Model 1 over images in the validation set. The distribution of prediction confidence for nominal images was more even than it was for the GS fail classifications, indicating that Model 1 typically classified GS fail images with a high level of confidence. Model 1 was less confident in classifying images as nominal due to the natural diversity of that image type.*

To take a closer look at how this model classifies images, we investigated specific examples of images that the model both correctly and incorrectly classified. Model 1 classified GS fail images with a high level of confidence, even when they were not extremely clear cases. In Figure 9 we can see examples where Model 1 classified images that were examples of GS fails with strong rolling, but also that the model correctly classified images with shorter streaks or fewer targets with nearly the same high level of confidence. Comparing these results to the prediction confidence for the images in Figure 10, we saw that the prediction confidence for nominal images was slightly lower overall, indicating that Model 1 had a stronger understanding of the features that make an image a GS fail.

In Figure 11, the images in the false positives category had clear GS fail features, indicating the features used for classification were consistent and as expected. Comparatively, the false negatives, shown in Figure 12, were likely images that were flagged as GS fails based on alerts from the internal HST AlertObs System, which is managed by the NASA Goddard Space Flight Center. This internal tracking system catalogs systematic failures

in the telescope's operations, which includes guide star acquisition failures. However, an observation taken during the period of a failure does not always contain the noticeable rolling feature, which was what we were interested in. In addition, since AlertObs does not automatically perform image quality checks, some observations flagged may be nominal. With that in mind, Model 1 and AlertObs combined may be enough to automatically flag all GS fails, and determine severity of rolling. The model may then be helpful to users with sparse observations by giving recommendations on which images may have been affected significantly enough to impact the quality of the science data within the image.

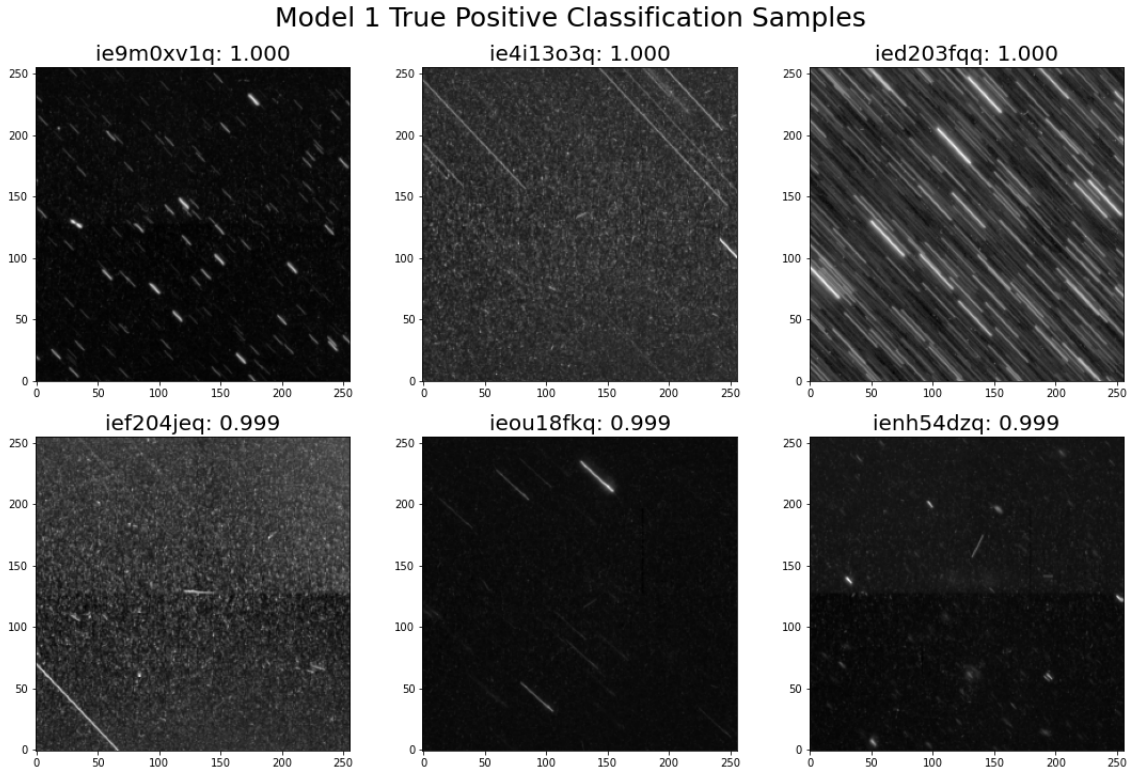

Fig. 9.—*True positive classifications made by Model 1 of images in the validation data set. Each image title contains the rootname of the image, followed by the prediction confidence of the model for that image. Images in this category included some GS failures with clear rolling features, such as the top left and top right images; as well as, some with less noticeable rolling, such as those in the bottom row. All of these images were classified by Model 1 with a high level of confidence, irrespective of the dramatic differences between the GS fail scenes.*

