Using Multiple Instance Learning for Explainable Solar Flare Prediction

C. Huwyler\textsuperscript{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 was 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; Kana-vati 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 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, it has been proposed that an attention-based MIL model 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 modeled 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 show difficulties in revealing how they consult the available 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 down-flow 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...
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. This interaction 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 wavelengths and spatio-temporal resolutions are 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.

We focus in this work on high-resolution spectral data from the near-ultraviolet regime, specifically on a read-out window around the Magnesium (Mg) II k and h line cores centered at 2796.34 and 2803.52 Å (Fig. 2 shows an example of a quiet-sun spectral profile). These data were collected by NASA’s Interface Region Imaging Spectrograph (IRIS) satellite (De Pontieu et al., 2014) that is designed to observe the transition region between the solar chromosphere and the corona, and it provides an unprecedented spectral (53 mÅ), spatial (0.33-0.4 arcsec) and temporal resolution (two-second cadence) along a vertical slit pointed onto a small patch on the Sun. 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 within a 25 minute window (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, considerable pre-processing is required to bring the dataset in a form suitable 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). Although some spectral shapes have been related to certain classes (e.g. according to Panos et al., 2018) some are
clear indicators for a flare, however they do not appear in all flares), it is generally unclear how individual spectra are responsible for the overall label of an observation.

Different Observation Settings. 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), as introduced in section 2.2.

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. 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 ~30,000 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 solar flare prediction given only weak labels using the MIL paradigm. To this end we treat each timestep of our 25 minute long observation time windows as an unsorted bag of individual spectra, thereby ignoring in this work any spatial structure (correlations between neighboring spectra) or the temporal dynamics (correlations between subsequent spectra) of the problem. 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. These criteria are only satisfied by 6 X-class flares and 13 M-class flares, resulting in a total of 19 PF observations (flare start times are indicated in table A.3). 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. Panos and Kleint (2020) justify the small time window of 25 minutes by the fact that IRIS observed on average only 25 minutes before each major flare. The assumption that each timestep has the same label as the whole selected observation time window 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. IRIS operates in two different observation modes: fixed sit-and-stare (always observing the same patch, co-moving with the rotation of the Sun) and moving n-step raster (sweeping over a larger region of the Sun in a repeated manner). Hence, IRIS spectral data comes in the form of spatio-temporal data cubes that contain one time axis (for sit-and-stare) or space-time axis (for n-step raster) \( t \), the \( y \)-axis as a spatial axis (along the vertical slit) and a spectral axis \( \lambda \).

All selected observations in tables (A.1 and A.2) include the same spectral readout window 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 \( \AA \) 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 under-exposed (less than 10 counts per bin) and clipped spectra

---

2 According to the IRIS flare list distributed on \( \text{iris.lmsal.com} \)

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 radiation observed mostly in the vertical direction with respect to the solar surface, while spectra on the solar limb are built up from radiation observed more horizontally and show different characteristics.
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 [Kerr, G. S. et al., 2015].

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 \texttt{irisreader} Python library.

### 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. To get a more robust estimate for the validation set error, we chose to divide the set of observation time windows into three cross-validation folds, stratified by flare class (no flare, M, X). Thereby we took care that no information can leak from the training to the validation set by making sure that spatially and temporally close observation time windows are restricted to the same fold (see observation groups indicated in tables \text{A.1} and \text{A.2}).

The selected observation time windows come at a various number of steps (see tables \text{A.1} and \text{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 works have asserted that Mg II \(h\&k\) profiles 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 ([Vernazza et al., 1981; Leenaarts et al., 2013a]; [Pereira et al., 2013]). 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 \textit{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 a \textit{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.

Before entering MIL more deeply, let us first clarify the connection of a few important MIL notions to our dataset:

- An \textit{instance} is a single Mg II \(h\&k\) intensity profile belonging to a pixel on the spectrograph’s slit.
- A \textit{bag} is the collection of all Mg II \(h\&k\) intensity profiles along the spectrograph’s slit measured for one particular timestep (exposure) of an observation.
- A model predicts at the \textit{instance-level}, when it predicts a quantity on each individual intensity profile, such as an instance label or an attention value.
- A model predicts at the \textit{bag-level}, when it predicts a quantity on a single timestep (exposure) of the full spectrograph’s slit, such as a bag label.

