Using Multiple Instance Learning for Explainable Solar Flare Prediction

C. Huwyler[1], M. Melchior

Institute for Data Science, Fachhochschule Nordwestschweiz (FHNW)
Bahnhofstrasse 6, 5210 Windisch
Switzerland

Abstract

In this work we leverage a weakly-labeled dataset of spectral data from NASA's IRIS satellite for the prediction of solar flares using the Multiple Instance Learning (MIL) paradigm. While standard supervised learning models expect a label for every instance, MIL relaxes this and only considers bags of instances to be labeled. This is ideally suited for flare prediction with IRIS data that consists of time series of bags of UV spectra measured along the instrument slit. In particular, we consider the readout window around the Mg II h&k lines that encodes information on the dynamics of the solar chromosphere. Our MIL models are not only able to predict whether flares occur within the next $\sim$25 minutes with accuracies of around 90%, but are also able to explain which spectral profiles were particularly important for their bag-level prediction. This information can be used to highlight regions of interest in ongoing IRIS observations in real-time and to identify candidates for typical flare precursor spectral profiles. We use k-means clustering to extract groups of spectral profiles that appear relevant for flare prediction. The recovered groups show high intensity, triplet red wing emission and single-peaked h and k lines, as found by previous works. They seem to be related to small-scale explosive events that have been reported to occur tens of minutes before a flare.

Keywords: Solar flares, Multiple Instance Learning

1. Introduction

The Multiple Instance Learning (MIL) paradigm has originated from drug activity prediction problems (Dietterich et al., 1997) and has recently gathered attention in the domain of image segmentation (e.g. Oquab et al. (2015); Pinheiro and Collobert (2015)) and especially in medical imaging (e.g. Wang et al. (2017); Sadafi et al. (2020); Liu et al. (2018); Sudharshan et al. (2019); Kanavati et al. (2021)). In the MIL setting, label information is not available for single instances, but only for groups or so called *bags* of multiple instances. As such, MIL is part of the weakly supervised learning domain (Zhou, 2018). The relationship between individual instances and the overall bag label is made explicit by modeling how instance-level information is aggregated to bag-level information (Carbonneau et al., 2018; Amores, 2013; Foulds and Frank, 2010). This allows MIL models to use the available bag-level label information to reveal the hidden instance-level labels. In the image segmentation domain, a pixel-based segmentation map is learned only from an image-level label. Recently, an attention-based MIL model has been proposed that shows supreme bag-level performance and allows to easily produce saliency maps that show on which instances the model focuses to make a particular bag-level prediction (Ilse et al., 2018).

Physics-based modeling of observed phenomena is usually done in a bottom-up fashion where the interactions between individual microscopic instances are modelled and then aggregated to a macroscopic quantity. The results on the macroscopic level are then compared to actual observations. In contrast, data-driven approaches are usually only able to model physical phenomena in a top-down fashion where the macroscopic quantities (such as bag labels) are known but no information about the states of the microscopic quantities (individual instances) is available. Often data-driven approaches only reveal poorly how they model the relationship between individual instances to reach their bag-level prediction. This frequently hinders such *black box* models to get accepted in the broader research community. To this end MIL is a very attractive candidate, since it makes the relationship between bags and instances manifest and is able to discover key instances in the decision process of the model.

In this work we apply the MIL paradigm to the problem of solar flare prediction. Solar flares are sudden and intense bursts of radiation from the Sun across the entire electromagnetic spectrum that are created by rapid downflow and subsequent deceleration of hot plasma triggered by reconnection of magnetic loops (Benz, 2016). They are often (but not necessarily) accompanied by the ejection of plasma into space through the solar corona (*coronal mass ejection*) and occur mainly in so-called *active regions* of the Sun. The magnitude of solar flares is conventionally

---

[1]Email: cedric.huwyler@fhnw.ch

indicated by the peak flux of the X-ray radiation emitted at wavelengths between 1-8 Å and grouped into the classes X, M, C, B and A with the peak flux decreasing on a logarithmic scale. The high amounts of radiation and charged particles from strong solar flares and coronal mass ejections interact with Earth's magnetosphere and can lead to severe geomagnetic storms that last for several hours and can have widespread effects on terrestrial infrastructure and the near-Earth space environment (Lanzerotti, 2007; Pulkkinen, 2007). The ability to predict strong solar flares of class X and M tens of minutes if not hours in advance would yield valuable preparation time.

The prediction of solar flares is an active topic in the space weather community (Ahmed et al., 2013; Bobra and Couvidat, 2015; Boucheron et al., 2015; Nishizuka et al., 2017; Liu et al., 2017; Jonas et al., 2018; Florios et al., 2018; Huang et al., 2018; Liu et al., 2019; Armstrong and Fletcher, 2019; Galvez et al., 2019; Chen et al., 2019; Ahmadzadeh et al., 2021). To this end, solar data recorded in different wavelenghts and spatio-temporal resolutions is gathered by a fleet of satellites and an array of telescopes. Prediction of solar flares was so far mainly undertaken with data from the *solar dynamics observatory* (SDO) (Pesnell et al., 2012) that continually observes the full solar disk at different ultraviolet (UV) passbands. Various levels of success are reported that are often difficult to compare because they are based on different pre-processing pipelines, different training and validation set splits and different performance metrics.
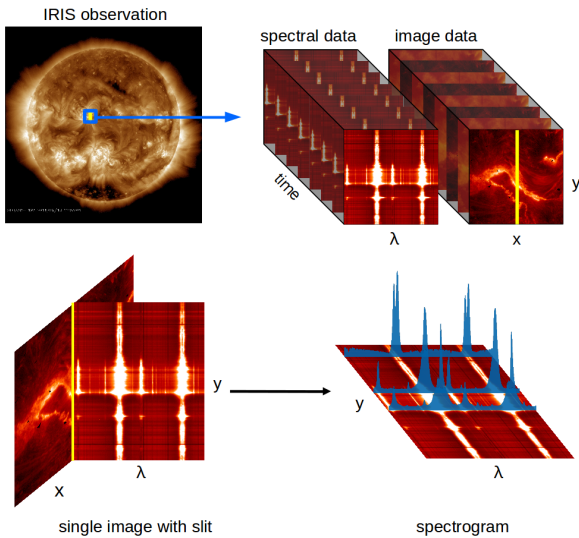

Figure 1: How IRIS records data.
Top: IRIS observes time series of small particular high-resolution patches of the solar surface according to a predefined observation schedule (left). It provides us with a spatio-temporal datacube of spectrograms along the slit accompanied by a data cube of images of the region around the slit (right).
Bottom: Spectrograms are only measured along the position of the slit (yellow) in the center of the recorded image (left). A spectrogram contains a spectrum for each *y*-pixel along the slit (right).
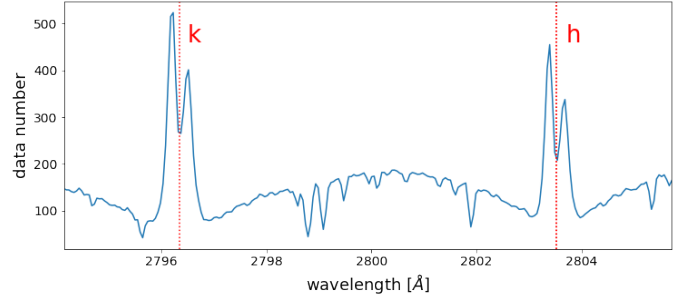

Figure 2: Typical example of a quiet-sun spectral profile in the Mg II readout window. The Mg II k and h lines (left and right peaks) are clearly visible and show the typical central core reversal.

We focus in this work on high-resolution spectral data from the near-ultraviolet regime, specifically on a readout window around the Magnesium (Mg) II k and h line cores centered at 2796.34 and 2803.52 Å (Fig. 2 shows an example quiet-sun spectral profile). This data was collected by NASA's *Interface Region Imaging Spectrograph* (IRIS) satellite that is designed to observe the transition region between the solar chromosphere and the corona and provides an unprecedented spectral and spatio-temporal resolution at a two-second cadence along a vertical slit pointed onto a small patch on the Sun (De Pontieu et al., 2014). In addition, also high-resolution images of the region around the slit (*slit-jaw images, SJI*) in up to four wavebands are recorded (Fig. 1). The Mg II h&k window encodes information on the dynamics of the entire chromosphere (Vernazza et al., 1981; Leenaarts et al., 2013a,b; Pereira et al., 2013), such as plasma velocities, temperatures and formation heights. A recent study found a set of unique spectral shapes in this window that clearly indicate the presence of a flare (Panos et al., 2018). A further study based on similar data has used specific expert-based features to show that flare prediction in the Mg II h&k readout window could in principle be possible (Panos and Kleint, 2020). In particular a decreasing ratio between the Mg k and h integrated intensities seems to indicate an upcoming flare already before its impulsive phase.

While spectral data from IRIS comes at an unprecedented spatio-temporal resolution in the UV regime, it is not perfectly suited for machine learning. In particular, three main challenges render the application of standard supervised machine learning models difficult:

Weak labels. When working with IRIS data, the individual spectra seem to provide a natural fundamental unit of data. However, labels are not available on the level of individual spectra, but only on the level of entire observations (e.g. non-flaring active region, pre-flare or flare). It is unclear how individual spectra are responsible for the overall label of an observation.

Heterogeneous data. In contrast to SDO that continually observes the full solar disk, IRIS focuses on a small field-

of-view (FOV) of up to 175 arcsec × 175 arcsec (see Fig. 1) that is defined according to an observation schedule. Limited by IRIS' downlink rate, a trade-off between spatial, temporal and spectral resolution has to be made by the observer, resulting in different settings for each observation. In addition, IRIS operates in different observing modes (fixed *sit-and-stare* or moving *n-step raster*).

Class imbalance. The typical class imbalance to encounter in observational data from the Sun is even more pronounced. Only 10 X-class and 57 M-class flares happened to be observed by IRIS over the last solar cycle[2]. A significant part of these flares do not qualify for our dataset (see criteria in section 2), reducing the available observations to 6 X-class and 13 M-class flares. This is a significantly low number in contrast to the total number of ~ 30000 observations since the commissioning of IRIS in 2013 and the total number of observed active regions that we estimate to be in the hundreds.

In this work we address the problem of prediction given only weak labels using the MIL paradigm. To this end we treat each timestep of an observation as an unsorted bag of individual spectra. This allows us to assign bag labels to active regions that are flaring or non-flaring in the near future. The resulting dataset can then be used to learn prediction models and to analyze what groups of spectral profiles dominate the decision process. In addition, the chosen approach uses a pooling function which allows to deal with differently sized bags of spectra. To our knowledge, we are the first to apply the MIL paradigm to the solar flare prediction problem.

2. Dataset

2.1. Selection of sample observations

Our dataset consists of 37 IRIS observations labeled either as (non-flaring) active region (AR) or pre-flare (PF) active region (listed in tables A.1 and A.2) with tens to hundreds of timesteps. Selecting representative observations for each class is a challenge because they need to satisfy a range of criteria: the FOV and especially the vertical slit should be centered on events that are characteristic for the class, the measured data should contain the Mg II h&k readout window and they should not be located too close to the solar limb[3]. These criteria are only satisfied by 6 X-class flares and 13 M-class flares, resulting in a total of 19 PF observations. To mitigate class imbalance we selected a representative sample of 18 AR observations that do not flare within the next couple of hours, identical

to the one used by Panos and Kleint (2020). Also similar to Panos and Kleint (2020), 25 minutes from each observation that are representative for the given class have manually been extracted, i.e. every timestep of the observation should encode some information about whether there is going to be a flare or not within the next 25 minutes. This allows us to broadcast the label of each observation (AR or PF) to each of its timesteps and enables us to expand the 37 labeled observations to roughly 10'000 labeled bags containing roughly 600 instances on average.

2.2. Data pre-processing

As introduced, IRIS' distinguishing feature is its spectrograph that refracts light along a vertical slit into its different wavelength components. The vertical slit (indicated in Fig. 1 with a yellow line) along which spectral data is measured is typically composed of a few hundred vertically aligned pixels. Hence, IRIS spectral data comes in the form of spatio-temporal data cubes that contain one time axis t, the y-axis as a spatial axis (along the vertical slit) and a spectral axis λ.