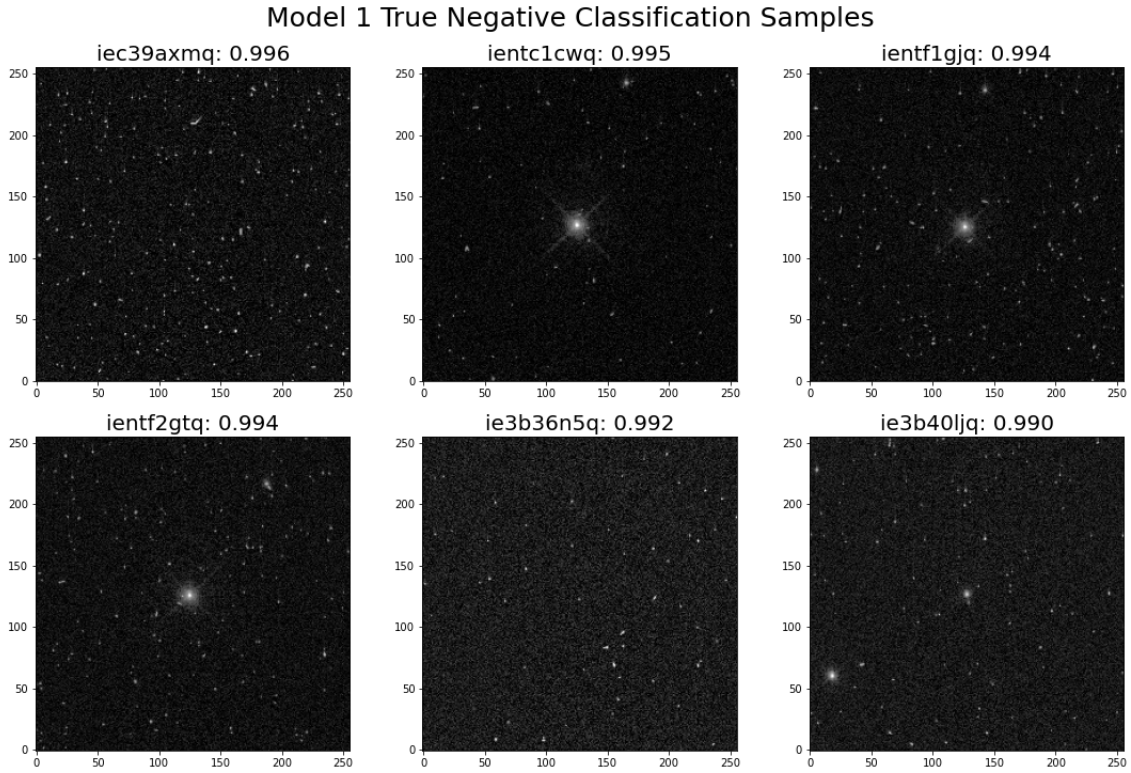

Fig. 10.—*True negative classifications made by Model 1 of images in the validation data set. Each image title contains the rootname of the image, followed by the prediction confidence of the model for that image. We saw a greater spread in the prediction confidence for nominal classifications by Model 1, which was also represented in Figure 8. The highest confidence nominal classifications by Model 1 typically included more simplistic images of one, or several stars, rather than galaxies or clusters.*