### 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 \textit{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.

To deal with a wider variety of problems, the standard MIL assumption can be relaxed by adopting \textit{e.g.} a \textit{presence-based}, a \textit{threshold-based} or a \textit{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 \(X = \{x_k\}_{k=1,...,n_X}\) where \(x_k \in \mathbb{R}^d\) 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 \(x_k\), resulting in vectors with \(d=240\) dimensions. We attach a label \(Y \in \{0, 1\}\) to each bag

[https://github.com/i4Ds/IRISreader](https://github.com/i4Ds/IRISreader)
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 models the conditional probability for the bag $X$ belonging to the positive class,

$$P(Y = 1|X) = S(X),$$

with

$$\hat{Y} = \begin{cases} 1 & S(X) \geq t \\ 0 & S(X) < t \end{cases},$$

where $t \in (0, 1)$ is a threshold value. We decompose $S$ into a function $f$ that operates on the instances $x_k$ of the bag $X$, 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

$$S = h_{\psi} \circ g_{\varphi} \circ f_\theta.$$  

(3)

The components $h_\psi$, $g_\varphi$, $f_\theta$ are represented by neural networks parameterized by weights $\theta$, $\varphi$ and $\psi$, that can be trained by minimizing a loss $L(Y, P(Y = 1|X))$. 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] *[50]*: *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 $p_k = f_\theta(x_k)$ for belonging to the positive class. The inductive bias is chosen that all learnable parameters are in $f_\theta$, 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 $g(p_k) = \frac{1}{m} \sum_k p_k$ 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):

$$S_r(X) = g(\{f_0(x_k)\}) = \frac{1}{r} \log \left( \frac{1}{n_X} \sum_k e^{r f_0(x_k)} \right),$$

(4)

where $r$ interpolates between the mean ($r \to 0$) and the maximum ($r \to \infty$) and can be considered as a tunable 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, the function $f$ is used to map the individual instances $x_k$ to an embedding $e_k \in \mathbb{R}^m$ that is most suitable for bag classification:

$$e_k = f_\theta(x_k).$$

(5)

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

$$S(X) = h_\psi(E) = h_\psi(g_\varphi(\{f_\theta(x_k)\})).$$

(6)

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):

\[ g_\phi(\{e_k\}) = \sum_{k=1}^{n_X} a_k e_k, \]

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

\[ a_k = \frac{\exp [w^T (\tanh (Ve_k) \odot \sigma (Ve_k))]}{\sum_{j=1}^{n_X} \exp [w^T (\tanh (Ve_j) \odot \sigma (Ve_j))]}, \]

with learned weights \( U, V \in \mathbb{R}^{l \times m} \) and \( w \in \mathbb{R}^l \) that can be summarized as a set of pooling parameters \( \phi = \{U, V, w\} \). 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 \( \sigma(\cdot) \). The free parameter \( l \) controls the complexity of the attention mechanism. In addition, we introduce a sharpening parameter \( \gamma \) to control the skewness of the \( a_k \) by setting \( a_k \rightarrow a_k^\gamma / (\sum_j a_j^\gamma) \), introducing the denominator to ensure that the \( a_k \) still sum to one.

The model has the freedom to assign below-average (<1/nX) or above-average (>1/nX) attention values to individual spectral profiles. For a later comparison of the learned attention values between spectral profiles across different observations (with different nX), the attention values can be brought to a common, normalized scale by defining \( \tilde{a}_k = n_X a_k \). Then, average attention equals to 1, below-average attention to values < 1 and above-average attention to values > 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 \( f_\theta \), \( g_\phi \) and \( h_\psi \)

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 regularization measure. We employed both the instance-based (further called ibMIL) and the attention-based (further called abMIL) MIL model as introduced in Sect. 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 \( \gamma \). In this work, we chose \( l = 10 \) and \( m = 10 \). From previous, unpublished experiences with autoencoders, we believe that a latent space with \( m = 10 \) dimensions is sufficiently high-dimensional to represent IRIS spectra (this is supported by Sadykov et al. (2021) who use a 4-dimensional space). The choice of \( l = 10 \) is somewhat more arbitrary, however firstly the abMIL models perform already extremely well with \( l = 10 \) (see section 4.1) and secondly we prevent the models from overfitting with early stopping (section 3.3.1). The particular choice of \( l \) and \( m \) could definitely be better justified with ablation experiments, however we leave this for later work as it does not change the qualitative results. The role of the parameters \( r \) and \( \gamma \) will be discussed later.

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 neighborhood 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 \( f_\theta \) followed by 20% of dropout. For ibMIL we then applied a single fully-connected node with sigmoid activation, while for the abMIL model we applied 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 (see Fig. 6). For implementation details we refer to appendix Appendix A.