All selected observations in tables A.1 and A.2 include spectral readout windows around the Mg II h and k line cores. As the exact boundaries of the Mg II h&k readout window differ from observation to observation, we extracted the specific window between 2794.14-2805.72 Å that is contained in all observations and linearly interpolated the available spectral bins to a common length of 240 equally-spaced bins. We removed spectra that contain missing values (at least one value < -100) or are underexposed (less than 10 counts per bin) and clipped spectra with negative values to zero. Furthermore, we normalized the spectral profiles to the range $[0, 1]$ and hence let our models focus more on the shape of spectra and on relative intensity differences. Note that the normalized profiles still provide access to a proxy of the absolute intensities of the Mg II h&k peaks through the scaled intensity of the continuum radiation between the two peaks. The h and k peak intensities are strong predictors since they are strongly correlated with chromospheric gas temperatures.

The selected and pre-processed slices are stored in three-dimensional spatio-temporal data cubes with dimensions $[t, y, s]$, where t is a time-step index, y is the y-pixel on the camera and s is the spectral bin. All the data preprocessing pipelines were set up with the `irisreader` Python library[4].

2.3. Sampling and cross-validation

Because the number of available observations is very limited and does most certainly not capture the full variability in the dynamics leading to a flare, splitting of the data into a training and validation set must be handled with care. We manually divided our selected set of observation time windows into different observation groups (as

---

[2]According to the IRIS flare list distributed on `iris.lmsal.com`

[3]Since spectra are the superposition of radiation absorbing and emitting layers of the solar atmosphere, spectra around the solar center are built up from vertical layers of plasma while spectra on the solar limb are built up from horizontal layers and show different characteristics.

[4]`https://github.com/i4Ds/IRISreader`

indicated in tables A.1 and A.2) that are then mapped to the training and validation sets. Observation time windows that are less than ten hours apart and spatially close are assigned to the same group to make sure that no information leaks from the training to the validation set in a later train-validation split and artificially decreases the validation error. We chose to randomly split the observation *groups* into three cross-validation folds stratified by flare class and train of three models, each with two folds as the training set and one as the validation set. This gives us access to a validation set performance for each observation.

The selected observation time windows come at various number of steps (see tables A.1 and A.2). We created a dataset that is balanced towards the classes and the groups of observations therein. To this end, the dataset is composed of 10'000 bags, equally split into the classes AR and PF. The bags for each class are then equally divided among the available observation groups and among the groups again equally distributed over the available observation time windows. If the number of bags to sample was smaller or equal to the number of available timesteps in an observation, we subsampled without replacement, otherwise we sampled with replacement and took oversampling from a few observations into account.

3. Method

Previous work has asserted that each Mg II h&k profile observed by IRIS can serve as a diagnostic for the physics in the solar chromosphere and can be related to different parameters of the solar atmosphere such as temperatures and velocities at different heights. In addition, it has been shown that particular types of profiles only appear around flares (Panos et al., 2018; Panos and Kleint, 2020). This leads us to the assumption that individual spectra can be treated as fundamental *instances* with predictive power for flare prediction with IRIS data. However, individual spectra are not labeled and class labels (AR / PF) only exist on the level of the individual timesteps. Each timestep is composed by a *bag* of a few hundred spectra along the instrument slit. Here, Multiple Instance Learning (MIL) offers a natural choice for investigating flare prediction, and, in particular, for identifying groups of spectra that are key to the prediction process.

3.1. Multiple Instance Learning (MIL)

In a standard classification setting, individual instances are thought to belong to either a positive or a negative class. In a MIL setting, this information is considered latent and hidden from us. Labels are only available at the *bag-level*, but not at the *instance-level*. In the so-called *standard MIL assumption* (Foulds and Frank, 2010), bags are labeled as negative if they only contain negative instances and positive, if they contain at least one positive

instance. This particular (asymmetric) way of propagating instance-level information to the bag-level puts a lot of emphasis on positively labeled instances.
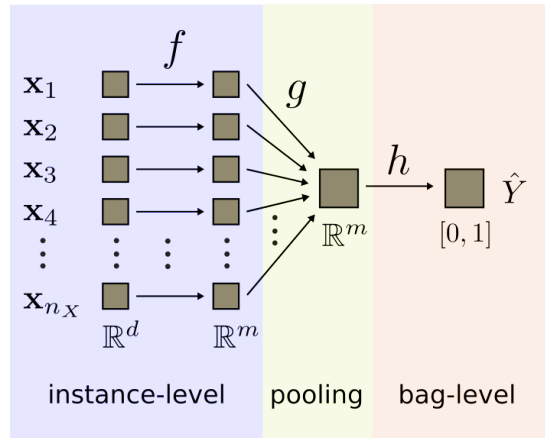

Figure 3: Typical setup of a MIL model: A bag X of instances $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1,\cdots,n_X}$ is individually mapped by an instance-level function $f : \mathbb{R}^d \to \mathbb{R}^m$, aggregated to a single bag-level representation by a pooling function $g : \mathbb{R}^m \times \cdots \times \mathbb{R}^m \to \mathbb{R}^m$ and converted to a probability by a bag-level classifier $h : \mathbb{R}^m \to [0,1]$. f, g and h can be represented by neural networks.

To deal with a wider variety of problems, the standard MIL assumption can be relaxed by adopting e.g. a *presence-based*, a *threshold-based* or a *collective* assumption (see Foulds and Frank (2010) for a review). The threshold-based and collective assumption relax the standard MIL assumption by allowing also a few positive instances in negatively labeled bags, what is often closer to reality.

In the following, we denote quantities at the bag-level with capital letters and quantities at the instance-level with lowercase letters. We denote a bag of instances as $X = \{\mathbf{x}_k\}_{k=1,...,n_X}$ where $\mathbf{x}_k \in \mathbb{R}^d$ is the representation of a single instance and n_X is the number of instances in bag X. Individual spectral profiles can be represented as vectors by assigning the i-th interpolated spectral bin to the i-th component of $\mathbf{x}_k$, resulting in vectors with $d=240$ dimensions. We attach a label $Y \in \{0,1\}$ to each bag X and choose 0 to encode non-flaring active regions (AR) and 1 to encode pre-flare active regions (PF).

MIL models are usually expressed with a *scoring function* S (Ilse et al., 2018) which predicts the label $\hat{Y}$ of bag X as

$$\hat{Y} = S(X), \qquad (1)$$

and is trained by minimizing a loss $L(Y, \hat{Y})$. We decompose S into a function f that operates on the instance-level, a permutation-invariant pooling function g that maps from instance to bag-level and a classifier h that operates on the bag-level (Fig. 3), hence

$$S = h_\psi \circ g_\varphi \circ f_\theta. \qquad (2)$$

The components h_ψ, g_φ, f_θ are represented by neural networks parameterized by weights θ, φ and ψ. The decomposition of S is at the heart of this MIL formulation, as it acknowledges that the bag label is an aggregate of the information residing in its different instances.

MIL approaches are usually subdivided into two main branches (Foulds and Frank, 2010): *instance-based* and *embedding-based*. Below we select and analyze two models, one of each branch.

3.1.1. Instance-based model

By design, instance-based models provide the ability to propagate bag labels down to the hidden instance labels. This helps to increase the interpretability of the decisions taken by the model and allows to isolate instances that are responsible for a positive bag label. Prediction at instance-level requires that f acts as a classifier and assigns a probability $p_k \in [0,1]$ to each instance by $p_k = f_\theta(\mathbf{x}_k)$ for belonging to the positive class. The inductive bias is chosen that all learnable parameters are in f_θ, while none are placed in g or h. This effectively restricts g to be a pooling function free of learnable parameters and h to be the identity function. The function g is chosen in accordance with the MIL assumption to be adopted: here one could choose $g(p_k) = \max_k p_k$ to satisfy the standard MIL assumption (*max pooling*, one single spectral profile in a bag can already indicate a flare) or $g(p_k) = \frac{1}{n_X} \sum_k p_k$ as a simple example that satisfies the collective assumption (*average pooling*, a flare requires a few strong or many weaker flare-indicating spectral profiles). While the first approach is hard to train (Pinheiro and Collobert, 2015) and does not reflect reality, the second does not account for the fact that a few particular spectra might actually be enough to indicate a flare and that a majority of active spectra cannot replace a few very active spectra. As the answer should lie somewhere between the mean and the maximum, we used a smooth interpolation between them with a free parameter r (Pinheiro and Collobert, 2015; Ramon and De Raedt, 2000):

$$\hat{Y}^{(r)} = g(\{f_\theta(\mathbf{x}_k)\}) = \frac{1}{r} \log\left(\frac{1}{n_X}\sum_k e^{r\,f_\theta(\mathbf{x}_k)}\right), \quad (3)$$

where r interpolates between the mean ($r \to 0$) and the maximum ($r \to \infty$) and can be considered as a tuneable hyperparameter that selects the number of highest probability instances to be considered for classifying the bag. This function is often called 'Log-Sum-Exp' in the literature (e.g. in Pinheiro and Collobert (2015)), however we find 'Log-Mean-Exp' more adequate.

3.1.2. Embedding-based model

Embedding-based approaches give up the requirement of predicting at the instance-level and instead try to achieve the best possible performance at the bag-level, without restricting f to return scalar probabilities. To this end they use f to map the individual instances $\mathbf{x}_k$ to an embedding $\mathbf{e}_k \in \mathbb{R}^m$ that is most suitable for bag classification:

$$\mathbf{e}_k = f_\theta(\mathbf{x}_k).$$

The instance-level embeddings are then aggregated to a single bag embedding $\mathbf{E} = g_\varphi(\{\mathbf{e}_k\})$ that is then used as input for a bag-level classifier network h_ψ with

$$\hat{Y} = h_\psi(\mathbf{E}) = h_\psi(g_\phi(\{f_\theta(\mathbf{x}_k)\})). \quad (4)$$

For this work we chose the attention-based pooling proposed by Ilse et al. (2018) that works under the *weighted* collective assumption (Foulds and Frank, 2010):

$$g_\phi(\{\mathbf{e}_k\}) = \sum_{k=1}^{n_X} a_k\,\mathbf{e}_k, \quad (5)$$

where the weights a_k with $\sum_{k=1}^{n_X} a_k = 1$ are computed by a gated attention mechanism

$$a_k = \frac{\exp\left[\mathbf{w}^T\left(\tanh\left(\mathbf{V}\mathbf{e}_k\right) \odot \sigma\left(\mathbf{U}\mathbf{e}_k\right)\right)\right]}{\sum_{j=1}^{n_X} \exp\left[\mathbf{w}^T\left(\tanh\left(\mathbf{V}\mathbf{e}_j\right) \odot \sigma\left(\mathbf{U}\mathbf{e}_j\right)\right)\right]}, \quad (6)$$

with learned weights $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{l \times m}$ and $\mathbf{w} \in \mathbb{R}^l$ that can be summarized as a set of pooling parameters $\phi = \{\mathbf{U}, \mathbf{V}, \mathbf{w}\}$. Because the tanh-activated network alone is not sufficiently non-linear to model complex relationships among instances, it is additionally gated by a sigmoid activation function $\sigma(\cdot)$ (Ilse et al., 2018). The free parameter l controls the complexity of the attention mechanism. In addition, we introduce a sharpening parameter γ to control the skewness of the a_k by setting $a_k \mapsto a_k^\gamma/(\sum_{j=1}^{n_X} a_j^\gamma)$, ensuring that the sum remains normalized to 1.

While with embedding-based models we give up the possibility of interpretability at the instance-level, the use of an attention mechanism gives us back some interpretability in form of the per-instance weights in the pooling function.