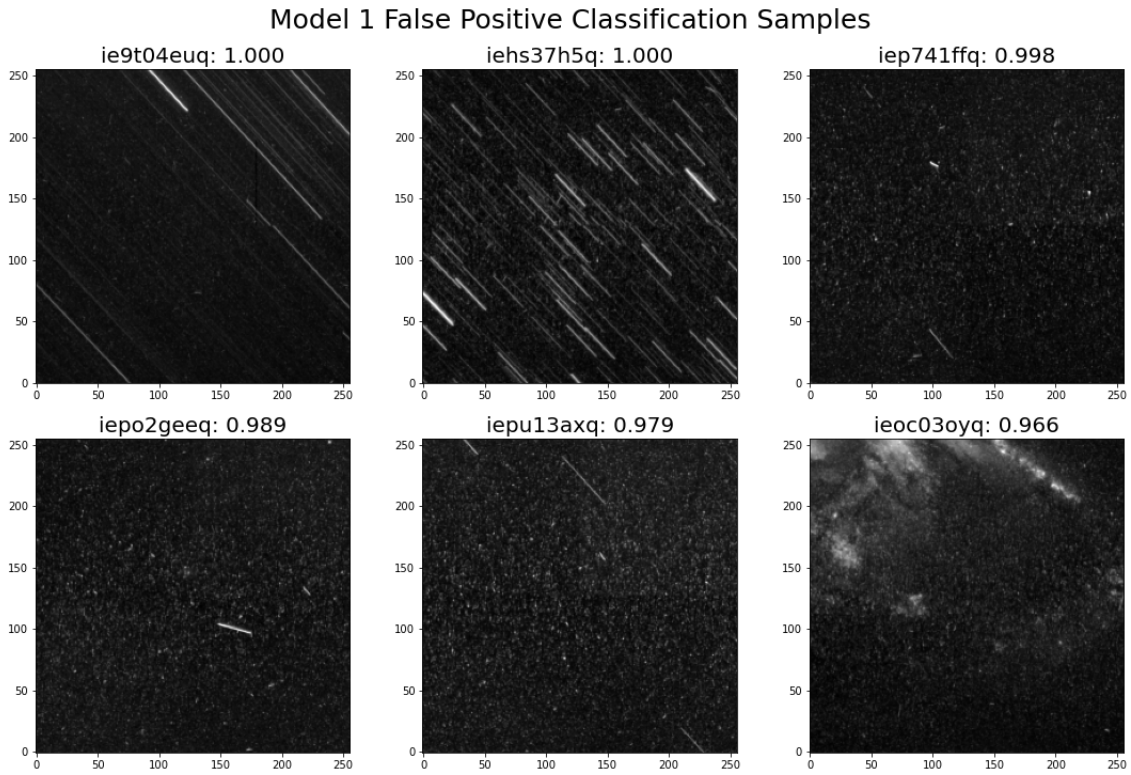

Fig. 11.—*False positive classifications made by Model 1 on images in the validation dataset. Each image title contains the rootname of the image, followed by the prediction confidence of the model for that image. The images in this set were images that were flagged as nominal in our data set, but were incorrectly classified as GS fails by the model. Typically these images included pieces of larger astronomical objects, or images that had one or more prominent cosmic rays in the image. However, the images in the top left, and center of this grid were classified with a confidence of 1.000, and could potentially be images that were incorrectly marked as nominal in our data set. If this is the case, then Model 1 could potentially help find other images that have been incorrectly flagged.*

