Draft version May 7, 2025
Typeset using LaTeX default style in AASTeX631

Determining the Best Method for Classifying Solar Flares Without Peak Energy Measurements

Aidan Floyd

ABSTRACT

In this paper I examine solar flare observation data from the Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI) dataset, taken from 2003 to 2004. I estimate the class of each flare in the dataset using the RHESSI X-ray flux counts and the known class of the most energetic flare in the dataset. I then plot the relationship between solar flare duration and total energy and utilize model fitting techniques to determine the most accurate model for flare classification. I find that the Bayesian Gaussian Mixture Model is the most accurate classification scheme. I also examine the relationship between total energy and peak X-ray flux by plotting the data in a log-log scale and fitting an optimized polynomial. The best fit polynomial is represented by the function $\log_{10} y = 0.08(\log_{10} x)^2 - 0.114\log_{10} x + 0.491$.

1. INTRODUCTION

Measureable solar flares are typically divided into 5 main classes: A, B, C, M, and X. Each class is then divided into subclasses such as C5 or M9 to denote the relative strength of flares within each class. This classification system is important when discussing solar flare predictions and their impact on Earth's environment; only M and X-class flares are energetic enough to have a noticeable effect on Earth. These flare types are also much more common during a solar maximum, or the period where there is the greatest amount of activity on the surface of the Sun.

Flare classification is typically based on the peak emissions of a solar flare in the soft X-ray (0.1 to 0.8 nm wavelength) spectrum. Flare classification is a logarithmic scale, with each class being 10 times more energetic than the last. Number subclassification is also based on peak flux and denotes the strength of the flare relative to other flares in the same class. For example, an X5-class flare would have a peak X-ray flux that is 5 times larger than an X1-class flare and 10 times larger than an M5-class flare. As such, accurate flare classification is reliant on taking observations of the soft X-ray flux of a flare. In this study I utilize various models to determine the accuracy of estimating flare classes using only flare duration and total energy rather than peak X-ray flux and determine the best model for accomplishing this task. I also examine the relationship between energy and peak flux.

To achieve this, I used flare data from the Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI). I chose this dataset due to its extensive record of flares across a large timescale. The dataset also includes measurements of duration, total energy, and peak soft X-ray flux for each flare event. This allows me to properly assess the accuracy of each model by comparing the results to the flares' actual peaks. For this study I took every RHESSI observation from January 1, 2003 to December 31, 2004. I chose this timescale because it is a solar maximum, giving me more flare events to model. The timescale also includes two energetic events, in October 2003 and July 2004, which I can use as a basis for calculating the true flare classifications.

2. METHODS

When determining the optimal classification method for flares given only duration and total energy, I first approximated the class of each flare in the dataset using RHESSI observations of peak soft X-ray emission. The RHESSI data listed peak energy observations in counts rather than standard flux units, so I approximated the flare classes using the known class of the highest value in the dataset. The highest value occurred on October 29, 2003, with a total count value of 9,662,335. This observation coincides with a known X45-class flare. By assuming that the highest value in the dataset represented this X45-class flare, I converted the RHESSI counts of each event to a flare class. I confirmed this conversion to be accurate by comparing my estimated classes to the known true classes of some large flares in the dataset, which I found were closely matching. I separated the flares into 5 separate classes, denoting A, B, C, M, and X-class flares.

I then utilized this estimation to examine 6 potential methods for flare classification. The classifiers I chose to examine for this study were K-Nearest Neighbors, Random Forest, Logistic Regression, Gaussian Naive Bayes, Bayesian Gaussian Mixture Model, and Quadratic Discriminant Analysis. For each classifier I used 5-fold cross-validation to determine the optimal fitting parameters. I then found a micro-averaged ROC curve for each method. To accomplish this I binarized the class data, with one flare class compared against all other flare classes, and calculated the ROC for that class. I repeated this process for each class, then found the average of all class ROC values. This method is helpful in determining the overall accuracy of each classification.