3.2. Architecture choices for f_θ, g_ϕ and h_ψ

Given the limited dataset with the limited number of observations, we chose rather simple architectures that are nevertheless complex enough to potentially overfit on the training data. Besides the use of a dropout layer (Srivastava et al., 2014), we only applied early stopping on the validation loss as a regularisation measure. We employed both the instance-based (further called *ibMIL*) and the attention-based (further called *abMIL*) MIL model as introduced in 3.1. The ibMIL model includes r as model-specific hyperparameter. The abMIL model involves the hyperparameters l (complexity of attention mechanism), m (dimension of embedding) and γ. In this work, we chose $l = 10$ and $m = 10$. The model could definitely profit from a fine-tuning of l and m, however we leave this for later work as it does not change the qualitative results. The role of the parameters r and γ will be discussed later.
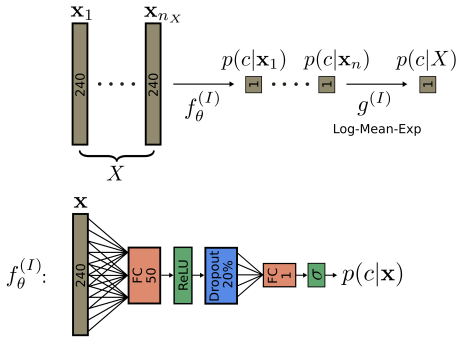
5

Figure 4: Chosen ibMIL model architecture.

Figure 5: Chosen abMIL model architecture.

Our simple model architectures consist only of fully-connected layers, as the most important spectral features can be extracted purely from counts at individual wavelengths and do not require the consideration of neighbourhood relations. Nevertheless, convolutional features both in the spectral and spatial domain could potentially improve the predictions and will be explored in future work. For both models we used a 50 nodes fully-connected (FC) hidden layer with ReLU activation in f_θ followed by 20% of dropout. For ibMIL we then applied a single fully-connected node with sigmoid activation, for abMIL another ReLU-activated layer with m nodes to map input instances to the m-dimensional embedding. The chosen model architectures are sketched in Figs. 4 (ibMIL) and 5 (abMIL).

3.3. Dealing with heterogeneous bag sizes

Because the amount of spectral pixels on the y-axis per timestep is only constant per observation but varies between different observations, our models need to deal with various bag sizes. At the theoretical level this is not a problem, as the pooling function g does not care about the number of entities it has to aggregate. However deep learning frameworks such as Keras make it much easier to deal with layers of constant input dimension. To this end we zero-pad all the timesteps in our sample to 1100 pixels and store the location of the original data as binary masks. For implementation details we refer to appendix Appendix A.

3.3.1. Model training

Instead of a binary cross-entropy loss, we chose a mean absolute error (MAE) loss

$$L(\hat{Y}, Y) = \frac{1}{n}|\hat{Y}_i - Y_i|. \tag{7}$$

We found that optimizing MAE loss leads to more stable training. MAE is generally known to perform better in settings with noisy labels (Karimi et al., 2020; Song et al., 2022) and we expect that this helps to better deal with outliers that our model cannot identify correctly.
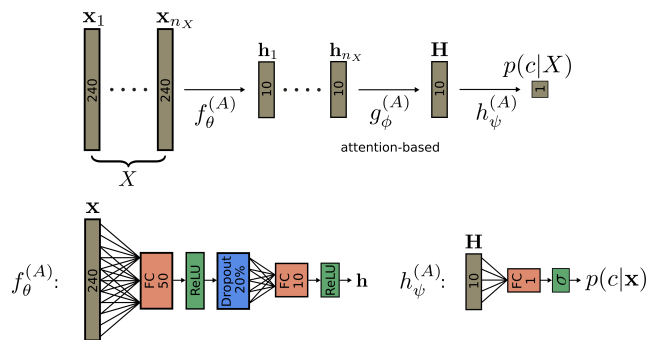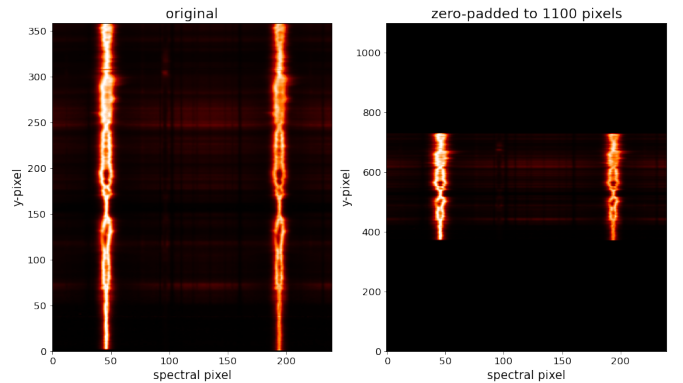

Figure 6: Sample timestep of our dataset in its original form (left) and after zero-padding to 1100 y-pixels (right).

All models were optimized with Adam with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. With a learning rate of 0.0001 we observed stable and efficient learning, for higher learning rates, the training started to oscillate and did not converge. Moreover, more stable training was reached when the input spectra were in addition centered around a mean of 0 and scaled to standard deviation 1. We trained each model with a batch size of 32, stopped the training as soon as the validation loss did not improve anymore for at least 20 epochs and stored the model parameters with the best validation loss.

The pre-processed dataset and the implemented models can be found on Github[5].

4. Results

In the following we present the results and outputs of the models described in section 3. We divide this section into two parts: *bag-level* and *instance-level* prediction. In the bag-level part we evaluate how well the two selected
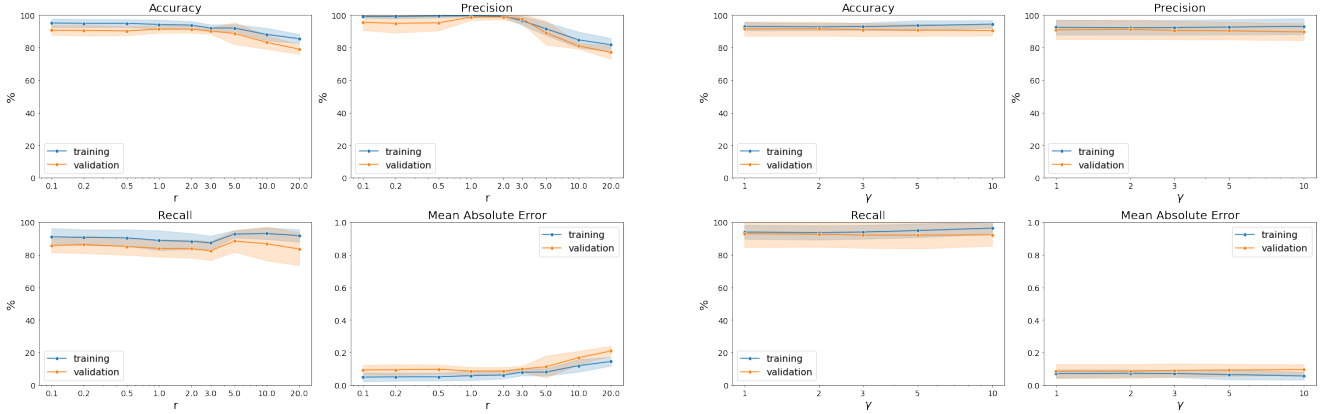
---

[5] https://github.com/chuwyler/irismil

6

Figure 7: Distributions of ibMIL (left) and abMIL (right) bag performance measures for different choices of the hyperparameters r or γ. Both models perform extremely well with accuracies $\gtrsim 90\%$. While the performance of the ibMIL model is sensitive to the choice of r and r seems to have an impact on the precision-recall trade-off, the performance of the abMIL model seems not to be affected by the choice of γ.

proof-of-concept models perform at the bag-level in predicting a flare within the next 25 minutes. In the instance-level part we discuss the core result of this work: how the predictions at instance-level can serve to increase the explainability of flare prediction models. In particular, we extract groups of spectra that are of special interest to the models and recover similar profiles as in previous works.

4.1. Prediction at bag-level

We trained and evaluated the models described in section 3.2 on the three cross-validation folds described in section 2 for different choices of the hyperparameters r or γ, respectively. In particular, we chose $r \in [1, 2, 3, 5, 10, 20]$ or $\gamma \in [1, 2, 3, 5, 10]$ and trained 10 models with different parameter initializations for each fold and choice of hyperparameter. The resulting values for accuracy, precision, recall and mean absolute error (MAE) averaged over all three folds are indicated in Tables A.4 (ibMIL) and A.5 (abMIL) and visualized in Fig. 3.3.1. Both models performed exceptionally well and show a peak validation accuracy on the dataset at $\gtrsim 90\%$ for the smallest chosen values of r or γ.

For the instance-based model, the validation accuracy drops from around 90% to around 80% when going from $r = 1$ (closer to average pooling) to $r = 20$ (closer to max pooling). The drop is associated with a loss in precision, which implies that more false positives are returned by the model for larger values of r. This makes sense, since higher r bring the model closer to maximum pooling where fewer spectra are considered for the bag-level decision. Here, it is much more likely that a non-flaring active region has a few flare-like spectra than that a pre-flare observation has only non-flare-like spectra. Interestingly, the attention-based model shows almost constant accuracy, precision and recall. We assume that the attention-based model is far less susceptible to changes in γ because it manages to adapt the learned embeddings due to its higher flexibility. Consequently, we only report instance-level results for $\gamma = 1$.

To better understand the behavior of our models, we report the distribution of predicted bag-level probabilities per observation in Fig. A.15 in the appendix. Thereby we considered for each observation only the learned models that had this particular observation in the validation fold. Probably due to its increased flexibility, the abMIL model distributes bag probabilities much more confidently than the ibMIL model. For both models, there are clear examples of false positives and false negatives, depending on the value chosen for the parameter r or γ.

As can be observed in Fig. 8, most of the false positives and false negatives seem to originate from a few observations, but not all. Clearly, some timesteps of PF observations may not contain information that would announce a flare and some timesteps from AR observations may contain information that would indicate a precursor signal for a flare (we suspect that this might be the reason why the MAE loss leads to a more stable training). We estimated the performance on the observation-level by specifying detection thresholds through 1) the required probability to count a bag as a pre-flare sign (*bag probability threshold*) and 2) the required number of bags with pre-flare signs to label an observation as pre-flare (*flare bag number*). In Fig. 8 we show the model performance at observation-level in terms of precision, recall and F1-score over the validation folds for different choices of bag probability thresholds and a minimum of 5 flare bags against the different chosen values for r and γ. The detection statistics seem not very sensitive to the chosen required number of flare bags. The results of the ibMIL model are sensitive to the choice of probability threshold and, similar to the performance on the bag-level, higher values of the parameter r correspond to lower precision (more false positives) and higher recall (less false negatives). In contrast, the abMIL model is insensitive to the choice of probability threshold and the parameter γ influences the abMIL precision-recall trade-off only mildly. Note that the abMIL precision is lower on the observation-level than on the bag-level. This is because

7

of the larger number of false positives, since the bag-level probabilities of some active regions are distributed very broadly (see Fig. A.15). The choice of an optimal decision threshold is not further discussed here due to the small amount of labeled data.

In Fig. 9 we plot the probability evolution of two selected full pre-flare observations for different values of r and γ together with the saliency maps (introduced in section 4.2 below). While we make use of only 25-minute long AR and PF slices from selected observations, we apply the trained models to the entire observations with lengths of up to several hours to evaluate how our models extrapolate. Intuitively, we would expect the pre-flare probability to increase when getting closer to a flare, however this is not observed, sometimes probabilities even drop before a flare (e.g. in the ibMIL prediction for the top observation in Fig. 9). Interestingly, some of the abMIL models and some of the ibMIL models seem extremely confident in their predictions already hours before the flare onset, despite the carefully designed validation scheme (see section 5 for a discussion).

4.2. Prediction at instance-level

The learned instance-level probabilities (ibMIL) and attention values (abMIL) can be used to identify spectral profiles that are of particular importance for the models when making their predictions. This may help to understand why the models make a particular prediction at the bag-level and consequently to improve theoretical models in heliophysics that try to explain and predict the emergence of solar flares. In particular, the recovered instance-level predictions can be used to produce saliency maps that provide an excellent quick-look tool to identify regions of interest (along the instrument slit) for flare prediction. Saliency maps for two selected full flare observations are provided in Fig. 9. Regions with activities that trigger the models are clearly highlighted. The attention values were brought to a comparable scale by multiplication with the number of instances per bag n_X, resulting in an attention value of 1 for average attention.