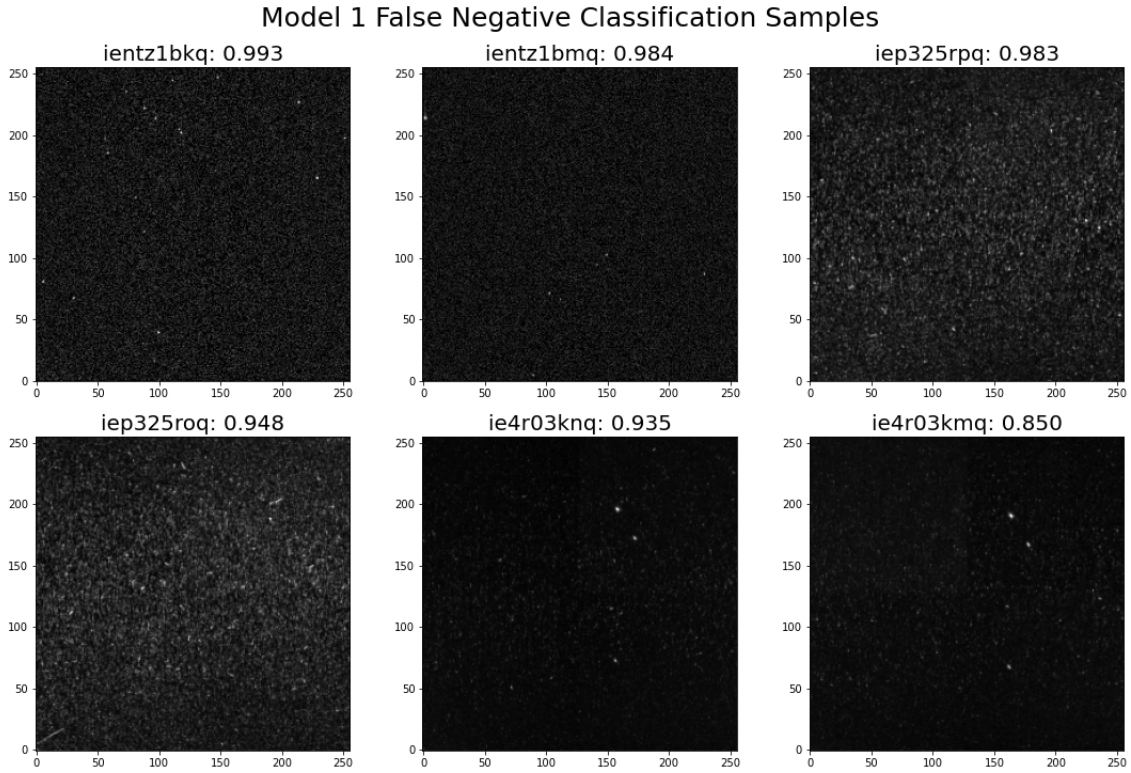

Fig. 12.—*False negative classifications made by Model 1 of images in the validation data set. Each image title contains the rootname of the image, followed by the prediction confidence of the model for that image. The images in this set were images that were flagged as GS fails in our data set, but were incorrectly classified as nominal by the model. Model 1 generally classified these images with a high level of confidence, and none of these images had significant GS fail features present. This result showed that many of the images in this category were images that the Quicklook team flagged based on the AlertObs System for HST. Despite being GS fails, the telescope did not drift far enough to generate streaked sources.*

4.3. Model 1 Deployment Test

To test Model 1's reliability in practice, we curated a final test set of WFC3/UVIS GO nominal and GS fail observations from June 1, 2022 to December 31, 2022, excluding moving targets, spatial scans, and grisms. The test set contained 2633 observations (2406 nominal; 227 GS fail). The true negative and false positive rates were consistent at 0.9 and 0.1, respectively. However, the false negative and true positive rates worsened to 0.25

and 0.75, respectively. The true negative and true positive samples were similar to their counterparts from the validation set, consisting of observations with clear nominal or GS fail features. The false positive samples consisted of observations with high sky backgrounds (e.g. nebulae, galaxies, star clusters) or with long cosmic rays. We also found some images with other anomalies, such as scattered light/dragon's breath and figure-8 ghosts, that may have been missed during initial Quicklooking, and thus were not tagged with having these anomalies (Gosmeyer, The Quicklook Team, 2017). The false negative samples either did not contain any noticeable rolling, or rolled at an angle other than 45 degrees with respect to the origin. Data augmentation is one solution for this problem, which we explore more in Section 4.5. Figure 13 illustrates some false positives and false negatives. Even though performance moderately decreased for GS fails, Model 1 can be a reliable tool for catching unusual observations and determining the severity of "rolled" objects within an observation.