When determining the relationship between total energy, duration, and peak soft X-ray emission, I first plotted the values. I plotted total energy and peak soft X-ray emission on a logarithmic scale, while duration was plotted on a linear scale. After viewing each plot, it was clear that energy and peak emission were much more closely linked than energy and duration or duration and peak emission. I focused my study on examining the energy-peak relationship by fitting a polynomial to the data. I utilized 5-fold cross-validation to determine the optimal degree of polynomial for best fit.

3. RESULTS

In my ROC curve analysis of the various classification schemes, I found that the best-performing model was the Bayesian Gaussian Mixture Model, with a total AUC value of 0.992. The second-best model for this analysis was Random Forest with an AUC value of 0.985. Of the 6 models studied, the worst model for this dataset was K-Nearest Neighbor with a value of 0.963. These results are shown in Figure 1.

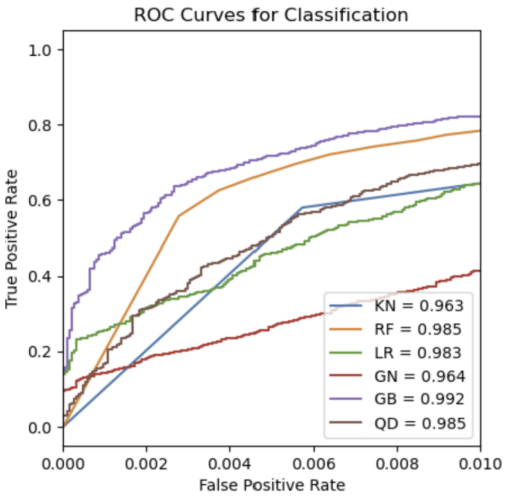

Figure 1. ROC curves for each classification scheme, labelled with measurements of the total area under the curve. KN denotes K-Nearest Neighbor, RF denotes Random Forest, LR denotes Logistic Regression, GN denotes Gaussian Naive Bayes, GB denotes Bayesian Gaussian Mixture Model, QD denoted Quadratic Discriminant Analysis. As seen by the total area under the curve, the Bayesian Guassian Mixture Model is the best fit for the data.

I compared the Bayesian Gaussian Mixture Model scheme with my calculated flare classes from the RHESSI peak emission measurements by plotting the duration-energy relationship twice, once utilizing my calculated classes and once utilizing the GMM estimation. These plots are shown in Figure 2. This side-by-side comparison shows that the GMM classification closely matches the true flare class distribution of the data, particularly on a larger scale, with some inaccuracies existing along the borders between flare classes. As expected, the model estimation seemed to smooth out the class borders overall, which is most clearly seen in the A-to-B and B-to-C class borders.

I then plotted the relationships both energy and duration have to the peak X-ray emission, and found that total energy and peak X-ray emission have an obvious correlation in a log-log scale. To determine the correlation, I fitted a polynomial to the data. I found that the best fit for this dataset consisted of a 2nd-degree polynomial, described by the following equation:

$$\log_{10} y = 0.08(\log_{10} x)^2 - 0.114\log_{10} x + 0.491 \tag{1}$$

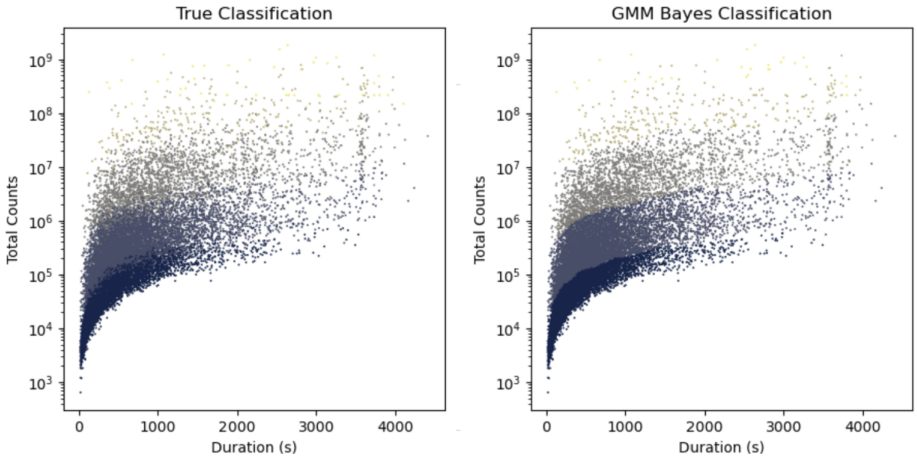

Figure 2. Left: Classifications of the dataset as a function of total energy and duration, calculated from the observed RHESSI peak X-ray counts. Right: Classifications of the dataset as a function of total energy and duration, estimated using Bayesian Gaussian Mixture Model analysis using optimized fitting parameters. The Bayesian analysis closely matches the class distribution present in the actual data, indicating a high level of reliability for this model.

This polynomial fit is shown in Figure 3.

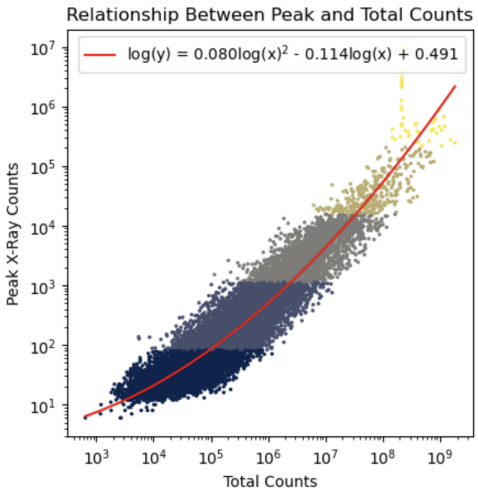

Figure 3. Relationship between total energy count and peak X-ray emission count, with a 2nd-degree best-fit polynomial.

4. DISCUSSION

The results show that a Bayesian Gaussian Mixture Model can predict a flare's class given only the duration and total energy with a high degree of accuracy. The model's hight AUC value of 0.992 compared to other potential models indicates that this model would be the preferred flare classification method for a dataset that does not have information on the peak soft X-ray flux. Most of the model's errors occurred at the border between flare classes, particularly for the less-energetic A and B flares. This indicates that the model's predictions for high or low subclasses within a class, such as a B1 or B9-class flare, should be viewed as less accurate, while central predictions such as B5 would be more accurate.

The results also show a clear log-log relationship between a flare's total energy and its peak soft X-ray flux, most accurately modeled by a 2nd-degree polynomial. The uncertainties of the polynomial fit are greater for low-energy flares and lesser for high-energy flares, so this polynomial could be used to predict the peak X-ray flux of high-energy flares with a relatively high degree of accuracy.

5. CONCLUSION

For future studies I would like to create confusion matrices for each potential classification scheme. The method used in this study compares the overall accuracy of each model, while a confusion matrix would allow me to examine the ability of each model to differentiate between specific flare classes. As shown in the figures above, the preferred Bayesian Gaussian Mixture Model had some difficulty with low-energy flares. Perhaps another model would be better at predicting low-energy flare classes, while the Bayesian model can be used for high-energy flares.

I would like to repeat this study over a solar minimum. This would allow me to determine if the energy and duration relationships present in this study are applicable to flares as a whole. It would also help to determine whether the ratio of low-peak to high-peak flares is the same for a solar maximum and minimum. If the ratios are similar this would indicate that solar activity levels only impact the frequency of flares as a whole rather than the average energy of each flare.

I would also like to examine the trends of peak X-ray energy, total energy, and total duration over a long period of time. This would allow me to determine the effect of the solar cycle on the accuracy of my model predictions and on the energy-peak relationship. Identifying patterns within the solar cycle would also be helpful in predicting the impact of flares on Earth in the near future.