4.2.1. Spectra of particular interest to flare prediction

To extract groups of spectra with particular importance for flare prediction, we used the well-known k-means clustering algorithm (MacQueen, 1967) as a vector quantization method and assessed the distribution of assigned probabilities and attention values per recovered group via mean and standard deviation. The k-means algorithm has been used before to cluster UV-spectra (Pietarila et al., 2007; Panos et al., 2018; Dalda et al., 2019; Woods et al., 2021). We follow this approach since it is simple and easily reproducible. We chose a number of $k = 128$ groups through evaluation of the within-cluster sum-of-squares distances for different choices of k, approximately identifying an elbow centered around $k \sim 80$ and choosing an upper limit to the elbow to make sure that also smaller groups are captured.
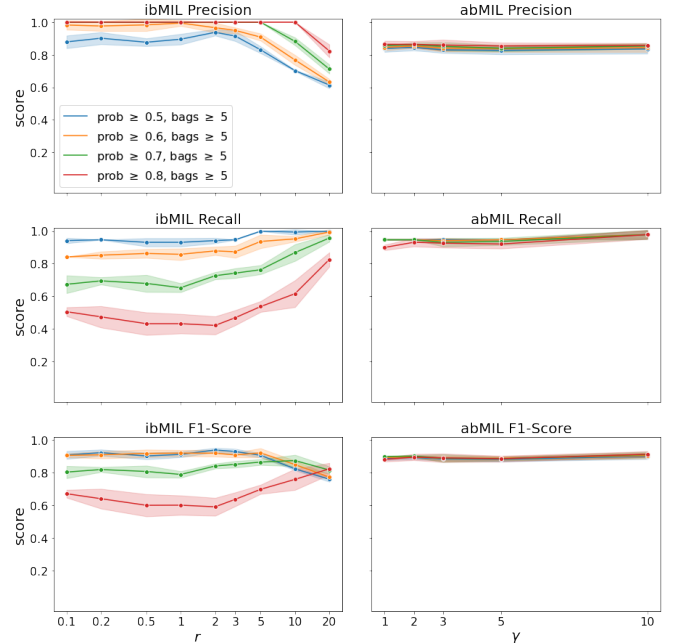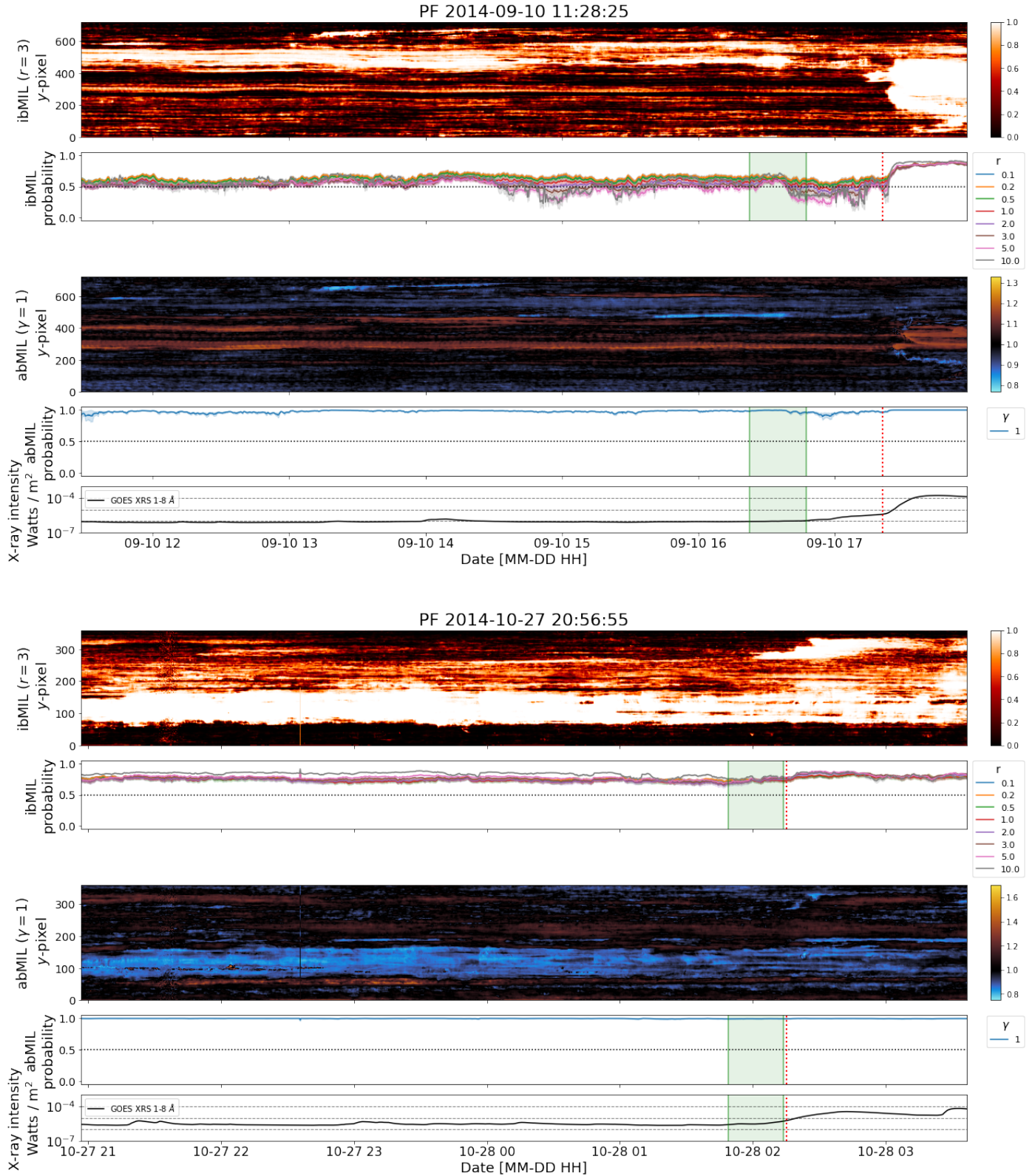

Figure 8: Mean precision, recall, F1-score and standard deviations on the observation-level for different values of the parameters r or γ and for different detection thresholds for ibMIL (left column) and abMIL (right column) models.

For the groups with mean probability $\gtrsim 0.7$ we discovered comparable overarching groups as in a pure k-means study by Woods et al. (2021) characterized as follows: single-peaked spectra with and without triplet red wing emission, double-peaked spectra of high intensity with and without triplet red wing emission and broad mostly double-peaked spectra. In Fig. 10, the top 24 spectral groups with highest mean of the probabilities assigned by our ibMIL model ($r = 1$) are visualized, sorted by mean probability in descending order. In addition to mean probability and standard deviation (p), we also indicate the number of spectra assigned to the group (n) and what percentage of the assigned spectra originates from pre-flare observations (pf). Practically all of the spectral profiles in groups with high mean probabilities have predominantly been sampled from pre-flare observations. Similarly, the groups with low mean probabilities are mainly composed of spectral profiles from non-flaring active regions (Fig. 12).

This is the core result of our work. The shallow ibMIL model managed to reproduce state-of-the-art results as a side product of the weakly-supervised prediction problem. As more deeply explored in Woods et al. (2021), the features characterizing the recovered overarching groups can be well related to the physics associated with solar flares:

- Emission in the two blended red wing triplets (appearing in red and purple groups) is typically associated with a rapid temperature increase in the lower chromosphere (Pereira et al., 2015) and is often observed in flares and other high-energy phenomena,

Figure 9: Saliency maps and probability evolutions for two selected full flare observations (#23 and #28, indicated by start date in the titles) created by the ibMIL (first and second row) and abMIL (third and forth row) models that performed best when the particular observation was in the validation set. The saliency maps feature the probabilities recovered by an ibMIL model (with $r = 3$) or the normalized attention values recovered by an abMIL model (with $\gamma = 1$). The start of the flare is indicated by a red dashed line and the observation time window that entered our dataset is shaded in green. The bottom plot features the X-ray intensity in 1-8 Å measured by the NASA's GOES satellite.
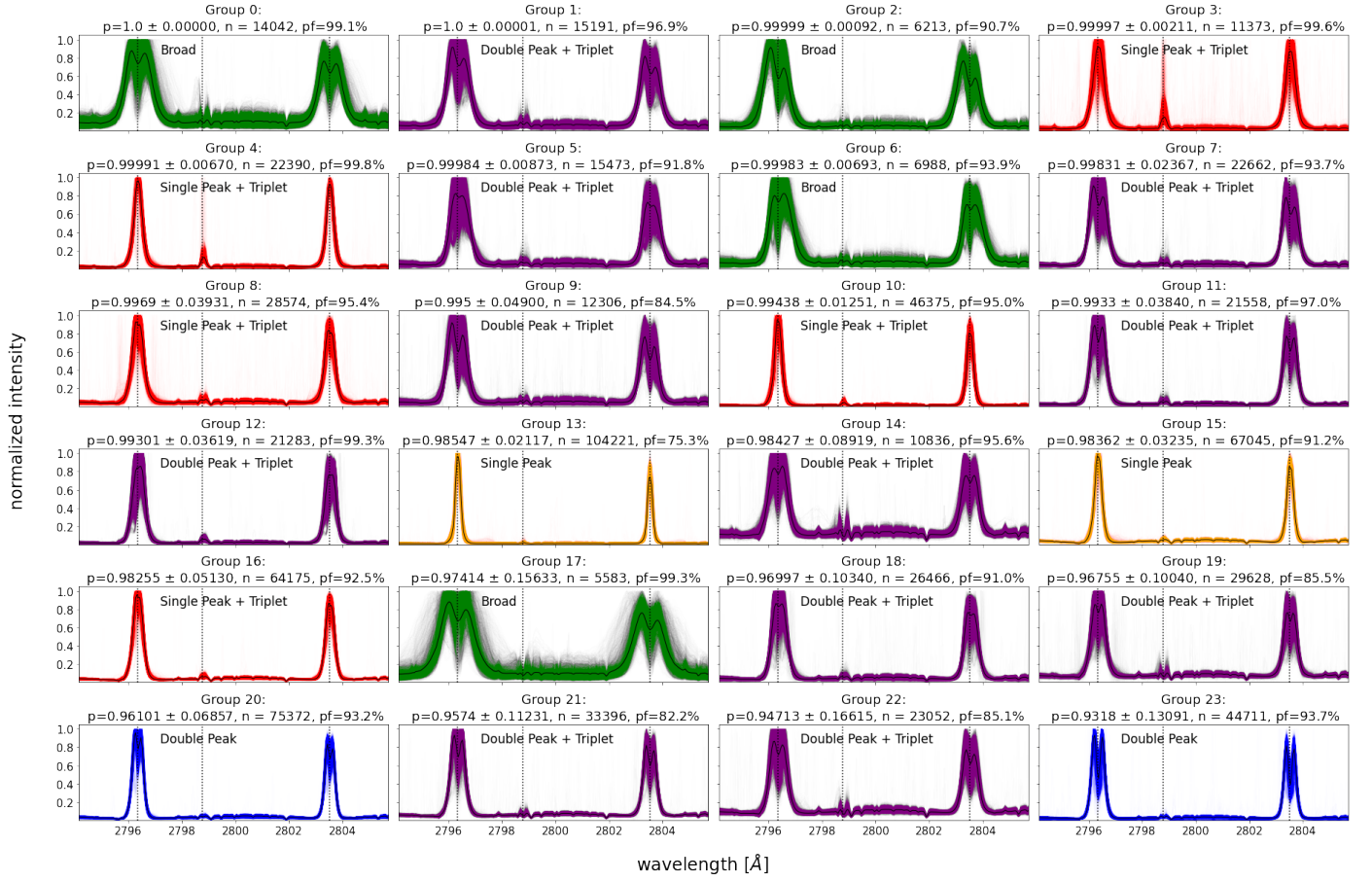
9

Figure 10: Top 24 spectral groups in terms of mean ibMIL probability ($r = 1$), sorted by mean probability in descending order. For each group the mean probability (p) and standard deviation ($\pm$), the number of assigned spectra (n) and what percentage of the assigned spectra originates from pre-flare observations (pf) are indicated. In addition to the group centroid (black), a sample of at maximum 10'000 assigned spectra is visualized (in color). The different colors stand for different characteristic overarching groups: broad double-peaked spectra (green), single-peaked spectra with (red) and without (orange) triplet red wing emission and double-peaked spectra with (purple) and without (blue) triplet red wing emission. The dotted vertical lines represent the positions of the k, triplet and h line cores respectively. All the recovered groups with probability ≥ 0.7 can be related to one of the above-mentioned overarching groups.

in particular also in explosive events accompanied by UV bursts (Vissers et al., 2015; Peter et al., 2014; Schmit, D. J. et al., 2014). In some cases (particularly in groups 14, 19, 22), even double-peaked triplet emission can be observed. According to Pereira et al. (2015), these are mostly caused by temperature profiles with several rapid variations in height rather than just one single dominant increase.