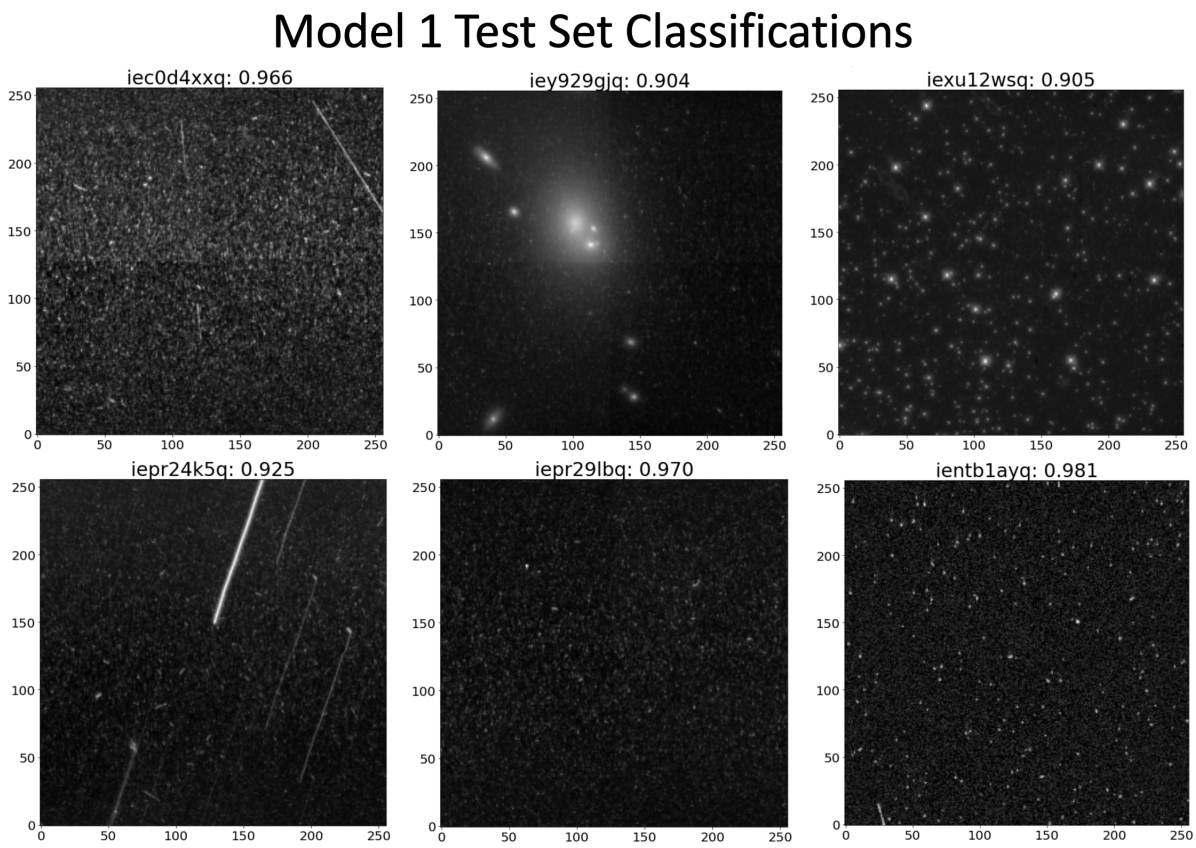

Fig. 13.—*False positive (top row) and false negative (bottom row) classifications made by Model 1 of images in the test set. Each image title contains the rootname of the image, followed by the prediction confidence of the model for that image. False positives included observations with long cosmic rays, galaxies, and missed anomalies such as figure-8 ghosts. False negatives included observations with rolling at angles other than 45 degrees with respect to the origin, or no noticeable rolling at all.*

4.4. Model 1 Rotational Tests

Lastly, we assessed Model 1 by taking an example of both nominal and GS fail images that the model classified with a high level of confidence (P>0.99), rotating the image to angles between [0,360], and predicting the classification probability on the rotated images. The results of this test for Model 1 are shown in Table 1 below.

Degree of Rotation	Nominal Image Classification Probability	Guide Star Failure Classification Probability
0	0.994	1.000
90	0.994	0.038
180	0.993	1.000
270	0.994	0.035

Table 1: *Prediction confidence of Model 1 when assessing the same nominal or guide star failure images rotated to different angles. The classification confidence for nominal images stays relatively stable for Model 1, but when a GS fail image is rotated to 90 or 270 degrees, it misclassified these images. This rotational dependence indicated that GS fails typically occur at the same angles in WFC3/UVIS data.*

This result illustrated that Model 1 did not struggle to correctly classify nominal images that were rotated, but did significantly struggle to classify the rotated GS fail image, which was expected. Due to the aging of Hubble's gyroscopes, GS fails typically occur at a 45 degree angle with respect to the detector. These results suggest that for the non-augmented data set, most of the GS fail images were in similar directions so the model was unable to identify them when they were in orientations that weren't well represented in the test data set.

4.5. Model 2 Results

Model 2 was trained on the augmented dataset. Loss and accuracy plots for this model, which are shown in Figure 14, showed that it was best at epoch 10 of training, so all of the results were assessed at this epoch of training. At this epoch, Model 2 reached about 85% accuracy. The confusion matrix for Model 2 in Figure 15 shows that it was fairly accurate in identifying GS fails, but that it only accurately classified 77% of the nominal images in the data set. This difference may originate from including grism and spatial scan images in Model 2's training data set. Based on these results, we found that training using augmented data to create a more balanced data set did not increase accuracy for both classes.

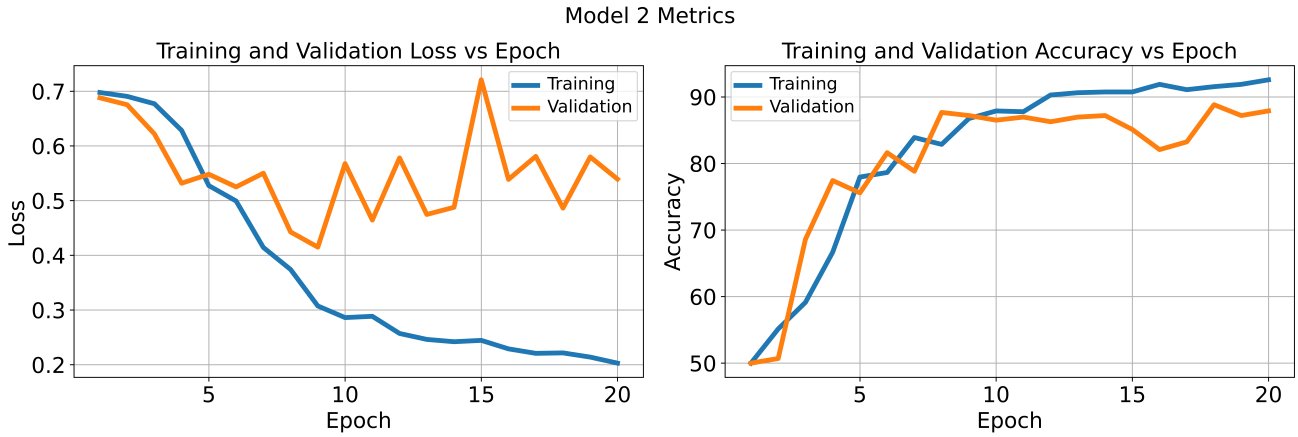

Fig. 14.—*Plots of epoch vs. loss (left) and accuracy (right) metrics for Model 2 training and validation. Epoch 10 was chosen as the stopping point for Model 2 because loss was relatively minimal in both training and validation near that epoch. Model 2 overfit our data about twice as fast as Model 1.*