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

3.3.1. Model training

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

$$L(S(X), Y) = \frac{1}{n} \sum_{i} |S(X_i) - Y_i|.$$  

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.

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."
Probably due to its increased flexibility, the abMIL model distributes bag probabilities much more distinctively 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 $\gamma$. As can be seen in Fig. 7, 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).

The results given in Fig. 7 are an aggregate of the 10'000 bags in our dataset derived from the 37 observations in tables A.1 and A.2. Since there is a considerable amount of correlation among the different bags, we estimate the performance of our models on the observation-level by first aggregating the bag results over the individual observations and then aggregating over all observations. In particular, we do this 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 $\gamma$. 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 $\gamma$ 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 of the larger number of false positives, since the bag-level probabilities of some active regions are distributed very broadly (see Fig. 7). 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 bag-level probability evolution of two selected full pre-flare observations for different values of $r$ and $\gamma$ together with instance-level saliency maps (introduced in section 4.2). 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), we currently do not have an explanation for this effect. 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 so-called saliency maps that provide an excellent quick-look tool to identify regions of interest (along the instrument slit) for flare prediction. To visualize the probabilities or attention values predicted by our models, we align the vertically stacked slit pixel predictions horizontally, producing a plot with a spatial y-axis and a temporal x-axis. In this way, regions of interest on the instrument slit during the evolution of the flare can be highlighted. Saliency maps for two selected full flare observations are provided in Fig. 9. Regions with activities that trigger the models are clearly visible.

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 solar spectra (Pietarila et al. 2007; Sánchez Almeida and Lites 2000; Vitiček and Sánchez Almeida 2011; Panos et al. 2018; Dalda et al., 2019; Woods et al. 2021). We follow this approach since it is simple and easily reproducible. To determine a suitable choice for the number of clusters k, we employed the ‘elbow rule’ (Thorndike 1953) by visual inspection of the different within-cluster sum-of-squares distances for different choices of k. An elbow could approximately be identified at around k = 80. To make sure that also distinguished groups with a smaller number of members can be captured, we chose to ‘over cluster’ the data with an intuitive choice of k = 128. It turns out that k = 128 is sufficient to prove the point of this paper, however in the future this decision could be better substantiated.

For the groups with mean probability > 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, in particular also in explosive events accompanied by UV bursts (Vissers et al., 2015; Pepper 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 a temperature distribution along the optical depth 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).
Figure 9: Saliency maps and bag-level 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 fourth 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 $r = 3$) and 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 NASA’s GOES satellite.
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 \(\geq 0.7\) can be related to one of the above-mentioned overarching groups.

- High peak intensity (corresponding to low continuum in our normalization 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 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 (as illustrated in Fig. A.19). While both the ibMIL and the abMIL model provide explanations for their predictions, the asymmetric explanations by the ibMIL model are better.
ter suited for the particular use-case of identifying possible flare precursor candidates.

5. Discussion

We have demonstrated 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 analyzing 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 $\gamma$ (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 $\gamma$ 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 $\gamma$ and propose 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 observed that k-means is not perfectly suited for the analysis of model predictions when working with normalized spectra: 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. Also, Doppler shifts and instrument artifacts such as cosmic ray spikes can move two spectra of otherwise similar shape to different, unrelated groups. In this work, we could alleviate these problems partially by choosing significantly more groups ($k = 128$) than proposed by the elbow rule ($k = 80$).

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 λhk 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 way to exploit the available weak labels on the observation-level by providing a model that links the data from individual spectra up to the observation or bag-level in a well-defined way. This allows training the model with weakly-labeled data and subsequent predictions on the observation or bag-level. Secondly, the specifically selected MIL models allow to propagate the weak observation labels down the other way to individual, instance-level spectral profiles (ibMIL) or at least allow the identification of spectral profiles that are of particular importance for the prediction of the model (abMIL).

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 (Woods et al., 2021) 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 as many available IRIS observations as possible 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.

Acknowledgements

We thank Brandon Panos for valuable comments. This research has been supported by SNSF NFP75 grant no. 407540_167158. IRIS is a NASA small explorer mission developed and operated by LMSAL with mission operations executed at NASA Ames Research center and major contributions to downlink communications funded by ESA and the Norwegian Space Centre. We would like to extend our gratitude to LMSAL for kindly providing us direct access to their IRIS data storage.