- Single-peaked Mg II h and k lines (appearing in red and orange groups) have been observed in flares and the umbræ of sunspots, where the line cores that are usually reversed in quiet-sun conditions (see Fig. 2) go into emission. In contrast to flares, the single-peaked profiles around sunspots are very narrow and do generally not show additional triplet red wing emission. The single-peaked profiles with strong triplet red wing emission from groups 3 and 4 seem to be similar to the universal flaring profile identified by Panos et al. (2018) that typically appears over flare

ribbons or also small-scale explosive events (Panos et al., 2018; da Costa and Kleint, 2017; Zhu et al., 2019; Kerr, G. S. et al., 2015).

- High peak intensity (corresponding to low continuum in our normalisation scheme) is correlated with high chromospheric gas temperatures (Kerr, G. S. et al., 2015) and observed in all of the recovered groups.

- Broad profiles (appearing in green group) are possibly associated with unresolved up and downflows of up to 200 km/s by da Costa and Kleint (2017).

The high variance in especially the 'broad' and 'double-peaked + triplet' groups suggests that the k-means algorithm with $k = 128$ groups does not provide the perfect granularity for all groups: Division of the double-peaked profiles into further subgroups reveals clusters of various peak ratios corresponding to up and downflows of plasma in the upper chromosphere (Leenaarts et al., 2013a) that
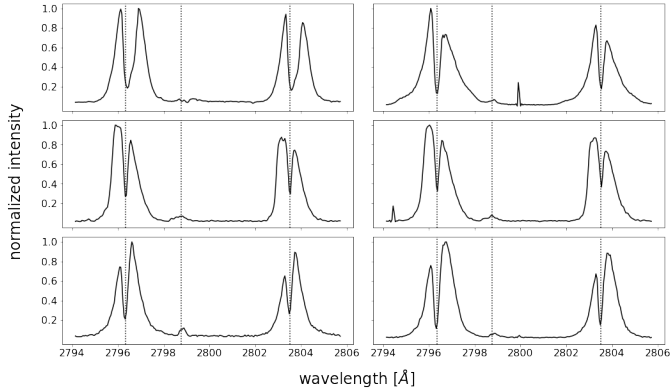
Figure 11: Some examples of broad irregular profiles with probability $p = 1.0$.
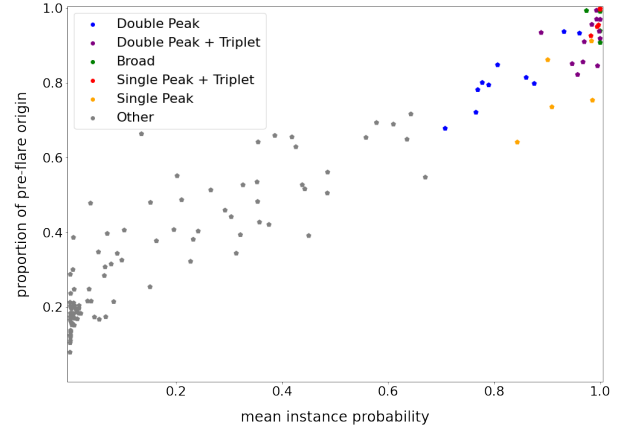
Figure 12: Plot of mean instance probability (ibMIL $r = 1$) against the proportion of how many spectra originate from pre-flare observations for all the recovered groups. The groups with mean instance probability ≥ 0.7 are colored similar to Fig. 10, the rest is denoted as 'Other' and colored in gray. There is a clear and almost linear relationship between mean instance probability and pre-flare proportion.

could be further investigated in future work. In addition, group 17 discloses extremely broad profiles (examples are visualized in Fig. 11) with assigned probabilities of practically 1.0, some with extremely deep central core reversals. These spectra are also discovered by Woods et al. (2021) and indicated as *irregular profiles with broad wings.*

The ibMIL model assigns high probabilities to spectral profiles that are highly correlated with an upcoming flare. As such, it rates how important individual spectra are for *only* a positive prediction (what we refer to as 'asymmetric explanation'). In contrast, the abMIL model assigns high attention values to spectral profiles that are of particular importance for the prediction in general, regardless whether it is positive or negative ('symmetric explanation'). Here, accordingly, spectral profiles can act as flare precursors or exactly the opposite - indicators that clearly no flare is going to take place within the next few hours. As a result, a naive sorting of the clusters by average attention weights does not provide the best flare precursor candidates as obtained in the ibMIL case. While both the ibMIL and the abMIL model provide explanations for their predictions, we believe that the asymmetric explanations by the ibMIL model are better suited for the particular use-case of identifying possible flare precursor candidates.

Given the recovered overarching groups, we are able to examine more deeply why the ibMIL model assigned high probabilities to particular pixels on the slit. As an example, we provide a visualization of the recovered groups for the 10 September 2014 flare in Fig. 13 (the corresponding probabilities are visualized in the upper part of Fig. 9). We notice that high importance is given to broad spectra (colored in green) surrounded by spectra with triplet red wing emission (colored in red) in the upper part of Fig. 13 within the pre-flare selection window. In Fig. 14, a slit-jaw image taken during this phase where the instrument slit in the center is overplotted with the corresponding group colors is displayed. It appears that the observed groups of spectra are caused by a larger mass flow in this region.

In Fig. A.16 in the appendix we list how often the

selected overarching groups of average ibMIL probability ≥ 0.7 appear in the different observation windows included in our dataset. For the pre-flare active region observations typically the frequencies of one or more of the groups are strongly enhanced, presumably the ones that are connected to the underlying processes that led to the flare. We could not identify a single group that is always strongly present in the pre-flare phase, though.

5. Discussion

We have seen that instance-based MIL models are more suited to identify possible flare precursors at the cost of bag-level performance, while embedding-based models show a very high bag-level performance but lack explainability. The missing explainability can partially be restored by analysing attention weights, however the identification of pre-flare precursors is far less easy than originally expected: a particular spectral profile can be considered as important by the model and be assigned a high attention value, however the model does not suggest a straightforward procedure to associate the spectral profile with a positive or negative bag-level class.

The parameters r (ibMIL) or γ (abMIL) control the skewness of the distribution of assigned instance-level probabilities or attention values, respectively. While high values of r lead the ibMIL model to focus only on the strongest predictors and significantly impact the precision-recall trade-off, the choice of γ has an almost negligible impact in case of the abMIL model. As stated previously, we believe that the reason for this is that the abMIL model can flexibly accommodate distortions resulting from modified parameters. We find that the results presented in this paper are qualitatively similar for different choices of r or γ and pro-
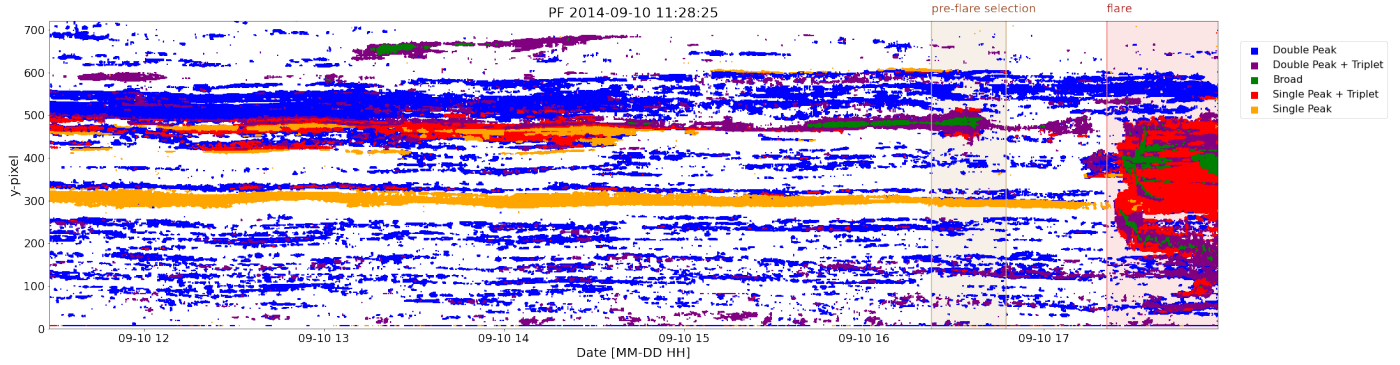
11

Figure 13: Groups recovered from the entire 10 September 2014 flare observation. The duration of the flare and the 25 minutes pre-flare timespan selected for our dataset are highlighted. The flaring part appears to be dominated by single-peaked spectra with triplet red wing emission and broad spectral profiles, however this statement is of limited validity, since the clustering was done only on pre-flare data and does not necessarily extend to spectral profiles observed in flares. In the pre-flare part, high importance is assigned to a mass flow with broad profiles and triplet red wing emission in the middle of the upper part of the slit. This is clearly visible in Fig. 14. The yellow region (single-peaked without triplet) coincides with the sunspot umbra in Fig. 14.
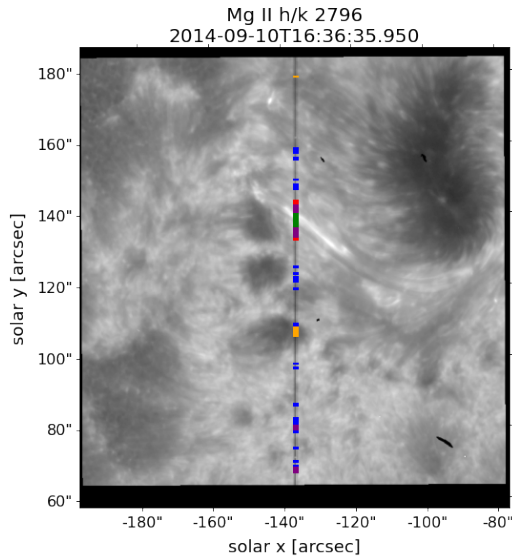

Figure 14: Slit-jaw image taken during the pre-flare selection of the 10 September 2014 flare observation. The instrument slit is indicated by a vertical line and the different overarching groups from Figs. 10 and 13 are highlighted. The mass-flow in the upper middle part is clearly visible and is captured by spectra from the broad (green) group, accompanied by spectra from the triplet emission groups (red, purple).

pose to use values that lead to the best bag-level validation set accuracy ($r = 1$ and $\gamma = 1$ in the case of our dataset).

With the chosen very simple models and evaluation strategy we see significant potential for further improvements. The models neglect the spatial structure (correlations between neighboring spectra) and the dynamics (correlations between successive spectra) - aspects that both might be crucial for accurate flare prediction and proper identification of flare-preceding patterns. In addition, our models are too shallow to capture more advanced features than just the positions of emission peaks, line widths and the continuum intensity. Deeper models could find more advanced features such as the the peak ratios in or between the k and h lines. The latter has been reported to be reduced before the onset of and during flares (Panos et al., 2018; Panos and Kleint, 2020). Finally, we observe that k-means is not perfectly suited for the analysis of the model predictions: Important very localized features such as triplet emission or peak ratios in the k or h line are washed out by the Euclidean distance measure even when the input spectra are standardized. Also, Doppler shifts and instrument artifacts such as cosmic ray spikes can move two spectra of otherwise similar shape to different, unrelated groups.

However, our models seem nevertheless to be able to discern between non-flaring active regions and pre-flare active regions with very high accuracy on the validation set, the abMIL model in particular. In some examples they predict flares with high confidence even hours before the flare (see e.g. the top observation in Fig. 9). We assume that such a high performance would not be expected in an operational scenario. For direct comparison to other flare prediction approaches, we would need to incorporate prior probabilities for AR and PF events. This is mainly due to the fact that our models were trained and validated on a balanced dataset. Moreover, in contrast to other publications that predict flares within the next 12 or 24 hours, we

only consider a time span of approximately half an hour before flare onset (25 minutes plus margin) for the evaluation of our models. Some works suggest that typical pre-flare signatures such as small-scale heating events might be much more common closer to the flare onset (Panos and Kleint, 2020; Woods et al., 2021).