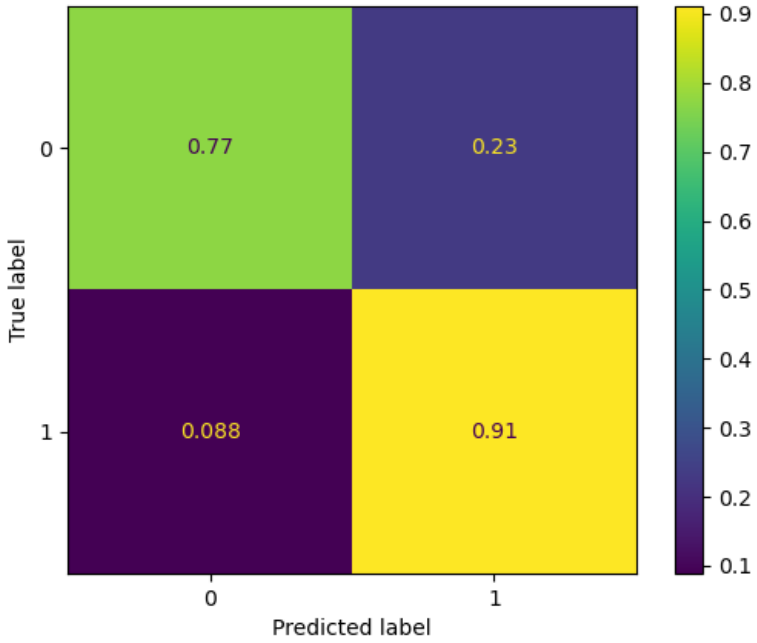

Fig. 15.—*The confusion matrix for Model 2 calculated over the validation data set. While Model 2 has a true positive rate equal to that of Model 1 (Figure 7), the true negative rate was significantly lower at 0.77. This result demonstrated that Model 2, which was trained on augmented data, had a much more difficult time correctly classifying nominal images. This difference may potentially be a result of including spatial scans and grism images in the training data set for this model.*

Since Model 2 was trained on the augmented data set, the rotational test was performed over the full 360 degree range. These results showed that Model 2 was much more accurate at classifying images within the validation set that had been rotated to varying angles since the classification confidence never dropped below 99%. Despite being much more accurate, there were still slight dips in classification confidence near 90 and 270, similar to the results from Model 1. However, Model 2, which trained on a data set containing rotated versions of the GS fail images, performed better when identifying rotated versions of the same image and further supports that the model generalized for all rotations.

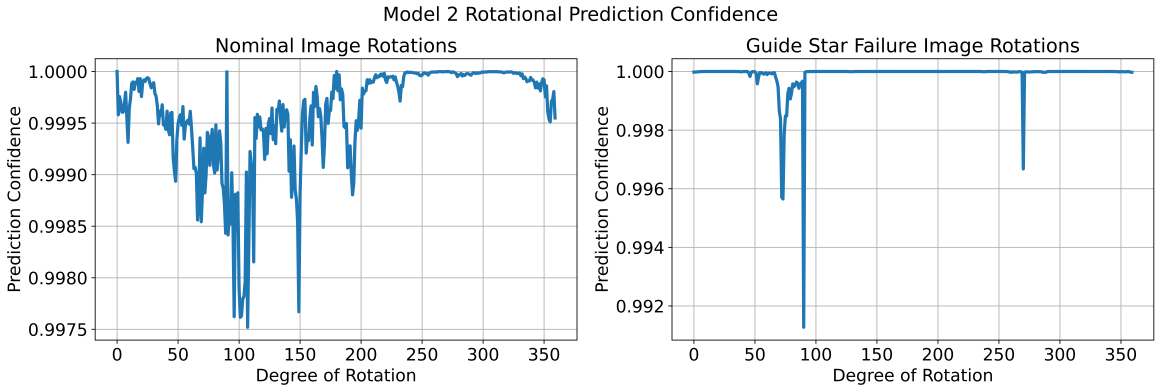

Fig. 16.—*Prediction confidence as a function of rotation angle. Note the range of the y-axis is between 0.99 and 1 for both plots (i.e. high confidence levels). Since Model 2 was trained on augmented data, it was possible to evaluate the images over the full range of possible rotation angles. These plots show that Model 2 was able to correctly classify rotated images with a much higher level of accuracy than Model 1 (Table 1), which misclassified the GS fail images when rotated to 90 and 270 degrees. Training on augmented data helped remove rotational bias that Model 1 experienced towards GS fails at a particular angle.*