References

Ahmadzadeh, A., Aydin, B., Georgoulis, M.K., Kempton, D.J., Mahajan, S.S., Angyryk, 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
Carboneau, 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., DuFurnier, B., Zhou, T., Wang, X., Zhu, H., Sun, Z., Gombosi, T.I., 2019. Identifying solar flare precursors using time series of sdo/hmi data.
Appendix A. Using mask 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 \( f_\theta \) 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 score

\[
  S(X) = \frac{1}{r} \log \left( \frac{1}{n_i} \sum_{k=1}^{n} M_{ik} e^{r P_{ik}} \right), \tag{A.2}
\]

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

Similarly, the abMIL approach computes the bag score 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 \( f_\theta \) 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 \( Q_{ik} = \exp \left[ \mathbf{w}^T \tanh \left( \mathbf{V e}_{ik} \odot \mathbf{U e}_{ik} \right)^{\gamma} \right] \) 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 $\gamma$ 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.
| #    | Start (min) | Duration (arcsec) | Observation Mode | FOV Center (arcsec) | Steps | Cadence (sec) | Slit Pixels | Group Validation | Fold |
|------|-------------|-------------------|------------------|---------------------|-------|--------------|-------------|------------------|------|
| 0    | 2014-03-29 21:05:36 | (686, -166) | Medium sit-and-stare | 155                | 9.6   | 1006         | 24.8        | 23               | 0    |
| 1    | 2014-04-08 04:57:17 | (313, -95)    | Medium coarse 8-step raster | 560               | 21.3  | 4          | 25.0        | 218              | 0    |
| 2    | 2014-04-09 06:15:31 | (-352, -198)   | Medium coarse 4-step raster | 288               | 16.8  | 388         | 24.8        | 208              | 0    |
| 3    | 2014-04-09 22:01:14 | (-777, -303)  | Medium coarse 8-step raster | 704               | 43.8  | 387         | 24.8        | 207              | 0    |
| 4    | 2014-04-09 22:02:40 | (-613, -322)  | Medium coarse 8-step raster | 478               | 3.1   | 387         | 24.8        | 206              | 0    |
| 5    | 2014-04-10 00:31:15 | (375, -26)    | Medium coarse 16-step raster | 388               | 33.6  | 387         | 24.9        | 205              | 0    |
| 6    | 2014-04-10 01:05:21 | (133, -99)    | Very large coarse 8-step raster | 560               | 21.3  | 4          | 25.0        | 204              | 0    |
| 7    | 2014-04-10 01:06:52 | (124, -99)    | Very large coarse 8-step raster | 777               | 203   | 388         | 24.8        | 203              | 0    |
| 8    | 2014-04-10 02:04:20 | (659, -293)   | Very large coarse 8-step raster | 777               | 203   | 388         | 24.8        | 202              | 0    |
| 9    | 2014-04-10 02:05:29 | (-219, 205)   | Large coarse 4-step raster | 560               | 21.3  | 4          | 25.0        | 201              | 0    |
| 10   | 2014-04-10 03:00:18 | (452, -332)   | Medium coarse 32-step raster | 777               | 200   | 388         | 24.8        | 200              | 0    |

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.
Figure A.16: Saliency maps and probability evolutions for two selected active region observations (#0 and #17) created by the ibMIL (first and second row) and abMIL (third and fourth row) models in a similar fashion as in Fig. All the models predict the top observation as a clear negative. In contrast, ibMIL models with \( r \geq 5 \) falsely predict the bottom observation to be pre-flare because of local brightenings observed around slit pixels 85 and 150. As previously explained, the saliency maps with the attention values are difficult to read.
Figure A.17: Groups recovered from the entire 03 March 2014 (top) and 17 October 2015 (middle) active region observations, visualized in a similar fashion as in Fig. 13. The timespan of 25 minutes selected for our dataset is highlighted ('active region selection'). Practically none of the pre-flare groups are identified in the top observation. The second observation shows more pre-flare-like activity around pixels 150 and 85 (as can be compared with Fig. A.16). Similar to the September 10 flare (Figs. 13 and 14), this might be due to mass flows and associated brightenings around these pixels, as illustrated in the slit-jaw image overplot in the bottom figure (taken at the position of the dashed black line in the middle plot).
Figure A.18: 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.
Figure A.19: Top 24 spectral groups in terms of mean normalized abMIL attention ($\gamma = 1$), sorted by mean attention in descending order. For each group the mean attention ($a$) 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 groups of spectra are the same as in the ibMIL case in Fig. 10, however the sorting is different. It is apparent that abMIL’s focus is on spectra from both pre-flare and non-flaring active regions.