Finally, finding a truly representative sample of non-flaring and pre-flare active regions is a challenging task in all flare prediction problems, since the distribution of the underlying population is not known well enough a priori. Even though our dataset was created in a joint effort between heliophysics and machine learning experts, there is always the potential danger for a selection bias that could impact the reported model performance estimates.

6. Conclusion and Outlook

We demonstrated the application of the Multiple Instance Learning (MIL) paradigm to predict flares given spectral Mg II h&k data recorded by the IRIS satellite and in particular to extract groups of spectral profiles that act as flare precursor signals. The reasons why we advocate the use of MIL for IRIS data are two-fold: Firstly, MIL offers a well-defined way to link the available weak labels on observation-level down to individual spectra. Secondly, the specifically selected MIL models allow us to propagate the observation level labels down to individual spectral profiles or at least allow the identification of spectral profiles that are of particular importance for the prediction of the model.

We evaluated both an instance-based approach (ibMIL) and an embedding-based approach incorporating an attention mechanism (abMIL) on a balanced dataset sampled from 18 non-flaring and 19 pre-flare active region observations. Both models performed reasonably well on the bag-level with accuracies of $\gtrsim 90\%$ in predicting whether or not an active region is going to produce a flare within the next $\sim$25 minutes. The abMIL model performed slightly better than the ibMIL model at the cost of reduced explainability of its results. For both models saliency maps of the computed probabilities or attention values provide insights into the decision process over the time evolution of the observation. For the core result of this paper, we used the instance-level predictions by the ibMIL model to extract a set of typical pre-flare spectral profiles. The recovered groups are consistent with the results of other works in the domain and seem to be related to changes in the temperature, velocity and density distribution induced by small-scale explosive events that have been reported to occur tens of minutes before a flare.

In view of developing a production-ready flare prediction model based on IRIS data, several aspects need to be explored:

Firstly, we plan to further extend our sample of non-flaring active region observations by evaluating our models on all available IRIS data and applying an active learning strategy. In addition, we intend to design an appropriate prior to extend the predictions of our models to more realistic class imbalance. Secondly, we propose to use models with substantially more capacity and suitable regularization to identify more complex pre-flare precursor features, such as asymmetries in the Mg II h and k lines and the Mg II k/h peak ratio. Thirdly, accurate flare prediction requires the identification of spatio-temporally evolving patterns. We plan to set up a pooling function that takes into account spatial structure through the use of convolutional neural networks. The dynamical aspect of the problem could be included by using an embedding-based approach and pooling the embeddings of the individual timesteps with a recurrent neural network. Fourthly, the inclusion of data from other spectral windows observed by IRIS, such as C II, Si IV or O IV will almost certainly provide additional and very valuable information (as explored in Panos et al. (2021) and Panos and Kleint (2021)). Fifthly, we plan to extend the tools to analyze our results, since k-means does not seem to be optimal to divide spectral profiles into groups. We plan to evaluate alternative clustering strategies (such as Deep Embedded Clustering (Xie et al., 2016)), possibly applied directly to the embeddings recovered by the MIL models.

We have shown that MIL provides a very valuable interpretable modeling approach for the prediction of solar flares. It looks very promising to also be applied to other solar data, such as images recorded by the HMI or AIA instruments on the SDO satellite.

References

Ahmadzadeh, A., Aydin, B., Georgoulis, M.K., Kempton, D.J., Mahajan, S.S., Angryk, R.A., 2021. How to train your flare prediction model: Revisiting robust sampling of rare events. ApjS 254, 23. doi:10.3847/1538-4365/abec88.

Ahmed, O.W., Qahwaji, R., Colak, T., Higgins, P.A., Gallagher, P.T., Bloomfield, D.S., 2013. Solar flare prediction using advanced feature extraction, machine learning, and feature selection. Solar Physics 283, 157–175. doi:10.1007/s11207-011-9896-1.

Amores, J., 2013. Multiple instance classification: Review, taxonomy and comparative study. Artificial Intelligence 201, 81–105. doi:10.1016/j.artint.2013.06.003.

Armstrong, J.A., Fletcher, L., 2019. Fast solar image classification using deep learning and its importance for automation in solar physics. Solar Physics 294, 80. doi:10.1007/s11207-019-1473-z.

Benz, A.O., 2016. Flare observations. Living Reviews in Solar Physics 14, 2. doi:10.1007/s41116-016-0004-3.

Bobra, M.G., Couvidat, S., 2015. Solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm. ApJ 798, 135. doi:10.1088/0004-637x/798/2/135.

Boucheron, L.E., Al-Ghraibah, A., McAteer, R.T.J., 2015. Prediction of solar flare size and time-to-flare using support vector machines. ApJ 812, 51. doi:10.1088/0004-637x/812/1/51.

Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G., 2018. Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition 77, 329 – 353. doi:10.1016/j.patcog.2017.10.009.

Chen, Y., Manchester, W.B., Hero, A.O., Toth, G., DuFumier, B., Zhou, T., Wang, X., Zhu, H., Sun, Z., Gombosi, T.I., 2019. Identifying solar flare precursors using time series of sdo/hmi images and sharp parameters. Space Weather 17, 1404–1426. doi:10.1029/2019SW002214.

da Costa, F.R., Kleint, L., 2017. A parameter study for modeling mg ii h and k emission during solar flares. ApJ 842, 82. doi:10.3847/1538-4357/aa6eaf.

Dalda, A.S., de la Cruz Rodríguez, J., Pontieu, B.D., Gošić, M., 2019. Recovering thermodynamics from spectral profiles observed by IRIS: A machine and deep learning approach. ApJ 875, L18. doi:10.3847/2041-8213/ab15d9.

De Pontieu, B., Title, A.M., Lemen, J.R., Kushner, G.D., Akin, D.J., Allard, B., Berger, T., Boerner, P., Cheung, M., Chou, C., Drake, J.F., Duncan, D.W., Freeland, S., Heyman, G.F., Hoffman, C., Hurlburt, N.E., Lindgren, R.W., Mathur, D., Rehse, R., Sabolish, D., Seguin, R., Schrijver, C.J., Tarbell, T.D., Wülser, J.P., Wolfson, C.J., Yanari, C., Mudge, J., Nguyen-Phuc, N., Timmons, R., van Bezooijen, R., Weingrod, I., Brookner, R., Butcher, G., Dougherty, B., Eder, J., Knagenhjelm, V., Larsen, S., Mansir, D., Phan, L., Boyle, P., Cheimets, P.N., DeLuca, E.E., Golub, L., Gates, R., Hertz, E., McKillop, S., Park, S., Perry, T., Podgorski, W.A., Reeves, K., Saar, S., Testa, P., Tian, H., Weber, M., Dunn, C., Eccles, S., Jaeggli, S.A., Kankelborg, C.C., Mashburn, K., Pust, N., Springer, L., Carvalho, R., Kleint, L., Marmie, J., Mazmanian, E., Pereira, T.M.D., Sawyer, S., Strong, J., Worden, S.P., Carlsson, M., Hansteen, V.H., Leenaarts, J., Wiesmann, M., Aloise, J., Chu, K.C., Bush, R.I., Scherrer, P.H., Brekke, P., Martinez-Sykora, J., Lites, B.W., McIntosh, S.W., Uitenbroek, H., Okamoto, T.J., Gummin, M.A., Auker, G., Jerram, P., Pool, P., Waltham, N., 2014. The interface region imaging spectrograph (iris). Solar Physics 289, 2733–2779. doi:10.1007/s11207-014-0485-y.

Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89, 31–71. doi:10.1016/S0004-3702(96)00034-3.

Florios, K., Kontogiannis, I., Park, S.H., Guerra, J.A., Benvenuto, F., Bloomfield, D.S., Georgoulis, M.K., 2018. Forecasting solar flares using magnetogram-based predictors and machine learning. Solar Physics 293, 28. doi:10.1007/s11207-018-1250-4.

Foulds, J., Frank, E., 2010. A review of multi-instance learning assumptions. The Knowledge Engineering Review 25, 1–25. doi:10.1017/S026988890999035X.

Galvez, R., Fouhey, D.F., Jin, M., Szenicer, A., Muñoz-Jaramillo, A., Cheung, M.C.M., Wright, P.J., Bobra, M.G., Liu, Y., Mason, J., Thomas, R., 2019. A machine-learning data set prepared from the NASA solar dynamics observatory mission. ApjS 242, 7. doi:10.3847/1538-4365/ab1005.

Huang, X., Wang, H., Xu, L., Liu, J., Li, R., Dai, X., 2018. Deep learning based solar flare forecasting model. i. results for line-of-sight magnetograms. ApJ 856, 7. doi:10.3847/1538-4357/aaae00.

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning, in: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning, PMLR. pp. 2127–2136. URL: http://proceedings.mlr.press/v80/ilse18a.html.

Jonas, E., Bobra, M., Shankar, V., Todd Hoeksema, J., Recht, B., 2018. Flare prediction using photospheric and coronal image data. Solar Physics 293, 48. doi:10.1007/s11207-018-1258-9.

Kanavati, F., Ichihara, S., Rambeau, M., Iizuka, O., Arihiro, K., Tsuneki, M., 2021. Deep learning models for gastric signet ring cell carcinoma classification in whole slide images. Technology in Cancer Research & Treatment 20, 15330338211027901. doi:10.1177/15330338211027901.

Karimi, D., Dou, H., Warfield, S.K., Gholipour, A., 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis 65, 101759. doi:10.1016/j.media.2020.101759.

Kerr, G. S., Simões, P. J. A., Qiu, J., Fletcher, L., 2015. Iris observations of the mg ii h and k lines during a solar flare. A&A 582, A50. doi:10.1051/0004-6361/201526128.

Lanzerotti, L.J., 2007. Space weather effects on communications. Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 247–268. doi:10.1007/978-3-540-34578-7_9.

Leenaarts, J., Pereira, T.M.D., Carlsson, M., Uitenbroek, H., De Pontieu, B., 2013a. The Formation of IRIS Diagnostics. I. A Quintessential Model Atom of Mg II and General Formation Properties of the Mg II h&k Lines. ApJ 772, 89. doi:10.1088/0004-637X/772/2/89.

Leenaarts, J., Pereira, T.M.D., Carlsson, M., Uitenbroek, H., De Pontieu, B., 2013b. The Formation of IRIS Diagnostics. II. The Formation of the Mg II h&k Lines in the Solar Atmosphere. ApJ 772, 90. doi:10.1088/0004-637X/772/2/90.

Liu, C., Deng, N., Wang, J.T.L., Wang, H., 2017. Predicting solar flares UsingSDO/HMI vector magnetic data products and the random forest algorithm. ApJ 843, 104. doi:10.3847/1538-4357/aa789b.

Liu, H., Liu, C., Wang, J.T.L., Wang, H., 2019. Predicting solar flares using a long short-term memory network. ApJ 877, 121. doi:10.3847/1538-4357/ab1b3c.

Liu, M., Zhang, J., Adeli, E., Shen, D., 2018. Landmark-based deep multi-instance learning for brain disease diagnosis. Medical Image Analysis 43, 157–168. doi:10.1016/j.media.2017.10.005.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations, in: Cam, L.M.L., Neyman, J. (Eds.), Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press. pp. 281–297.

Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Watari, S., Ishii, M., 2017. Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms. ApJ 835, 156. doi:10.3847/1538-4357/835/2/156.

Oquab, M., Bottou, L., Laptev, I., Sivic, J., 2015. Is object localization for free? - weakly-supervised learning with convolutional neural networks, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 685–694. doi:10.1109/CVPR.2015.7298668.

Panos, B., Kleint, L., 2020. Real-time Flare Prediction Based on Distinctions between Flaring and Non-flaring Active Region Spectra. ApJ 891, 17. doi:10.3847/1538-4357/ab700b.

Panos, B., Kleint, L., 2021. Exploring mutual information between IRIS spectral lines. II. calculating the most probable response in all spectral windows. ApJ 915, 77. doi:10.3847/1538-4357/ac00c0.

Panos, B., Kleint, L., Huwyler, C., Krucker, S., Melchior, M., Ullmann, D., Voloshynovskiy, S., 2018. Identifying Typical Mg II Flare Spectra Using Machine Learning. ApJ 861, 62. doi:10.3847/1538-4357/aac779.