4.6. Model 3 Results

Model 3 was trained to help us understand how including spatial scans and grism images in the data set affected model performance. The confusion matrix for Model 3, shown in Figure 17, illustrates that Model 3 correctly identified nominal images at a rate of 95%, but could only identify GS fails at a rate of 79%, which was much lower than Model 1. These results illustrated that including spatial scans and grism images in the training set caused the model to be less accurate in identifying GS fails and complicated distinguishing GS fails from nominal images.

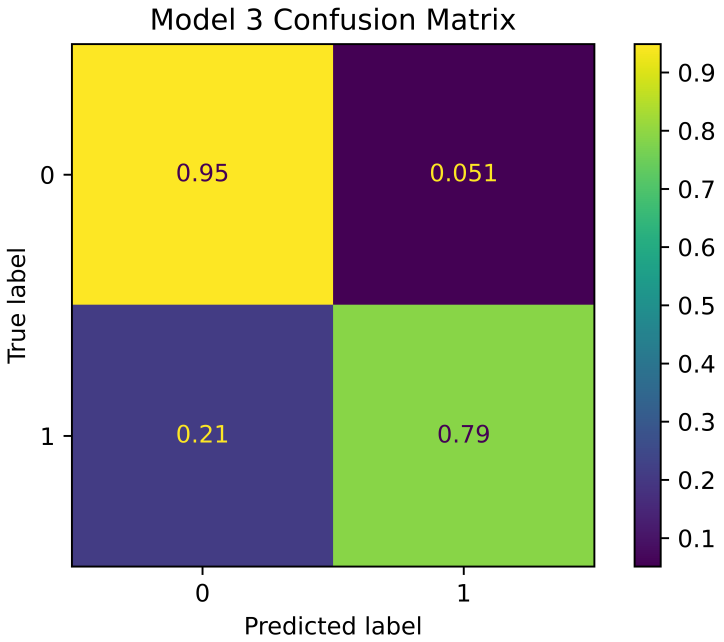

Fig. 17.—*The confusion matrix over the validation set for Model 3 at epoch 10 of training. This model trained on non-augmented data that included spatial scans and grism images, and had a higher true negative rate than Model 1 at 0.95. However, Model 3's true positive rate was much lower than Model 1's at 0.79, indicating that learning the features of spatial scans and grism images decreased performance on learning the features of GS fails.*

Similarly to Model 1, Model 3 also struggled to correctly classify images when they were rotated. The results shown in Table 2 show similar drops in prediction confidence seen in Model 1, which was expected.

Degree of Rotation	Nominal Image Classification Probability	Guide Star Failure Classification Probability
0	0.975	1.000
90	0.971	0.116
180	0.974	1.000
270	0.972	0.078

Table 2:*Prediction confidence of Model 3 when shown the same nominal or guide star failure images rotated to different angles. Model 3 was also trained on non-augmented data, and had difficulty correctly classifying GS failure images when rotated to 90 and 270 degrees similar to Model 1.*

5. Discussion

The models trained in this work showed strong results overall, but were not perfect. We plan to continue using machine learning algorithms to improve upon this work and help automate anomaly detection processes on Hubble, as well as on other future observatories. Potential further work includes exploring data augmentation methods, and a deeper look into the patterns and features that the models use to classify images.

5.1. Data Augmentation

The most interesting area to explore in further work would be to develop a more effective data augmentation method. Since the imbalance between GS fail and nominal images was so significant, we attempted to use augmented data to improve the model's performance, as described in Section 2.3. In this process, we cropped the augmented images to [180x180] pixels, which partially cut down the size of the images, but allowed us to create any new augmented versions of the images. Cutting the image size down may have affected the model's ability to train on them effectively. Some potential ways to solve this problem include:

- Rotate and crop the images prior to resizing them within the data processing pipeline (see Section 2.2), rather than after the images have already been processed.
- Use only four augmented copies by only rotating the images to 90, 180, or 270 degrees to avoid the need to crop images after rotating them.
- Incorporate synthetic data into the data sets. A method that convincingly smears nominal images into "GS fails" or simulates them is nontrivial, but could provide a way to increase data set sizes.

5.2. Further Assessment

Aside from data augmentation, a deeper analysis of our model on specific classification edge cases is desirable. For example, evaluating how the model performs on images that include other anomalies, such as figure-8 ghosts or satellite trails, would capture greater insight on reliability in operation. It may also be of interest to adjust the classification threshold of our models to see if this will help decrease the number of false negatives that would occur. In addition, computer vision evaluation techniques such as saliency maps would assist us in understanding exactly what features in an image the models use for classification

(Simonyan et al., 2014). The saliency maps may help reaffirm the conclusions we drew in Section 4 about how our models classified images. These two assessments would help give us a more well-rounded understanding of the functional performance of our models.

6. Conclusions

WFC3 has collected an abundance of amazing scientific data over the course of its lifetime onboard HST. However, as current observatories continue to age, and as newer observatories begin to produce substantial amounts of data, the current method of manually flagging guide star failure anomalies will no longer be sufficient. To address this gap, we trained machine learning models to identify WFC3/UVIS images affected by guide star failures. We developed a data processing and augmentation pipeline to reduce and standardize our data sets. Below is a summary of the models:

- Model 1, our best model, was trained on a non-augmented data set that did not include spatial scans or grism images. Model 1 performed with an accuracy of 90% and 83% with respect to the validation and test data.
- Misclassifications from Model 1 typically included images that were incorrectly flagged as nominal, or images that were flagged as guide star failures using Hubble's AlertObs System.
- Model 2 was trained on an augmented data set and had a Model 2 had a true positive rate of 0.91, which was on par with the results of Model 1, but a true negative rate of 0.77, which is significantly lower than that of Model 1. The lower true negative rate for Model 2 showed that the augmentation methods used in this project may not have been suitable to help the CNN train more effectively, although it did help to reduce the error seen in rotational tests of the model significantly.
- Model 3 trained on a non-augmented data set that included spatial scans and grism images. In the evaluations, it had a higher true negative rate than Model 1, at 0.95, but a much lower true positive rate at 0.79 suggesting that including spatial scans and grism images made this model focus more on learning the features of a nominal image than the features of a GS fail during training.

These results suggest that the use of augmented data and the inclusion of spatial scans and grism images in our data sets made it difficult for the models to learn the features of an image affected by guide star failure. We discussed potential improvements for our data augmentation methods, and further assessments to be performed on the models, such as:

- Rotating and cropping images prior to resizing them during the data processing pipeline, to prevent the loss of data.
- Creating fewer augmented images by restricting the rotation angles.
- Use saliency maps to better understand the ways that our models are classifying images.

We plan to continue implementing machine learning algorithms to improve the future of astronomical operations, and we recommend observers use the models available on GitHub to evaluate their own observations.

Acknowledgements

We thank Amanda Pagul and Joel Green for their careful review of the report. We also thank Sylvia Baggett for her review and initial conception of this work. In addition, we thank the STScI Instruments Division for the 2023 Internship Program, which this work took place. Lastly, we thank STScI's HST Machine Learning Group DeepHST for their discussion and recommendations in support of this work.

References

Dauphin F., Medina J.V., McCullough P.R., 2021, "WFC3 IR Blob Classification with Machine Learning", (WFC3-ISR 2021-08)

Dauphin F., Montes M., Easmin N., et al., 2022, "WFC3/UVIS Figure-8 Ghost Classification using Convolutional Neural Networks", (WFC3-ISR 2022-03).

Gosmeyer C.M., The Quicklook Team, 2017, "WFC3 Anomalies Flagged by the Quicklook Team", (WFC3-ISR 2017-22)

Sahu K., et al., 2021, "Wide Field Camera 3 Data Handbook, Version 5.0", (WFC3 DHB)

Kingma D.P., Ba J., 2014, "Adam: A Method for Stochastic Optimization", (arXiv: 1412.6980)

Lukic V., Bruggen M., Banfield J.K., et al., 2018, "Radio Galaxy Zoo: Compact and extended radio source classification with deep learning", (arXiv: 1801.04861)

Marinelli M., Dressel L. 2024. "Wide Field Camera 3 Instrument Handbook, Version 16.0" (WFC3 IHB)

Maslej-Kresnakova V., Bouchefry K., Butka P. 2021. "Morphological classification of compact and extended radio galaxies using convolutional neural networks and data augmentation techniques" (arXiv: 2107.00385)

Paillassa M., Bertin E., Bouy H. 2020. "MaxiMask and MaxiTrack: Two new tools for identifying contaminants in astronomical images using convolutional neural networks" (arXiv: 1907.08298)

Wang Y., Li M., Pan Z., Zheng J. 2019. "Pulsar Candidates Classification with Deep Convolutional Neural Networks" (arXiv: 1909.05301)

Simonyan K., Vedaldi A., Zisserman A., 2014, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", (arXiv: 1312.6034)