Panos, B., Kleint, L., Voloshynovskiy, S., 2021. Exploring mutual information between IRIS spectral lines. i. correlations between spectral lines during solar flares and within the quiet sun. ApJ 912, 121. doi:10.3847/1538-4357/abf11b.

Pereira, T.M.D., Carlsson, M., Pontieu, B.D., Hansteen, V., 2015. The formation of iris diagnostics. iv. the mg ii triplet lines as a new diagnostic for lower chromospheric heating. ApJ 806, 14. doi:10.1088/0004-637x/806/1/14.

Pereira, T.M.D., Leenaarts, J., De Pontieu, B., Carlsson, M., Uitenbroek, H., 2013. The Formation of IRIS Diagnostics. III. Near-ultraviolet Spectra and Images. ApJ 778, 143. doi:10.1088/0004-637X/778/2/143.

Pesnell, W.D., Thompson, B.J., Chamberlin, P.C., 2012. The solar dynamics observatory (sdo). Solar Physics 275, 3–15. doi:10.1007/s11207-011-9841-3.

Peter, H., Tian, H., Curdt, W., Schmit, D., Innes, D., Pontieu, B.D., Lemen, J., Title, A., Boerner, P., Hurlburt, N., Tarbell, T.D., Wuelser, J.P., Martínez-Sykora, J., Kleint, L., Golub, L., McKillop, S., Reeves, K.K., Saar, S., Testa, P., Kankelborg, C., Jaeggli, S., Carlsson, M., Hansteen, V., 2014. Hot explosions in the cool atmosphere of the sun. Science 346, 1255726. doi:10.1126/science.1255726.

Pietarila, A., Socas-Navarro, H., Bogdan, T., 2007. Spectropolarimetric observations of the CaII λ8498 and λ8542 in the quiet sun. ApJ 663, 1386–1405. doi:10.1086/518714.

Pinheiro, P.O., Collobert, R., 2015. From image-level to pixel-level labeling with convolutional networks, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1713–1721. doi:10.1109/CVPR.2015.7298780.

Pulkkinen, T., 2007. Space weather: Terrestrial perspective. Living Reviews in Solar Physics 4, 1. doi:10.12942/lrsp-2007-1.

Ramon, J., De Raedt, L., 2000. Multi instance neural networks, in: Proceedings of the ICML-2000 workshop on attribute-value and relational learning, pp. 53–60.

Sadafi, A., Makhro, A., Bogdanova, A., Navab, N., Peng, T., Albarqouni, S., Marr, C., 2020. Attention based multiple instance learning for classification of blood cell disorders, in: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, Cham. pp. 246–256.

Schmit, D. J., Innes, D., Ayres, T., Peter, H., Curdt, W., Jaeggli, S., 2014. Molecular absorption in transition region spectral lines. A&A 569, L7. doi:10.1051/0004-6361/201424432.

Song, H., Kim, M., Park, D., Lee, J.G., 2022. Learning from noisy labels with deep neural networks: A survey. IEEE transactions on neural networks and learning systems PP. doi:DOI:10.1109/TNNLS.2022.3152527.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Sudharshan, P., Petitjean, C., Spanhol, F., Oliveira, L.E., Heutte, L., Honeine, P., 2019. Multiple instance learning for histopathological breast cancer image classification. Expert Systems with Applications 117, 103–111. doi:10.1016/j.eswa.2018.09.049.

Vernazza, J.E., Avrett, E.H., Loeser, R., 1981. Structure of the solar chromosphere. III. Models of the EUV brightness components of the quiet sun. ApjS 45, 635–725. doi:10.1086/190731.

Vissers, G.J.M., van der Voort, L.H.M.R., Rutten, R.J., Carlsson, M., Pontieu, B.D., 2015. Ellerman bombs at high resolution. iii. simultaneous observations with iris and sst. ApJ 812, 11. doi:10.1088/0004-637x/812/1/11.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 3462–3471doi:10.1109/CVPR.2017.369.

Woods, M.M., Dalda, A.S., Pontieu, B.D., 2021. Unsupervised machine learning for the identification of preflare spectroscopic signatures. ApJ 922, 137. doi:10.3847/1538-4357/ac2667.

Xie, J., Girshick, R., Farhadi, A., 2016. Unsupervised deep embedding for clustering analysis, in: Balcan, M.F., Weinberger, K.Q. (Eds.), Proceedings of The 33rd International Conference on Machine Learning, PMLR, New York, New York, USA. pp. 478–487. URL: https://proceedings.mlr.press/v48/xieb16.html.

Zhou, Z., 2018. A brief introduction to weakly supervised learning. National Science Review 5, 44–53.

Zhu, Y., Kowalski, A.F., Tian, H., Uitenbroek, H., Carlsson, M., Allred, J.C., 2019. Modeling mg ii h, k and triplet lines at solar flare ribbons. ApJ 879, 19. doi:10.3847/1538-4357/ab2238.

Appendix A. Using masking to account for heterogeneous bag sizes

Once all the bags are zero-padded to $n = 1100$ pixels, the whole dataset can be represented by the data cube $\mathbf{X} \in \mathbb{R}^{b \times n \times d}$ and the vector $\mathbf{Y} \in \mathbb{R}^b$, where $b = 10'000$ represents the number of bags, $n = 1100$ the number of pixels and $d = 240$ the number of spectral bins. The masks are represented by a matrix $\mathbf{M} \in \mathbb{R}^{b \times n}$ and the number of (non-zero) instances per bag is

$$ n_i = \sum_{k=1}^{n} M_{ik}. \tag{A.1} $$

The ibMIL instance-level classifier f_θ learns the matrix of probabilities $\mathbf{P} \in \mathbb{R}^{b \times n}$ given the matrix $\mathbf{X}$ which is then pooled to compute the bag probability

$$ \hat{Y}_i = \frac{1}{r} \log \left(\frac{1}{n_i} \sum_{k=1}^{n} M_{ik} e^{r \, P_{ik}} \right), \tag{A.2} $$

where the multiplication with M_{ik} makes sure that only masked probabilities enter the bag aggregate.

Similarly, the abMIL approach computes the bag probability $\hat{Y}_i$ from a matrix of bag embeddings $\mathbf{E} \in \mathbb{R}^{b \times m}$, where m is the latent dimension of the bag embeddings, that is aggregated as

$$ E_{ij} = \sum_{k=1}^{n} e_{ikj} A_{ik}, \tag{A.3} $$

where $\mathbf{e} \in \mathbb{R}^{b \times n \times m}$ represents the individual embeddings computed by f_θ and $\mathbf{A} \in \mathbb{R}^{b \times n}$ represents the attention values learned by the pooling g. These have to be appropriately masked:

$$ A_{ik} = \frac{M_{ik} Q_{ik}}{\sum_{j=1}^{n} M_{ij} Q_{ij}} \tag{A.4} $$

with $Q_{ik} = \exp \left[\mathbf{w}^T \tanh \left(\mathbf{V} \, \mathbf{e}_{ik} \right) \odot \sigma \left(\mathbf{U} \, \mathbf{e}_{ik} \right) \right]^\gamma$ and the vectors $\mathbf{e}_{ik} \in \mathbb{R}^m$.
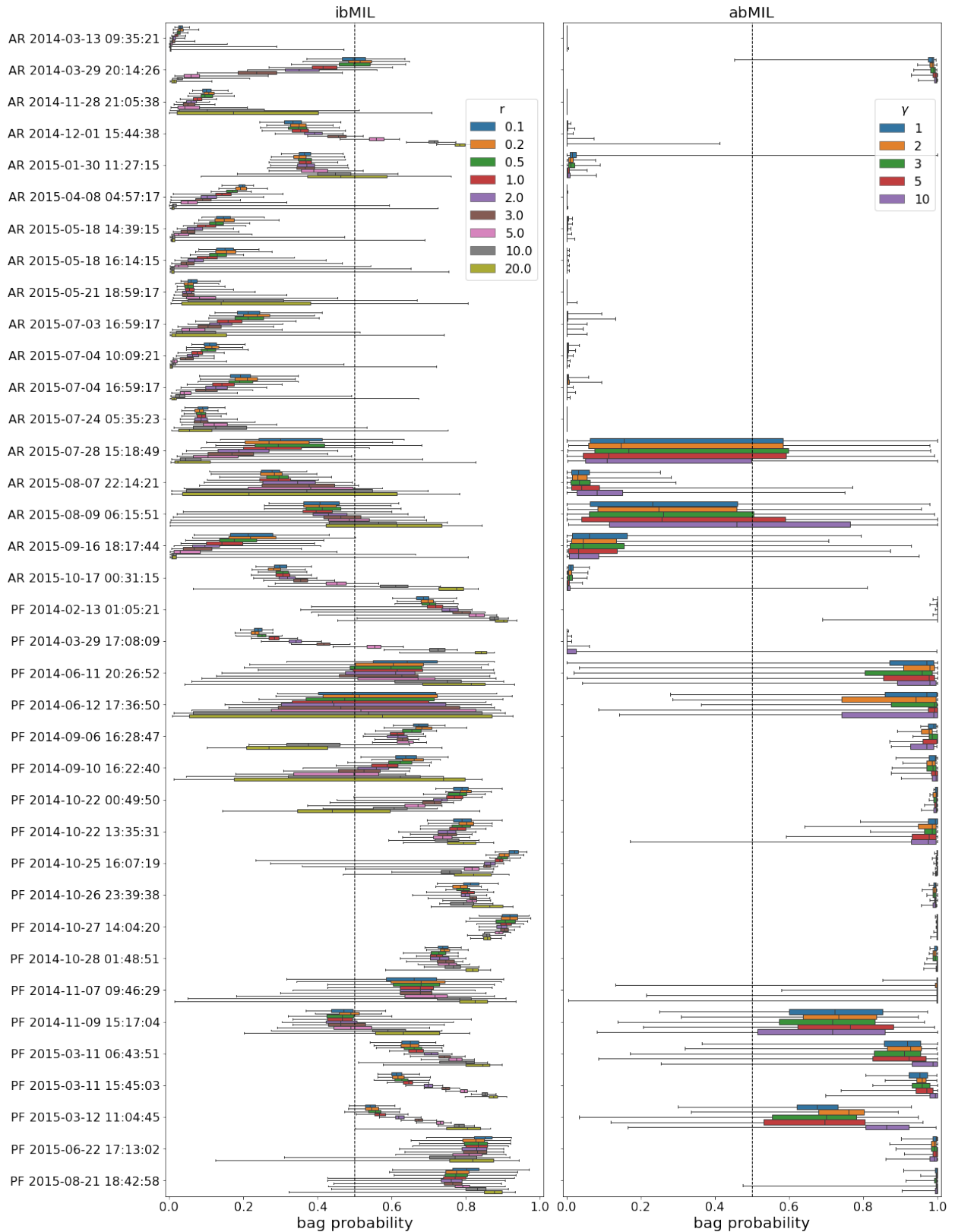
Figure A.15: Recovered ibMIL and abMIL validation bag probability distributions for individual observations. The non-flaring active region (AR) observations are located in the top half and the pre-flare observations (PF) in the bottom-half. The individual distributions are indicated with a box plot whose whiskers stretch to the minimum and maximum of the distribution. The free parameter r has a particular impact on the distribution of ibMIL bag probabilities, while the impact of γ on the abMIL bag probabilities is less pronounced. Note that the two misclassified observations from 29 March 2014 are from different locations on the Sun.

16

#	Start	Duration (min)	Observation Mode	FOV Center (arcsec)	Steps	Cadence (sec)	Slit Pixels	Group	Validation Fold
0	2014-03-13 09:35:21	24.9	Very large sit-and-stare	(522, 23)	162	9.3	548	0	0
1	2014-03-29 20:14:26	24.8	Medium sit-and-stare	(686, -166)	90	16.8	401	1	1
2	2014-11-28 21:05:38	24.9	Very large sit-and-stare	(-34, -322)	157	9.6	1096	2	1
3	2014-12-01 15:44:38	24.8	Large sit-and-stare	(-79, -329)	155	9.7	779	3	1
4	2015-01-30 11:27:15	20.3	Very large dense 4-step raster	(-726, 161)	240	20.2	548	4	2
5	2015-04-08 04:57:17	24.9	Very large sit-and-stare	(46, -118)	285	5.3	548	5	0
6	2015-05-18 14:39:15	25.0	Large coarse 4-step raster	(298, -98)	280	21.4	777	6	0
7	2015-05-18 16:14:15	25.0	Large coarse 4-step raster	(313, -95)	280	21.1	777	6	0
8	2015-05-21 18:59:17	24.9	Very large sit-and-stare	(-386, 398)	277	5.4	1095	7	2
9	2015-07-03 16:59:17	24.8	Large sparse 8-step raster	(-189, 203)	272	43.8	777	8	1
10	2015-07-04 10:09:21	24.9	Very large sit-and-stare	(64, 174)	165	9.1	548	9	1
11	2015-07-04 16:59:17	24.8	Large sparse 8-step raster	(17, 202)	272	43.9	777	8	1
12	2015-07-24 05:35:23	24.8	Large sit-and-stare	(537, 203)	163	9.2	388	10	2
13	2015-07-28 15:18:49	24.9	Medium coarse 4-step raster	(-229, -289)	164	36.5	207	11	2
14	2015-08-07 22:14:21	24.7	Large coarse 8-step raster	(541, 125)	160	74.3	777	12	0
15	2015-08-09 06:15:51	25.0	Large coarse 8-step raster	(230, -370)	160	75.3	777	13	0
16	2015-09-16 18:17:44	24.8	Medium coarse 16-step raster	(-578, -346)	704	33.6	207	14	0
17	2015-10-17 00:31:15	24.9	Large sit-and-stare	(-561, -232)	478	3.1	388	15	2

Table A.1: Selected non-flare active regions (AR) in our sample

#	Start	Duration (min)	Observation Mode	FOV Center (arcsec)	Steps	Cadence (sec)	Slit Pixels	Group	Validation Fold
18	2014-02-13 01:05:21	25.1	Large coarse 8-step raster	(133, -99)	280	41.7	773	16	1
19	2014-03-29 17:08:09	24.8	Very large coarse 8-step raster	(507, 282)	160	71.9	1094	17	2
20	2014-06-11 20:26:52	25.0	Medium coarse 8-step raster	(-777, -305)	560	21.3	203	18	2
21	2014-06-12 17:36:50	25.0	Medium coarse 8-step raster	(-691, -303)	560	21.4	203	19	0
22	2014-09-06 16:28:47	17.8	Large sit-and-stare	(-704, -299)	114	9.5	775	20	0
23	2014-09-10 16:22:40	24.9	Large sit-and-stare	(-121, 125)	160	9.4	776	21	1
24	2014-10-22 00:49:50	24.6	Large sit-and-stare	(-343, -316)	91	16.4	776	22	1
25	2014-10-22 13:35:31	23.8	Very large coarse 8-step raster	(-249, -298)	88	131.2	1094	23	1
26	2014-10-25 16:07:19	25.0	Large sit-and-stare	(406, -319)	280	5.4	387	24	2
27	2014-10-26 23:39:38	24.8	Large sit-and-stare	(637, -287)	93	16.2	387	25	2
28	2014-10-27 14:04:20	6.3	Large coarse 8-step raster	(727, -293)	120	25.6	387	26	0
29	2014-10-28 01:48:51	24.8	Large sit-and-stare	(787, -270)	93	16.2	387	27	0
30	2014-11-07 09:46:29	25.1	Large coarse 16-step raster	(-659, 224)	976	24.3	386	28	1
31	2014-11-09 15:17:04	8.6	Large coarse 4-step raster	(-219, 205)	56	37.4	775	29	1
32	2015-03-11 06:43:51	24.9	Large coarse 8-step raster	(-420, -193)	160	75.0	776	30	0
33	2015-03-11 15:45:03	24.8	Large coarse 4-step raster	(-352, -198)	288	16.8	388	30	0
34	2015-03-12 11:04:45	22.4	Large sit-and-stare	(-190, -190)	258	5.2	388	30	0
35	2015-06-22 17:13:02	24.8	Large sparse 16-step raster	(70, 187)	704	33.1	387	31	2
36	2015-08-21 18:42:58	25.2	Medium dense 32-step raster	(-452, -337)	480	102.1	207	32	2

Table A.2: Selected pre-flare active regions (PF) in our sample

#	Flare Start (min)	Flare Class (GOES)
18	2014-02-13 01:31:59	M1.8
19	2014-03-29 17:35:37	X1.0
20	2014-06-11 20:53:00	M3.9
21	2014-06-12 18:02:58	M1.0
22	2014-09-06 16:50:04	M1.1
23	2014-09-10 17:21:02	X1.6
24	2014-10-22 01:16:06	M8.7
25	2014-10-22 14:01:43	X1.6
26	2014-10-25 16:41:18	X1.0
27	2014-10-27 00:06:06	M1.3
28	2014-10-27 14:11:59	X2.0
29	2014-10-28 02:15:03	M3.4
30	2014-11-07 10:12:49	M1.0
31	2014-11-09 15:27:04	M2.3
32	2015-03-11 07:10:10	M1.8
33	2015-03-11 16:11:00	X2.1
34	2015-03-12 11:38:18	M1.4
35	2015-06-22 17:39:00	M6.5
36	2015-08-21 19:09:55	M1.1

Table A.3: Flare start times and magnitudes for the pre-flare active regions in our sample

r	train acc	val acc	train prec	val prec	train rec	val rec
0.1	95.1% ± 2.7%	90.7% ± 2.9%	99.1% ± 0.9%	95.4% ± 4.8%	91.1% ± 5.2%	85.7% ± 4.3%
0.2	94.9% ± 2.6%	90.6% ± 3.3%	99.1% ± 1.3%	94.9% ± 5.8%	90.7% ± 4.8%	86.3% ± 5.4%
0.5	94.9% ± 2.4%	90.2% ± 2.8%	99.4% ± 0.9%	95.2% ± 4.9%	90.4% ± 5.1%	85.1% ± 5.3%
1.0	94.2% ± 2.9%	91.4% ± 2.4%	99.5% ± 0.5%	98.7% ± 1.9%	88.8% ± 6.1%	83.9% ± 5.3%
2.0	93.8% ± 2.3%	91.3% ± 2.3%	99.3% ± 0.6%	98.9% ± 1.6%	88.3% ± 4.9%	83.7% ± 5.6%
3.0	92.1% ± 2.1%	90.1% ± 1.8%	96.5% ± 2.7%	97.7% ± 3.0%	87.4% ± 4.2%	82.5% ± 5.8%
5.0	92.0% ± 2.3%	88.7% ± 6.7%	91.6% ± 4.0%	89.1% ± 7.3%	92.7% ± 2.2%	88.4% ± 6.6%
10.0	88.1% ± 3.9%	83.2% ± 4.1%	84.8% ± 4.9%	81.0% ± 1.8%	93.1% ± 3.5%	86.8% ± 10.5%
20.0	85.5% ± 2.7%	79.0% ± 3.0%	81.7% ± 4.0%	77.1% ± 4.0%	91.8% ± 3.8%	83.6% ± 10.1%

Table A.4: ibMIL model results for training and validation accuracy (acc), precision (prec) and recall (rec) for different choices of r.

γ	train acc	val acc	train prec	val prec	train rec	val rec
1	93.0% ± 2.6%	91.5% ± 4.4%	92.4% ± 4.7%	90.9% ± 5.8%	93.9% ± 4.3%	93.0% ± 8.3%
2	92.7% ± 2.2%	91.5% ± 4.3%	92.3% ± 4.4%	91.1% ± 6.0%	93.6% ± 4.4%	92.6% ± 7.8%
3	92.9% ± 2.2%	91.0% ± 4.1%	92.4% ± 4.6%	90.6% ± 5.6%	94.0% ± 4.2%	92.1% ± 7.9%
5	93.5% ± 3.1%	90.8% ± 3.7%	92.6% ± 4.6%	90.3% ± 5.4%	94.9% ± 4.2%	92.1% ± 8.3%
10	94.4% ± 2.4%	90.5% ± 3.2%	93.0% ± 4.9%	89.6% ± 5.2%	96.4% ± 3.8%	92.3% ± 6.8%

Table A.5: abMIL model results for training and validation accuracy (acc), precision (prec) and recall (rec) for different choices of γ.

			Single Peak [%] + Triplet	Single Peak [%]	Broad [%]	Double Peak [%] + Triplet	Double Peak [%]	Rest [%]
AR	2014-03-13 09:35:21	——	0.000	0.000	0.000	0.090	0.151	99.759
AR	2014-03-29 20:14:26	——	0.000	0.003	0.000	1.174	3.270	95.553
AR	2014-11-28 21:05:38	——	0.001	0.845	0.000	0.070	0.498	98.586
AR	2014-12-01 15:44:38	——	0.309	14.005	0.005	1.165	3.912	80.604
AR	2015-01-30 11:27:15	——	0.186	2.322	0.001	0.591	9.618	87.282
AR	2015-04-08 04:57:17	——	0.049	0.012	0.000	0.292	3.789	95.858
AR	2015-05-18 14:39:15	——	0.001	0.020	0.000	0.234	1.835	97.910
AR	2015-05-18 16:14:15	——	0.002	0.026	0.002	0.185	1.891	97.896
AR	2015-05-21 18:59:17	——	0.018	0.048	0.010	0.565	0.359	99.000
AR	2015-07-03 16:59:17	——	0.393	0.274	0.000	0.501	8.484	90.348
AR	2015-07-04 10:09:21	——	0.000	0.019	0.005	0.294	2.264	97.418
AR	2015-07-04 16:59:17	——	0.262	0.029	0.000	0.697	9.057	89.955
AR	2015-07-24 05:35:23	——	0.002	0.010	0.000	0.009	0.753	99.227
AR	2015-07-28 15:18:49	——	0.007	0.848	0.053	1.150	4.580	93.361
AR	2015-08-07 22:14:21	——	0.197	6.705	0.001	0.214	1.613	91.270
AR	2015-08-09 06:15:51	——	1.467	6.625	0.194	2.518	7.181	82.015
AR	2015-09-16 18:17:44	——	0.101	0.107	0.000	0.550	4.224	95.019
AR	2015-10-17 00:31:15	——	1.948	0.064	0.494	4.892	4.481	88.121
PF	2014-02-13 01:05:21	——	16.954	5.729	6.471	13.918	9.109	47.819
PF	2014-03-29 17:08:09	——	1.112	0.529	1.985	7.246	2.177	86.950
PF	2014-06-11 20:26:52	——	2.644	1.169	0.849	7.625	12.527	75.186
PF	2014-06-12 17:36:50	——	5.366	3.796	0.989	10.798	12.897	66.154
PF	2014-09-06 16:28:47	——	1.661	9.639	0.000	1.850	25.626	61.225
PF	2014-09-10 16:22:40	——	0.456	2.359	0.971	5.769	13.002	77.443
PF	2014-10-22 00:49:50	——	3.943	10.274	0.063	5.293	32.431	47.997
PF	2014-10-22 13:35:31	——	4.330	20.697	0.656	6.193	20.779	47.346
PF	2014-10-25 16:07:19	——	10.567	30.471	0.054	7.103	18.893	32.911
PF	2014-10-26 23:39:38	——	13.469	24.890	0.000	4.472	14.415	42.753
PF	2014-10-27 14:04:20	——	16.904	17.713	0.295	30.599	19.172	15.317
PF	2014-10-28 01:48:51	——	1.572	0.006	0.173	30.184	24.033	44.032
PF	2014-11-07 09:46:29	——	4.416	4.264	1.835	10.773	11.432	67.280
PF	2014-11-09 15:17:04	——	2.601	5.998	0.018	2.156	15.322	73.905
PF	2015-03-11 06:43:51	——	5.921	13.976	0.973	11.086	11.206	56.837
PF	2015-03-11 15:45:03	——	10.731	16.083	0.627	8.914	4.709	58.936
PF	2015-03-12 11:04:45	——	4.165	20.639	0.003	3.853	5.719	65.620
PF	2015-06-22 17:13:02	——	14.361	17.065	0.227	8.947	13.283	46.117
PF	2015-08-21 18:42:58	——	2.697	1.225	1.470	38.331	8.122	48.155

Figure A.16: Frequency distributions of the recovered overarching groups observed in the different observation windows included in our dataset. The frequencies of one or more of the selected groups are strongly enhanced in pre-flare active region observations.